Étude de la diversité et de la structure des communautés virales à l'échelle des agro-écosystèmes. Le modèle épidémiologique des mastrévirus des Poaceae à La Réunion

Sohini Claverie

UNIVERSITÉ DE LA RÉUNION
Faculté des Sciences et Technologies

UMR Peuplements Végétaux et Bio-agresseurs en Milieu Tropical
CIRAD – Université de La Réunion

École doctorale Sciences Technologies Santé (STS, ED 542)

THÈSE

Pour l'obtention du grade de
DOCTEUR DE L'UNIVERSITÉ DE LA REUNION

Discipline du doctorat : Biologie Moléculaire

Étude de la diversité et de la structure des communautés virales à l'échelle des agro-écosystèmes

Le modèle épidémiologique des mastrévirus des Poaceae à La Réunion

---

Présentée et soutenue publiquement par

Sohini CLAVERIE

Le 2 Juin 2020, devant le jury composé de :

Mylène OGLIASTRO	Directrice de Recherches, INRAe, Montpellier	Rapportrice
Sébastien MASSART	Professeur, Université de Liège, Gembloux	Rapporteur
Nathalie BECKER	Maître de conférence, HDR, MNHN, Paris	Examinatrice
Camille LEBARBENCHON	Maître de conférence, Université de La Réunion	Examinateur
Stéphane POUSSIER	Professeur, Université de La Réunion	Examinateur
Jean-Michel LETT	Chercheur HDR, CIRAD, Réunion	Directeur de Thèse
Pierre LEFEUVRE	Chercheur, CIRAD, Réunion	Invité

POLE RECHERCHE
Ecoles **D**octorales

LETTRE D'ENGAGEMENT DE NON-PLAGIAT

Je, soussigné(e) Sohini CLAVERIE, en ma qualité de doctorant(e) de l'Université de La Réunion, déclare être conscient(e) que le plagiat est un acte délictueux passible de sanctions disciplinaires. Aussi, dans le respect de la propriété intellectuelle et du droit d'auteur, je m'engage à systématiquement citer mes sources, quelle qu'en soit la forme (textes, images, audiovisuel, internet), dans le cadre de la rédaction de ma thèse et de toute autre production scientifique, sachant que l'établissement est susceptible de soumettre le texte de ma thèse à un logiciel anti-plagiat.

Fait à Saint Pierre, le 20 Avril 2020

Signature : 

Extrait du Règlement intérieur de l'Université de La Réunion
(validé par le Conseil d'Administration en date du 11 décembre 2014)

Article 9. Protection de la propriété intellectuelle – Faux et usage de faux, contrefaçon, plagiat

L'utilisation des ressources informatiques de l'Université implique le respect de ses droits de propriété intellectuelle ainsi que ceux de ses partenaires et plus généralement, de tous tiers titulaires de tels droits.
En conséquence, chaque utilisateur doit :
- utiliser les logiciels dans les conditions de licences souscrites ;
- ne pas reproduire, copier, diffuser, modifier ou utiliser des logiciels, bases de données, pages Web, textes, images, photographies ou autres créations protégées par le droit d'auteur ou un droit privatif, sans avoir obtenu préalablement l'autorisation des titulaires de ces droits.

La contrefaçon et le faux
Conformément aux dispositions du code de la propriété intellectuelle, toute représentation ou reproduction intégrale ou partielle d'une œuvre de l'esprit faite sans le consentement de son auteur est illicite et constitue un délit pénal.
L'article 444-1 du code pénal dispose : « Constitue un faux toute altération frauduleuse de la vérité, de nature à causer un préjudice et accomplie par quelque moyen que ce soit, dans un écrit ou tout autre support d'expression de la pensée qui a pour objet ou qui peut avoir pour effet d'établir la preuve d'un droit ou d'un fait ayant des conséquences juridiques ».
L'article L335_3 du code de la propriété intellectuelle précise que : « Est également un délit de contrefaçon toute reproduction, représentation ou diffusion, par quelque moyen que ce soit, d'une œuvre de l'esprit en violation des droits de l'auteur, tels qu'ils sont définis et réglementés par la loi. Est également un délit de contrefaçon la violation de l'un des droits de l'auteur d'un logiciel (…) ».

Le plagiat est constitué par la copie, totale ou partielle d'un travail réalisé par autrui, lorsque la source empruntée n'est pas citée, quel que soit le moyen utilisé. Le plagiat constitue une violation du droit d'auteur (au sens des articles L 335-2 et L 335-3 du code de la propriété intellectuelle). Il peut être assimilé à un délit de contrefaçon. C'est aussi une faute disciplinaire, susceptible d'entraîner une sanction.
Les sources et les références utilisées dans le cadre des travaux (préparations, devoirs, mémoires, thèses, rapports de stage…) doivent être clairement citées. Des citations intégrales peuvent figurer dans les documents rendus, si elles sont assorties de leur référence (nom d'auteur, publication, date, éditeur…) et identifiées comme telles par des guillemets ou des italiques.
Les délits de contrefaçon, de plagiat et d'usage de faux peuvent donner lieu à une sanction disciplinaire indépendante de la mise en œuvre de poursuites pénales.

Remerciements

Ainsi s’achève mes 3 ans et demi de thèse et je l’espère le début de ma carrière de chercheur, consécration de mes études initiales de Biochimie. Durant ces 3 ans et demi de thèse de par mon age, j’ai eu droit (souvent) à la question fatidique « Alors Sohini, c’est quand que tu fais un bébé ? » ma réponse fut toujours « le seul accouchement prévu est celui de ma thèse » . Oui ça y est...le bébé est la !

Tout d’abord, je tiens à remercier les membres du jury, Mylène Ogliastro, Sébastien Massart, Nathalie Becker, Camille Lebarbenchon et Stéphane Poussier pour avoir accepté de faire partie de mon jury de thèse et pour le temps consacré à évaluer ce travail. Par ailleurs, je souhaite également remercier les membres de mon comité de thèse : Frédéric Chiroleu, Denis Filloux, Emmanuel Jacquot, Gael Thébaud et Philippe Roumagnac pour m’avoir aiguillée tout au long de ce projet.

Je remercie également mes différents financeurs (la région Réunion, l’Université de La Réunion, le CIRAD et la fondation agropolis) sans qui ce travail aurait vu le jour ainsi que Bernard Reynaud de m’avoir chaleureusement accueillie et fait confiance au 3P. Merci également à nos collaborateurs Darren Martin pour son aide et ses conseils sur la phylogénie, la recombinaison ainsi que son accueil à Cape Town. Merci également à Arvind Varsani, pour son implication et de m’avoir si bien formée par le passé en laboratoire à Christchruch.

Je voudrais bien évidemment exprimer toute ma gratitude aux deux personnes qui ont porté elles aussi ce petit bébé et sans qui cela aurait été possible.
Un immense merci à Jean-Michel de m’avoir guidée, pour sa confiance avant et durant ce projet. Merci de m’avoir donnée l’opportunité de faire cette thèse, de voyager aux quatre coins du monde ainsi que pour les sessions ride au ski !
Un immense merci également à mon encadrant Pierre, pour son implication sans faille, son suivi tout au long de cette thèse, sa rigueur, son exigence, son œil affûté et sa patience inébranlable. Merci de m’avoir initiée au monde diabolique de la bioinformatique, j’ai fait tellement de progrès grâce à toi, merci!
Je vous remercie tous les deux pour votre temps, votre investissement, votre patience et votre pédagogie. Merci pour les nombreuses heures de discussion, de présentation, de corrections, de lecture et relecture. Je pense que tous deux ont perdu quelques cheveux durant ma thèse, ahahah !
Merci également au dernier membre de notre équipe de géminilovers, Mumu pour son aide précieuse au labo et d’avoir continué à y croire même si clairement parfois on avait envie de pleurer tellement ça ne marchait pas ! La malédiction du clonage a encore frappé !

Je tiens à remercier tous mes collègues au Cirad, la liste est si longue que j'aurai trop peur d'oublier certains ! Vous avez tous participé de près comme de loin à cette thèse, par votre aide ou par votre bonne humeur au quotidien. J'ai tellement apprécié travailler avec chacun de vous, c'est si important d'avoir une bonne ambiance au travail, et cela a été possible grâce à vous tous !

Merci notamment à Virginie Ravigné, Benoit Facon et le Chichi(roleu) pour leur aide en statistiques, sous R et également pour leur aide en écologie. Merci également à Mathieu Rouget, pour son aide précieuse sur les concepts d'écologie, « Dieu » sait à quel point j'étais perdue parfois ahahah ! Merci d'avoir été si pédagogue et d'avoir pris le temps de m'expliquer.

Merci à Stéphane sans qui je ne serais au CIRAD à l'heure actuelle. Merci de m'avoir poussée, alors que j'étais étudiante en Biochimie, à continuer dans mon projet de devenir enseignant-chercheur ainsi que de m'avoir permis de faire ce premier stage au CIRAD il y a déjà 7 ans, ce qui a marqué le début de cette grande aventure ! Le CIRAD ma deuxième (première ?) maison !

Merci à Isabelle Robène, Nathalie Becker et Hélène Delatte pour leur bonne humeur au quotidien, les recettes de cuisine ainsi que pour les petites discussions qui m'ont fait tellement de bien !

Merci également à tous les techniciens, qui ont fait que mes longues heures au labo ont été si agréables, rythmées par des discussions, des rires, des pleurs, la découverte de choix musicaux particuliers ainsi que leur aide au labo et sur le terrain !

Merci à mes deux collègues du bureau d'en face, Jacques et Cyril, ce combo merveilleux ! Merci Jacques pour ces discussions sur tellement de sujets différents et si passionnants. Cela m'a permis de m'évader plus d'une fois ! Merci Cyril, collègue et ami, d'avoir été là dans les bons comme dans les mauvais moments, surtout dans ces phases de doute qui après nos discussions s'estompaient! Un immense merci à toi !

Merci à mon collègue préféré (ahahah), mon roux préféré, ma belle biche fournissant tellement de bonbons haribo (le diabète me guette !). Merci de m'avoir tellement fait rire, un bol d'air frais ! Olivier, surtout reste comme tu es, c'est si épique (oui épique !).

Merci aux collègues partageant mon bureau. Tout d'abord Sylvia et Noura, les ralstogirls de folie, pour ce début de thèse de folie et ensuite Anziz et Cathu. Ces années auprès de vous, cela a été unique! Merci pour votre soutien et pour cette amitié ! Bon courage à Anziz et Cathu ainsi que tous mes thésards-amis pour leur fin de thèse, vous allez y arriver, je suis de tout cœur avec vous !

Ce début de thèse a été marqué par la rencontre d'une personne formidable, dans un train quittant Montpellier. Cette personne assise en face de moi avec des perles de Tahiti, se rendait elle aussi à La Réunion, elle aussi pour une thèse (au CIRAD en plus !). Ma jumelle de thèse, Alizée ! « Seigneur » que je suis tellement contente que tu sois entrée dans ma vie ce jour la, si tu savais !

Cette amitié j'espère quelle durera toute ma vie, en fait non j'espère pas je le sais c'est tout ! Cette thèse aurait été si différente sans toi, elle l'a été d'autant plus belle grâce à toi ! A tous ces fous rires, ces vidéos envoyées avec des filtres, ces messages, ces conversations, ces pleurs, tous ces moments de doute mais surtout tous ces moments de soutien inébranlable ! J'ai tellement aimé nos petites discussions en face de nos portes de bureau ou dans nos voitures après 20h juste devant le portail du CIRAD, ahahah, c'est si mémorable ! Je suis à 300 % avec toi pour cette fin de thèse! Merci d'avoir été si présente pour moi !

Un immense merci à Sarah, collègue de promo, binôme, coloc, et surtout amie incommensurable ! Cette thèse sans toi aussi n'aurait pas du tout, mais du tout été pareil! Merci tellement pour toute ton aide au boulot comme dans la vie de tous les jours! Merci pour ce soutien indéfectible avant, pendant et après la thèse!!!!! Merci pour toutes les conversations également devant la porte de bureau, dans le parking du CIRAD et devant le portail du CIRAD. Avec Alizée, c'est vraiment notre marque de fabrique ahahah ! Je suis si nostalgique en repensant à nos trois voitures juste après la fermeture du CIRAD pour des sessions « papotage » ! Merci Sarah pour ta force de caractère qui a été une force pour moi ! Maintenant que je vais de nouveau avoir une vie, on va pouvoir se voir de nouveau comme on adore le faire ! Ça me manque tant...Merci encore pour tout Sarah ! Ces années de thèse avec Alizée et toi ont été si drôles à vos côtés ! Merci pour ces moments de bêtises (et parfois de sérieux) que l'on a partagés toutes les trois !

Un grand merci à mes amis d'enfance Ambre, Véro, Gazou, Lau, Oli, Perrine, Maryline, Sandrine, Erika, Séverine, Candy, Fred, Leila, Romain, Nash, Laurent et tous les autres de m'avoir soutenue, d'avoir cru en moi et d'avoir accepté d'être beaucoup moins présente pour eux depuis le début de ma thèse, vous faites partie de ma plus grande fierté, sachez le !

Je tiens profondément à remercier également ma belle famille, le clan Marot pour leur soutien, les moments de partages, de jeux ainsi que les bons petits plats de ma belle mère Danièle (merci énormément !!!) surtout durant cette dernière ligne droite !

Cette dernière partie de remerciements, je la consacre évidemment à ma famille proche: mes quatre parents, ma sœur et l'homme qui partage ma vie ! Lilyi, merci d'avoir accepté que je sois moins présente dans ta vie. Même si j'étais peu présente depuis le début de la thèse, tu ne peux pas savoir à quel point je suis fière de toi, fière d'être ta sœur, je profite pour te le dire ! J'aimerai à l'avenir être plus disponible pour toi et je le ferai, promis !

James que dire, ou plutôt par ou commencer ? Tu n'as connu que la Sohini durant la thèse, « Seigneur » sacrée expérience ahahah ! Mais malgré ces trois années de stress, de pression, tu as su voir qui j'étais, tu m'as aidée, tu m'as incitée à me dépasser, à être plus forte et j'espère que tu es fier de moi car j'ai tout donné ! Merci d'avoir su accepter que la thèse fasse partie intégrante de ma vie, car c'était mon projet personnel et non le tien et pourtant tu l'as porté indirectement. Merci d'avoir été si compréhensif, de m'avoir laissée travailler tous les week-ends, finir à 20h au CIRAD… sans jamais broncher. Construire une relation de couple dans un tel contexte, où

son conjoint ne fait que bosser et passe très peu de temps avec l'autre n'est pas chose évidente… Tout le monde ne comprend pas et pourtant toi tu l'as fait ! Merci infiniment d'avoir fait partie de ce projet et de partager ma vie.

Comme on dit le meilleur pour la fin ! Mes parents… merci infiniment à mes parents qui ne sont pas deux mais quatre! Merci tout d'abord à ma belle mère et à mon beau père d'avoir été la, pour leur bonne humeur au quotidien. Merci Emma pour les bons petits plats, les discussions toujours aussi agréables et le soutien incontestable dans ces moments de doute et de découragement que j'ai eus. Merci Jean François d'avoir accepté que je réquisitionne Maman de temps en temps et merci d'être mon coiffeur attitré depuis que je suis en thèse !

Maman… sincèrement, merci, pour toutes ces fois ou j'ai appelé, parfois à des heures improbables, juste pour parler car j'en avais tout simplement besoin. Merci pour tous les messages de soutien, de courage que tu m'as envoyé au quotidien. Cette thèse a été tellement rythmée avec des problèmes de santé divers et variés… merci d'être venue chez les médecins avec moi, merci d'avoir été là avant et après les examens et les opérations. Merci pour tous ces appels quotidiens durant mes 1h30 de trajet de voiture. C'est devenu notre petit rituel de s'appeler avant ou après le boulot, et j'en garde toujours un agréable souvenir, des moments de légèreté...

Papa… mon mentor depuis toujours, celui qui m'a toujours poussée à me dépasser et à réaliser mes rêves à 300 %, même les plus fous ! Merci d'avoir toujours cru en moi, même lorsque que du haut de mes 11 ans je t'ai dit que je voulais être chercheur ! Tu as toujours été la pour moi, la thèse n'est qu'un exemple parmi tant d'autres ! Merci pour tous tes messages (même ceux ou tu as eu des silences radio de ma part sans m'en tenir rigueur ahahah), tous tes appels dans les bons comme dans les pires moments. Merci pour ton soutien sans limite et pour tous les conseils que tu m'as donnés. L'éducation que tu m'as transmise, les valeurs que tu as toujours défendues ont été un pilier pour moi notamment à travers cette thèse...

Papa, Maman, merci pour tout l'amour dans lequel vous m'avez faite grandir. Sans vous, je ne serais pas arrivée jusque-là ! Vous êtes ma force ! Merci de m'avoir soutenue et motivée. Merci d'avoir cru en moi et d'y croire encore. Je suis si fière des personnes que vous êtes et j'espère que vous serez encore plus fiers de la personne que je suis.

Un immense merci à toutes les personnes qui de près ou de loin m'ont accompagnée durant ces années de thèse.

Je dédie cette thèse à la fillette de 11 ans que j'étais et que je suis toujours au fond de moi…

PS : je suis restée fidèle à moi-même : une pipelette qui fait beaucoup de répétitions. Je ne saurai compter le nombre de mercis dans cette partie ahahah !

Sommaire

Liste des acronymes viraux selon le comité international de taxonomie virale

ACSV :	*Axonopus compressus streak virus*
ALCV :	*Alfalfa leaf curl virus*
AYVSGA :	*Ageratum yellow vein Singapore alphasatellite*
BCSMV :	*Bromus catharticus striate mosaic virus*
BDV :	*Barley dwarf virus*
BYDV :	*Barley yellow dwarf virus*
CLCuEA :	*Cotton leaf curl Egypt alphasatellite*
CLCuGeA :	*Cotton leaf curl Gezira alphasatellite*
CLCuMA :	*Cotton leaf curl Multan alphasatellite*
CLCuMuA :	*Cotton leaf curl Multan alphasatellite*
CLCuSAA :	*Cotton leaf curl Saudi Arabia alphasatellite*
ClLCrA :	*Cleome leaf crumple alphasatellite*
CMMGA :	*Cassava mosaic Madagascar alphasatellite*
CpCAV :	*Chickpea chlorosis Australia virus*
CpCDPV :	*Chickpea chlorotic dwarf Pakistan virus*
CpCDV :	*Chickpea chlorotic dwarf virus*
CpCV :	*Chickpea chlorosis virus*
CpRLV :	*Chickpea redleaf virus*
CpYDV :	*Chickpea yellow dwarf virus*
CpYV :	*Chickpea yellows virus*
CSMV :	*Chloris striate mosaic virus*
CTV :	*Citrus tristeza virus*
CYDV :	*Cereal yellow dwarf virus*
CYMA :	*Cucurbit yellow mosaic alphasatellite*
DCSMV :	*Digitaria ciliaris striate mosaic virus*
DDSMV :	*Digitaria didactyla striate mosaic virus*
DfaA :	*Dragonfly associated alphasatellite*
DSV :	*Digitaria streak virus*
EACMV :	*East African cassava mosaic virus*
EcmLV :	*Euphorbia caput-medusae latent virus*
EIAV :	*Eleusine indica associated virus*
EMSV :	*Eragrostis minor streak virus*
ESV :	*Eragrostis streak virus*

EuYMA : *Euphorbia yellow mosaic alphasatellite*
EuYMV : *Euphorbia yellow mosaic virus*
GLCuA : *Guar leaf curl alphasatellite*
GmusSLA : *Gossypium mustelinum symptomless alphasatellite*
MeCMA : *Melon chlorotic mosaic alphasatellite*
MeRAV : *Melinis repens associated virus*
MiSV : *Miscanthus streak virus*
MSDV : *Maize streak dwarfing virus*
MSMV : *Maize striate mosaic virus*
MSRV : *Maize streak Reunion virus*
MSV : *Maize streak virus*
ODV : *Oat dwarf virus*
SSV : *Sugarcane streak virus*
SWSV : *Sugarcane white streak virus*
ToLCuBuA : *Tomato leaf curl Buea alphasatellite*
TYDV : *Tobacco yellow dwarf virus*
TYLCV-IL : *Tomato yellow leaf curl virus Israel*
TYLCV-Mld : *Tomato yellow leaf curl virus mild*
USV : *Urochloa streak virus*
WDIV : *Wheat dwarf India virus en Inde*
WDV : *Wheat dwarf virus*
WfaGA 1 : *Whitefly associated Guatemala alphasatellite 1*
WfaGA 2 : *Whitefly associated Guatemala alphasatellite 2*
WfaPRA 1 : *Whitefly associated Puerto Rico alphasatellite 1*
SSMV-2 : *Sporobolus striate mosaic virus 2*
SSRV : *Sugarcane streak Reunion virus*
SSV : *Sugarcane streak virus*
SWSV : *Sugarcane white streak virus*
ToLCuBuA : *Tomato leaf curl Buea alphasatellite*
TYDV : *Tobacco yellow dwarf virus*
TYLCV-IL : *Tomato yellow leaf curl virus Israel*
TYLCV-Mld : *Tomato yellow leaf curl virus mild*
USV : *Urochloa streak virus*
WDIV : *Wheat dwarf India virus en Inde*
WDV : *Wheat dwarf virus*
WfaGA 1 : *Whitefly associated Guatemala alphasatellite 1*

WfaGA 2 : *Whitefly associated Guatemala alphasatellite 2*

WfaPRA 1 : *Whitefly associated Puerto Rico alphasatellite 1*

Liste des abréviations

- **Organismes et institutions**

3P :	Pôle de Protection des Plantes
BGPI :	Biologie et Génétique des Interactions Plante-Parasite
CIRAD :	Centre de coopération Internationale en Recherche Agronomique pour le Développement
GenBank :	Banque de données génomiques des Etats-Unis d'Amérique
ICTV :	International Committee on Taxonomy of Viruses
NCBI :	National Center of Biotechnology Information
UMR/PVBMT :	Unité Mixte de Recherche/Peuplements Végétaux et Bio-agresseurs en Milieu Tropical

- **Autres abréviations**

***Acc*I :**	Enzyme de restriction produit par la souche bactérienne *Acinetobacter calcoaceticus*
ADN :	Acide Désoxyribonucléique
ADN-A :	Composant A du génome des bégomovirus
ADN-B :	Composant B du génome des bégomovirus
ADNdb :	Acide désoxyribonucléique double brin
ADNsb :	Acide désoxyribonucléique simple brin
AfSV :	African streak virus
ARN :	Acide ribonucléique
ARNi :	Acide ribonucléique interférence
ARNdb :	Acide ribonucléique double brin
ARNsb :	Acide ribonucléique simple brin
***BamH*I :**	Enzyme de restriction produit par la souche bactérienne *Bacillus amyloliquefaciens*
cccDNA :	Forme super-enroulée de l'ADN viral
CMD :	Cassava mosaic disease
CMGs :	Cassava mosaic geminiviruses
CP :	Protéine de capside
CR :	Common region
CRESS-DNA :	Circular Rep-encoding ssDNA
CRT :	Cyclic reversible termination

dN :	Taux de mutations non synonymes
dNTP :	désoxynucléotide triphosphate
dS :	Taux de mutations synonymes
ELISA :	Enzyme Linked Immuno Sorbent Assay
emPCR :	Emulsion polymerase chain reaction
IR :	Intergenic Region ou Région intergénique
***Kpn*I :**	Enzyme de restriction produit par la souche bactérienne *Klebsiella pneumoniae*
LIR :	Long intergenic region
MID :	Multiplex identifier
ML :	Maximum-likelihood
MP :	Protéine de Mouvement
MSD :	Maize streak disease
***Nco*I :**	Enzyme de restriction produit par la souche bactérienne *Nocardia corallina*
***Nde*I :**	Enzyme de restriction produit par la souche bactérienne *Neisseria denitrificans*
NGS :	Next Generation Sequencing
NSP :	Nuclear Shuttle Protein
ONT :	Oxford Nanopore Technologies
ORF :	Open Reading Frames ou Cadre ouverte de lecture
Ori :	Origine de réplication
OTU :	Operational Taxonomic Unit
PacBio :	Pacific bioscience
pb :	Paire de bases
PCR :	Réaction de Polymérisation en Chaine
***PstI* :**	Enzyme de restriction produit par la souche bactérienne Providencia stuartii
RA :	Random Amplification
RCA :	Rolling Circle Amplification
RCR :	Rolling Circle Replication ou Réplication en Cercle Roulant
RDR :	Recombination-Dependent Replication ou Réplication dépendante de la Recombinaison
Ren :	Replication Enhancer ou Protéine activatrice de la réplication
Rep :	Protéine de Réplication
***Sac*I :**	Enzyme de restriction produit par la souche bactérienne *Streptomyces achromogenes*

***Sal*I :**	Enzyme de restriction produit par la souche bactérienne Streptomyces albus
SDT :	Sequence Demarcation Tool
SGS :	Second Generation Sequencing
SIR :	Short intergenic region
siRNA :	Small interfering Ribonucleic Acid
SMRT :	Single-molecule real-time sequencing
SOOI :	Sud-ouest de l'océan Indien
ssDNA :	Single-stranded circular DNA
TGS :	Third generation sequencing
TrAP :	Transcriptional Activator Protein ou Protéine activatrice de la transcription
VANA :	Virion-associated nucleic acids
VEM :	Vector enabled metagenomics
ZMW :	zero-mode waveguide

Liste des figures

- **Introduction**

- Article 4

- ## Annexes

Liste des tableaux

- **Introduction**

- Review

- **Chapitre 1**

- Article 1

- **Chapitre 2**

- Article 3

- **Annexes**

Introduction

Sommaire de l'introduction générale

Introduction

1. Contexte général

« Des virus partout – partout des virus » tel pourrait être l'adage résumant l'**ubiquité** des virus dans l'environnement. La présence de virus partout sur le globe dans des habitats aux conditions environnementales diverses, parfois très contrastées, voire extrêmes tels que les lacs et sols de l'Antarctique (Rastrojo & Alcamí, 2018), le sable du désert du Sahara (Prigent *et al.*, 2005) ou encore les lacs alcalins ou hyper-salins (Atanasova *et al.*, 2012 ; Jiang *et al.*, 2004), rendent compte de leur grande diversité génétique et de leur forte capacité d'adaptation.

Les virus sont les entités les plus **abondantes** de la biosphère (Breitbart & Rohwer, 2005 ; Suttle, 2005, 2007). On estime ainsi que la présence de particules virales dans les écosystèmes marins serait de l'ordre du milliard par litre d'eau et que la masse de carbone de ces dernières équivaudrait à celle de 75 millions de baleines bleues (Suttle, 2005). Les virus se caractérisent également par leur grande **diversité génétique** et **biologique** tant dans leur plasticité phénotypique, que par leurs stratégies d'infection, ainsi qu'à travers leurs larges et diverses gammes d'hôtes incluant les bactéries (Weinbauer & Rassoulzadegan, 2004), les archées (Rangishvili *et al.,* 2006), les plantes (García-Arenal *et al.*, 2001), les champignons (Chu *et al.*, 2002), les animaux (Grubman & Baxt, 2004) et le corps humain (Breitbart *et al.*, 2003). Il existe un large spectre d'**interactions** entre les **virus**, leurs **hôtes** et leurs **vecteurs de transmission**. Les interactions virus-hôtes varient d'associations antagonistes (parasitisme et pathogénicité) à mutualistes (Rohwer & Thurber, 2009 ; Roossinck, 2011 ; Suttle, 2007). Néanmoins, depuis la découverte du premier virus le tobacco mosaic virus (TMV) par Ivanovski en 1892 et Beijerinck en 1898 (Bos, 1999), les études sur la diversité virale ce sont majoritairement focalisées sur les **virus pathogènes** responsables de maladies en raison de leur recrudescence et de leur impact majeur sur la santé humaine (Ebola chez l'Homme ; Coltart *et al.*, 2017), animale (la grippe aviaire chez les animaux ; Alexander, 2007) ou sur le rendement des plantes d'intérêt agronomique (la striure du maïs chez les plantes ; Bosque-Pérez,

2000). Or, il est acquis aujourd'hui que les virus responsables de maladies ne représentent qu'une **fraction minoritaire** de la diversité virale présente (Cooper & Jones, 2006 ; Rosario & Breitbart, 2011 ; Wren *et al.*, 2006).

Afin de combler ce **manque d'exhaustivité** dans l'étude de la diversité virale, des techniques dites de **métagénomiques** ont été développées (Handelsman *et al.*, 1998) permettant d'analyser sans *a priori* l'ensemble des virus (**virome**) d'un échantillon environnemental. En parallèle, des avancées méthodologiques associées au séquençage haut-débit (Roossinck *et al.*, 2015) ont permis l'avènement de la métagénomique virale (**viromique**, Koonin & Dolja, 2018). Ces études ont mis en évidence l'immense diversité virale existante et suggère que ces virus ont un rôle essentiel dans le fonctionnement des écosystèmes. Aussi, si ces études ont enrichi nos connaissances, elles ont aussi révélé un ensemble de séquences nucléotidiques, appelé « matière noire », pour lesquelles l'attribution d'une quelconque ressemblance avec des virus connus voir même d'une fonction reste difficile (Rosario & Breitbart, 2011). La présence majoritaire de cette matière noire dans les résultats de nombreuses études métagénomiques souligne de nouveau que beaucoup reste à être découvert en virologie.

A la fois l'ubiquité, l'abondance, la grande diversité génétique et biologique des virus et le large spectre interactions virus - vecteur - hôte - environnement suggèrent que les virus font partie intégrante des écosystèmes d'où l'importance d'étudier les communautés virales dans leur globalité. Ainsi, en inventoriant de façon précise et exhaustive la diversité virale à l'échelle des écosystèmes, il sera alors possible de caractériser la **structure des communautés virales** (*i.e.* l'assemblage de diverses populations d'espèces virales distinctes) et de déterminer leur **dynamique** (interactions, évolution, adaptation) au sein des écosystèmes.

2. Les phytovirus

2.1. Généralités sur les phytovirus

2.1.1. L'origine des phytovirus

L'utilisation du terme phytovirus pourrait suggérer que les virus de plantes forme un groupe cohésif de virus apparentés. En réalité, les virus de plantes sont très divers dans leurs origines et histoires évolutives. Trois scenarios principaux sur l'origine multiple (**polyphylétique**) des virus de plante ont été avancés avec (i) l'existence d'un **ancêtre commun** antérieur à la divergence des eucaryotes (cas des picornavirus par exemple), (ii) l'évolution grâce aux **transferts horizontaux** de gènes (cas des bunyavirus par exemple) et (iii) l'**origine parallèle** d'éléments génétiques apparentés (cas des circovirus et des géminivirus par exemple) (**Figure 1** ; Dolja & Koonin, 2011). L'analyse des origines possible des différents groupes de phytovirus suggère que ces trois voies d'évolution ont été importantes dans l'histoire et que leurs contributions varient d'un groupe viral à l'autre.

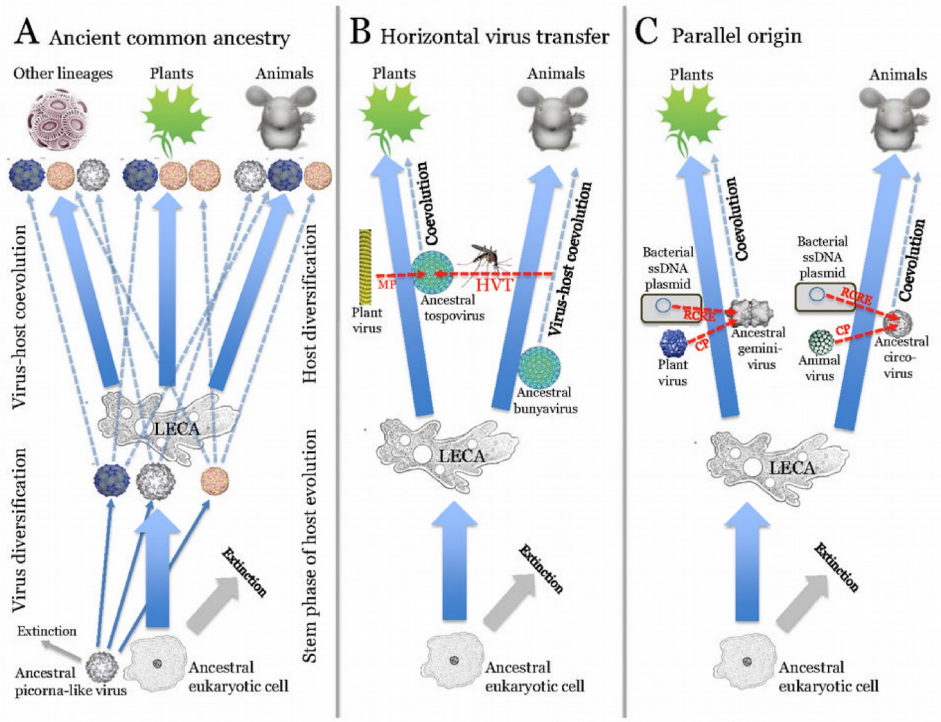


Figure 1. Représentation d'exemples associés aux trois scénarios proposés pour expliquer les origines des phytovirus. Les abréviations correspondent à : LECA, le dernier ancêtre commun eucaryote ; MP, la protéine de mouvement, HVT, le transfert viral horizontal ; CP, la protéine de capside ; RCRE, l'endonucléase de réplication en cercle roulant (Dolja & Koonin, 2011).

2.1.2. Les particularités des phytovirus

Malgré la diversité des histoires évolutives des virus infectant les plantes, certaines spécificités sont généralement associées à ceux-ci (Astier *et al.*, 2001), telles que (i) la présence majoritaire de virus **non enveloppés**, (ii) la présence fréquente de **plusieurs composants** du génome viral, (iii) l'association avec des molécules **satellites**, (iv) la présence de gènes codant pour une **protéine de mouvement** ainsi que (v) leur **rare transmission par contact direct**.

Chez les virus à plusieurs composants génomiques, on parle de « **virus segmenté** », lorsque les différents composants du génome viral sont tous encapsidés au sein d'un virion unique, et de « **virus multipartite** », lorsque les composants génomiques sont encapsidés dans des particules distinctes (Varsani *et al.*, 2018). La plupart des virus multipartites connus infectent les plantes (90% des espèces et des genres). Ils ont été décrits dans 13 des 24 familles de phytovirus connues (Lucía-Sanz & Manrubia, 2017), comme par exemple les bégomovirus de la famille des *Geminiviridae* (Fontes *et al.*, 1994) ou encore l'ensemble des virus appartenant aux *Partitiviridae* et *Bromoviridae* (Lucía-Sanz & Manrubia, 2017). Les données actuelles ne supportent aucune explication convaincante sur les raisons qui expliqueraient la prévalence de virus multipartites chez les phytovirus (Sicard *et al.*, 2016). La grande majorité des molécules satellites a elle aussi été identifiée en association avec des phytovirus monopartites (*i.e.* ne présentant qu'un seul composant génomique). Ces composés co-transmis avec le virus assistant peuvent modifier l'accumulation et la pathogénicité de ce dernier (Gnanasekaran & Chakraborty, 2018).

Une autre particularité chez les phytovirus est la présence fréquente de gènes codant pour une ou plusieurs protéines de mouvement. Les protéines de mouvements en se liant à un système de translocation interne, induisent un élargissement de la taille limite d'exclusion du plasmodesme facilitant ainsi le transport de virions et/ou génomes viraux (Benitez-Alfonso *et al.*, 2010 ; Lucas, 2006).

La dernière caractéristique des phytovirus est la rare transmission par contact direct, contrairement à de nombreux virus infectant les animaux. Bien que les

phytovirus puissent être transmis via des plantes parasites, des greffes ou encore par le biais de sols et eaux contaminés, le mode de transmission le plus efficace est la transmission par vecteurs, par le pollen et les graines (Jones, 2018). Néanmoins, la transmission vectorielle semble être la plus courante, avec une transmission par arthropodes majoritaire (principalement des hémiptères) mais également par nématodes ou champignons (Hogenhout *et al.*, 2008 ; Nault, 1997 ; Tamada & Kondo, 2013) .

2.2. Une diversité sous-estimée et biaisée

Les virus ont été classés en sept groupes en fonction de la **nature** et de la **structure** de leurs **acides nucléiques** (ADN simple ou double brin, ARN simple brin positif ou négatif, ou ARN double brin) et de leurs **stratégies de réplication** (Baltimore, 1971). Dans chacun de ces groupes, les virus sont classés de manière hiérarchique et similaire à la classification utilisée pour les autres organismes vivants, classant les virus par ordre, famille, sous-famille, genre et espèce. La classification des espèces et des souches virales est également basée sur (i) des critères de similarité des séquences nucléotidiques et l'utilisation de seuil de distinction taxonomique, (ii) la gamme d'hôte naturel, (iii) la localisation cellulaire et tissulaire, (iv) la pathogénie et (v) les propriétés physicochimiques et antigéniques. Depuis 2017, une classification phylogénétique basée uniquement sur l'information de la séquence génomique des virus a été approuvée par l'ICTV avec l'usage de rangs taxonomiques supérieurs, notamment pour les virus à ARN fondée sur la phylogénie d'un marqueur universel l'ARN polymérase ARN-dépendante (Kuhn, 2019). Par ailleurs, l'ICTV a récemment accepté les données partielles issues du séquençage haut débit afin d'enrichir la taxonomie virale (Simmonds *et al.*, 2017).

Comme pour l'ensemble des virus, les connaissances actuelles de la diversité des phytovirus souffrent d'un biais important liés au fait que la grande majorité des recherches menées jusqu'ici se sont focalisées sur les virus pathogènes de cultures et de plantes ornementales (**Figure 2** ; Cooper & Jones, 2006 ; Roossinck, 2011 ; Wren *et al.*, 2006).

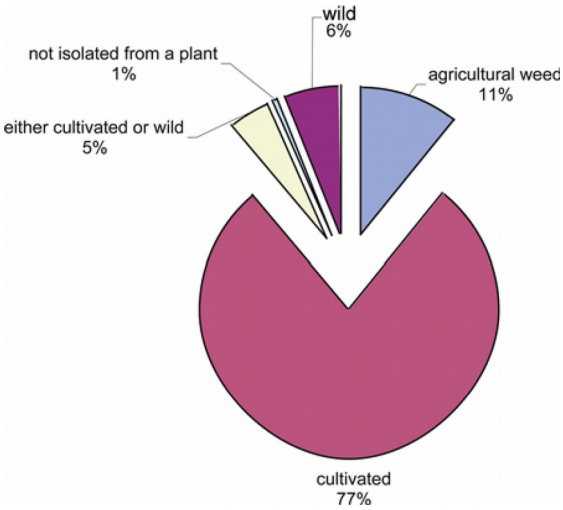


Figure 2. Proportion des plantes sources à partir desquelles les phytovirus ont été initialement isolés (Wren *et al.*, 2006).

De plus, bien que l'infection virale puisse être visuellement non identifiable (Remold, 2002), il a été souvent supposé dans le passé que l'**absence de symptômes apparents** rendait compte d'une **absence de virus** au sein de la plante (Prendeville *et al.*, 2012 ; Remold, 2002). Les symptômes d'infection virale sont en effet parfois difficiles à distinguer des stress environnementaux. La vision faussée de la phytovirosphère excluant les plantes sauvages ou non cultivées (adventices, friches etc.) a entraîné non seulement un biais dans la connaissance de la **diversité virale globale** des écosystèmes végétaux mais également une **sous-estimation** du rôle potentiel du milieu sauvage en tant que **réservoir de biodiversité** ainsi que son implication dans la compréhension de l'émergence des phytovirus pathogènes.

Très peu de plantes sauvages infectées présentent les symptômes caractéristiques de maladies virales connues et seul un diagnostic du virus lui-même peut rendre compte d'une infection (Roossinck *et al.*, 2010, 2015). Par exemple, une étude concernant cinq virus infectant 21 populations sauvages de *Cucurbita pepo* a révélé que 80% des plantes infectées ne présentaient pas de symptômes visibles (Prendeville *et al.*, 2012). A l'aide de techniques de séquençage de nouvelle génération (ou NGS pour *Next generation sequencing*), plusieurs études visant à décrire la diversité des virus infectant les plantes sauvages ont démontré que (i) de nombreux virus restent à découvrir et à caractériser (« matière noire »), (ii) les taux d'infection des

plantes sauvages sont généralement élevés, (iii) les infections virales en l'**absence de symptômes** sont communes et majoritaires et (iv) la diversité virale est **plus abondante** au sein du compartiment sauvage que du compartiment cultivé (Bernardo *et al.*, 2018 ; Muthukumar *et al.*, 2009 ; Roossinck *et al.*, 2010). Ces études ont notamment montré que 70% des plantes sauvages analysées au Costa Rica (Roossinck *et al.*, 2010), 26 % en Oklahoma (Muthukumar *et al.*, 2009) et entre 26 et 36 % en France et en Afrique du Sud (Bernardo *et al.*, 2018) étaient infectées par des virus.

2.3. Les processus évolutifs impliqués dans la diversification et l'adaptation des phytovirus

La capacité des virus à s'adapter à de nouveaux hôtes et environnements (succès de la transmission et de l'infection, détournement de la machinerie cellulaire de l'hôte, contournement de résistance etc.) dépend fortement de leur capacité à générer de la variabilité génétique (Acosta-Leal *et al.*, 2011 ; Sanjuán & Domingo-Calap, 2016). Il existe trois moteurs moléculaires générant de la diversité génétique à savoir la mutation, la recombinaison et le réassortiment ainsi que trois forces évolutives modulant cette variabilité génétique avec la dérive génétique, la sélection et la migration (Escriu, 2017 ; Moya *et al.*, 2004).

2.3.1. Les moteurs moléculaires générant de la variabilité génétique

2.3.1.1. La mutation

La mutation résulte de la **copie imparfaite** du matériel génomique du parent vers la génération suivante ou suite à des phénomènes physiques ou chimiques (les rayonnements UV par exemple) modifiant les bases nucléotidiques (Acosta-Leal *et al.*, 2011). Les mutations existent sous trois formes : l'**insertion** et la **délétion** correspondant respectivement à l'introduction ou la perte d'un ou plusieurs nucléotides, ainsi que la **substitution** rendant compte d'un remplacement d'un nucléotide par un autre. La mutation peut être **bénéfique** (apport d'un avantage sélectif),

neutre (sans effet) ou **délétère** (apport d'un désavantage). Du fait de la grande compacité des génomes viraux et de la forte épistasie (*i.e.* effet d'une mutation dans un gène sur l'expression d'un autre gène au sein du même génome ; Escriu, 2017), il est considéré que la majorité des mutations sont délétères voire létales (Domingo-Calap *et al.*, 2009 ; Sanjuán *et al.*, 2004). Cependant, certaines mutations délétères chez un hôte peuvent être bénéfiques chez un autre (Siobain Duffy *et al.*, 2006). Ce phénomène est appelé la **pléiotropie antagoniste** (Whitlock, 1996).

Le **taux de substitution** d'un virus représente le nombre de mutations par position nucléotidique et par unité de temps et illustre la vitesse à laquelle un virus évolue. Le taux de substitution est étroitement lié au **taux de mutation** qui correspond à la probabilité qu'une modification de l'information génétique soit transmise à la génération suivante (Sanjuán & Domingo-Calap, 2016). Le taux de mutation varie grandement parmi les virus et est largement lié au **taux d'erreur** de la polymérase mais peut être également modulé selon les mécanismes de réplication, la nature, la taille et la structure du génome, le microenvironnement cellulaire ou encore l'accès à la réparation post-réplicative par exemple (Sanjuan *et al.*, 2010 ; Sanjuán & Domingo-Calap, 2016).

Initialement, on a distingué les virus à « **évolution lente** » tels que les virus à ADN (en particulier ceux à ADNdb) et les virus à « **évolution rapide** », généralement les virus à ARN. Les taux de mutation de ces deux groupes de virus s'étendent de 10^{-3} à 10^{-5} erreurs par nucléotide et par réplication pour les virus à ARN contre 10^{-6} à 10^{-8} pour les virus à ADN (Garcia-Diaz & Bebenek, 2007 ; Jenkins *et al.*, 2002). Cependant, les estimations des taux de substitution ont révélé des espèces virales à ADN simple brin pouvant évoluer aussi rapidement que les virus à ARN. C'est le cas par exemple, des géminivirus et des parvovirus qui présentent des taux de substitution compris entre 1 et 3 x 10^{-4} substitution par site et par an (**Figure 3** ; Duffy & Holmes, 2008 ; Harkins *et al.*, 2009). L'ensemble de ces estimations semblent remettre en question le paradigme sur la vitesse d'évolution rapide des virus à ARN et lente des virus à ADN. De plus, il a été constaté que plus la période d'observation était longue, plus le taux d'évolution diminuait (Aiewsakun & Katzourakis, 2016 ; Simmonds *et al.*, 2019).

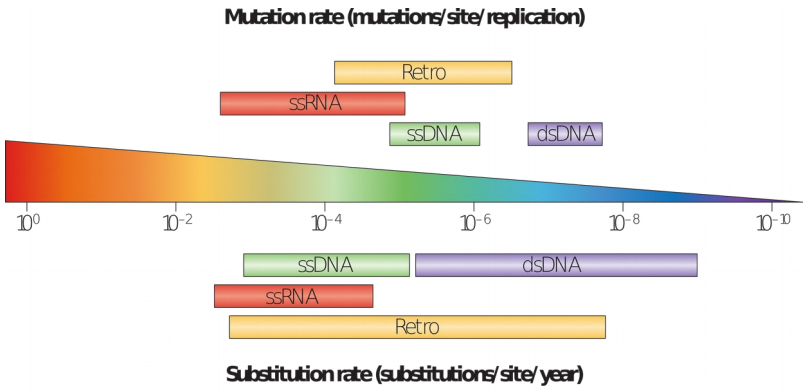


Figure 3. Comparaison entre les taux de substitution et de mutation chez les différents groupes de virus (Duffy *et al.*, 2008).

Par exemple, l'analyse d'ARN anciens du barley stripe mosaic virus (BSMV) a révélé des taux de substitutions plus bas (soit un taux moyen de 3,9 x 10^{-5} substitution/site/an comparé aux analyses de séquences contemporaines de BSMV (soit un taux de substitution moyen de 7,3 x 10^{-4} substitution/site/an ; Smith *et al.*, 2014). De telles différences d'estimation entre les mesures du taux d'évolution à court et long terme seraient probablement dues au fait que dans les mesures à court terme de nombreuses mutations délétères ou transitoirement bénéfiques sont prises en compte alors qu'elles seront purgées à long terme (Lythgoe *et al.*, 2017).

2.3.1.2. La recombinaison

La recombinaison est un processus évolutif majeur qui rend compte de la formation de **molécules nucléotidiques chimères** à partir d'échanges de matériel génétique issu de génomes parentaux (**Figure 4** ; Martin *et al.*, 2011 ; Owor *et al.*, 2007). La recombinaison associée à la mutation donne accès à un **polymorphisme** beaucoup plus important que par mutation seule (Crameri *et al.*, 1998 ; Stemmer, 1994). La recombinaison a déjà été observée entre virus de même espèce (Bousalem *et al.*, 2000 ; Zhou *et al.*, 1997), d'espèces différentes (Monci *et al.*, 2002 ; Vigne *et al.*, 2008), de genres différents (Hernández-Zepeda *et al.*, 2013 ; Moonan *et al.*, 2000) et même entre familles différentes (Saunders & Stanley, 1999). Il existe plusieurs types de recombinaison tels que (i) la **recombinaison homologue**, durant laquelle une portion d'un génome est remplacée par une portion homologue d'un autre génome et (ii) la **recombinaison non-homologue** qui implique le

réarrangement génomique par duplication, insertion ou délétion. Différents mécanismes sont associés à la création de molécules nucléotidiques chimères tels que par exemple un ré-attachement des complexes de réplication et du brin en cours de synthèse, à des matrices distinctes suite à un détachement prématuré des complexes de réplication. Chez les virus de la famille des *Geminiviridae*, les conflits entre les complexes enzymatiques de transcription et de réplication ou des cassures monocaténaires dans le brin matrice pourraient être impliqués (Owor *et al.*, 2007). La recombinaison peut alors se produire selon le mécanisme connu sous le nom de « **réplication dépendante de la recombinaison** » (RDR) qui correspond à l'un des deux mécanismes de la réplication chez les géminivirus ou encore le bactériophage T4 (Jeske *et al.*, 2001; Kreuzer *et al.*, 1995). La recombinaison permet à un virus d'acquérir une grande variabilité génétique, créant de nouveaux arrangements au sein du génome qui sont à l'origine de **nouvelles souches** (cas de la souche Ougandaise de l'East African cassava mosaic virus impliquée dans la forme sévère de la maladie de la mosaïque du manioc en Afrique ; Zhou *et al.*, 1997), de **nouvelles espèces** (cas du African cassava mosaic Burkina Faso virus issu de la recombinaison entre un ancêtre des isolats de l'African cassava mosaic virus, du tomato leaf curl Cameroun virus et du cotton leaf curl Gezira virus ; Tiendrébéogo *et al.*, 2012) ou encore de **nouveaux genres**, (cas des becurtovirus qui sont issus de la recombinaison entre bégomovirus et curtovirus ; Hernández-Zepeda *et al.*, 2013).

2.3.1.3. Le réassortiment

Les virus segmentés et les virus multipartites ont également accès à un mode d'évolution unique, une forme de recombinaison génétique appelée réassortiment des composants génomiques (qualifiée également de « **pseudo recombinaison** »). Le réassortiment, durant lequel des composants du génome sont échangés entre souches ou espèces, implique soit la **co-encapsidation** (virus segmentés), soit la **co-transmission** (virus multipartites) de nouvelles combinaisons de composants génomiques dérivés d'au moins deux virus parentaux (**Figure 4** ; Varsani *et al.*, 2018). Le réassortiment des composants du génome est courant dans divers phytovirus à ARN (Gu *et al.*, 2007 ; Qiu & Moyer, 1999 ; White *et al.*, 1995) et à ADN (Hu

et al., 2007 ; Pita *et al.*, 2001). Par ailleurs, les réassortiments sont possibles entre deux **espèces virales différentes** (cas de l'ADN-A du bean dwarf mosaic virus et de l'ADN-B du tomato mottle virus ; Hou *et al.*, 1998 ou encore de l'ARN 3 du cucumber mosaic virus et les ARN 1 et 2 du peanut stunt virus ; White *et al.*, 1995) ou **deux souches différentes** (cas de l'ADN-A de la souche UG2 et l'ADN-B de la souche UG3 de l'EACMV ; Pita *et al.*, 2001 ou encore le cas de réassortiment entre les souches du Pacifique et de l'Asie du banana bunchy top virus ; Hu *et al.*, 2007).

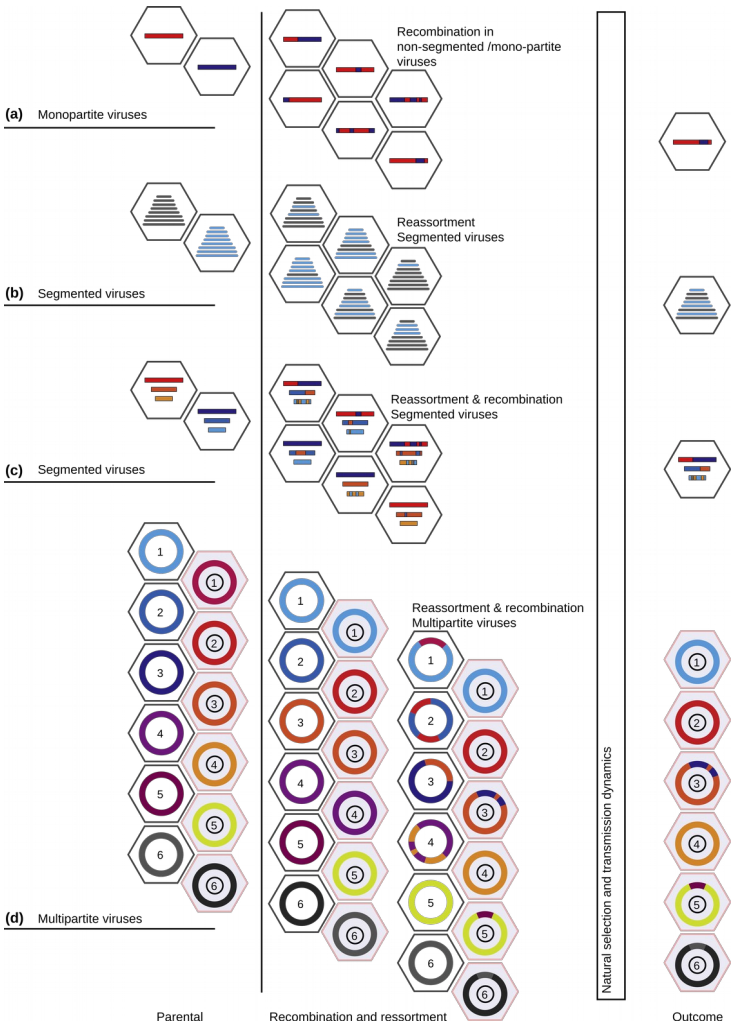


Figure 4. Représentation (a) de la recombinaison entre virus monopartites, (b) du réassortiment entre virus segmentés, et de la recombinaison et du réassortiment entre (c) virus segmentés et (d) virus multipartites (Varsani *et al.*, 2018).

De plus, le réassortiment a été associé au contournement de résistance (Qiu & Moyer, 1999), à la capacité d'adaptation à l'hôte (Hou *et al.*, 1998) ou encore à une virulence accrue (Gu *et al.*, 2007 ; Pita *et al.*, 2001). Contrairement à la recombinaison et au réassortiment entre virus segmentés, le réassortiment entre virus multipartites peut survenir par un brassage des composants lors de la transmission et ne nécessite pas que les virus parentaux co-infectent la même cellule (Varsani *et al.*, 2018).

2.3.2. Les forces évolutives modulant la variabilité génétique

2.3.2.1. La dérive génétique et la sélection

Les nouveaux variants génétiques (génotypes viraux) ainsi générés par la mutation, la recombinaison et/ou le réassortiment, sont soumis à deux forces majeures : la **dérive génétique** et la **sélection**. Ces deux **forces évolutives** déterminent la fréquence de distribution des variants génétiques au sein de la population virale. En effet, toutes les modifications au sein du génome ne perdurent pas systématiquement à la génération suivante et seules certaines d'entre elles perdureront au sein de la population virale. Afin de comprendre ces deux forces évolutives, il convient dans un premier temps de définir le concept de *fitness* également appelé valeur adaptative ou sélective. La ***fitness*** décrit la capacité de reproduction d'un individu ou d'un variant génétique avec lequel elle contribue à la génération suivante dans un environnement donné (Escriu, 2017).

La dérive génétique est un processus induisant des **fluctuations aléatoires** dans les fréquences des variants au sein d'une population d'une génération à l'autre (**Figure 5** ; Charlesworth, 2009). Ce processus évolutif agit indifféremment sur tous les variants d'une population et ne dépend pas de la *fitness* des individus. L'intensité de la dérive génétique dépend de l'intensité des **goulots d'étranglement** (*i.e.* une réduction de la taille de la population ; **Figures 5b et 5c**). Dans des cas extrêmes, une très forte dérive génétique peut conduire à une perte de certains variants à la génération suivante voire même à l'extinction d'une population (Zwart & Elena, 2015).

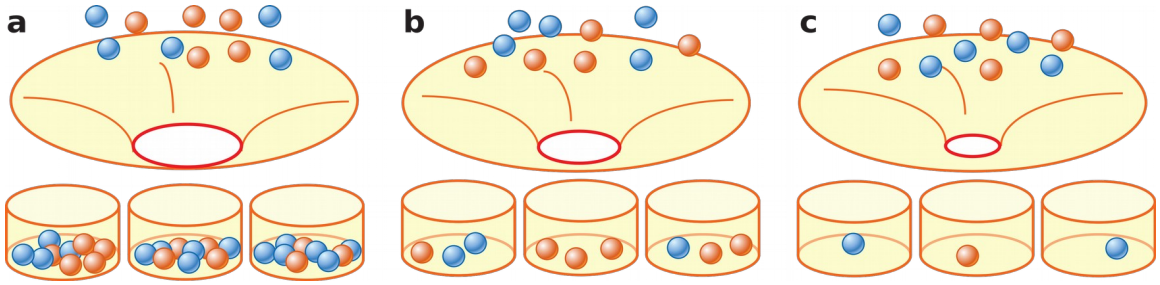


Figure 5. Représentation de la dérive génétique. Les billes de couleurs correspondent aux variants viraux et les boites représentent les individus infectés. Plus le goulot d'étranglement est important (b et c), moins le nombre de variants viraux est important et plus la variation de la fréquence des variants viraux au sein des individus infectés devient grande (Zwart *et al.*, 2015).

Chez les phytovirus par exemple, la dérive génétique se produit au cours de différentes étapes de l'infection virale, telles que (i) la **colonisation** des organes de la plantes (French & Stenger, 2003 ; H. Li & Roossinck, 2004 ; Sacristan *et al.*, 2003), (ii) la **transmission** par des vecteurs (Ali *et al.*, 2006 ; Betancourt *et al.*, 2008) ou par contact entre feuilles non infectées et infectées (Sacristan *et al.*, 2011) ainsi que (iii) durant les **interactions** entre virus co-infectants la même plante (Fraile *et al.*, 1997 ; Gutierrez & Gutierrez, 1999). Par exemple, les estimations actuelles des virus transmis selon le mode circulant non propagatif (cas du tomato yellow leaf curl Israel et mild transmis par *Bemisia tabaci* ; Péréfarres *et al.*, 2014) et selon le mode non circulant (cas du potyvirus Y et du cucumber mosaic virus transmis par puceron (Betancourt *et al.*, 2008 ; Moury *et al.*, 2004) ont suggéré que seules une à deux particules virales participent à la génération suivante. Ces travaux rendent compte de l'importance du goulot d'étranglement et de la dérive génétique à chaque évènement de transmission horizontale par insecte vecteur.

Les effets de la dérive génétique et de la sélection sont souvent difficiles à dissocier, car la sélection entraîne également une diminution de la diversité de la population et peut augmenter la diversité entre les populations si celles-ci sont soumises à des pressions de sélection différentes (García-Arenal *et al.*, 2003). Contrairement à la dérive génétique, la sélection n'agit pas indifféremment sur tous les variants d'une population mais favorise les **variants viraux les mieux adaptés** (avec une plus grande *fitness*). Cette force est généralement quantifiée avec le **coefficient de sélection**, défini

comme la différence entre deux variants et qui peut être associé à l'analyse des mutations. Une des mesures possibles est l'évaluation du ratio **dN/dS**, où dN correspond au taux de mutations **non-synonymes** et dS le taux de mutations **synonymes** (Escriu, 2017).

2.3.2.2. La migration

La migration est une force évolutive permettant des flux de gènes au sein d'une métapopulation. La métapopulation est un concept écologique définit par Levins en 1969 correspondant à un ensemble de populations d'individus d'une même espèce séparées spatialement ou temporellement et étant interconnectées par les migrations d'individus. La migration des phytovirus est très variable selon leur mode de transmission ou encore l'intervention humaine dans leur dispersion (e.g. transport de matériel contaminé).

3. L'écologie virale

La prise en compte de l'écologie pour l'étude des phytovirus est récente et a été stimulé par (i) la recrudescence des études sur le compartiment sauvage révélant une large diversité virale ainsi que (ii) par la nécessité de comprendre l'influence de la diversité et de la structure des systèmes de cultures sur la propagation des phytovirus (Malmstrom *et al.*, 2011 ; Power, 2008 ; Roossinck, 2013).

L'écologie virale des plantes se concentre principalement sur l'étude de ce qu'on appelle la «trinité écologique indissociable» couramment représenté par le triangle de la maladie (**Figure 6**). Ce concept repose sur les différents acteurs du cycle d'infection virale à savoir les virus, les vecteurs et les hôtes, et leurs interactions (Malmstrom *et al.*, 2011). Il illustre l'importance de chacun des acteurs dans la résultante de leurs interactions et in fine dans l'expression potentielle de maladies. L'écologie virale cherchera à examiner (i) les rôles écologiques des phytovirus et de leurs vecteurs au sein des écosystèmes globaux (cultivés et non cultivés) et (ii) l'influence réciproque des caractéristiques de l'écosystème sur la distribution et l'évolution des phytovirus (Malmstrom *et al.*, 2011).

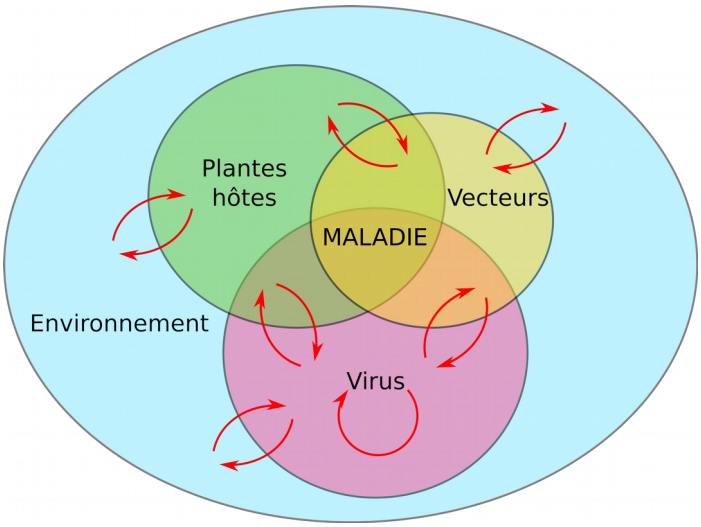


Figure 6. Représentation schématique du triangle de la maladie. Les cercles représentent chaque acteur au sein de l'environnement et les flèches rouges correspondent aux interactions (d'après Islam *et al.*, 2017).

3.1. Les interactions à l'échelle des organismes

3.1.1. Les interactions virus-vecteur

«Etre transmis ou disparaître» tel est le dilemme pour tous les virus. En effet, les virus étant des **parasites obligatoires stricts**, la transmission à de nouveaux hôtes est l'un des processus les plus importants en matière d'écologie virale (Cooper & Jones, 2006 ; Malmstrom *et al.*, 2011).
On distingue deux modes de transmission : la **transmission verticale** et la **transmission horizontale** (Astier *et al.*, 2007).

La transmission verticale correspond à la transmission du parent à la descendance (voie interne) alors que la transmission horizontale implique la transmission à de nouveaux individus (voie externe) qui fait généralement intervenir un troisième partenaire, **le vecteur**. Ces modes de transmission affectent l'évolution des relations plantes – virus – vecteurs. L'importance relative de ces modes de transmissions reste néanmoins inconnue pour la grande majorité des phytovirus.

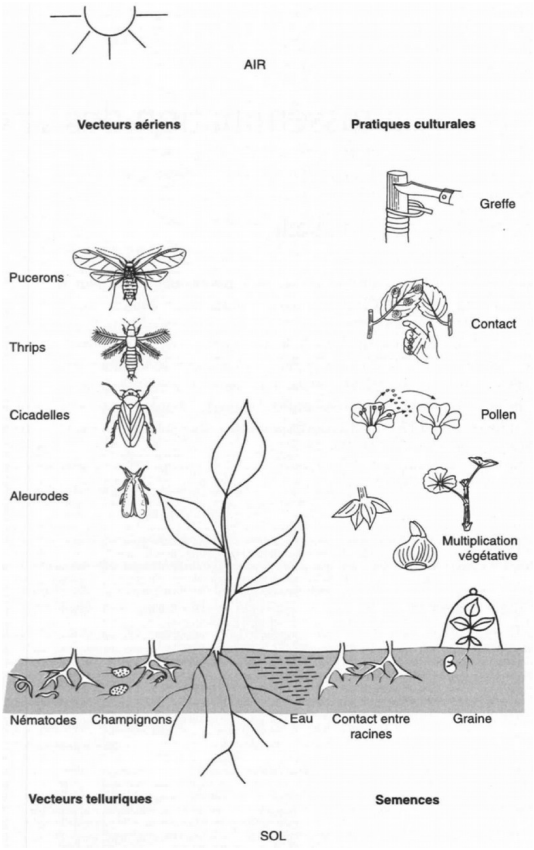


Figure 7. Représentation des différents modes de transmission des phytovirus (Astier, 2007).

3.1.1.1. La transmission verticale

Dans le cas des phytovirus, la transmission verticale regroupe la transmission par la **graine** (cas du bean common mosaic virus ; Morales, 1987), par le **pollen** (également impliqué dans la transmission horizontale dans certains cas ; Mink, 1993) et par **multiplication végétative** (boutures, tubercules, bulbes) (**Figure 7**). Il est important de noter que dans le cas de la transmission verticale, il est essentiel que les virus ne présentent pas de caractère antagoniste pour la plante, ce qui *in fine* nuirait à leurs propres transmissions (Hamelin *et al.*, 2017). Ce serait en particulier le cas pour les virus dits persistants et asymptomatiques (Roossinck *et al.*, 2010), qui ne sont pas associés à des symptômes de maladie chez la plante (virus cryptiques), sont transmis verticalement et peuvent induire des effets positifs sur celle-ci (Roossinck, 2010). Là encore, le même biais de connaissance lié à l'étude des virus responsables de maladie existe et ces virus cryptiques restent encore largement méconnus.

3.1.1.2. La transmission horizontale

Parmi les nombreux modes de transmission horizontale, les transmissions par **vecteurs** et par **contact direct** ont été les plus largement étudiées (**Figure 7**). Certaines voies de transmission horizontale comme par exemple par le pollen via les abeilles ou le nectar ont été négligées et leur impact semble sous-estimé (Jones, 2018).

La transmission virale **par contact direct** inclut par exemple la transmission par contact au niveau du feuillage (Sacristan *et al.*, 2011), par les greffes naturelles entre racines (Hunter *et al.*, 1958), par les eaux de ruissellement ou sols contaminés (Li *et al.*, 2016 ; Mehle *et al.*, 2018). Toutefois, les plantes étant par nature immobile, la transmission horizontale des phytovirus se fait surtout par l'intermédiaire de vecteurs. Il existe différents vecteurs de type **téllurique** (nématodes et champignons) ou **aérien** (pucerons, thrips, aleurodes, cicadelles, acariens, etc.). Ces différents vecteurs sont capables de transmettre des phytovirus sur des distances plus ou moins longues allant de quelques centimètres à plusieurs kilomètres (Bragard *et al.*, 2013). Parmi ces vecteurs, ce sont les insectes piqueurs suceurs qui ont été les plus étudiés.

La transmission par insecte vecteur est généralement classée selon quatre catégories en fonction (i) du temps nécessaire à l'insecte pour se nourrir avant de contracter le virus, (ii) de sa durée de vie virulifère et infectieux, (iii) de la durée pendant laquelle il doit s'alimenter pour transmettre le virus et (iv) de la durée durant laquelle le virus est présent chez le vecteur (Bragard *et al.*, 2013).

La transmission par les insectes a lieu selon deux modes distincts : la transmission **non circulante** et la transmission **circulante**. La transmission non circulante se caractérise par la localisation du virus au niveau des pièces buccales du vecteur. Cette localisation implique une interaction spécifique et reversible des particules virales avec les composants moléculaires des pièces buccales (Uzest *et al.*, 2007). Les périodes d'acquisition et d'inoculation sont généralement courtes allant de quelques secondes à quelques minutes pour les virus dits à transmission **non persistante** et de quelques heures à

quelques jours pour les virus à transmission **semi-persistante** (Bragard *et al.*, 2013). La transmission dite circulante se caractérise par la diffusion du virus depuis le système digestif jusqu'aux glandes salivaires. Certains virus circulants dits **propagatifs** ou **multipliants** se répliquent également au sein de leur vecteur (Bragard *et al.*, 2013).

3.1.2. Les interactions virus-plante

L'infection par un virus peut avoir un effet néfaste (qualitatif et quantitatif) sur la *fitness* de la plante hôte. L'effet qualitatif sur l'hôte est exprimé par la **pathogénicité** du virus (capacité d'un agent pathogène à provoquer une maladie chez un hôte particulier) et l'effet quantitatif d'un virus est défini par la **virulence** (degré de dommage causé à l'hôte ; Aurora Fraile & García-Arenal, 2010 ; Woolhouse *et al.*, 2002). Ainsi, plus un virus est virulent, plus les effets nocifs sur son hôte seront importants. En conséquence, les plantes ont développé des stratégies de défenses efficaces à la fois pour éviter ou limiter les infections et pour réduire son coût (Agnew *et al.*, 2000). L'hôte à travers l'utilisation de ces stratégies affecte à son tour la *fitness* du virus. Il existe deux mécanismes de défense chez les plantes à savoir (i) la **résistance** (capacité de l'hôte à limiter la multiplication et la propagation virale) et (ii) la **tolérance** (capacité de l'hôte à diminuer les dommages causés par l'infection virale). De telles relations entre les virus et les plantes rendent compte d'une **co-évolution ancienne** entre ces derniers (Fraile & García-Arenal, 2010).

3.1.2.1. Le continuum symbiotique des phytovirus

En écologie virale, les virus sont considérés comme des symbiotes. La symbiose est définie comme deux entités dissemblables vivant l'une avec l'autre dans une relation intime. Ces symbiotes s'inscrivent dans un continuum d'interactions entre **mutualisme** (où les deux entités bénéficient de la relation) et **antagonisme** (où l'une des entités bénéficient de la relation aux dépens de l'autre ; **Figure 8** ; Roossinck *et al.*, 2015).

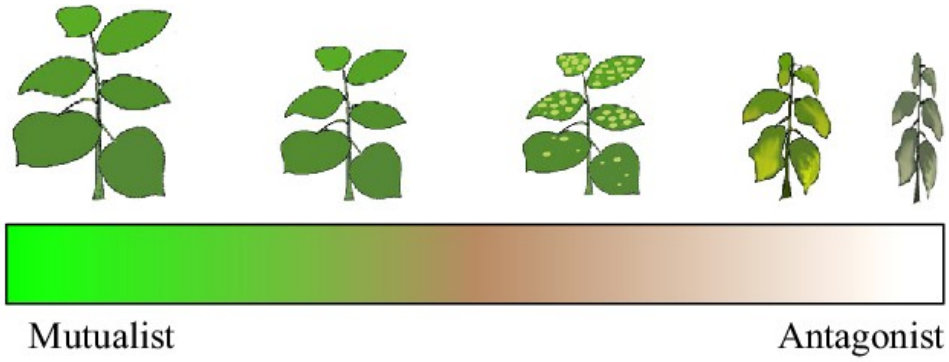


Figure 8. Représentation du continuum symbiotique des phytovirus. L'infection virale peut être bénéfique (gauche de la frise) ou néfaste pour la plante infectée (droite de la frise ; Roossinck, 2015).

Les virus peuvent évoluer d'un mode de vie à l'autre en fonction des conditions environnementales (**Figure 8**). Par exemple, si un virus peut être un agent pathogène dans des conditions normales, il peut être bénéfique en cas de stress (Bao & Roossinck, 2013). Parmi les virus mutalistes, certains permettent par exemple de s'adapter aux changements environnementaux extrêmes (cas du tobacco mosaic virus et cucumber mosaic virus induisant une résistance à la sécheresse et/ou au froid ; Xu *et al.*, 2008) alors que d'autres sont associés à leurs hôtes depuis si longtemps que la ligne de démarcation entre hôte et virus s'est estompée (relation entre les polydnavirus et les guêpes parasitoides ; Webb *et al.*, 2006).

Bien que la plupart des études en virologie s'est concentrée sur les virus antagonistes (pathogènes) des cultures, des analyses plus larges ont révélé des interactions plus complexes et stables, en particulier dans le cas des communautés de plantes sauvages. En effet, la plupart des virus infectant les plantes sauvages n'induisent pas de symptômes apparents (Prendeville *et al.*, 2012 ; Roossinck, 2011, 2012 ; Stobbe & Roossinck, 2014). Bien qu'il existe des exemples de virus provoquant des maladies chez les plantes sauvages, affectant la taille de la population de plantes ou la composition des plantes d'un écosytème par exemple (Malmstrom *et al.*, 2005 ; Power *et al.*, 2011 ; Prendeville *et al.*, 2012 ; Rodelo-Urrego *et al.*, 2013 ; Rúa *et al.*, 2011), il a été proposé que les virus des plantes sauvages sont le plus souvent des commensaux (*i.e.* flore virale non pathogène), voire des mutualistes, et que l'agriculture aurait favorisée l'émergence de virus pathogènes responsables de maladies (Roossinck, 2011 ; Wren *et al.*, 2006 ; Xu *et al.*, 2008).

3.1.2.2. Les stratégies d'adaptation aux hôtes

Face à la multitude de plantes hôtes présents dans un environnement, les phytovirus ont la possibilité d'adopter deux stratégies d'adaptation distinctes : soit infecter et se répliquer de façon optimale dans une ou quelques espèces de plantes apparentées (cas des phytovirus dits « **spécialistes** »), soit infecter une plus large gamme d'espèces de plantes (cas des phytovirus dits « **généralistes** »). En se spécialisant à un seul ou quelques hôtes, les virus peuvent réduire les risques de concurrence au prix d'un accès à l'ensemble des ressources disponibles (stratégie « *Master of some* » avec une adaptation optimale permettant d'exploiter au mieux la ressource ; Futuyma & Moreno, 1988). En revanche, les virus généralistes auront l'avantage de pouvoir exploiter une plus large gamme d'espèces de plantes, avec une *fitness* plus élevée à l'échelle de l'ensemble des hôtes mais qui présente un coût du fait d'un niveau d'adaptation moindre à chaque espèce (« *Jack of all trades, master of none* » ; Elena *et al.*, 2009).

Les modèles théoriques soutiennent l'existence d'un **compromis adaptatif** (*adaptative trade-off*) entre les modes de vie généraliste et spécialiste. Il repose notamment sur le fait qu'un phytovirus ne peut pas maximiser sa *fitness* chez toutes les plantes disponibles au sein d'un environnement. Ainsi, l'adaptation à un nouvel hôte se traduit par une augmentation de la *fitness* vis à vis de cet hôte mais implique une réduction de la *fitness* vis à vis de l'hôte d'origine (Elena *et al.*, 2014 ; McLeish *et al.*, 2018). Même si les virus généralistes ont plus d'opportunité de transmission et de survie, du fait de leur gamme d'hôtes plus large, les concepts théoriques autour de la sélection naturelle suggèrent que la spécialisation serait favorisée (Woolhouse *et al.*, 2001).

Néanmoins, le généralisme reste une stratégie très commune parmi les phytovirus. Le maintien d'une large gamme d'hôte dans les environnements naturels plutôt que la spécialisation vers un ou quelques hôtes s'expliquerait en partie par l'hétérogénéité de la densité d'hôtes et de vecteurs dans l'espace et le temps. En effet, un environnement constant favoriserait généralement la spécialisation alors qu'un environnement mixte et variable

privilégierait plutôt le généralisme (Elena & Lenski, 2003). De manière intéressante, sous certaines conditions, des généralistes seraient ainsi capables d'avoir une *fitness* égale voire supérieure aux spécialistes (Remold, 2012). Ces généralistes dits « sans-coûts » ont souvent été observés chez les virus d'animaux, mais un seul cas a pour l'instant été décrit chez les virus de plantes avec le tobacco etch virus (Bedhomme *et al.*, 2012).

3.1.2.3. Les phénomènes de co-infection

Lorsqu'une plante est simultanément infectée par plusieurs virus, on parle alors de co-infection ou d'infection mixte. Les co-infections peuvent avoir un impact sur la *fitness* des virus présents, entraînant une variété d'interactions allant de la **synergie** à des **réactions antagonistes** au sein de l'hôte (Power, 2008). Une interaction synergique a un effet facilitant sur au moins un des partenaires viraux. Parmi les interactions facilitantes, on peut citer par exemple l'augmentation du taux de transmission (cas de la souche *Mild* du tomato yellow leaf curl virus (souche à virulence réduite) qui est mieux transmise en présence de la souche IL de ce même virus (souche dite sévère ; Péréfarres *et al.*, 2014) ou encore une amélioration dans la capacité à infecter certains types cellulaires (Sánchez-Navarro *et al.*, 2006). L'effet de la co-infection sur les symptômes est variable. Il peut entrainer leurs exacerbations et conduire dans certains cas à des épidémies plus sévères, comme observé avec la forme sévère de la maladie de la mosaïque du manioc dans le cas de la synergie entre virus du même genre (Legg & Fauquet, 2004) ou la maladie de la nécrose mortelle du maïs (*maize lethal necrosis disease*) dans le cas de la combinaison entre des virus appartenant à des familles différentes (Redinbaugh & Stewart, 2018).

Le co-infection peut également aboutir à une atténuation des symptômes. Dans ce cas, les interactions entre virus sont dites antagonistes et peuvent impliquer des phénomènes appelés la **protection croisée** (appelée également « exclusion de surinfection » ou « interférence homologue ») et l'**exclusion mutuelle** (également connue sous le terme de « suppression mutuelle» ou «suppression concurrentielle mutuelle » ; Syller, 2012). La protection croisée survient lorsqu'une infection antérieure par un virus

(qualifié de protecteur) interfère voir empêche l'infection ultérieure par un virus homologue (cas de souches peu virulentes de citrus tristeza virus qui confère une protection contre la forme sévère de CTV ; Folimonova *et al.*, 2010). Les cas d'exclusion mutuelle, moins caractérisés et étudiés chez les plantes que chez l'humain, peuvent être observés au sein de la plante hôte, se traduisant par la ségrégation spatiale des différents virus (Dietrich & Maiss, 2003). À ce jour, on ignore quelle est l'importance de ces interactions virus-virus dans les écosystèmes naturels et dans quelle mesure elles pourraient influencer les épidémies virales (Jeger *et al.*, 2006).

3.1.3. Les interactions virus-vecteur-plante

3.1.3.1. L'influence des phytovirus sur les interactions vecteur-hôte

La longue co-évolution entre virus et vecteur a également abouti à la mise en place de mécanismes regroupés sous le terme de « manipulation de la transmission » dans lesquels des modifications de la physiologie de la plante ou du vecteur aboutissent à l'augmentation de l'efficacité de transmission. Ces manipulations de la fréquence et de la nature des interactions vecteur-hôte par les phytovirus pour favoriser leur dissémination se produisent (i) de façon **indirecte** avec par exemple la modification de l'attractivité de la plante et (ii) de facon **directe** due à la présence du phytovirus au sein du vecteur (Dáder *et al.*, 2017).

La modification des signaux de l'hôte par les phytovirus inclue par exemple la modification des caractéristiques visuelles et olfactives (Blanc & Michalakis, 2016 ; Fereres & Moreno, 2009) ou encore de la composition en acides aminés, en sucres ou en métabolites (Gadhave *et al.*, 2019 ; Mauck *et al.*, 2014). En plus du renforcement de l'attraction initiale du vecteur pour les plantes infectées, les phytovirus peuvent également influencer le temps d'alimentation des vecteurs via par exemple la modification nutritionnelle des plantes en fonction de leur stratégie de transmission (non circulant ou circulant). Ainsi, les virus non circulants tendraient à diminuer la qualité nutritionnelle des plantes hôtes de manière à encourager les vecteurs à se

nourrir durant des périodes courtes et inversement pour les virus circulants, optimisant ainsi l'acquisition virale (Dáder *et al.*, 2017 ; Mauck *et al.*, 2012).

Par ailleurs, les phytovirus peuvent également induire des modifications des défenses induites par les plantes telles que par exemple, le dépôt de callose, les modifications du phloème, la libération de l'acide jasmonique ou la synthèse des composés volatils, rendant les plantes peu attrayantes ou inaccessibles aux vecteurs (Casteel *et al.*, 2014 ; Mauck *et al.*, 2014). De telles modifications influencent non seulement l'orientation, les comportements alimentaires et la dispersion des vecteurs (Ingwell *et al.*, 2012 ; Mauck *et al.*, 2014), mais dans certains cas améliorent la croissance, la fécondité, la survie ou la longévité de ces derniers, et donc leur *fitness* favorisant ainsi la transmission phytovirale (Casteel *et al.*, 2014 ; Fereres & Moreno, 2009 ; Mauck *et al.*, 2012).
Ces preuves de manipulation des phytovirus vis à vis de leurs vecteurs afin de promouvoir une dissémination efficace mettent en évidence des interactions virus-vecteurs-hôtes complexes rendant compte d'une **co-évolution** des phytovirus avec certains vecteurs et certaines espèces hôtes (Mauck *et al.*, 2012).

3.1.3.2. L'influence des vecteurs sur la gamme d'hôte des phytovirus

Les virus des plantes ont généralement un nombre limité de vecteurs efficaces, dont les caractéristiques intrinsèques (dynamique de population, préférences alimentaires...) auront une forte influence sur l'écologie du virus (Elena *et al.*, 2014). En effet, la gamme d'hôte d'un virus transmis par un vecteur est limitée par la gamme d'hôte préférentielle de ce même vecteur (Power, 2008). De ce fait, toute expansion de la gamme d'hôte du vecteur pourrait entraîner un élargissement de la gamme d'hôte des virus transmis par ce vecteur (Goldbach & Peters, 1994 ; Harrison & Robinson, 1999).

En outre, il a été démontré que parmi les virus transmis par vecteurs, 58,4 % le sont par une seule espèce de vecteur alors que seul 9,9 % des virus sont connus comme infectant une seule espèce de plante (Power & Flecker,2003)

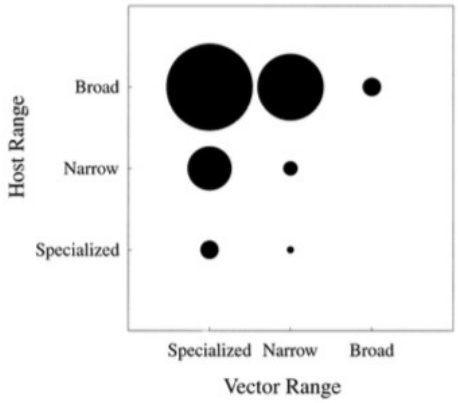


Figure 9. Relation entre la gamme de plantes hôtes et la gamme de vecteurs des phytovirus transmis par vecteurs (Power & Flecker, 2003).

En revanche, aucun virus ne possède une gamme étroite de plantes hôtes s'il possède une large gamme d'espèces vectrices (**Figure 9**). Ces schémas de spécificité de vecteur et d'hôte suggèrent que la distribution du virus est d'avantage limitée par la spécificité des relations virus-vecteur que par la spécificité des relations plante-virus. Cela pourrait s'expliquer par des pressions de sélection plus importante imposée par la nécessité d'avoir des vecteurs efficaces que celle imposée par les défenses de la plante hôte.

3.2. La dynamique des phytovirus à l'échelle du paysage

3.2.1. Le rôle de la biodiversité des plantes

Etant donné que la dynamique de maladie dépend entre autre des interactions entre l'agent pathogène et les plantes hôtes au sein d'un écosystème, l'hypothèse que la **biodiversité** et les changements de sa composition auraient un impact sur la **dynamique** des agents pathogènes notamment sur l'apparition, l'incidence et la sévérité des maladies a été avancée (**Figure 10** ; Keesing *et al.*, 2010).

Deux hypothèses majeures relient la biodiversité au risque de maladie à savoir l'effet d'amplification et l'effet dilution. L'**effet d'amplification** prédit que la diversité en espèces de plantes sera positivement corrélée au risque de maladie (Keesing *et al.*, 2006). En effet, la diversité entraînera une augmentation de l'abondance des hôtes potentiels pour l'agent pathogène. A contrario, l'**effet dilution** prédit une corrélation négative entre la biodiversité et le risque de maladie. Une réduction de la diversité en hôtes pourrait entraîner une augmentation de l'abondance de l'hôte principal (hôte focale) colonisant les habitats disponibles (Keesing *et al.*, 2006).

Figure 10. Schématisation des relations entre le risque de maladies, les activités humaines et le taux de biodiversité (Roossinck & Garcia-Arenal, 2015).

La réduction de la biodiversité dans les agro-écosystèmes, en termes de richesse spécifique végétale peut notamment être associée aux activités humaines (e.g. l'introduction de plantes invasives hôtes ou encore l'intensification et l'extensification de l'agriculture). Une étude sur la prévalence de virus sur un piment sauvage du Mexique a ainsi mis en évidence que le risque d'infection virale augmente avec le niveau d'anthropisation, se traduisant par une diminution de la diversité en espèces végétales et de la diversité génétique de l'hôte. Par ailleurs, les effets de la diversité sur le risque de maladie seraient liés à la gamme d'hôte de l'agent pathogène. Ainsi, un effet d'amplification nécessiterait un agent pathogène plutôt généraliste, alors que l'effet dilution serait probablement plus observé chez un pathogène spécialiste (Pagán *et al.*, 2012).

3.2.2. L'impact des activités humaines

L'analyse des causes associées aux maladies infectieuses a révélé que l'émergence de maladie virale serait dans 74% des cas dues au moins partiellement aux activités humaines, 16% aux vecteurs, 5% au climat et 5% à la recombinaison (**Figure 11** ; Anderson *et al.*, 2004).

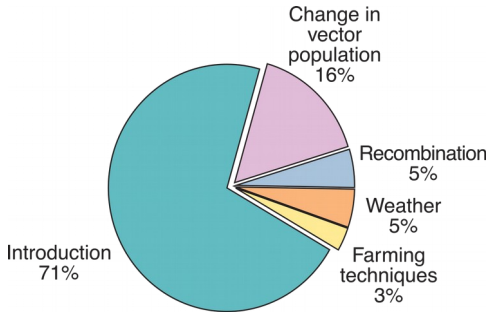


Figure 11. Proportion des maladies émergentes associés à des phytovirus en fonction du facteur jugé le plus important dans cette émergence (Anderson *et al.*, 2004).

Ces activités humaines se résument principalement par l'**introduction** de plantes loin de leur aire d'origine (Anderson *et al.*, 2004 ; Cooper & Jones, 2006 ; Faillace *et al.*, 2017 ; Fargette *et al.*, 2006 ; Jones, 2009). L'introduction de **plantes exotiques** (plante non indigène ou non native libérée intentionnellement ou accidentellement dans une nouvelle aire géographique ; Pyšek *et al.*, 2009) de façon accidentelle ou volontaire par l'homme peut facilement s'accompagner de l'introduction de vecteurs et de phytovirus (nouveaux ou déjà existants) vers de nouvelles zones géographiques (Mitchell & Power, 2003). De plus, ces introductions peuvent participer à la **dissémination** de phytovirus déjà présents dans l'aire géographique (Plus de détails dans la partie 3.2.3. Le cas particulier des plantes invasives et exotiques). Brièvement, les activités humaines créent des conditions favorables à la dissémination de virus introduits vers des **plantes indigènes** (plantes natives à la zone géographique) et/ou de virus indigènes vers des plantes exotiques (Jones, 2009). On peut citer le cas du MSV, décrit comme provoquant une maladie de type « nouveau contact » chez le maïs (Bosque-Pérez, 2000). Le maïs, originaire d'Amérique, a été introduit en Afrique au 16ème siècle. Ce n'est qu'après son introduction que le MSV, probablement originaire de plantes indigènes de la famille des Poaceae, aurait alors émergé sur cette culture (Fargette *et al.*, 2006 ; Martin *et al.*, 2001).

Par ailleurs, les pratiques agricoles humaines tels que le bouturage, le greffage, l'utilisation d'outils contaminés, de fertilisants ou encore l'introduction de matériel contaminé à multiplication végétative participent à la dissémination virale modifiant ainsi la dynamique des phytovirus. C'est le cas du SWSV qui a vu son aire de distribution s'élargir via l'introduction de plants de canne à sucre contaminés (Candresse *et al.*, 2014). Les activités humaines ont eu non seulement des effets directs sur l'écologie virale mais aussi des effets indirects associés par exemple au changement climatique (Anderson *et al.*, 2004).

3.2.3. Le cas particulier des plantes invasives et exotiques

Une **plante exotique** est une plante non indigène (ou non native) libérée intentionnellement ou accidentellement dans une nouvelle aire géographique alors qu'une **plante invasive** est une plante qui par sa prolifération produit des changements significatifs de structure ou de fonctionnement des écosystèmes (Richardson *et al.*, 2000). Une plante invasive peut être soit une plante exotique naturalisée (*i.e.* se reproduisant régulièrement dans sa nouvelle aire géographique et se maintenant à long terme dans celle-ci) soit une plante indigène (Richardson *et al.*, 2000). L'installation dans un nouvel environnement et l'accroissement de la densité des plantes invasives entraînent une augmentation des contacts avec les autres plantes et offrent ainsi de plus grandes opportunités de transmission virale. Par ailleurs, les plantes invasives peuvent héberger des phytovirus exotiques qui vont soient diminuer soient augmenter les effets des invasions biologiques (Faillace *et al.*, 2017 ; Rúa *et al.*, 2011). En effet, les plantes invasives exotiques peuvent laisser dans leurs aires d'origines leurs ennemis naturels (virus pathogènes, prédateurs...) leur permettant d'être plus robustes dans le nouvel environnement (phénomène appelé ***pathogen release*** correspondant à un relâchement de la pression des pathogènes ; Mitchell & Power, 2003). En outre, les plantes invasives (exotiques ou indigènes) peuvent également héberger des phytovirus pouvant être transmis aux plantes indigènes (Rua *et al.*, 2011).

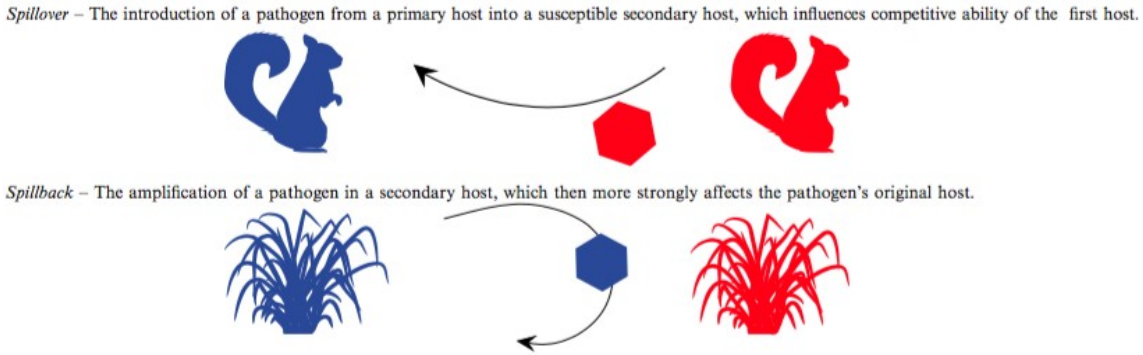


Figure 12. Représentation des phénomènes de *spillover* et de *spillback* (Faillace *et al.*, 2017).

Ce phénomène appelé ***pathogen spillover*** se caractérise par la dissémination d'un pathogène issu d'une plante exotique vers une plante indigène (**Figure 12** ; Faillace *et al.*, 2017 ; Power & Mitchell, 2004). *A contrario*, dans le cas d'un ***pathogen spillback,*** les phytovirus indigènes peuvent se disséminer au sein des plantes exotiques. Celles-ci amplifieraient alors les phytovirus indigènes conduisant à des effets plus intenses sur leurs hôtes indigènes d'origine (**Figure 12** ; Faillace *et al.,* 2017). Des travaux sur des Poaceae annuelles invasives exotiques dans les prairies de Californie, ont mis en évidence ce phénomène de *pathogen spillback*. Ainsi, les Poaceae exotiques tolérantes aux barley yellow dwarf virus (BYDV) et cereal yellow dwarf virus (CYDV) ont servi de réservoir de pathogènes, induisant une augmentation de la prévalence de ces virus dans les communautés de Poaceae indigènes (Malmstrom *et al.*, 2005). Ces pathogènes impactant plus les Poaceae indigènes ont alors permis aux Poaceae exotiques de devenir invasives ne laissant subsister que 1 % des Poaceae indigènes (Borer *et al.*, 2007; Malmstrom *et al.*, 2005). Par ailleurs, les plantes invasives (exotiques et indigènes) peuvent également avoir une influence sur l'écologie des vecteurs. Par exemple, l'augmentation de l'incidence du BYDV et CYDV dans les prairies de Californie semble avoir été médiée par des effets indirects sur les populations de pucerons. En effet, les pucerons préférant les Poaceae exotiques, ont vu leur fécondité s'accroître, accentuant ainsi la transmission du BYDV et CYDV par ces vecteurs (Malmstrom *et al.*, 2005).

3.2.4. L'agro-écosystème : une interface dynamique

Les activités humaines favorisent la promiscuité entre les écosystèmes naturels et cultivés (Alexander *et al.*, 2014). Cette interface entre écosystèmes sauvages et agricoles est appelée « interface agro-écologique » ou « agro-écosystème » et est constituée de **plantes cultivées**, de **plantes adventices** (plantes non cultivées associées aux cultures) et de **plantes non cultivées** qu'elles soient **exotiques** ou **indigènes** (Burdon & Thrall, 2008). Cette interface est au **carrefour** de deux systèmes très différents avec d'un côté les systèmes cultivées au sein desquels on observe une majorité de cultures **annuelles** managées par l'homme, et de l'autre, des écosystèmes sauvages principalement dominés par des plantes **pérennes** avec une **forte compétition** interplante (Alexander *et al.*, 2014).

Les phytovirus font partie intégrante des agro-écosystèmes qu'ils soient inféodés aux plantes sauvages ou aux plantes cultivées (Hogenhout *et al.*, 2008 ; Roossinck *et al.*, 2010). La grande diversité de plantes hôtes disponible au sein des agro-écosystèmes offre de nombreuses opportunités d'interaction entre les virus et les plantes, et d'élargissement de leur gamme d'hôte (Shates *et al.*, 2019). L'aptitude des virus à infecter des plantes sauvages et cultivées favoriserait le **maintien** de virus pathogènes des cultures au sein de l'environnement (Jones, 2009 ; Alexander *et al.*, 2014). Les plantes pérennes auraient un rôle prépondérant dans ce maintien, en raison de leur durée de vie prolongée par rapport à celle des cultures annuelles et de leur probabilité plus grande d'être infectées ou co-infectées (Alexander *et al.*, 2014).

L'interaction entre les virus initialement inféodés aux plantes sauvages et confrontés à de nouveaux hôtes introduits et présents à large échelle (*i.e.* les plantes cultivées) semble avoir été essentielle à l'émergence de nombreuses maladies de plantes (Fargette *et al.*, 2006 ; Jones, 2009). Les facteurs identifiés comme prépondérants dans ces cas d'émergence incluent des facteurs moléculaires comme la mutation, la recombinaison, et la dérive génétique, notamment durant la transmission, et des facteurs biologiques notamment l'élargissement des gammes de plantes hôtes suite à l'introduction de vecteurs polyphages (Burdon & Thrall, 2008 ; Jones, 2009 ;

Moury *et al.*, 2007). Par conséquent, l'étude des communautés virales à l'échelle des agro-écosystèmes devrait inclure idéalement la collecte et l'analyse par métagénomique de toutes les plantes et tous les vecteurs de l'agro-écosystème étudié. Cette stratégie « globale et intégrée » devrait permettre de déterminer par quels mécanismes et à quelle fréquence la gamme d'hôte et la virulence de certains virus sont modifiés (Alexander *et al.*, 2014; Roossinck & García-Arenal, 2015).

4. La métagénomique virale

Bien que le terme métagénomique n'ait été défini qu'en 1998 (Handelsman *et al.*, 1998), les approches métagénomiques sont apparues à la fin des années 1980. La **métagénomique** est l'analyse de l'ensemble des acides nucléiques présents dans un échantillon complexe. Les méthodologies de métagénomique s'affranchissent de la mise en culture et ne nécessite aucune connaissance préalable sur les échantillons à étudier. Les stratégies de métagénomique utilisent généralement les méthodes de séquençage dites de nouvelle génération (appelées NGS pour *Next Generation Sequencing*) qui permettent de séquencer massivement l'ADN.

De nombreuses études métagénomiques ont déjà révélé un large champ d'application, que ce soit pour la caractérisation des communautés microbiennes humaines et animales (Chiu, 2013 ; Lecuit & Eloit, 2013), des écosystèmes aquatiques (Dinsdale *et al.*, 2008) et des écosystèmes terrestres (Roossinck *et al.*, 2010). L'application de la métagénomique à la virologie, discipline appelée métaviromique, a permis de véritables bouleversements conceptuels avec l'affirmation de l'immense diversité et de l'ubiquité des virus (Koonin & Dolja, 2018). Le développement des protocoles basés sur la métagénomique et le séquençage à haut-débit représente aujourd'hui une approche incontournable pour identifier précisément et de façon quasi-exhaustive le **virome** au sein de son environnement biotique (**pathobiome**) (Roossinck *et al.*, 2015 ; Vayssier-Taussat *et al.*, 2014).

4.1. Les acides nucléiques cibles

La métagénomique virale est basée sur une première étape clé de purification des acides nucléiques viraux afin d'éviter la surreprésentation des acides nucléiques non viraux pouvant représenter la grande majorité des acides nucléiques totaux (Delwart, 2007).

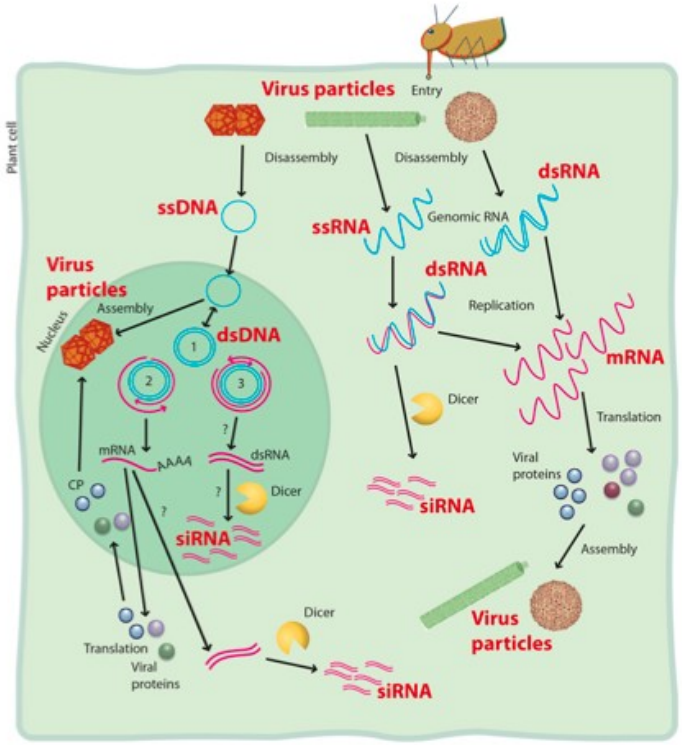


Figure 13. Représentation schématique d'une cellule de plante indiquant les différents compartiments où sont présents les différents acides nucléiques associés aux virus (Roossinck *et al.*, 2015).

Pour se faire, les approches de métagenomiques virales cibles quatre classes d'acides nucléiques : (i) les ARNs totaux ou ribodéplétés, ou les ADNs totaux, (ii) les acides nucléiques associés aux virions purifiés à partir des particules virales, (iii) les ARNs double brins et (iv) les petits ARNs issus du mécanisme de *silencing* (**Figure 13** ; Roossinck *et al.*, 2015). Il est important de garder à l'esprit que chacune de ces approches exclues certains groupes de virus et que certaines étapes peuvent biaiser les résultats. En outre, certaines études ont combiné plusieurs de ces techniques afin de mitiger ces limitations (Candresse *et al.*, 2014).

4.1.1. Les ARNs ou ADNs totaux

L'extraction des ARNs ou ADNs totaux constitue la méthode la plus simple en métagénomique virale et s'est révélée très efficace pour la découverte de nouveaux phytovirus au sein de plantes individuelles ainsi que dans des échantillons de plantes multiplexés (*pools* d'échantillons ; Roossinck *et al.*, 2015). Cependant, le principal défaut de cette méthode est qu'une proportion très élevée des séquences nucléiques obtenues n'est pas d'origine virale. Afin de pallier à ce handicap, des stratégies d'**enrichissement des acides nucléiques viraux** ont été développés. Pour la stratégie ARN, l'enrichissement est basé sur la **purification des ARN messagers par déplétion** des ARN ribosomiques de la plante hôte (Massart *et al.*, 2014). Dans le cas des virus à ADN circulaires tels que les géminivirus, la combinaison d'une étape d'**amplification en cercle roulant** a permis d'augmenter considérablement la proportion de séquences virales au sein des extraits d'échantillons de plantes (Idris *et al.*, 2014 ; Wyant *et al.*, 2012), d'insectes vecteurs (Rosario *et al.*, 2015, 2016) ou environnementaux (Dayaram *et al.*, 2016 ; Kraberger *et al.*, 2015).

4.1.2. Les acides nucléiques associés aux virions purifiés

Les approches de métagénomique basées sur les acides nucléiques associés aux virions (*virion-associated nucleic acids*, ou VANA) repose sur la purification de particules virales. Généralement, quatre étapes sont nécessaires, à savoir (i) une étape initiale de **centrifugation**, suivie (ii) d'une **filtration** puis (iii) du traitement de l'échantillon par **nucléases** afin d'éliminer les acides nucléiques non encapsidés, et enfin (iv) l'**extraction** des acides nucléiques associés aux virions protégées dans les capsides (Hall *et al.*, 2014). L'un des avantages de cette approche réside dans le fait qu'elle permet la détection simultanée des virus à ARN et ADN (Candresse *et al.*, 2014) mais ne permet pas d'isoler les virus non encapsidées. Cette approche à d'ailleur prouvé son efficacité dans la découverte de nouveaux ARN et ADN viraux dans des échantillons de tissus animaux, humains ou encore chez les plantes (Breitbart & Rohwer, 2005 ; Candresse *et al.*, 2014 ; Jones *et al.*, 2005).

4.1.3. Les ARNs double brin

Exception faite des petits ARN double brins associés au mécanisme d'ARN interférence (Conférer paragraphe 4.1.4) et en abondance dans les plantes (Pooggin, 2018), la forme double brin des ARN est une forme très caractéristique des virus, faisant d'elle une cible de choix pour des études de métagénomique virale (Roossinck *et al.*, 2010). Si certains virus ont leur génome composé d'ARN double brin, il peut aussi être généré durant la réplication des virus à ARN simple brin. Par contre, il est théoriquement impossible de détecter les virus à ADN (Roossinck *et al.*, 2010) à l'aide de cette approche. L'isolation des ARNdb repose généralement sur une étape d'**extraction des acides nucléiques** par l'utilisation de phénol et de chloroforme couplée à une étape de séparation des ARNdb et ARNsb à l'aide d'une étape de chomatographie sur colonne de cellulose (Dodds, 1984).

4.1.4. Les petits ARNs issus du mécanisme de *silencing*

Les petits ARNs résultent du mécanisme d'ARN *silencing* ou ARN interférence (ARNi). C'est un mécanisme induit chez la plante par la présence d'ARNdb aboutissant au clivage de ces ARNdb en petits ARNs de 21 à 24 nucléotides (Mlotshwa *et al.*, 2008). Le travail pionnier de Kreuze *et al. (*2009) sur patate douce a permis de démontrer que le séquençage haut débit des petits ARNs permet d'identifier des virus à ARN et ADN connus et inconnus, ainsi que leurs satellites viraux, et de reconstruire partiellement ou entièrement leurs génomes. Depuis cette étude pionnière, cette approche a été utilisée afin de détecter de nombreux phytovirus à ARN et à ADN aussi bien connus qu'inconnus. Généralement, trois étapes sont nécessaires avant la construction de la *librairie* de séquençage avec (i) l'**extraction** des ARNs totaux, (ii) la **migration** des extraits d'ARN sur gel d'acrylamide suivi de l'extraction et de la **purification** des ARNs ayant une taille entre **20 à 30 nucléotides** et (iii) la **ligation d'adapateurs** aux extrémités 3' et 5' avant **transcription inverse** (Kreuze *et al.*, 2009). Au-delà du côté laborieux associé à sa mise en œuvre et l'absence de multiplexage, l'utilisation de cette approche pourrait être limitante pour détecter les virus déclenchant faiblement ou supprimant la réponse ARNi.

4.2. Les techniques de séquençage haut débit

Les techniques de séquençage haut débit communément appelées *next generation sequencing* (NGS) ont permis de dépasser les **limitations du séquençage classique Sanger** (séquençage de première génération) et ont abouti au séquençage rapide et à **moindre coût** de génomes d'organismes provenant de l'ensemble de l'arbre du vivant. Ces technologies ont révolutionné le séquençage massif en parallèle ou séquençage haut débit. Elles ont abouti à l'explosion du nombre de séquences disponibles dans les bases de données nucléotidiques et permis un bond en avant dans la connaissance de la diversité et du fonctionnement du vivant (Barba *et al.*, 2013). On distingue les techniques NGS de seconde (SGS) et de troisième (TGS) générations. Alors que le séquençage est réalisé après amplification de l'ADN pour les techniques de seconde génération, les techniques de troisième génération permettent le séquençage directe des molécules d'ADN. Un des avantage associé à cette dernière génération est de permettre la lecture de longs fragments d'ADN (classiquement plusieurs dizaine de kilobases; Kraft & Kurth, 2019).

4.2.1. Les NGS de seconde génération

De nombreuses technologies NGS de seconde génération ont été développés ces 20 dernières années, dont les plus connues et les plus utilisées sont le 454/Roche, Illumina/Solexa et Ion Torrent/Life Technologies (**Table 1**). La plupart de ces technologies repose sur trois étapes communes à savoir (i) la **préparation des banques d'ADNs**, (ii) l'**amplification** des molécules d'ADN et (iii) le **séquençage** proprement dit. Globalement, ces différentes techniques se distinguent tant par ces trois étapes clés que par la taille des séquences obtenues (*reads*), la vitesse de séquençage et le taux d'erreur (Barba *et al.*, 2013).

Table 1. Récapitulatif des caractéristiques des principales technologies NGS (Barba *et al.*, 2013).

Sequencing platform	Amplification method	Sequencing chemistry	Read length (bp)	Sequencing Speed/h	Maximum Output Per run	Accuracy (%)	M^1 I^2 D^3
454 (Roche)	Emulsion PCR	Pyrosequencing	400–700	13 Mbp	700 Mbp	99.9	0.10, 0.3, 0.02 [23]
Illumina (Illumina)	Bridge PCR	Reversible terminators	100–300	25 Mbp	600 Gbp	99.9	0.12, 0.004, 0.006 [23]
SOLiD (Life Technologies)	Emulsion PCR	Ligation	75–85	21–28 Mbp	80–360 Gbp	99.9	Error is higher than Illumina [24]
PacBio (Pacific Biosciences)	No amplification Single molecule real-time (or SMRT)	Fluorescently labeled nucleotides	4, 000–5,000	50–115 Mbp	200 Mb–1 Gbp	95	1, 2, 12 [25]
Helicos (Helicos Biosciences)	No amplification Single molecule	Reversible terminators	25–55	83 Mbp	35 Gbp	97	Error is in the range of few percent but higher than 454 and Illumina and biased toward InDels [24]
Ion Torrent (Life Technologies)	Emulsion PCR	Detection of released H	100–400	25 Mb–16 Gbp	100 Mb–64 Gbp	99	M, 0.06, I + D 1.38 [26]
Nanopore (Oxford Technologies)	No amplification Single molecule		Very long reads up to 50 kbp	150 Mbp	Tens of Gbp	96	

M^1 = Mismatch bases; I^2 =Insertion; D^3 = Deletion.

4.2.1.1. La préparation des banques d'ADNs

Dans le cas de virus à ARN, la première étape sera la préparation d'**ADN complémentaire** (ADNc) par transcription inverse. Les opérations suivantes, communes à tous les types d'échantillons, consisteront à la construction de banques d'ADN, communément connues sous le nom de bibliothèques (ou *librairies* en anglais) avec l'ajout d'adaptateurs de séquençage, c'est à dire des fragments d'ADN spécifiques aux types de séquenceur et permettant à l'opération de séquençage de se dérouler (Van Dijk *et al.*, 2014 ; **Figure 14**).
Globalement deux méthodologies distinctes sont utilisées pour l'ajout des adaptateurs de séquençage. La première consiste en une **fragmentation** mécanique (sonication, nébulisation ou cisaillement hydrodynamique par exemple) ou enzymatique, suivi d'un **traitement enzymatique** (via le fragment de Klenow, l'ADN polymérase T4 ou la nucléase de haricot mungo par exemple) pour obtenir des ADNs présentant des extrémités franches avant ligation des adaptateurs (Alnasir & Shanahan, 2015). La seconde méthodologie, classiquement appelée « **tagmentation** », consiste à réaliser la **fragmentation** et l'**ajout des adaptateurs** en **une seule étape** à l'aide d'un transposase.

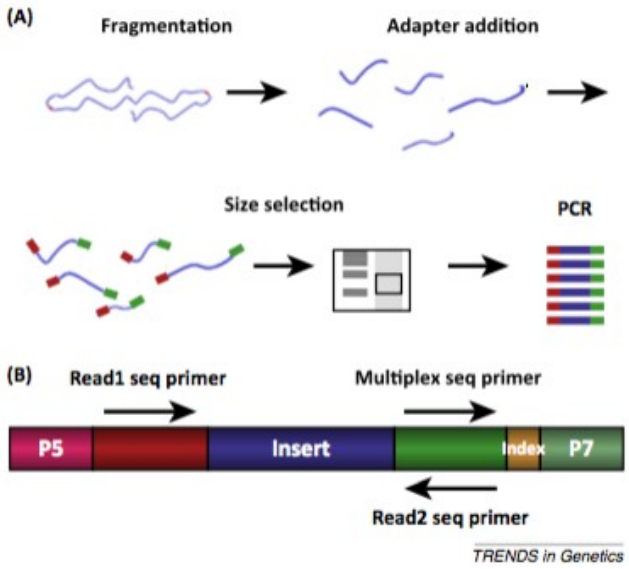


Figure 14. (A) Représentation simplifiée des différents étapes de la préparation d'une banque d'ADN (B) et architecture standard des ADNs d'une banque d'ADNs dans le cas de séquençage Illumina. Les adaptateurs sont indiqués en rouge et vert à deux tons, l'insert est en bleu. Un index ou un code à barres est indiqué en jaune. Cet index est propre à chaque banque permettant ainsi de distinguer les ADNs de banques différentes. Les amorces de séquençage sont indiquées par des flèches. Les extrémités 'P5' et 'P7' correspondent aux séquences utilisées pour la fixation et l'amplification des ADNs. Dans d'autres technologies, les noms de ces séquences terminales sont généralement définies par A et B (Van Dijk *et al.*, 2014).

4.2.1.2. Les différentes stratégies d'amplification

Dans la plupart des technologies SGS, l'étape de séquençage est précédée d'une étape d'amplification des ADNs fragmentés qui permettra d'obtenir plusieurs milliers de copies du même fragment d'ADN (amplicons) nécessaire au séquençage. La méthode d'amplification des ADNs fragmentés diffère selon les plateformes de séquençage NGS utilisées (**Table 1**). On distingue deux types d'amplification à savoir (i) la **PCR en émulsion** ou emPCR (454, SoliD, Ion Torrent) et (ii) la **PCR sur phase solide** (Illumina ; Barba *et al.*, 2013 ; Metzker, 2010). L'émulsion de l'emPCR est composée de gouttelettes aqueuses appelées micelles dans un milieu hydrophobe. Chaque micelle contient une bille recouverte d'adaptateurs complémentaires à ceux présents sur les fragment d'ADN et tous les composants nécessaires à l'amplification. L'emPCR est optimisée pour avoir le plus grand nombre de micelles contenant un fragment unique d'ADN de l'échantillon de départ. Ainsi en fin de réaction, chaque bille sera recouverte par plusieurs milliers de copies provenant du même fragment d'ADN. La PCR sur phase solide repose sur la présence

d'adaptateurs complémentaires fixés sur un support solide appelé *flowcell*. Dans le cas des PCR par pont sur phase solide par exemple, après liaison des ADNs fragmentés aux adaptateurs, l'extrémité libre de chaque fragment s'hybride à un adaptateur complémentaire situé à proximité formant une structure en pont. Ce pont d'amplification initie alors la synthèse du brin complémentaire. En procédant à des cycles multiples de ce type d'amplification en phase solide, on obtient plusieurs milliers de copies de la même séquence d'ADN clustérisées sur une plaque.

4.2.1.3. Les techniques de séquençage de seconde génération

Il existe principalement deux techniques permettant le séquençage de petits *reads* à savoir (i) le **séquençage par ligation** (SBL ; Goodwin *et al.*, 2016) et (ii) le **séquençage par synthèse** (SBS). Le séquençage par ligation (SOLiD) repose sur un système de cycle de ligation et clivage d'une sonde couplée à un fluorochrome. La sonde est constituée d'une ou deux bases connues (*one-base-encoded probe* ou *two-bases-encoded probe*), suivies de bases universelles (s'appariant avec n'importe quelles bases) permettant après interprétation des fluorochromes de reconstituer la séquence nucléotidique (Goodwin *et al.*, 2016).

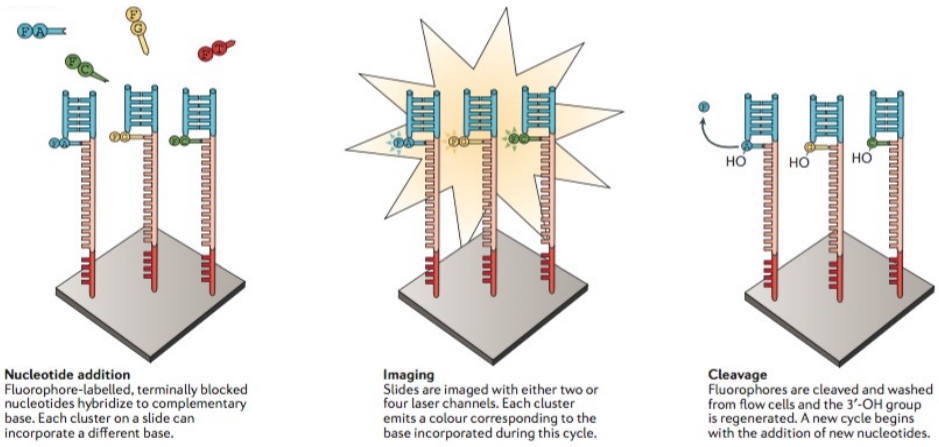


Figure 15. Représentation des étapes de séquençage par synthèse utilisant l'approche de terminaison réversible cyclique Illumina. L'ADN polymérase et les dNTPs marqués par un fluochrome (F) spécifique à chaque base sont ajouté au *flowcell*. A chaque cycle, un unique nucléotide est incorporé au brin en cours de séquençage. Après l'incorporation, les nucléotides libres sont éliminés et les fluorochromes sont excités par lasers. Le fluorochrome émet alors à une longueur d'onde spécifique du nucléotide incorporé. Les fluorochromes sont ensuite clivés et éliminés et le groupe 3'-OH du dernier nucléotide incorporé est régénéré afin de permettre un nouveau cycle d'incorporation (Goodwin *et al.*, 2016).

Dans les approches SBS, une polymérase est utilisée pour le séquençage et un signal, tel qu'un fluorophore ou une modification de la concentration ionique, rend compte de l'incorporation d'un nucléotide dans le brin en cours de synthèse. Il existe deux méthodes de SBS à savoir (i) la **terminaison réversible cyclique** (CRT) et (ii) **l'addition d'un type unique de nucléotide** (SNA).

Chez Illumina par exemple, à chaque cycle, un mix comprenant l'ensemble des quatre dNTPs marqués par quatre fluorochromes différents est ajouté. Après l'incorporation d'un unique dNTP complémentaire à la séquence, les dNTPs non liés sont éliminés et la nature du dNTP incorporé est identifiée par mesure de fluorescence. Afin de permettre la poursuite de l'élongation, le fluochrome est clivé et le groupement 3'-OH modifié est régénéré (**Figure 15** ; Metzker, 2010).

La méthode SNA regroupant les techniques de **pyroséquençage 454** et **Ion Torrent**, repose sur l'incorporation d'un mix contenant **qu'un seul type de dNTP**. Ainsi, quatre mix différents sont présentés successivement au fragment d'ADN à séquencer et seul le dNTP complémentaire sera incorporé. Dans le cas de séquences homopolymériques (succession de plusieurs nucléotides identiques), plusieurs dNTPs sont incorporés simultanément induisant l'émission d'un signal (du à la luciférase pour 454 ou l'émission d'un ion H+ pour Ion Torrent) proportionnel au nombre de nucléotides ajoutés (**Figure 16** ; Goodwin *et al.*, 2016).

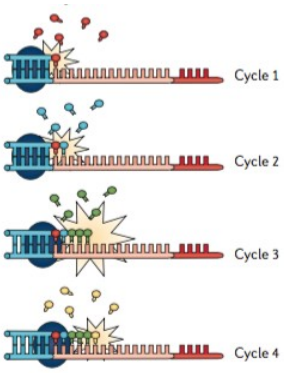


Figure 16. Représentation des étapes de séquençage par synthèse utilisant l'approche d'addition séquentielle de chaque nucléotide. Un seul type de dNTP est présent pendant chaque cycle et plusieurs dNTP du même type peuvent être incorporés au cours d'un cycle induisant une augmentation de l'intensité du signal émis (Goodwin *et al.*, 2016).

4.2.1.4. Les limitations des techniques NGS de seconde génération

Malgré le fait que les techniques NGS ont permis de réelles avancées en matière de diagnostic et de caractérisation de virus à la fois connus et inconnus, il est important de garder à l'esprit que ces techniques font face à **plusieurs limites**. L'une de celles-ci concerne l'ensemble des étapes requises pour la préparation des *librairies*. Ces étapes impliquant souvent une amplification de l'ADN, peuvent être génératrices de **produits chimériques** (*i.e.* des recombinants artificiels ; Lasken & Stockwell, 2007), de **biais quantitatifs** en modifiant les fréquences relatives des virus présents (Gallet *et al.*, 2017) ou de contamination inter-échantillons (Sinha *et al.*, 2017).

L'étape de séquençage est également génératrice de contamination inter-échantillon (exemple du *index hopping/switching* lors du séquençage Illumina Costello *et al.*, 2018) ou d'**erreurs dans la séquence**. On peut citer par exemple les erreurs dues à la polymérase, à la présence de régions homopolymères pour le pyroséquençage 454 (signaux d'intensité trop élevée ou trop faible entrainant une sous ou surestimation du nombre de nucléotides ; Huse *et al.*, 2007), à une défaillance du clivage du fluorochrome ou encore à une différence insuffisante entre les spectres d'émission pour le séquençage Illumina (mauvaise interprétation des signaux d'émission ; Dohm *et al.*, 2008).

Du fait de ces biais et erreurs, il apparaît nécessaire d'ajouter des témoins de séquençage (témoins positifs, mélanges de matrices d'ADN en quantité connues, témoins négatifs…) mais aussi d'appliquer des **seuils de détection** définis pour maximiser la sensibilité de détection sans entraîner un taux élevé de **faux positifs**. Par ailleurs, la petite taille des *reads* obtenus peut rendre difficile l'assemblage de ces *reads* en plus larges séquences (appelées *contigs*) tout comme leur simple assignation taxonomique.

4.2.2. Le séquençage de troisième génération (TGS)

Afin de pallier aux limitations de tailles de *reads* associées aux SGS, de nouvelles techniques de séquençage dites de troisième génération (TGS) ont été développées. Actuellement, les TGS permettent de générer des séquences de « grandes longueurs » (*long reads*) de plusieurs dizaines de kilobases. Deux plateformes de séquençage permettent du TGS à savoir (i) **Pacific Biosciences** (PacBio) qualifié de séquençage en temps réel d'une molécule unique (*single-molecule real-time sequencing* ; SMRT ; Eid *et al.*, 2009) et plus récemment (ii) **Oxford Nanopore Technologies** (ONT) notamment avec le MinION (Branton *et al.*, 2009).

Le SMRT de Pacific Biosciences est basé sur le **suivi de l'activité d'une ADN polymérase** fixée dans un puit appelé *zero-mode waveguide* (ZMW). Chaque type de nucléotide porte un fluorochrome spécifique sur son groupe phosphate. Ainsi, à chaque incorporation d'un nucléotide à l'ADN matrice dans le site actif de la polymérase, une impulsion de fluorescence est émise et enregistrée par le capteur ZMW. Le flurochrome est ensuite clivé par la polymérase conduisant à un arrêt du signal d'émission et permettant l'incorporation du nucléotide suivant (Eid *et al.*, 2009 ; Goodwin *et al.*, 2016). Cette technique de **séquençage en temps réel** est un processus assez **rapide** (quelques heures) ayant un **taux d'erreur** compris entre **10** et **15%** (Lee *et al.*, 2014). Toutefois, les spécificités de cette méthodologie permettent, grâce à un grand nombre de lectures des mêmes fragments, d'obtenir par assemblage des génomes de grande qualité avec une précision supérieure à 99,99%. Le potentiel de cette technique de séquençage a été démontré notamment à travers le séquençage de régions génomiques complexes chez le chimpanzé et l'humain (Chaisson *et al.*, 2015 ; Huddleston *et al.*, 2014) et de génomes de bactéries (Chin *et al.*, 2013).

Le séquençage par nanopore repose sur la **mesure de la variation de champ électrique** durant la translocation de la molécule d'ADN à travers un pore. Ces variations de champ électrique sont alors converties en séquences d'ADN (Goodwin *et al.*, 2016). Cette technique présente l'avantage principal d'être accessible à l'aide d'un séquenceur de **petite taille** et de **faible coût**

appelée MinION (Lee *et al.*, 2014). Les *reads* produits sont aussi longs que ceux obtenus à l'aide de la plateforme de Pacific Biosciences mais les **taux d'erreur** varient entre **10** et **30 %** (Lee *et al.*, 2014). Cependant, des **algorithmes de correction d'erreur** et une chimie en constante évolution permettent de compenser ces difficultés (Lee *et al.*, 2014). Le séquençage nanopore a été utilisé avec succès pour la détection et/ou le séquençage de nombreux virus chez l'Homme (notamment Ebola ; Quick *et al.*, 2016), les animaux (le cowpox virus par exemple ; Kilianski *et al.*, 2015) et les plantes (le yam mild mosaic virus par exemple ; Filloux *et al.*, 2018).

4.3. Identification et classification des séquences virales

Les approches métagénomiques donnent accès à de grandes quantités de données brutes. Afin de pouvoir filtrer et contrôler la qualité des séquences et de les assigner à des virus, le développement et l'utilisation d'outils bioinformatiques sont devenus indispensables. Malgré le caractère très haut-débit des méthodes *NGS*, l'assignation taxonomique des séquences obtenues reste une étape difficile. En effet, les *reads* produits sont souvent de faible longueur. Même s'il est théoriquement possible d'identifier des virus sur la base de lectures courtes, l'identification s'avère complexe d'autant plus que ces virus peuvent appartenir à des groupes mal décrits voir complètement inconnus (Krishnamurthy & Wang, 2017 ; Scholz *et al.*, 2012). La majorité des séquences obtenues (60 à 95%) dans le cadre d'étude de métaviromique reste alors bien souvent non-identifée et se retrouve dans la catégorie appelé *dark matter* (ou matière noire en français ; Roux *et al.*, 2015).

Afin de palier à ces difficultés, une étape dite d'assemblage des *reads* (assemblage *de novo*) est souvent réalisée avant l'assignation taxonomique. L'assemblage est un processus permettant de grouper les *reads* présentant des zones identiques et chevauchantes en une séquence consensus appelée contigue (*contig*) (**Figure 17**). Ces *contigs* sont des séquences plus longues que celles du jeu de données brutes et permettent idéalement l'obtention de génomes entiers. Cependant, du fait de la difficulté à séquencer certaines régions génomiques et des biais spécifiques des assembleurs disponibles, l'obtention de génomes entiers reste complexe.

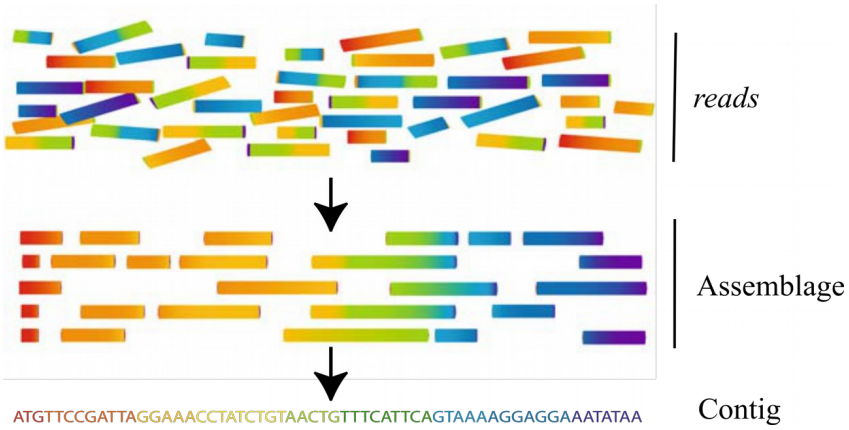


Figure 17: **Représentation schématique du principle d'assemblage *reads*** (D'après Commins *et al.*, 2009).

Ainsi, suivant les caractéristiques des assembleurs et les paramètres utilisés, la proportion de faux négatifs (assemblage non réalisé) ou faux positifs (assemblage de *contigs* chimériques) sera variable. L'obtention d'un génome correctement assemblé (peu de *contigs*, *contigs* de grande taille) nécessite donc de choisir une méthode d'assemblage appropriée à la stratégie de séquençage choisie initialement.

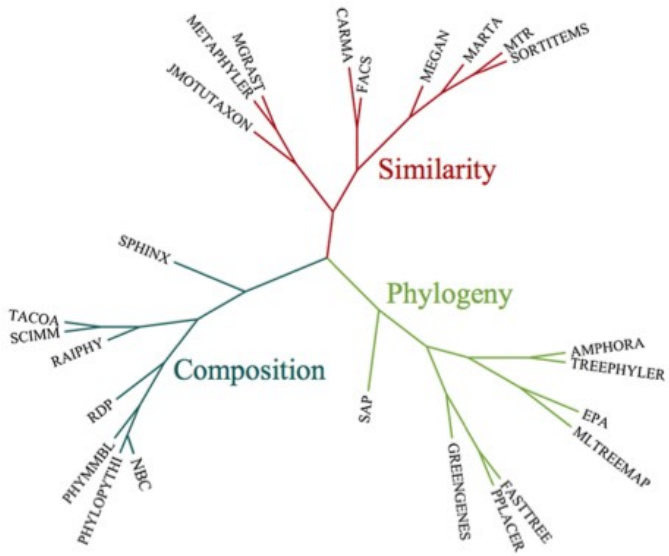


Figure 18: **Représentation de la similarité entre différents programmes d'assignation des *reads*** (Bazinet & Cummings, 2012). Trois classes de méthodes d'assignation existent. Elles sont basées soit sur la similarité entre les séquences (en rouge), la composition des séquences (en bleu) ou leur phylogénie (en vert).

Que ce soit sur des *reads* ou des *contigs*, l'étape suivante consistera généralement à identifier l'organisme dont proviennent les séquences, c'est

l'**assignation taxonomique**. S'il existe une multitude de stratégies d'assignation taxonomique (**Figure 18** ; Bazinet & Cummings, 2012), toutes perdent en performance lorsqu'il s'agit d'assigner des *reads* courts de génomes viraux. Historiquement, les assignations ont été faites sur la base de similarité de séquences (BLAST et ses dérivés). Néanmoins, ces dernières années, de nouvelles procédures basées sur la composition en « k-mer » (*i.e.* sous séquences de longueur k) ou encore sur le positionnement phylogénétique ont vu le jour (**Figure 18** ; Bazinet & Cummings, 2012). Cette dernière classe de méthode basée sur la phylogénie permet d'obtenir une classification basée sur un modèle probabiliste (tels que ceux utilisés pour la reconstruction phylogénétique par maximum de vraisemblance) dans un contexte phylogénétique (Matsen *et al.*, 2010). Une fois les séquences assignées, elles sont généralement regroupées selon un certain pourcentage d'identité, en *cluster*, *virotype* ou *OTU* (*Operational Taxonomic Unit*). Il est alors possible de procéder à des mesures de diversité de la composition taxonomique d'un échantillon et à la comparaison de plusieurs d'entre eux. Cependant, dans le cas de séquences ne présentant pas de similarités proches avec les séquences des bases de données, l'assignation devient alors problématique et complexe (Massart *et al.*, 2014).

4.4. De la métagénomique spatiale à la caractérisation moléculaire des phytovirus

Cette partie a fait l'objet d'un chapitre de review publié en 2018 : Claverie, S., Bernardo, P., Kraberger, S., Hartnady, P., Lefeuvre, P., Lett, J. M., Galzi, S., Filloux, D., Harkins, G. W., Varsani, A., Martin, D. P., Roumagnac, P. (2018). From spatial metagenomics to molecular characterization of plant viruses: A geminivirus case study. *In Advances in virus research* (Vol. 101, pp. 55-83). Academic Press.

Cet article retrace l'émergence de la métagénomique virale et ses apports en termes de connaissances sur la diversité et la distribution des phytovirus à l'échelle de l'écosystème mais également en termes de compréhension de l'écologie et de l'évolution des phytovirus.

CHAPTER THREE

From Spatial Metagenomics to Molecular Characterization of Plant Viruses: A Geminivirus Case Study

Sohini Claverie[*], Pauline Bernardo[†,‡,§], Simona Kraberger[¶], Penelope Hartnady[‖], Pierre Lefeuvre[*], Jean-Michel Lett[*], Serge Galzi[‡,§], Denis Filloux[‡,§], Gordon W. Harkins[#], Arvind Varsani[¶,], Darren P. Martin[‖], Philippe Roumagnac[‡,§,1]**

[*]CIRAD, UMR PVBMT, F-97410 St Pierre, La Réunion, France
[†]Department of Plant Pathology, Ohio State University, OARDC, Wooster, OH, United States
[‡]CIRAD, UMR BGPI, Montpellier, France
[§]BGPI, Univ. Montpellier, CIRAD, INRA, Montpellier SupAgro, Montpellier, France
[¶]The Biodesign Center for Fundamental and Applied Microbiomics, Center for Evolution and Medicine, School of Life Sciences, Arizona State University, Tempe, AZ, USA
[‖]Computational Biology Division, Department of Integrative Biomedical Sciences, Institute of Infectious Diseases and Molecular Medicine, University of Cape Town, Observatory, South Africa
[#]South African Medical Research Council Bioinformatics Unit, South African National Bioinformatics Institute, University of the Western Cape, South Africa
[**]Structural Biology Research Unit, Department of Clinical Laboratory Sciences, University of Cape Town, Rondebosch, Cape Town, South Africa
[1]Corresponding author: e-mail address: philippe.roumagnac@cirad.fr

Contents

Advances in Virus Research, Volume 101
ISSN 0065-3527
https://doi.org/10.1016/bs.aivir.2018.02.003

Abstract

The number of plant viruses that are known likely remains only a vanishingly small fraction of all extant plant virus species. Consequently, the distribution and population dynamics of plant viruses within even the best-studied ecosystems have only ever been studied for small groups of virus species. Even for the best studied of these groups very little is known about virus diversity at spatial scales ranging from an individual host, through individual local host populations to global host populations. To date, metagenomics studies that have assessed the collective or metagenomes of viruses at the ecosystem scale have revealed many previously unrecognized viral species. More recently, novel georeferenced metagenomics approaches have been devised that can precisely link individual sequence reads to both the plant hosts from which they were obtained, and the spatial arrangements of these hosts. Besides illuminating the diversity and the distribution of plant viruses at the ecosystem scale, application of these "geometagenomics" approaches has enabled the direct testing of hypotheses relating to the impacts of host diversity, host spatial variations, and environmental conditions on plant virus diversity and prevalence. To exemplify how such top-down approaches can provide a far deeper understanding of host–virus associations, we provide a case-study focusing on geminiviruses within two complex ecosystems containing both cultivated and uncultivated areas. Geminiviruses are a highly relevant model for studying the evolutionary and ecological aspects of viral emergence because the family *Geminiviridae* includes many of the most important crop pathogens that have emerged over the past century. In addition to revealing unprecedented degrees of geminivirus diversity within the analyzed ecosystems, the geometagenomics-based approach enabled the focused in-depth analysis of the complex evolutionary dynamics of some of the highly divergent geminivirus species that were discovered.

1. INTRODUCTION

Although viruses are the most abundant and diverse entities on earth and are arguably "the most successful form of life" (Wasik and Turner, 2013), they are still primarily regarded as agents of disease. As exemplified by other chapters of this book focusing on a range of host types, it is now apparent, however, that viruses are core components of global ecosystems within which the spectrum of virus–host interactions include parasitism and pathogenicity at one extreme and symbiosis and mutualism at the other

(Roossinck, 2011b, 2015; Suttle, 2007). The ecological roles of viruses and the benefits that they can bestow on host species have mainly been overlooked because of how difficult it is to detect and observe them. This difficulty has historically restricted most virological studies to the small minority of pathogenic virus species that are associated with diseases of cultivated plants, domesticated animals, and humans.

The beginnings of plant virology—and virology per se—date back to the late 19th century when the first virus ever discovered, later named tobacco mosaic virus (Family *Virgaviridae*, Genus *Tobamovirus*), was reported by Ivanovski in 1892 and Beijerinck in 1898 (Bos, 1999). Since these two pioneering studies, approximately 4400 other virus species have been described of which approximately 1400 infect plants (10th Report of the International Committee on Taxonomy of Viruses, ICTV). The vast majority of plant viruses are transmitted by a vector organism that has the capacity to directly inject virus particles into the cytoplasm of plant cell types that are amenable to virus replication.

While most known plant virus species have single-stranded (ss) RNA genomes (Zaitlin and Palukaitis, 2000), the biggest emerging worldwide threat to crops are arthropod-borne viruses with ssDNA genomes in the family *Geminiviridae*. Examples of important emergent diseases that are caused by geminiviruses include maize streak disease (Harkins et al., 2009b), cassava mosaic disease (Patil and Fauquet, 2009), cotton leaf curl disease (Briddon and Markham, 2000), and tomato leaf curl disease (Accotto et al., 2000).

Geminiviruses display substitution rates that are as high as those of viruses with ssRNA genomes and they are known to readily recombine both within and between species. Viruses within this genus are therefore predisposed to rapid adaptive evolution (Duffy and Holmes, 2008; Harkins et al., 2009a; Lefeuvre et al., 2010; Monjane et al., 2011). They are therefore good, and relevant, models for studying plant virus evolution and emergence.

With the increasing availability of geminivirus genome sequences, population-scale studies looking in-depth at genetic diversity within individual crop-infecting geminivirus species have recently been reported (Kraberger et al., 2017b; Pande et al., 2017). Besides revealing ample evidence of recombination, these studies have also determined that the origins of much of the recombinationally acquired genetic material are likely geminiviruses belonging to species that have never been characterized. In cases where parental viruses belong to species that have been characterized, many of these "parental" virus species are known to primarily infect uncultivated hosts. Because of their focus on individual host and virus species, however,

such studies cannot directly address hypotheses relating to the potentially complex, but highly relevant, evolutionary, and ecological factors underlying the origin and spread of such recombinants.

Crucially, the development and application of sequence-nonspecific virus discovery approaches (Roossinck et al., 2015; Rosario et al., 2012b) has tremendously increased the rate at which novel highly divergent geminiviruses have been discovered within uncultivated plants, including *Euphorbia caput-medusae* (Bernardo et al., 2013), cleome (Fontenele et al., 2017), wild rice (Kraberger et al., 2017a), uncultivated grapevines (Perry et al., 2018), and *Plantago lanceolata* (Susi et al., 2017). The picture of global-geminivirus diversity that is emerging from these virus discovery studies is that the economically relevant geminivirus crop pathogens represent a minute fraction of the geminiviruses that could potentially emerge from uncultivated species to threaten crop production in the future.

In this chapter, we initially review the ongoing development of viral metagenomics-based approaches, including the most recent iterations of the georeferenced metagenomics approaches. Besides providing insights into the distribution and prevalence of plant viruses across landscapes, these approaches have enabled the quantitative estimation of the genetic diversity within global plant virus populations. We then focus on a geminivirus case study to highlight how a top-down georeferenced metagenomics approach can both illuminate the extant of geminivirus diversity within an ecosystem and reveal new host-geminivirus associations, paving the way toward a better ecosystem scale understanding of host–parasite dynamics and the emergence of new crop diseases.

2. METAGENOMICS-BASED APPROACHES: THE WITHOUT A PRIORI INVESTIGATION REVOLUTION

2.1 Pioneering Metagenomics Studies

Although the term "metagenomics" was first used by Jo Handelsman and coworkers in 1998 to describe "the genomic analysis of a microorganism's population in an environmental sample" (Handelsman et al., 1998), the conceptual and technical framework upon which metagenomics is based was first introduced more than a decade earlier. In the early 1980s, Pace and colleagues used 5S and 16S rRNA culture-independent analyses to describe microbial communities in environmental samples (Pace et al., 1985). This culture-independence readily provided reasonably unbiased views of microbial populations within environmental samples—populations in which approximately 99% of species could not be cultured (Amann et al., 1995).

The multitude of culture-independent microbe discovery and quantification techniques that were devised subsequent to the seminal work of Pace et al. (1985)—techniques that are now collectively referred to as metagenomics approaches—enabled the large-scale discovery, identification, and study of microbial taxa within thousands of the Earth's myriad environments (Alberti et al., 2017; Gilbert and Dupont, 2011; Temperton and Giovannoni, 2012).

2.2 Viral Metagenomic Studies: Virus Discovery and Dark Matter

Although metagenomics was first used to characterize bacterial, archaeal, and fungal communities, it is increasingly being employed in virology investigations (Delwart, 2007). Whereas there are universally conserved genes in bacteria, archaea, and fungi that allow the direct comparison of all members of these kingdoms, there are no genes that are conserved across all viruses that can be used as canonical phylogenetic markers.

To circumvent this problem, protocols have been developed to enable the specific but mostly sequence-independent amplification of virus sequences. In their seminal work, Allander et al. (2001) used a technique based on DNase sequence-independent single primer amplification. This technique allowed the first metagenomic study of uncultured marine viral communities (Breitbart et al., 2002) and marked the emergence of metagenomics in virology as a means of identifying new viruses (Breitbart and Rohwer, 2005). The strategy used was based on (i) the enrichment of viruses in an environmental sample using a combination of differential filtration and ultracentrifugation and (ii) the amplification of viral genomes with random primers.

A growing number of methodologies, based primarily on high-throughput sequencing technologies, are now available to interrogate the genetic diversity within virus populations without any prior knowledge of what species the populations contain. Using these approaches, novel viral communities have been studied in a wide range of environments including those within the human body (Breitbart et al., 2003; Lecuit and Eloit, 2014), in the oceans (Breitbart et al., 2004, Brum et al., 2015, Suttle, 2005; see chapter "Viruses in marine ecosystems: From open waters to coral reefs" by Weynberg), in fresh water lakes (Dayaram et al., 2016, Roux et al., 2012; see chapters "Water-mediated transmission of plant, animal, and human viruses" by Mehle et al. and "Viruses in polar lake and soil ecosystems" by Rastrojo and Alcamí), and in terrestrial ecosystems (Ng et al., 2011b; Roossinck, 2012).

Along with the discovery of hundreds of individual new viral species, the rich viral diversity that has been identified in such studies emphasizes that

viruses are likely essential components within ecosystems ranging in scale from that of the human gut (Scarpellini et al., 2015; Stecher et al., 2012) to that of the planet-wide oceanic planktonic or bacterial ecosystems (Alberti et al., 2017; Sunagawa et al., 2015). Furthermore, these studies have shown that up to 60%–99% of the sequences recovered are "dark matter" because they can only tentatively be classified as being of viral origin due to their lacking any convincingly detectable homology to any nucleotide sequences with known origins that are currently deposited within large publically nucleotide sequence databases such as GenBank (Rosario and Breitbart, 2011). Although our inability to definitively classify these "dark matter" sequences can hamper ecological and evolutionary studies of viromes, their mere presence strongly suggests that global viral diversity could far exceed that which is presently known (Rosario and Breitbart, 2011).

One of the main reasons for suspecting that many of these dark matter sequences are of viral origin is that even when metagenomics-derived sequence reads do display homology to known virus sequences, the degree of similarity of these sequences to their best-matching viral homologs within the sequence databases is frequently close to the threshold of what would be considered "convincingly detectable homology." For example, a recent analysis of replication-associated protein (Rep) sequences of ssDNA viruses demonstrated the extent of how much diversity exists even within a well-known evolutionarily conserved virus protein. Phylogenetic analysis of Rep sequences from 659 circular Rep-encoding ssDNA viruses (CRESS-DNA viruses) that had been discovered in various metagenomics studies revealed that the vast majority of these clustered outside of the four CRESS-DNA virus families that are presently recognized by the International Committee for Virus Taxonomy (Simmonds et al., 2017), and that most of the viruses from which these sequences were derived are likely representative of tens, if not hundreds, of presently uncharacterized family level CRESS-DNA virus lineages.

2.3 Plant Viral Metagenomics: Looking Beyond Cropping Systems

As is true for viruses more generally, there are profound gaps in our understanding of the nature and diversity of plant viruses. Existing knowledge is biased by a focus on only a small subset of plant–virus interactions. Since the 19th century discovery of tobacco mosaic virus, most plant virus studies have been motivated by an imperative to manage the diseases of crop and ornamental plants. Nevertheless, paradoxical metagenomics-based reports

on uncultivated plant species of high frequencies of viral infections but low levels of disease symptoms (Muthukumar et al., 2009; Roossinck et al., 2010), and reconstructions of the emergence histories of important crop pathogens suggesting that many of these likely originated in uncultivated hosts, have prompted a reassessment of how viral diseases of plants originate (Fargette et al., 2006). Also, the discovery that many different plant virus species can coexist and potentially interact either within an individual host, within multiple individuals of a single host species, or within different host species within the same plant community (Vayssier-Taussat et al., 2014), has seriously undermined the plausibility of the "one virus—one disease" model. An alternative model is that viral diseases are the outcome of evolutionary and ecological processes operating at the level of whole viral communities, or viromes, and that diseases are a consequence of otherwise benign viruses inflicting excessive harm on particular host plants (Vayssier-Taussat et al., 2014).

It is in fact becoming increasingly evident that the emergence of disease causing viruses involves processes operating at scales ranging from individual molecular interactions within host cells to the interactions of host and potential host organisms across landscapes. Hence, it has been posited that the long-term coevolution of plants and their viruses within resilient natural ecosystems may lead to mutual adaptation and low disease burdens (Cooper and Jones, 2006; Jones, 2009). However, the reality is probably more complex because the influence of viruses on natural hosts can range from fitness reducing (e.g., Kelley, 1994; Remold, 2002; Malmstrom et al., 2005) to fitness enhancing (Márquez et al., 2007; Xu et al., 2008). By contrast, ecosystems that have been disturbed, for example by human activities such as agriculture, can potentially foster an environment where there are no strong evolutionary constraints preventing the emergence of viruses and/or vectors that inflict considerable damage on certain host species (Fargette et al., 2006; Roossinck and Garcia-Arenal, 2015).

Given that most of the information that we presently have on plant virus diversity is derived from viruses belonging to the small number of species that cause recognizable diseases on crop and ornamental plants (Malmstrom et al., 2011; Wren et al., 2006), we still cannot fully understand why and how plant viral diseases emerge. Recent viral metagenomic studies have therefore started to look beyond cropping systems and substantial methodological advances have been made enabling the study of plant viral diversity and evolution at the scale of entire ecosystems (Roossinck et al., 2015; Stobbe and Roossinck, 2014).

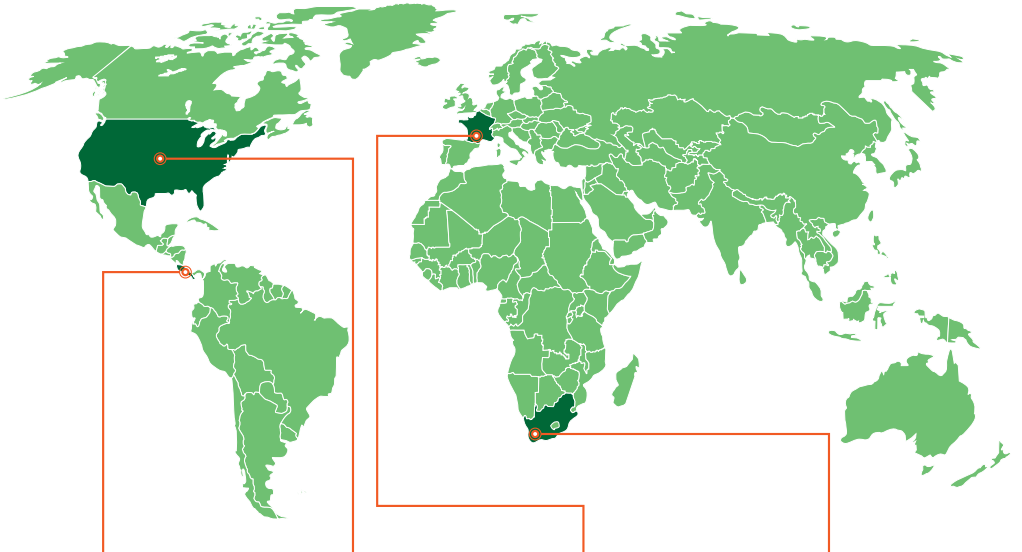
2.4 Georeferenced Metagenomics Approaches: Exploring Contemporary Plant Viral Dynamics Within a Defined Spatiotemporal Framework

Ecology can be loosely defined as the study of how organisms interact with each other and with the environment. While determining the inventory of virus species that infect individual plants within an ecosystem could facilitate the inference of probable host ranges, comparing the genetic composition of virus populations in different hosts could enable the inference of viral community structure. Furthermore, comparing the species composition of viral communities and the genetic structures of virus populations before and after particular environmental perturbations, could provide insight about how viral populations respond to such perturbations.

In the context of understanding viral emergence, the scale at which viral ecology studies are particularly pertinent is that of agroecological interfaces that span cropping systems and natural environments (Alexander et al., 2014). At these interfaces, changes in environmental factors (e.g., the abundances of host and vector species) are expected to be more abrupt than in natural environments. Crucially, we suggest that the environmental heterogeneity at these interfaces will potentially leave imprints both on the community structures of virus populations in these ecosystems, and on the individual rapidly evolving genomes from which these populations are composed.

In 2010, Roossinck et al. developed a viral metagenomics approach that could precisely link individual sequence reads from bulked multiplexed sequencing reactions to information on the individual hosts from which the sequences were sampled (Roossinck et al., 2010). This paved the way toward quantitatively studying viral diversity, prevalence, and spatial distributions across landscapes. Using this approach, pioneering "ecogenomics" studies in natural environments indicated that ~70% of sampled plants in a Costa Rican forest (Roossinck et al., 2010) and ~25% of plants in an Oklahoma prairie (Muthukumar et al., 2009) harbored identifiable plant viral sequences (Fig. 1).

"Geometagenomics" is a subsequent enhancement of the "ecogenomics" approach, which has been developed to further consider the spatial arrangements of plant samples, and the precise environmental contexts of individual sampling sites (Bernardo et al., 2018). The geometagenomics experimental design involves systematically defined sampling locations within a predefined georeferenced grid that might contain, for example, 100 sampling points (or geonodes) at 500-m spacing (10 nodes × 10 nodes; Bernardo et al., 2018).



	Costa Rica (Area de Conservacion Guanacaste)	USA (Tall Grass Prairie Preserve - Oklahoma)	France (Camargue)	South Africa (Cape Floristic Region)
Year	2010	2010	2010 and 2012	2010
Ecosystem	Dry forest, cloud forest, and rain forest	Tallgrass prairie	Mediterranean wetlands/crops	Fynbos/crops
Approach	Ecogenomics	Ecogenomics	Geometagenomics	Geometagenomics
Plant types	Uncultivated	Uncultivated	Cultivated vs uncultivated	Cultivated vs uncultivated
Plant diversity	High	Low	Low vs high	Low vs high
Virus prevalence	70%	25%	26%, 2010 36%, 2012	28%
Reference	Roossinck et al. (2010)	Muthukumar et al. (2009)	Bernardo et al. (2017)	Bernardo et al. (2017)

Fig. 1 Summary of the published plant virus ecogenomics and geometagenomics studies that have been conducted over the past decade.

This georeferenced approach yields geographically tagged DNA amplicons from both known and unknown virus species with both RNA and DNA genomes, and further allows viral sequences to be linked to specific host plants at a specific location and time. The a priori choice of sampling points allows the identification of reference ecosystems that would be most appropriate for determining, for example, the impacts of agriculture on viral demographics and evolution within natural endangered ecosystems, or the transmission rates of viruses between cultivated and uncultivated plants in areas where crop fields border natural vegetation.

To date, the geometagenomics approach has been applied to the study of only two ecosystems (Fig. 1): one in the Western Cape region of South Africa and the other in the Rhône River Delta region in France (Bernardo et al., 2018). The South African site included endangered evergreen shrubland (renosterveld) and beach scrub (strandveld) situated beside fertile plains planted with barley, wheat, and other crops (Bernardo et al., 2018). The French site included saline steppes and xero-halophitic meadows surrounded by areas under intensive rice, wheat, and alfalfa cultivation (Bernardo et al., 2018). The objectives of this study were to assess whether (i) plant-associated virus communities were more prevalent but less diverse in cultivated areas and (ii) viruses from particular known families were significantly associated with cultivated or uncultivated areas.

These French and South African geometagenomics studies revealed that 26%–36% of the plant samples throughout the georeferenced sampling grids in the two countries contained plant and/or fungal virus sequences and that viral prevalence in these ecosystems was similar to that observed in the Oklahoma prairie ecosystem (Muthukumar et al., 2009) (Fig. 1). Importantly, these and other estimates of plant viral prevalence should be taken as lower bounds of the actual prevalences within particular environments. This is because there are two biases that lead viral prevalence to be underestimated, both of which have impacted all plant virus ecogenomics and geometagenomics studies to date. The first bias arises because viral genomes may not be isolated with the same efficiency from all host species, and so some viruses may be missed. The second bias arises because sequenced genomic fragments from viruses that are closely related to previously characterized plant viruses are easier to identify as being of plant viral origin than are those from viruses that are only distantly related (or completely unrelated) to any previously characterized taxon. As a result of these two biases, many viruses that are present within analyzed plant samples will either not be present at high enough titers to be isolated and sequenced, or, if they are isolated and sequenced, will be so distantly related to

currently known viruses that they will not be identifiable as being virus derived and will be categorized as dark matter of indeterminate origin.

In the single geometagenomics study so far published (Bernardo et al., 2018), the detected proportion of infected plant samples (a proxy for virus prevalence) was higher for cultivated host plants than for uncultivated ones. Likewise, it was found that overall virus prevalence was generally higher in cultivated crop areas than in natural vegetation areas. Furthermore, viruses from some virus families showed strong associations with either cultivated or uncultivated areas. These results are consistent with hypotheses that generally predict increased pathogen prevalence as host abundance increases (Agrawal et al., 2006; Keesing et al., 2010). It is, for example, known that the prevalence of a pathogen might be increased through nonrandom biodiversity loss caused by changes to the physical or biotic environment (Lacroix et al., 2014; Pagan et al., 2012), through increasing the absolute abundance of hosts and/or pathogen-transmitting vectors (Mitchell et al., 2002; Roche et al., 2012), or through increasing the probability that susceptible hosts and vectors will encounter each other (Allan et al., 2003).

In the endangered South African Fynbos Biome, Bernardo et al. (2018) further found that the prevalence of plant viruses was significantly higher in exotic plants than in indigenous plants. Invasive exotic plants are considered to be the most severe threat currently facing the Fynbos Biome because competition between these exotic and indigenous plants reduces the diversity of the indigenous plants (Downey and Richardson, 2016). An additional threat posed by exotic plants may be their capacity to increase pathogen loads in the indigenous species with which they compete (Borer et al., 2007; Malmstrom et al., 2005). On the other hand, disease-causing plant viruses might also, over time, cause declines in the density and distribution of some exotic species and, ultimately, facilitate the recovery of indigenous species, as is possible for pathogens more generally (Flory and Clay, 2013).

2.5 Vector-Enabled Metagenomics (VEM): Exploring the Diversity of Plant Viruses Using Their Insect Vectors

Using ecogenomics or geometagenomics to characterize virus populations at the scale of whole ecosystems will likely yield only limited information about the transmission dynamics of viruses among plants within an ecosystem. VEM has therefore been developed to complement plant-focused metagenomics studies. VEM approaches generally investigate the diversity of plant viruses within their insect vectors. With VEM, insects are sampled

and, as with the ecogenomics and geometagenomics approaches, viral particles are purified and viral nucleic acids are sequenced (Ng et al., 2011a).

The VEM approach exploits the natural ability of highly mobile insect vectors to accumulate viruses from many plant species over time and space within an ecosystem (Ng et al., 2011a). The VEM approach can even be expanded to include the sampling of insect predators (such as dragonflies) that feed on plant virus transmission vectors (Rosario et al., 2012a). The VEM approach has been used on several insect vectors and insect vector predators such as leafhoppers (Fontenele et al., 2018; Kamali et al., 2017), earwigs (Godinho et al., 2017), whiteflies (Ng et al., 2011a; Rosario et al., 2014, 2015, 2016), and dragonflies (Rosario et al., 2013) to identify novel and known plant viruses. Besides enabling the discovery of dozens of previously unknown viruses, the VEM approach has prompted reevaluation of the geographical ranges of known viruses. For example, a VEM study focusing on Puerto Rican dragonflies identified the first mastrevirus genome from the Americas (Rosario et al., 2013)—a region long believed to be devoid of this important geminivirus genus. Since then, additional mastreviruses have been reported from the Americas using VEM or metagenomics-based approaches to analyze switchgrass (Agindotan et al., 2015), sugarcane (Boukari et al., 2017; Candresse et al., 2014), leafhoppers and maize (Fontenele et al., 2018).

3. DISCOVERY OF NEW VIRAL TAXA USING METAGENOMICS-BASED APPROACHES: THE GEMINIVIRUS TEST CASE

3.1 Establishing Taxonomic Groups From Metagenomics Datasets

One current limitation of viral metagenomics studies is the difficulty of assigning virus-derived sequencing reads to established taxonomic groups such as species, genera, and families. This is due in part to the fact that there is no robust means by which small genome fragments can be taxonomically classified with a high degree of accuracy. Therefore, most studies attempt to group sequences, usually using an arbitrarily defined degree of pairwise nucleotide identity, into so-called operational taxonomic units (OTUs) that might or might not reflect widely accepted taxonomic groups such as strains, species, genera, or families (Roossinck, 2012). This approach works well with commonly used pairwise sequence similarity searching approaches (e.g., BLASTn or BLASTx) and tentative classifications that are made using such approaches can be further refined using phylogenetic analyses (Bernardo et al., 2018).

In the geometagenomics study of Bernardo et al. (2018), 78% of detected plant virus OTUs shared 27%–75% identity (median = 49%) with isolates of known plant virus species, suggesting that these OTUs might represent novel species within 19 of the 22 plant virus families that are currently recognized by the ICTV (King et al., 2012; Roossinck, 2011a). Interestingly, ~81% of these apparently novel viruses were found within uncultivated plants, supporting the notion that the presently known crop-infecting plant virus species represent just a relatively small fraction of all terrestrial plant-infecting virus species (Wren et al., 2006). Interestingly, six OTUs from the Bernardo et al. (2018), study potentially represented highly divergent geminivirus species. Since many of the most important pathogens of cultivated crops that have emerged over the past 50 years have been geminiviruses, here we will look in more detail at these six OTUs and use classical molecular approaches to fully characterize them.

3.2 The Discovery and Characterization of Novel Geminiviruses

Four out of these six OTUs were from four different South African uncultivated endemic plant species (*E. caput-medusae*, *Limeum africanum*, *Exomis microphylla*, and *Polygala garcinii*). Whereas one of the two remaining OTUs that were discovered in France was identified in an uncultivated species (*Juncus maritimus*) the other was from a cultivated species (*Medicago sativa*, alfalfa).

While the first step of a top-down georeferenced metagenomics approach involves extensively sampling viral sequences from large numbers of plants each with known geographical coordinates, the second step involves the use of standard molecular approaches such as the cloning and full genome sequencing of the identified viruses. Therefore, DNA was extracted from geminivirus-infected samples from each of four species (*L. africanum*, *E. microphylla*, *P. garcinii*, and *J. maritimus*) using the DNeasy Plant Mini Kit (Qiagen, Germany). Extracted DNA was used as a template for PCR amplification of the complete genomes of the novel geminiviruses using pairs of abutting primers (designed based on assembled reads from the metagenomics step; Table 1), each of which spanned a restriction enzyme site. Amplicons from *L. africanum* were then cloned as described by Susi et al. (2017), whereas those from the other three plants were cloned as described by Bernardo et al. (2013). The cloned genomes were Sanger-sequenced using primer walking. Note that genomes recovered from *M. sativa* (alfalfa leaf curl virus, ALCV) and *E. caput-medusae* (*E. caput-medusae* latent virus, EcmLV)

Table 1 List of Abutting Primers Used for PCR Amplification of the Complete Genomes of the Four Novel Geminivirus Detected in the Western Cape Region of South Africa and the Rhône Delta Region of France

Host Plant	Virus[a]	Accession Number	Abutting Primers
Juncus maritimus	JmaV	MG001958	5′-TGGTACCAATTGAGCCGCTC-3′
			5′-GGGTACCCAAGGGCAATTTTG-3′
Polygala garcinii	PgaV	MG001959	5′-ATACTTCTAGATCAATTCGCTTCCTG-3′
			5′-CGGTCTCTAGACTGCGGTTGCACAA-3′
Exomis microphylla	EmaV	MG001960	5′-AGTCGACCTCTGGCTATCTC-3′
			5′-TGTCGACGAATGTGACCGTTA-3′
Limeum africanum	LaaV	MG001961	5′-TTCTTGAACTGAAGATAGGCCTCCTCTTC-3′
			5′-GGAGGCCCAGAAACTCCTACAGAA-3′

[a]JmaV, juncus maritimus-associated virus (2740 nts); PgaV, polygala garcinii-associated virus (2798 nts); EmaV, exomis microphylla-associated virus (2974 nts); and LaaV, limeum africanum-associated virus (2963 nts).

have already been reported on Bernardo et al. (2013) and Roumagnac et al. (2015) and were both assigned to the same new genus, *Capulavirus* (Varsani et al., 2017).

The four newly determined genome sequences were for juncus maritimus-associated virus (JmaV; MG001958; 2740 nts) from *J. maritimus*, polygala garcinii-associated virus (PgaV; MG001959; 2798 nts) from *P. garcinii*, limeum africanum-associated virus (LaaV; MG001961; 2963 nts) from *L. africanum*, and exomis microphylla-associated virus (EmaV; MG001960, 2974 nts) from *E. microphylla* (Fig. 2).

While three of these sequences have the characteristic geminivirus virion strand replication origin "TAATTATAC" nonanucleotide motif, the sequence from *E. microphylla* contains a "TAAGATTCC" nonanucleotide that has so far been found in two other geminiviruses: the eragrovirus eragrovirus curvula streak virus and the becurtovirus beet curly top Iran virus. The arrangement of open reading frames (ORFs) within the four novel geminivirus genomes are similar to those described previously for other geminiviruses (Fig. 2).

The four novel geminivirus genomes were aligned using MUSCLE (Edgar, 2004) with representative sequences from the nine geminivirus genera that are currently recognized by the ICTV and four sequences from novel geminiviruses that have not yet been assigned to genera by the ICTV. The alignment was used to infer a neighbor-joining (NJ) phylogenetic tree with a Jukes-Cantor nucleotide substitution model and 1000 bootstrap iterations in MEGA7 (Kumar et al., 2016). The NJ phylogenetic tree was midpoint rooted, and branches with <60% support were collapsed using TreeGraph2 (Stover and Muller, 2010).

The capsid protein (CP) and Rep amino acid sequences of the representative geminiviruses and the four new sequences reported here were aligned using MUSCLE (Edgar, 2004). The resulting alignments were used to infer maximum-likelihood phylogenetic trees using PHYML 3.0 (Guindon et al., 2010) using the rtREV + G and rtREV + G + I amino acid substitution models for CP and Rep, respectively, inferred to be the best-fitting models using ProtTest (Abascal et al., 2005). The trees were rooted with CP and Rep sequences of representative genomoviruses (Krupovic et al., 2016). Branches with <0.8 aLRT branch support were collapsed using TreeGraph2 (Stover and Muller, 2010). All pairwise identities (genome, CP, and Rep) were determined using SDT v1.2 (Muhire et al., 2014).

While the genome-wide pairwise nucleotide identities of the four new genomes were <60% to all other known geminiviruses (Fig. 3), phylogenetic

Fig. 2 Genome organizations of viruses belonging to the different geminivirus genera. *Abbreviations*: TYLCV, tomato yellow leaf curl virus, BGMV, bean golden mosaic virus, BCTIV, beet curly top Iran virus, EcmLV, Euphorbia caput-medusae latent virus, BCTV, beet curly top virus, GRBV, grapevine red blotch virus, ECSV, Eragrostis curvula streak virus, MSV, maize streak virus, TPCTV, tomato pseudo-curly top virus, TCTV, turnip curly top virus, CCDaV, citrus chlorotic dwarf associated virus, MMDav, mulberry mosaic dwarf associated virus, AGmV, apple geminivirus, TaGV1, tomato associated geminivirus 1, EmaV, Exomis microphylla associated virus, JmaV, Juncus maritimus associated virus, LaaV, Limeum africanum associated virus and PgaV, Polygala garcinii associated virus.

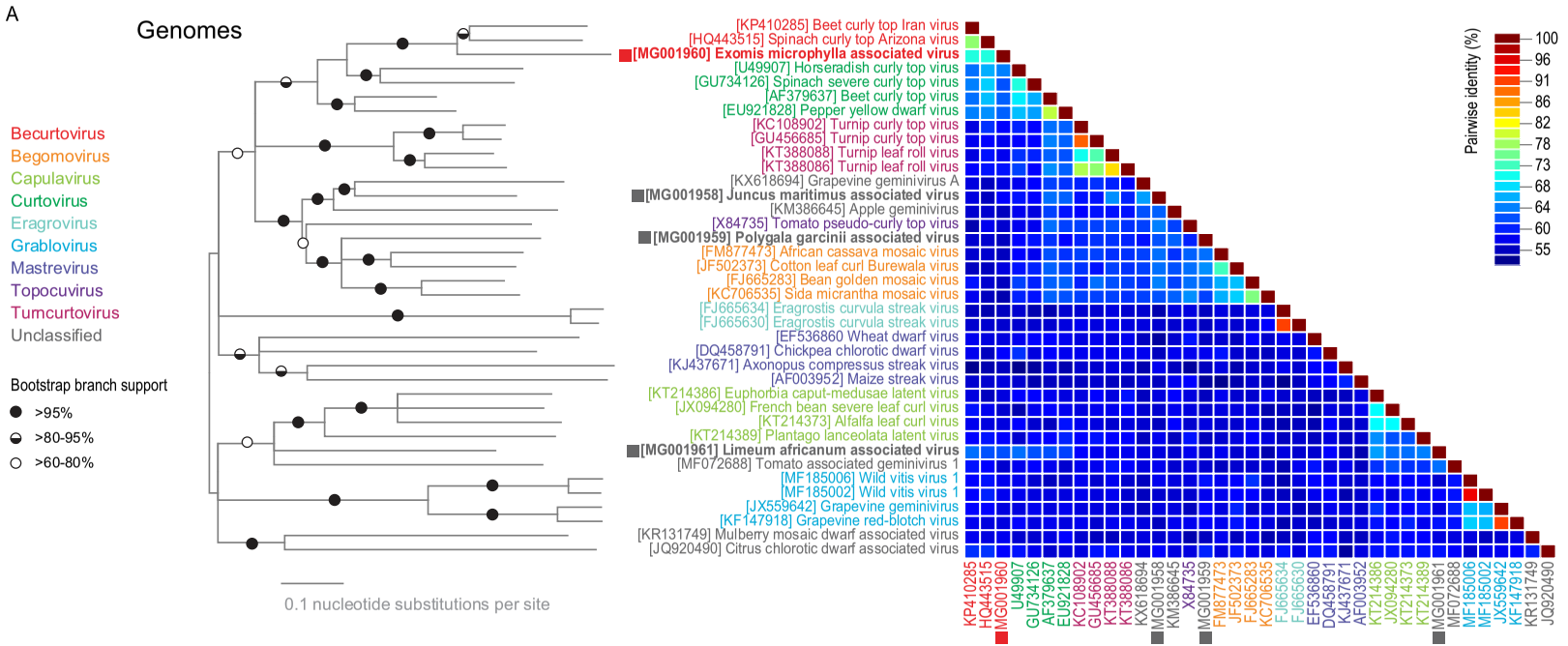


Fig. 3 Phylogenetic and nucleotide pairwise identity analyses. (A, *left*) Unrooted neighbor-joining tree inferred from aligned full-genome nucleotide sequences of representative isolates from the various geminivirus genera. (*Branch support*) *Solid dot*, >95% bootstrap support. *Half-filled dot*, 80%–95%, or *unfilled dots*, 60%–80%. *Branches with less than 60% bootstrap support* have been collapsed. The four novel geminiviruses discovered in France and South Africa are indicated with *red* and *gray squares*, to the *right*. (A, *right*) Genome-wide pairwise identities of representative isolates from various geminivirus genera as determined using SDT v1.2. (B) Maximum-likelihood phylogenetic trees, respectively, applying the rtREV + G and rtREV+G + I amino acid substitution models for CP (B, *left*) and Rep (B, *right*) of representative geminivirus genera. *Numbers associated with branches* indicate percentage aLRT support for these branches. *Both trees* are rooted with CP and Rep sequences of representative genomoviruses (not shown). *Branches with less than 80% aLRT support* have been collapsed.

(Continued)

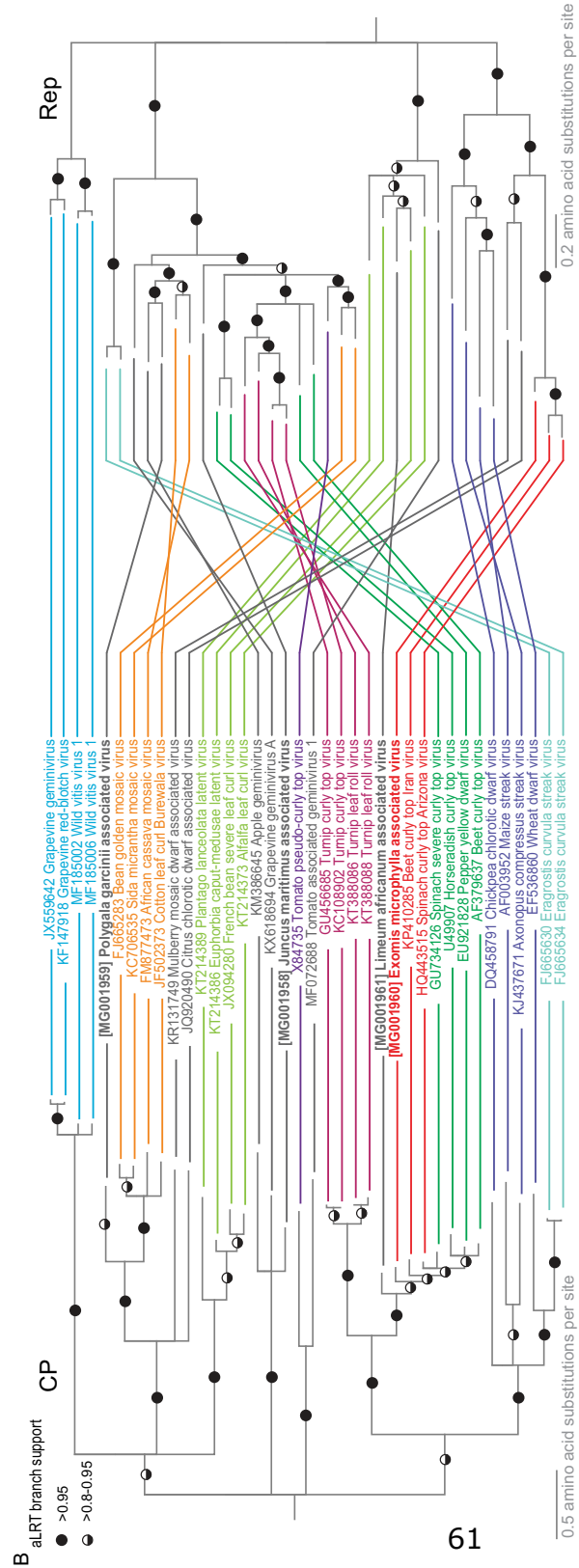

B
aLRT branch support
>0.95
>0.8-0.95
CP
Rep
JX559642 Grapevine geminivirus
KF147918 Grapevine red-blotch virus
MF185002 Wild vitis virus 1
MF185006 Wild vitis virus 1
[MG001959] Polygala garcinii associated virus
FJ665283 Bean golden mosaic virus
KC706535 Sida micrantha mosaic virus
FM877473 African cassava mosaic virus
JF502373 Cotton leaf curl Burewala virus
KR131749 Mulberry mosaic dwarf associated virus
JQ920490 Citrus chlorotic dwarf associated virus
KT214389 Plantago lanceolata latent virus
KT214386 Euphorbia caput-medusae latent virus
JX094280 French bean severe leaf curl virus
KT214373 Alfalfa leaf curl virus
KM386645 Apple geminivirus
KX618694 Grapevine geminivirus A
[MG001958] Juncus maritimus associated virus
X84735 Tomato pseudo-curly top virus
MF072688 Tomato associated geminivirus 1
GU456685 Turnip curly top virus
KC108902 Turnip curly top virus
KT388086 Turnip leaf roll virus
KT388088 Turnip leaf roll virus
[MG001961] Limeum africanum associated virus
[MG001960] Exomis microphylla associated virus
KP410285 Beet curly top Iran virus
HQ443515 Spinach severe curly top virus
GU734126 Spinach severe curly top virus
U49907 Horseradish curly top virus
EU921828 Pepper yellow dwarf virus
AF379637 Beet curly top virus
DQ458791 Chickpea chlorotic dwarf virus
AF003952 Maize streak virus
KJ437671 Axonopus compressus streak virus
EF538660 Wheat dwarf virus
FJ665630 Eragrostis curvula streak virus
FJ665634 Eragrostis curvula streak virus
0.5 amino acid substitutions per site
0.2 amino acid substitutions per site

Fig. 3—Cont'd

analysis of the genomes (Fig. 3) revealed that PgaV, EmaV, LaaV, and JmaV, respectively, cluster most closely with begomoviruses, becurtoviruses, capulaviruses, and an unclassified genus-level geminivirus lineage containing the two viruses, apple geminivirus, and grapevine geminivirus A (Al Rwahnih et al., 2016; Liang et al., 2015). EmaV is the first becurtovirus-like virus to be identified in Africa (Fig. 3).

PgaV and LaaV do not cluster with the sequences of any particular geminivirus genus but instead branch basal to the begomovirus and capulavirus clades, respectively. This might indicate that the full breadth of geminivirus diversity encompassed within the capulavirus and curtovirus genera is broader than previously known. If, however, it is discovered that these viruses have unique transmission vectors, it might be more appropriate to accommodate these within a new geminivirus genus (as was done for the capulaviruses; Varsani et al., 2017).

Consistent with previous reports (Briddon et al., 1996; Stanley et al., 1986), there is substantial incongruence between the phylogenies of the CP and Rep proteins (Fig. 3). It is plausible that this incongruence is a consequence of intergenus recombination events having played a part in the origin of some of the geminivirus genera (Bernardo et al., 2013; Briddon et al., 2010; Hernandez-Zepeda et al., 2013; Klute et al., 1996; Varsani et al., 2009). While both the Rep and CP sequences of PgaV cluster with begomovirus sequences and those of EmaV cluster with becurtovirus sequences, the Rep of LaaV clusters with capulaviruses and its CP clusters with becurtoviruses. Similarly, the Reps of JmaV, apple geminivirus, and grapevine geminivirus A are begomovirus-like but phylogenetically their CPs do not cluster closely with those of any other geminiviruses (Fig. 3).

The South African and French geometagenomics studies (Bernardo et al., 2018) therefore revealed an unexpected degree of geminivirus diversity within an area encompassing just $\sim 20\,km^2$ of the Earth's surface. Although nine geminivirus genera are now accepted by ITCV, when the samples in these studies were first collected in 2010–12, only four geminivirus genera were known (Zerbini et al., 2017). It is noteworthy that despite six completely novel geminivirus sequences having been identified in this study, not a single previously known geminivirus species was found. The six novel geminivirus sequences have an average pairwise Hamming distance (a proxy of the identity score) of 0.396 to one another, which is almost as high as the average pairwise Hamming distances (0.405) of the genome sequences representing the breadth of previously known geminivirus diversity (Fig. 3).

3.3 Divergent Newly Discovered Geminiviruses Such as the Capulaviruses Can Reveal Much About the Ecology and Evolution of Geminiviruses

Although geminiviruses have been well studied since the 1970s (Bock et al., 1974; Duffus and Gold, 1973), the discovery of several new major geminivirus lineages using metagenomic approaches, and the subsequent biological characterization of isolates representing some of these lineages, has illuminated several gaps in our knowledge regarding the ecology and evolution of this family (Agindotan et al., 2015; Bernardo et al., 2013; Boukari et al., 2017; Candresse et al., 2014; Fontenele et al., 2017; Kraberger et al., 2017a; Kreuze et al., 2009; Rosario et al., 2013; Susi et al., 2017). For example, a more in-depth characterization of the capulaviruses discovered in the French and South African metagenomics surveys has prompted the reevaluation of our understanding of geminivirus pathogenicity. Indeed, as with other recently discovered geminiviruses found in uncultivated hosts (Kraberger et al., 2017a; Perry et al., 2018), the capulaviruses EcmLV and *P. lanceolata* latent virus (PlLV) do not cause obvious infection symptoms in the uncultivated hosts from which they were initially isolated (Bernardo et al., 2013; Susi et al., 2017). If these viruses do indeed have decreased degrees of pathogenicity this could have arisen due to transmission being hampered within natural environments (Keesing et al., 2006), such that selection has favored less pathogenic virus variants which do not shorten their host's life span and can therefore be transmitted over longer time periods. It is plausible that geminiviruses such as EcmLV and PlLV are examples of viruses that, over hundreds or thousands of years, have become well adapted to infecting particular host species within particular natural environments.

It is noteworthy, however, that EcmlV has a broad experimental host range that includes tomato and *Nicotiana benthamiana*, in which this virus causes severe symptoms (Bernardo et al., 2013). This result possibly exemplifies the role that uncultivated plant species and their viruses might play during the emergence of new diseases following the introduction of exotic crop species into an ecosystem. The number of "new encounter" situations, where introduced cultivated plants first come in contact with viruses adapted to indigenous uncultivated plants, has increased tremendously with the progressive intensification and globalization of agriculture over the past two centuries (Jones, 2009). It is plausible that, if they are brought into contact with a susceptible cultivated host species, large numbers of presently unnoticed virus species (i.e., primarily those asymptomatically infecting

their natural hosts) such as EcmLV and PlLV, are at this moment poised to emerge as serious agricultural pathogens (Elena et al., 2014; Webster et al., 2007).

Unlike EcmLV and PlLV, the capulavirus discovered infecting the cultivated host, *M. sativa*, ALCV, is already potentially an emerging pathogen (Bernardo et al., 2016). ALCV causes severe disease symptoms in *M. sativa* and is transmissible by *Aphis craccivora* (Roumagnac et al., 2015)—an aphid species with a broad host range and almost worldwide distribution (CIE, 1983). The simple fact that ALCV is transmissible by an aphid rather than by any of the other previously known geminivirus transmission vectors (whiteflies, leafhoppers, and treehoppers; Zerbini et al., 2017) increases the risk that it could emerge as a serious pathogen.

The discovery that ALCV is aphid-transmitted supports the hypothesis that geminivirus coat protein sequences as a whole may be codiverging with geminivirus vector species (Rybicki, 1994). Under this evolutionary scenario, the most recent common ancestor of all the present day geminivirus vector species may have transmitted the most recent common ancestor of all the extant geminiviruses. As these vectors diversified over millions of years into aphids, treehoppers, leafhoppers, and whiteflies, the different geminivirus coat protein genes may have adapted over the same time scale to retain the capacities of their cognate viruses to be transmitted by these distinct insect lineages (Rybicki, 1994).

4. FROM METAGENOMICS TO BIOLOGICAL AND MOLECULAR CHARACTERIZATION OF PLANT VIRUSES: A CONCEPTUAL FRAMEWORK FOR IMPROVING OUR UNDERSTANDING OF VIRAL ECOLOGY AND EVOLUTION

Plant virus metagenomics studies have so far emphasized both that many more plant-infecting viruses remain undiscovered than those which are currently known, and that the geographical and host ranges of known viruses may be far broader than was previously believed. As exemplified by the geminivirus test case, a systematic top-down georeferenced metagenomics approach combining the power of the NGS-based approach for detecting and identifying viral sequences, and the reliability of classical molecular approaches for obtaining full-length genomes of newly discovered viruses, can represent a crucial step toward a better understanding of the ecology

and evolution of viruses that are important to humans. In the near future, metagenomics-based approaches (including RCA, ecogenomics, geometagenomics, and VEM), should be widely applicable to the epidemiological surveillance of viral pathogens, the routine diagnosis of viral diseases, and the large-scale spatiotemporal modeling of virus molecular evolution.

However, before the full potential of plant virus metagenomics is achieved, several technical issues need to be resolved. The first issue is that, with current technology, the full virus genomes and large viral sequence contigs that are assembled during metagenomics studies may be chimeras of reads from different viral genomes and, as such, may not actually exist. This problem reduces the usefulness of metagenomics sequences in phylogenetic-based inference of molecular clock rates, positive selection, geographical dispersal patterns, and population fluctuations. The second technical issue that must be addressed is the accurate differentiation of sequences: both between exogenous and endogenous virus sources, and between sequences derived from different components of segmented virus genomes. Consequently, the characterization of viruses discovered by metagenomics approaches at present still relies on full genome cloning from source material and Sanger sequencing.

Fortunately, third-generation sequencing techniques that are capable of much longer-read lengths (>15 kilobases on average) are under development by companies such as Pacific Biosciences and Oxford Nanopore (Goodwin et al., 2016). By negating the need to assemble full genomes from fragmentary short sequence reads, these new sequencing techniques should address both of the issues mentioned above. This will pave the way toward the direct use of metagenomics datasets in determining both the roles of plant viruses in global ecosystems, and which environmental conditions might prevent or precipitate the evolution of benign viruses into crop pathogens (Green et al., 2017; Shendure et al., 2017).

ACKNOWLEDGMENTS

Thank you to Bradley White for drawing Fig. 1. D.P.M., G.W.H., and A.V. have received research grants from the National Research Foundation of South Africa. P.R. has received an EU grant FP7-PEOPLE-2013-IOF (No. PIOF-GA-2013-622571). S.C., P.L., and J.M.L. were supported by the European Union (ERDF), the Conseil Régional de La Réunion and CIRAD. S.C. was a recipient of a PhD fellowship from the Agropolis Fondation (E-Space) and CIRAD. D.P.M., G.W.H., A.V., S.C., P.L., J.M.L., and P.R. declare that they have no conflict of interest.

REFERENCES

Abascal, F., Zardoya, R., Posada, D., 2005. Prottest: selection of best-fit models of protein evolution. Bioinformatics 21, 2104–2105.

Accotto, G.P., Navas-Castillo, J., Noris, E., Moriones, E., Louro, D., 2000. Typing of tomato yellow leaf curl viruses in Europe. Eur. J. Plant Pathol. 106, 179–186.

Agindotan, B.O., Domier, L.L., Bradley, C.A., 2015. Detection and characterization of the first north American Mastrevirus in switchgrass. Arch. Virol. 160, 1313–1317.

Agrawal, A.A., Lau, J.A., Hamback, P.A., 2006. Community heterogeneity and the evolution of interactions between plants and insect herbivores. Q. Rev. Biol. 81, 349–376.

Al Rwahnih, M., Alabi, O.J., Westrick, N.M., Golino, D., Rowhani, A., 2016. Description of A novel monopartite geminivirus and its defective subviral genome in grapevine. Phytopathology. Phyto07160282r.

Alberti, A., Poulain, J., Engelen, S., Labadie, K., Romac, S., Ferrera, I., Albini, G., Aury, J.M., Belser, C., Bertrand, A., Cruaud, C., Da Silva, C., Dossat, C., Gavory, F., Gas, S., Guy, J., Haquelle, M., Jacoby, E., Jaillon, O., Lemainque, A., Pelletier, E., Samson, G., Wessner, M., Acinas, S.G., Royo-Llonch, M., Cornejo-Castillo, F.M., Logares, R., Fernandez-Gomez, B., Bowler, C., Cochrane, G., Amid, C., Ten Hoopen, P., De Vargas, C., Grimsley, N., Desgranges, E., Kandels-Lewis, S., Ogata, H., Poulton, N., Sieracki, M.E., Stepanauskas, R., Sullivan, M.B., Brum, J.R., Duhaime, M.B., Poulos, B.T., Hurwitz, B.L., Pesant, S., Karsenti, E., Wincker, P., Team, G.T., Consortium, T.O., 2017. Viral to metazoan marine plankton nucleotide sequences from the Tara oceans expedition. Sci. Data 4, 170093.

Alexander, H.M., Mauck, K.E., Whitfield, A.E., Garrett, K.A., Malmstrom, C.M., 2014. Plant-virus interactions and the agro-ecological interface. Eur. J. Plant Pathol. 138, 529–547.

Allan, B.F., Keesing, F., Ostfeld, R.S., 2003. Effect of forest fragmentation on Lyme disease risk. Conserv. Biol. 17, 267–272.

Allander, T., Emerson, S.U., Engle, R.E., Purcell, R.H., Bukh, J., 2001. A virus discovery method incorporating DNase treatment and its application to the identification of two bovine parvovirus species. Proc. Natl. Acad. Sci. U.S.A 98, 11609–11614.

Amann, R.I., Ludwig, W., Schleifer, K.H., 1995. Phylogenetic identification and in-situ detection of individual microbial-cells without cultivation. Microbiol. Rev. 59, 143–169.

Bernardo, P., Golden, M., Akram, M., Naimuddin, Nadarajan, N., Fernandez, E., Granier, M., Rebelo, A.G., Peterschmitt, M., Martin, D.P., Roumagnac, P., 2013. Identification and characterisation of a highly divergent Geminivirus: evolutionary and taxonomic implications. Virus Res. 177, 35–45.

Bernardo, P., Muhire, B., Francois, S., Deshoux, M., Hartnady, P., Farkas, K., Kraberger, S., Filloux, D., Fernandez, E., Galzi, S., Ferdinand, R., Granier, M., Marais, A., Monge Blasco, P., Candresse, T., Escriu, F., Varsani, A., Harkins, G.W., Martin, D.P., Roumagnac, P., 2016. Molecular characterization and prevalence of two Capulaviruses: alfalfa leaf curl virus from France and Euphorbia caput-medusae latent virus from South Africa. Virology 493, 142–153.

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

Bock, K.R., Guthrie, E.J., Woods, R.D., 1974. Purification of maize streak virus and its relationship to viruses associated with streak diseases of sugarcane and Panicum maximum. Ann. Appl. Biol. 77, 289–296.

Borer, E.T., Hosseini, P.R., Seabloom, E.W., Dobson, A.P., 2007. Pathogen-induced reversal of native dominance in a grassland community. Proc. Natl. Acad. Sci. U. S. A. 104, 5473–5478.
Bos, L., 1999. Beijerinck's work on tobacco mosaic virus: historical context and legacy. Philos. Trans. R. Soc. Lond. B Biol. Sci. 354, 675–685.
Boukari, W., Alcala-Briseno, R.I., Kraberger, S., Fernandez, E., Filloux, D., Daugrois, J.H., Comstock, J.C., Lett, J.M., Martin, D.P., Varsani, A., Roumagnac, P., Polston, J.E., Rott, P.C., 2017. Occurrence of a novel Mastrevirus in sugarcane germplasm collections in Florida, guadeloupe and reunion. Virol. J. 14, 146.
Breitbart, M., Rohwer, F., 2005. Here a virus, there a virus, everywhere the same virus? Trends Microbiol. 13, 278–284.
Breitbart, M., Salamon, P., Andresen, B., Mahaffy, J.M., Segall, A.M., Mead, D., Azam, F., Rohwer, F., 2002. Genomic analysis of uncultured marine viral communities. Proc. Natl. Acad. Sci. U. S. A. 99, 14250–14255.
Breitbart, M., Hewson, I., Felts, B., Mahaffy, J.M., Nulton, J., Salamon, P., Rohwer, F., 2003. Metagenomic analyses of an uncultured viral community from human feces. J. Bacteriol. 185, 6220–6223.
Breitbart, M., Felts, B., Kelley, S., Mahaffy, J.M., Nulton, J., Salamon, P., Rohwer, F., 2004. Diversity and population structure of a near-shore marine-sediment viral community. Proc. Biol. Sci. 271, 565–574.
Briddon, R.W., Markham, P.G., 2000. Cotton leaf curl virus disease. Virus Res. 71, 151–159.
Briddon, R.W., Bedford, I.D., Tsai, J.H., Markham, P.G., 1996. Analysis of the nucleotide sequence of the treehopper-transmitted Geminivirus, tomato pseudo-curly top virus, suggests a recombinant origin. Virology 219, 387–394.
Briddon, R.W., Heydarnejad, J., Khosrowfar, F., Massumi, H., Martin, D.P., Varsani, A., 2010. Turnip curly top virus, a highly divergent Geminivirus infecting turnip in Iran. Virus Res. 152, 169–175.
Brum, J.R., Ignacio-Espinoza, J.C., Roux, S., Doulcier, G., Acinas, S.G., Alberti, A., Chaffron, S., Cruaud, C., De Vargas, C., Gasol, J.M., Gorsky, G., Gregory, A.C., Guidi, L., Hingamp, P., Iudicone, D., Not, F., Ogata, H., Pesant, S., Poulos, B.T., Schwenck, S.M., Speich, S., Dimier, C., Kandels-Lewis, S., Picheral, M., Searson, S., Tara Oceans, C., Bork, P., Bowler, C., Sunagawa, S., Wincker, P., Karsenti, E., Sullivan, M.B., 2015. Ocean plankton. 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, D.P., Varsani, A., Roumagnac, P., 2014. Appearances can be deceptive: revealing a hidden viral infection with deep sequencing in a plant quarantine context. PLoS One 9, E102945.
CIE, 1983. Distribution Maps of Plant Pests. Cab International, Wallingford, UK.
Cooper, I., Jones, R.A., 2006. Wild plants and viruses: under-investigated ecosystems. Adv. Virus Res. 67, 1–47.
Dayaram, A., Galatowitsch, M.L., Arguello-Astorga, G.R., Van Bysterveldt, K., Kraberger, S., Stainton, D., Harding, J.S., Roumagnac, P., Martin, D.P., Lefeuvre, P., Varsani, A., 2016. Diverse circular replication-associated protein encoding viruses circulating in invertebrates within a Lake ecosystem. Infect. Genet. Evol. 39, 304–316.
Delwart, E.L., 2007. Viral metagenomics. Rev. Med. Virol. 17, 115–131.
Downey, P.O., Richardson, D.M., 2016. Alien plant invasions and native plant extinctions: a six-threshold framework. AoB Plants 8.
Duffus, J.E., Gold, A.H., 1973. Infectivity neutralization used in serological tests with partially purified beet curly top virus. Phytopathology 63, 1107–1110.

Duffy, S., Holmes, E.C., 2008. Phylogenetic evidence for rapid rates of molecular evolution in the single-stranded Dna Begomovirus tomato yellow leaf curl virus. J. Virol. 82, 957–965.
Edgar, R.C., 2004. Muscle: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 32, 1792–1797.
Elena, S.F., 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, J.M., 2006. Molecular ecology and emergence of tropical plant viruses. Annu. Rev. Phytopathol. 44, 235–260.
Flory, S.L., Clay, K., 2013. Pathogen accumulation and long-term dynamics of plant invasions. J. Ecol. 101, 607–613.
Fontenele, R.S., Lamas, N.S., Lacorte, C., Lacerda, A.L.M., Varsani, A., Ribeiro, S.G., 2017. A novel Geminivirus identified in tomato and cleome plants sampled in Brazil. Virus Res. 240, 175–179.
Fontenele, R.S., Alves-Freitas, D.M.T., Silva, P.I.T., Foresti, J., Silva, P.R., Godinho, M.T., Varsani, A., Ribeiro, S.G., 2018. Discovery of the first maize-infecting Mastrevirus in the Americas using a vector-enabled metagenomics approach. Arch. Virol. 163, 263–267.
Gilbert, J.A., Dupont, C.L., 2011. Microbial metagenomics: beyond the genome. Ann. Rev. Mar. Sci. 3, 347–371.
Godinho, M.T., Paula, D.P., Varsani, A., Ribeiro, S.G., 2017. Genome sequence of cauliflower mosaic virus identified in earwigs (Doru Luteipes) through a metagenomic approach. Genome Announc. 5, e00043-17.
Goodwin, S., Mcpherson, J.D., Mccombie, W.R., 2016. Coming of age: ten years of next-generation sequencing technologies. Nat. Rev. Genet. 17, 333–351.
Green, E.D., Rubin, E.M., Olson, M.V., 2017. The future of Dna sequencing. Nature 550, 179–181.
Guindon, S., Dufayard, J.F., Lefort, V., Anisimova, M., Hordijk, W., Gascuel, O., 2010. New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of Phyml 3.0. Syst. Biol. 59, 307–321.
Handelsman, J., Rondon, M.R., Brady, S.F., Clardy, J., Goodman, R.M., 1998. Molecular biological access to the chemistry of unknown soil microbes: a new frontier for natural products. Chem. Biol. 5, R245–249.
Harkins, G.W., Delport, W., Duffy, S., Wood, N., Monjane, A.L., Owor, B.E., Donaldson, L., Saumtally, S., Triton, G., Briddon, R.W., Shepherd, D.N., Rybicki, E.P., Martin, D.P., Varsani, A., 2009a. Experimental evidence indicating that Mastreviruses probably did not co-diverge with their hosts. Virol. J. 6, 104.
Harkins, G.W., Martin, D.P., Duffy, S., Monjane, A.L., Shepherd, D.N., Windram, O.P., Owor, B.E., Donaldson, L., Van Antwerpen, T., Sayed, R.A., Flett, B., Ramusi, M., Rybicki, E.P., Peterschmitt, M., Varsani, A., 2009b. Dating the origins of the maize-adapted strain of maize streak virus, Msv-a. J. Gen. Virol. 90, 3066–3074.
Hernandez-Zepeda, C., Varsani, A., Brown, J.K., 2013. Intergeneric recombination between a new, spinach-infecting curtovirus and a new geminivirus belonging to the genus becurtovirus: first new world exemplar. Arch. Virol. 158, 2245–2254.
Jones, R.A.C., 2009. Plant virus emergence and evolution: origins, new encounter scenarios, factors driving emergence, effects of changing world conditions, and prospects for control. Virus Res. 141, 113–130.
Kamali, M., Heydarnejad, J., Pouramini, N., Masumi, H., Farkas, K., Kraberger, S., Varsani, A., 2017. Genome sequences of beet curly top Iran virus, oat dwarf virus, turnip curly top virus, and wheat dwarf virus identified in leafhoppers. Genome Announc. 5, e01674-16.

Keesing, F., Holt, R.D., Ostfeld, R.S., 2006. Effects of species diversity on disease risk. Ecol. Lett. 9, 485–498.
Keesing, F., Belden, L.K., Daszak, P., Dobson, A., Harvell, C.D., Holt, R.D., Hudson, P., Jolles, A., Jones, K.E., Mitchell, C.E., Myers, S.S., Bogich, T., Ostfeld, R.S., 2010. Impacts of biodiversity on the emergence and transmission of infectious diseases. Nature 468, 647–652.
Kelley, S.E., 1994. Viral pathogens and the advantage of sex inthe perennial grass Anthoxanthum odoratum. Philos. Trans. R. Soc. Lond. B Biol. Sci. 346 (1317), 295–302.
King, A.M.Q., Adams, M.J., Carstens, E.B., Lefkowitz, E.J., 2012. In: King, A.M.Q., Adams, M.J., Carstens, E.B., Lefkowitz, E.J. (Eds.), Virus taxonomy. classification and nomenclature of viruses. Ninth Report of the International committee on Taxonomy of Viruses. Elsevier, Amsterdam, Boston, Heidelberg, London, New York, Oxford, Paris, San Diego, San Francisco, Singapore, Sydney, Tokyo.
Klute, K.A., Nadler, S.A., Stenger, D.C., 1996. Horseradish curly top virus is a distinct subgroup II geminivirus species with rep and C4 genes derived from a subgroup III ancestor. J. Gen. Virol. 77 (Pt 7), 1369–1378.
Kraberger, S., Geering, A.D.W., Walters, M., Martin, D.P., Varsani, A., 2017a. Novel Mastreviruses identified in Australian wild rice. Virus Res. 238, 193–197.
Kraberger, S., Saumtally, S., Pande, D., Khoodoo, M.H.R., Dhayan, S., Dookun-Saumtally, A., Shepherd, D.N., Hartnady, P., Atkinson, R., Lakay, F.M., Hanson, B., Redhi, D., Monjane, A.L., Windram, O.P., Walters, M., Oluwafemi, S., Michel-Lett, J., Lefeuvre, P., Martin, D.P., Varsani, A., 2017b. Molecular diversity, geographic distribution and host range of monocot-infecting Mastreviruses in Africa and surrounding Islands. Virus Res. 238, 171–178.
Kreuze, J.F., 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.
Krupovic, M., Ghabrial, S.A., Jiang, D., Varsani, A., 2016. Genomoviridae: a new family of widespread single-stranded DNA viruses. Arch. Virol. 161, 2633–2643.
Kumar, S., Stecher, G., Tamura, K., 2016. Mega7: molecular evolutionary genetics analysis version 7.0 for bigger datasets. Mol. Biol. Evol. 33, 1870–1874.
Lacroix, C., Jolles, A., Seabloom, E.W., Power, A.G., Mitchell, C.E., Borer, E.T., 2014. Non-random biodiversity loss underlies predictable increases in viral disease prevalence. J. R. Soc. Interface 11, 20130947.
Lecuit, M., Eloit, M., 2014. The diagnosis of infectious diseases by whole genome next generation sequencing: a new era is opening. Front. Cell. Infect. Microbiol. 4.
Lefeuvre, P., Martin, D.P., Harkins, G., Lemey, P., Gray, A.J., Meredith, S., Lakay, F., Monjane, A., Lett, J.M., Varsani, A., Heydarnejad, J., 2010. The spread of tomato yellow leaf curl virus from the middle east to the world. PLoS Pathog. 6, E1001164.
Liang, P., Navarro, B., Zhang, Z., Wang, H., Lu, M., Xiao, H., Wu, Q., Zhou, X., Di Serio, F., Li, S., 2015. Identification and characterization of a novel geminivirus with monopartite genome infecting apple trees. J. Gen. Virol. 96, 2411–2420.
Malmstrom, C.M., Mccullough, A.J., Johnson, H.A., Newton, L.A., Borer, E.T., 2005. Invasive annual grasses indirectly increase virus incidence in California native perennial bunchgrasses. Oecologia 145, 153–164.
Malmstrom, C.M., Melcher, U., Bosque-Perez, N.A., 2011. The expanding field of plant virus ecology: historical foundations, knowledge gaps, and research directions. Virus Res. 159, 84–94.
Márquez, L.M., Redman, R.S., Rodriguez, R.J., Roossinck, M.J., 2007. A virus in a fungus in a plant: three-way symbiosis required for thermal tolerance. Science 315, 513–515.
Mitchell, C.E., Tilman, D., Groth, J.V., 2002. Effects of grassland plant species diversity, abundance, and composition on foliar fungal disease. Ecology 83, 1713–1726.

Monjane, A.L., Harkins, G.W., Martin, D.P., Lemey, P., Lefeuvre, P., Shepherd, D.N., Oluwafemi, S., Simuyandi, M., Zinga, I., Komba, E.K., Lakoutene, D.P., Mandakombo, N., Mboukoulida, J., Semballa, S., Tagne, A., Tiendrebeogo, F., Erdmann, J.B., Van Antwerpen, T., Owor, B.E., Flett, B., Ramusi, M., Windram, O.P., Syed, R., Lett, J.M., Briddon, R.W., Markham, P.G., Rybicki, E.P., Varsani, A., 2011. Reconstructing the history of maize streak virus strain a dispersal to reveal diversification hot spots and its origin in southern Africa. J. Virol. 85, 9623–9636.

Muhire, B.M., Varsani, A., Martin, D.P., 2014. Sdt: a virus classification tool based on pairwise sequence alignment and identity calculation. PLoS One 9, E108277.

Muthukumar, V., Melcher, U., Pierce, M., Wiley, G.B., Roe, B.A., Palmer, M.W., Thapa, V., Ali, A., Ding, T., 2009. Non-cultivated plants of the tallgrass prairie preserve of northeastern Oklahoma frequently contain virus-like sequences in particulate fractions. Virus Res. 141, 169–173.

Ng, T.F.F., Duffy, S., Polston, J.E., Bixby, E., Vallad, G.E., Breitbart, M., 2011a. Exploring the diversity of plant DNA viruses and their satellites using vector-enabled metagenomics on whiteflies. PLoS One 6, E19050.

Ng, T.F.F., Willner, D.L., Lim, Y.W., Schmieder, R., Chau, B., Nilsson, C., Anthony, S., Ruan, Y.J., Rohwer, F., Breitbart, M., 2011b. Broad surveys of DNA viral diversity obtained through viral metagenomics of mosquitoes. PLoS One 6, e20579.

Pace, N.R., Stahl, D.A., Lane, D.J., Olsen, G.J., 1985. Analyzing natural microbial populations by rRNA sequences. ASM News 51, 4–12.

Pagan, I., Gonzalez-Jara, P., Moreno-Letelier, A., Rodelo-Urrego, M., Fraile, A., Pinero, D., Garcia-Arenal, F., 2012. Effect of biodiversity changes in disease risk: exploring disease emergence in a plant-virus system. PLoS Pathog. 8, E1002796.

Pande, D., Madzokere, E., Hartnady, P., Kraberger, S., Hadfield, J., Rosario, K., Jaschke, A., Monjane, A.L., Owor, B.E., Dida, M.M., Shepherd, D.N., Martin, D.P., Varsani, A., Harkins, G.W., 2017. The role of Kenya in the trans-African spread of maize streak virus strain a. Virus Res. 232, 69–76.

Patil, B.L., Fauquet, C.M., 2009. Cassava mosaic geminiviruses: actual knowledge and perspectives. Mol. Plant Pathol. 10, 685–701.

Perry, K.L., Mclane, H., Thompson, J.R., Fuchs, M., 2018. A novel grablovirus from non-cultivated grapevine (vitis Sp.) in north America. Arch. Virol. 163, 259–262.

Remold, S.K., 2002. Unapparent virus infection and host fitness in three weedy grass species. J. Ecol. 90 (6), 967–977.

Roche, B., Dobson, A.P., Guegan, J.F., Rohani, P., 2012. Linking community and disease ecology: the impact of biodiversity on pathogen transmission. Philos. Trans. R. Soc., B 367, 2807–2813.

Roossinck, M.J., 2011a. The big unknown: plant virus biodiversity. Curr. Opin. Virol. 1, 63–67.

Roossinck, M.J., 2011b. The good viruses: viral mutualistic symbioses. Nat. Rev. Microbiol. 9, 99–108.

Roossinck, M.J., 2012. Plant virus metagenomics: biodiversity and ecology. Annu. Rev. Genet. 46, 359–369.

Roossinck, M.J., 2015. Move over, bacteria! Viruses make their mark as mutualistic microbial symbionts. J. Virol. 89, 6532–6535.

Roossinck, M.J., Garcia-Arenal, F., 2015. Ecosystem simplification, biodiversity loss and plant virus emergence. Curr. Opin. Virol. 10c, 56–62.

Roossinck, M.J., Saha, P., Wiley, G., Quan, J., White, J., Lai, H., Chavarria, F., Shen, G., Roe, B., 2010. Ecogenomics: using massively parallel pyrosequencing to understand virus ecology. Mol. Ecol. 19, 81–88.

Roossinck, M.J., Martin, D.P., Roumagnac, P., 2015. Plant virus metagenomics: advances in virus discovery. Phytopathology 105, 716–727.

Rosario, K., Breitbart, M., 2011. Exploring the viral world through metagenomics. Curr. Opin. Virol. 1, 1–9.
Rosario, K., Dayaram, A., Marinov, M., Ware, J., Kraberger, S., Stainton, D., Breitbart, M., Varsani, A., 2012a. Diverse circular ssdna viruses discovered in dragonflies (Odonata: Epiprocta). J. Gen. Virol. 93, 2668–2681.
Rosario, K., Duffy, S., Breitbart, M., 2012b. A field guide to eukaryotic circular single-stranded Dna viruses: insights gained from metagenomics. Arch. Virol. 157, 1851–1871.
Rosario, K., Padilla-Rodriguez, M., Kraberger, S., Stainton, D., Martin, D.P., 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.
Rosario, K., Capobianco, H., Ng, T.F., Breitbart, M., Polston, J.E., 2014. Rna viral metagenome of whiteflies leads to the discovery and characterization of a whitefly-transmitted Carlavirus in North America. PLoS One 9, E86748.
Rosario, K., Seah, Y.M., Marr, C., Varsani, A., Kraberger, S., Stainton, D., Moriones, E., Polston, J.E., Duffy, S., Breitbart, M., 2015. Vector-enabled metagenomic (Vem) surveys using whiteflies (Aleyrodidae) reveal novel Begomovirus species in the new and oldworlds. Virus 7, 5553–5570.
Rosario, K., Marr, C., Varsani, A., Kraberger, S., Stainton, D., Moriones, E., Polston, J.E., Breitbart, M., 2016. Begomovirus-associated satellite DNA diversity captured through vector-enabled metagenomic (Vem) surveys using whiteflies (Aleyrodidae). Virus 8, 36.
Roumagnac, P., Granier, M., Bernardo, P., Deshoux, M., Ferdinand, R., Galzi, S., Fernandez, E., Julian, C., Abt, I., Filloux, D., Mesleard, F., Varsani, A., Blanc, S., Martin, D.P., Peterschmitt, M., 2015. Alfalfa leaf curl virus: an aphid-transmitted geminivirus. J. Virol. 89, 9683–9688.
Roux, S., Enault, F., Robin, A., Ravet, V., Personnic, S., Theil, S., Colombet, J., Sime-Ngando, T., Debroas, D., 2012. Assessing the diversity and specificity of two freshwater viral communities through metagenomics. PLoS One 7.
Rybicki, E.P., 1994. A phylogenetic and evolutionary justification for 3 genera of Geminiviridae. Arch. Virol. 139, 49–77.
Scarpellini, E., Ianiro, G., Attili, F., Bassanelli, C., De Santis, A., Gasbarrini, A., 2015. The human gut microbiota and virome: potential therapeutic implications. Dig. Liver Dis. 47 (12), 1007.
Shendure, J., Balasubramanian, S., Church, G.M., Gilbert, W., Rogers, J., Schloss, J.A., Waterston, R.H., 2017. DNA sequencing at 40: past, present and future. Nature 550, 345–353.
Simmonds, P., Adams, M.J., Benko, M., Breitbart, M., Brister, J.R., Carstens, E.B., Davison, A.J., Delwart, E., Gorbalenya, A.E., Harrach, B., Hull, R., King, A.M., Koonin, E.V., Krupovic, M., Kuhn, J.H., Lefkowitz, E.J., Nibert, M.L., Orton, R., Roossinck, M.J., Sabanadzovic, S., Sullivan, M.B., Suttle, C.A., Tesh, R.B., Van Der Vlugt, R.A., Varsani, A., Zerbini, F.M., 2017. Consensus statement: virus taxonomy in the age of metagenomics. Nat. Rev. Microbiol. 15, 161–168.
Stanley, J., Markham, P.G., Callis, R.J., Pinner, M.S., 1986. The nucleotide sequence of an infectious clone of the geminivirus beet curly top virus. EMBO J. 5, 1761–1767.
Stecher, B., Denzler, R., Maier, L., Bernet, F., Sanders, M.J., Pickard, D.J., Barthel, M., Westendorf, A.M., Krogfelt, K.A., Walker, A.W., Ackermann, M., Dobrindt, U., Thomson, N.R., Hardt, W.D., 2012. Gut inflammation can boost horizontal gene transfer between pathogenic and commensal Enterobacteriaceae. Proc. Natl. Acad. Sci. U. S. A. 109, 1269–1274.
Stobbe, A.H., Roossinck, M.J., 2014. Plant virus metagenomics: what we know and why we need to know more. Front. Plant Sci. 5, 150.
Stover, B.C., Muller, K.F., 2010. Treegraph 2: combining and visualizing evidence from different phylogenetic analyses. BMC Bioinf. 11, 7.

Sunagawa, S., Coelho, L.P., Chaffron, S., Kultima, J.R., Labadie, K., Salazar, G., Djahanschiri, B., Zeller, G., Mende, D.R., Alberti, A., Cornejo-Castillo, F.M., Costea, P.I., Cruaud, C., D'ovidio, F., Engelen, S., Ferrera, I., Gasol, J.M., Guidi, L., Hildebrand, F., Kokoszka, F., Lepoivre, C., Lima-Mendez, G., Poulain, J., Poulos, B.T., Royo-Llonch, M., Sarmento, H., Vieira-Silva, S., Dimier, C., Picheral, M., Searson, S., Kandels-Lewis, S., Tara Oceans, C., Bowler, C., De Vargas, C., Gorsky, G., Grimsley, N., Hingamp, P., Iudicone, D., Jaillon, O., Not, F., Ogata, H., Pesant, S., Speich, S., Stemmann, L., Sullivan, M.B., Weissenbach, J., Wincker, P., Karsenti, E., Raes, J., Acinas, S.G., Bork, P., 2015. Ocean Plankton. Structure and function of the global ocean microbiome. Science 348, 1261359.

Susi, H., Laine, A.L., Filloux, D., Kraberger, S., Farkas, K., Bernardo, P., Frilander, M.J., Martin, D.P., Varsani, A., Roumagnac, P., 2017. Genome sequences of a capulavirus infecting *Plantago lanceolata* in the Aland Archipelago of Finland. Arch. Virol. 162, 2041–2045.

Suttle, C.A., 2005. Viruses in the sea. Nature 437, 356–361.

Suttle, C.A., 2007. Marine viruses—major players in the global ecosystem. Nat. Rev. Microbiol. 5, 801–812.

Temperton, B., Giovannoni, S.J., 2012. Metagenomics: microbial diversity through a scratched lens. Curr. Opin. Microbiol. 15, 605–612.

Varsani, A., Shepherd, D.N., Dent, K., Monjane, A.L., Rybicki, E.P., Martin, D.P., 2009. A highly divergent south African geminivirus species illuminates the ancient evolutionary history of this family. Virol. J. 6, 36.

Varsani, A., Roumagnac, P., Fuchs, M., Navas-Castillo, J., Moriones, E., Idris, A., Briddon, R.W., Rivera-Bustamante, R., Zerbini, F.M., Martin, D.P., 2017. Capulavirus and Grablovirus: two new genera in the family *Geminiviridae*. Arch. Virol. 162 (6), 1819–1831.

Vayssier-Taussat, M., Albina, E., Citti, C., Cosson, J.F., Jacques, M.A., Lebrun, M.H., Le Loir, Y., Ogliastro, M., Petit, M.A., Roumagnac, P., Candresse, T., 2014. Shifting the paradigm from pathogens to Pathobiome: new concepts in the light of meta-omics. Front. Cell. Infect. Microbiol. 4, 29.

Wasik, B.R., Turner, P.E., 2013. On the biological success of viruses. Annu. Rev. Microbiol. 67, 519–541.

Webster, C.G., Coutts, B.A., Jones, R.A.C., Jones, M.G.K., Wylie, S.J., 2007. Virus impact at the Interface of an ancient ecosystem and a recent agroecosystem: studies on three legume-infecting Potyviruses in the southwest Australian floristic region. Plant Pathol. 56, 729–742.

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, P., Chen, F., Mannas, J.P., Feldman, T., Sumner, L.W., Roossinck, M.J., 2008. Virus infection improves drought tolerance. New Phytol. 180 (4), 911–921.

Zaitlin, M., Palukaitis, P., 2000. Advances in understanding plant viruses and virus diseases. Annu. Rev. Phytopathol. 38, 117–143.

Zerbini, F.M., Briddon, R.W., Idris, A., Martin, D.P., Moriones, E., Navas-Castillo, J., Rivera-Bustamante, R., Roumagnac, P., Varsani, A., Ictv Report, C., 2017. Ictv virus taxonomy profile: Geminiviridae. J. Gen. Virol. 98, 131–133.

Ricklefs, R.E., Miller, G.L., 1999. Ecology, fourth ed. WH Freeman and Company, New York City, NY.
Rico, P., Ivars, P., Elena, S.F., Hernandez, C., 2006. Insights into the selective pressures restricting Pelargonium flower break virus genome variability: evidence for host adaptation. J. Virol. 80, 8124–8132.
Roche, B., Drake, J.M., Brown, J., Stallknecht, D.E., Bedford, T., Rohani, P., 2014. Adaptive evolution and environmental durability jointly structure phylodynamic patterns in Avian influenza viruses. PLoS Biol. 12, e1001931.
Rodelo-Urrego, M., Pagán, I., González-Jara, P., Betancourt, M., Moreno-Letelier, A., Ayllón, M.A., et al., 2013. Landscape heterogeneity shapes host-parasite interactions and results in apparent plant-virus codivergence. Mol. Ecol. 22, 2325–2340.
Roossinck, M.J., 2015. Plants, viruses and the environment: ecology and mutualism. Virology 479, 271–277.
Roossinck, M.J., García-Arenal, F., 2015. Ecosystem simplification, biodiversity loss and plant virus emergence. Curr. Opin. Virol. 10, 56–62.
Roossinck, M.J., Saha, P., Wiley, G.B., Quan, J., White, J.D., Lai, H., Chavarria, F., Shen, G., Roe, B.A., 2010. Ecogenomics: using massively parallel pyrosequencing to understand virus ecology. Mol. Ecol. 19, 81–88.
Roossinck, M.J., Martin, D.P., Roumagnac, P., 2015. Plant virus metagenomics: advances in virus discovery. Phytopathology 105, 716–727.
Rosenzweig, M.L., 1995. Species Diversity in Space and Time. Cambridge University Press, Cambridge, UK.
Rosindell, J., Cornell, S.J., 2013. Universal scaling of species-abundance distributions across multiple scales. Oikos 122, 1101–1111.
Rottstock, T., Joshi, J., Kummer, V., Fischer, M., 2014. Higher plant diversity promotes higher diversity of fungal pathogens, while it decreases pathogen infection per plant. Ecology 95, 1907–1917.
Sacristán, S., Fraile, A., Malpica, J.M., Garcia-Arenal, F., 2005. An analysis of host adaptation and its relationship with virulence in *Cucumber mosaic virus*. Phytopathology 95, 827–833.
Sallam, M.F., Xue, R.D., Pereira, R.M., Koehler, P.G., 2016. Ecological niche modelling of mosquito vectors of West Nile virus in St. John's County, Florida, USA. Parasit. Vectors 9, 371.
Samy, A.M., Peterson, A.T., 2016. Climate change influences on the global potential distribution of bluetongue virus. PLoS One 11, e0150489.
Schulze-Lefert, P., Panstruga, R., 2011. A molecular evolutionary concept connecting nonhost resistance, pathogen host range, and pathogen speciation. Trends Plant Sci. 16, 117–125.
Seabloom, E.W., Hosseini, P.R., Power, A.G., Borer, E.T., 2009. Diversity and composition of viral communities: coinfection of barley and cereal yellow dwarf viruses in California grasslands. Am. Nat. 173, E79–E98.
Seabloom, E.W., Borer, E.T., Mitchell, C.E., Power, A.G., 2010. Viral diversity and prevalence gradients in north American Pacific Coast grasslands. Ecology 91, 721–732.
Seabloom, E.W., Borer, E.T., Lacroix, C., Mitchell, C.E., Power, A.G., 2013. Richness and composition of niche-assembled viral pathogen communities. PLoS One 8, e55675.
Seabloom, E.W., Borer, E.T., Gross, K., Kendig, A.E., Lacroix, C., Mitchell, C.E., et al., 2015. The community ecology of pathogens: coinfection, coexistence and community composition. Ecol. Lett. 18, 401–415.
Sexton, J.P., Montiel, J., Shay, J.E., Stephens, M.R., Slatyer, R.A., 2017. Evolution of ecological niche breadth. Annu. Rev. Ecol. Evol. Syst. 48, 183–206.
Shaw, A.K., Peace, A., Power, A.G., Bosque-Pérez, N., 2017. Vector population growth and condition-dependent movement drive the spread of plant pathogens. Ecology 98, 2145–2157.

Shipley, L.A., Forbey, J.S., Moore, B.D., 2009. Revisiting the dietary niche: when is a mammalian herbivore a specialist? Integr. Comp. Biol. 49, 274–290.
Simmonds, P., Adams, M.J., Benkő, M., Breitbart, M., Brister, J.R., Carstens, E.B., et al., 2017. Consensus statement: virus taxonomy in the age of metagenomics. Nat. Rev. Microbiol. 15, 161–168.
Simpson, J.E., Hurtado, P.J., Medlock, J., Molaei, G., Andreadis, T.G., Galvani, A.P., et al., 2012. Vector host-feeding preferences drive transmission of multi-host pathogens: West Nile virus as a model system. Proc. Biol. Sci. 279, 925–933.
Soberón, J., Peterson, A.T., 2005. Interpretation of models of fundamental ecological niches and species' distributional areas. Biodivers. Inform. 2, 1–10.
Sokos, C., Andreadis, K., Papageorgiou, N., 2015. Diet adaptability by a generalist herbivore: the case of brown hare in a Mediterranean agroecosystem. Zool. Stud. 54, 27.
Solé, R.V., Montoya, M., 2001. Complexity and fragility in ecological networks. Proc. Biol. Sci. 268, 2039–2045.
Stafford, C.A., Walker, G.P., Ullman, D.E., 2011. Infection with a plant virus modifies vector feeding behavior. Proc. Natl. Acad. Sci. U.S.A. 108, 9350–9355.
Staniczenko, P., Sivasubramaniam, P., Suttle, K.B., Pearson, R.G., 2017. Linking macroecology and community ecology: refining predictions of species distributions using biotic interaction networks. Ecol. Lett. 20, 693–707.
Stoffels, R.J., Clarke, K.R., Closs, G.P., 2005. Spatial scale and benthic community organisation in the littoral zones of large oligotrophic lakes: potential for cross-scale interactions. Freshw. Biol. 50, 1131–1145.
Stow, C., Allen, C., Garmestani, A., 2007. Evaluating discontinuities in complex systems: toward quantitative measures of resilience. Ecol. Soc. 12, 26.
Strauss, A.T., Civitello, D.J., Cáceres, C.E., Hall, S.R., 2015. Success, failure and ambiguity of the dilution effect among competitors. Ecol. Lett. 18, 916–926.
Streicker, D.G., Turmelle, A.S., Vonhof, M.J., Kuzmin, I.V., McCracken, G.F., Rupprecht, C.E., 2010. Host phylogeny constrains cross-species emergence and establishment of rabies virus in bats. Science 329, 676–679.
Streicker, D.G., Fenton, A., Pedersen, A.B., 2013. Differential sources of host species heterogeneity influence the transmission and control of multihost parasites. Ecol. Lett. 16, 975–984.
Su, Q., Pan, H., Liu, B., Chu, D., Xie, W., Wu, Q., et al., 2013. Insect symbiont facilitates vector acquisition, retention, and transmission of plant virus. Sci. Rep. 3, 1367.
Sugihara, G., May, R.M., 1990. Applications of fractals in ecology. Trends Ecol. Evol. 5, 79–86.
Suzán, G., García-Peña, G.E., Castro-Arellano, I., Rico, O., Rubio, A.V., Tolsá, M.J., et al., 2015. Metacommunity and phylogenetic structure determine wildlife and zoonotic infectious disease patterns in time and space. Ecol. Evol. 5, 865–873.
Swei, A., Ostfeld, R.S., Lane, R.S., Briggs, C.J., 2011. Impact of the experimental removal of lizards on Lyme disease risk. Proc. Biol. Sci. 278, 2970–2978.
Tardy, O., Massé, A., Pelletier, F., Mainguy, J., Fortin, D., 2014. Density-dependent functional responses in habitat selection by two hosts of the raccoon rabies virus variant. Ecosphere 5, 1–16.
Tatineni, S., Robertson, C.J., Garnsey, S.M., Dawson, W.O., 2011. A plant virus evolved by acquiring multiple nonconserved genes to extend its host range. Proc. Natl. Acad. Sci. U.S.A. 108, 17366–17371.
Taylor, L.H., Mackinnon, M.J., Read, A.F., 1998. Virulence of mixed-clone and single-clone infections of the rodent malaria *Plasmodium chabaudi*. Evolution 52, 583–591.
Tellier, A., Brown, J.K.M., 2011. Spatial heterogeneity, frequency-dependent selection and polymorphism in host-parasite interactions. BMC Evol. Biol. 11, 319.

Thébault, E., Fontaine, C., 2010. Stability of ecological communities and the architecture of mutualistic and trophic networks. Science 329, 853–856.
Thrall, P.H., Burdon, J.J., Bever, J.D., 2002. Local adaptation in the Linum marginale–Melampsora lini host-pathogen interaction. Evolution 56, 1340–1351.
Thrall, P.H., Hochberg, M.E., Burdon, J.J., Bever, J.D., 2007. Coevolution of symbiotic mutualists and parasites in a community context. Trends Ecol. Evol. 22, 120–126.
Thrall, P.H., Barrett, L.G., Dodds, P.N., Burdon, J.J., 2016. Epidemiological and evolutionary outcomes in gene-for-gene and matching allele models. Front. Plant Sci. 6, 1084.
Toju, H., Yamamichi, M., Guimarães Jr., P.R., Olesen, J.M., Mougi, A., Yoshida, T., et al., 2017. Species-rich networks and eco-evolutionary synthesis at the metacommunity level. Nat. Ecol. Evol. 1, 0024.
Tollenaere, C., Susi, H., Laine, A.L., 2016. Evolutionary and epidemiological implications of multiple infection in plants. Trends Plant Sci. 21, 80–90.
Turner, M.G., 1989. Landscape ecology: the effect of pattern on process. Annu. Rev. Ecol. Syst. 20, 171–197.
Vassilakos, N., Simon, V., Tzima, A., Johansen, E., Moury, B., 2016. Genetic determinism and evolutionary reconstruction of a host jump in a plant virus. Mol. Biol. Evol. 33, 541–553.
Vayssier-Taussat, M., Albina, E., Citti, C., Cosson, J.F., Jacques, M.A., Lebrun, M.H., et al., 2014. Shifting the paradigm from pathogens to pathobiome: new concepts in the light of meta-omics. Front. Cell. Infect. Microbiol. 4, 29.
Vellend, M., 2010. Conceptual synthesis in community ecology. Q. Rev. Biol. 85, 183–206.
Viana, M., Mancy, R., Biek, R., Cleaveland, S., Cross, P.C., Lloyd-Smith, J.O., et al., 2014. Assembling evidence for identifying reservoirs of infection. Trends Ecol. Evol. 29, 270–279.
Viana, M., Cleaveland, S., Matthiopoulos, J., Halliday, J., Packer, C., Craft, M.E., et al., 2015. Dynamics of a morbillivirus at the domestic–wildlife interface: canine distemper virus in domestic dogs and lions. Proc. Natl. Acad. Sci. U.S.A. 112, 1464–1469.
Wang, Y., Hajimorad, M.R., 2016. Gain of virulence by *Soybean mosaic virus* on *Rsv4*-genotype soybeans is associated with a relative fitness loss in a susceptible host. Mol. Plant Pathol. 17, 1154–1159.
Warton, D.I., Blanchet, F.G., O'Hara, R.B., Ovaskainen, O., Taskinen, S., Walker, S.C., et al., 2015. So many variables: joint modeling in community ecology. Trends Ecol. Evol. 30, 766–779.
Weaver, S.C., Barrett, A.D., 2004. Transmission cycles, host range, evolution and emergence of arboviral disease. Nat. Rev. Microbiol. 2, 789.
Weitz, J.S., Poisot, T., Meyer, J.R., Flores, C.O., Valverde, S., Sullivan, M.B., et al., 2013. Phage-bacteria infection networks. Trends Microbiol. 21, 82–91.
Welti, E.A., Joern, A., 2015. Structure of trophic and mutualistic networks across broad environmental gradients. Ecol. Evol. 5, 326–334.
Whitlock, M.C., 1996. The red queen beats the jack-of-all-trades: the limitations on the evolution of phenotypic plasticity and niche breadth. Am. Nat. 148, S65–S77.
Wiens, J.A., 1989. Spatial scaling in ecology. Funct. Ecol. 3, 385–397.
Wilke, C.O., Forster, R., Novella, I.S., 2006. Quasispecies in time-dependent environments. Curr. Top. Microbiol. Immunol. 299, 33–50.
Winkler, I.S., Mitter, C., Scheffer, S.J., 2009. Repeated climate-linked host shifts have promoted diversification in a temperate clade of leaf-mining flies. Proc. Natl. Acad. Sci. U.S.A. 106, 18103–18108.
Wintermantel, W.M., Cortez, A.A., Anchieta, A.G., Gulati-Sakhuja, A., Hladky, L.L., 2008. Co-infection by two criniviruses alters accumulation of each virus in a host-specific manner and influences efficiency of virus transmission. Phytopathology 98, 1340–1345.

Wolinska, J., King, K.C., 2009. Environment can alter selection in host–parasite interactions. Trends Parasitol. 25, 236–244.
Woolhouse, M., Gaunt, E., 2007. Ecological origins of novel human pathogens. Crit. Rev. Microbiol. 33, 231–242.
Woolhouse, M.E., Gowtage-Sequeria, S., 2005. Host range and emerging and reemerging pathogens. Emerg. Infect. Dis. 11, 1842.
Woolhouse, M.E., Taylor, L.H., Haydon, D.T., 2001. Population biology of multihost pathogens. Science 292, 1109–1112.
Woolhouse, M.E., Haydon, D.T., Antia, R., 2005. Emerging pathogens: the epidemiology and evolution of species jumps. Trends Ecol. Evol. 20, 238–244.
Yoder, J.B., Clancey, E., Des Roches, S., Eastman, J.M., Gentry, L., Godsoe, W., et al., 2010. Ecological opportunity and the origin of adaptive radiations. J. Evol. Biol. 23, 1581–1596.
Young, H.S., Parker, I.M., Gilbert, G.S., Guerra, A.S., Nunn, C.L., 2017. Introduced species, disease ecology, and biodiversity–disease relationships. Trends Ecol. Evol. 32, 41–54.
Zhan, J.S., Thrall, P.H., Burdon, J.J., 2014. Achieving sustainable plant disease management through evolutionary principles. Trends Plant Sci. 19, 570–575.
Ziebell, H., Murphy, A.M., Groen, S.C., Tungadi, T., Westwood, J.H., Lewsey, M.G., Moulin, M., Kleczkowski, A., Smith, A.G., Stevens, M., Powell, G., 2011. Cucumber mosaic virus and its 2b RNA silencing suppressor modify plant-aphid interactions in tobacco. Sci. Rep. 1, 187.

FURTHER READING

Laine, A.L., 2008. Temperature-mediated patterns of local adaptation in a natural plant–pathogen metapopulation. Ecol. Lett. 11, 327–337.

5. Les géminivirus et le modèle épidémiologique des mastrévirus des Poaceae

5.1. Généralités sur les géminivirus

Les virus de la famille des *Geminiviridae* sont responsables de nombreuses **maladies émergentes** à travers diverses régions du monde notamment au sein des **zones tropicales** et **sub-tropicales**. Ces phytovirus infectant une **vaste gamme d'hôtes** (plantes monocotylédones et dicotylédones) induisent des **impacts majeurs** sur le rendement de **nombreuses cultures maraîchères** (tomate, piment, poivron, haricot), **vivrières** (maïs, manioc, patate douce) et **économiques** (blé, canne à sucre, coton) (Varma & Malathi, 2003). De **nombreux symptômes** sont associés aux infections par les géminivirus tels que la présence de mosaïques et/ou de striures, de jaunissement, de nécrose, de retard de croissance ou encore de déformation et d'enroulement foliaire. Le génome de ces virus est constitué d'une (monopartite) ou de deux (bipartite) molécules d'ADN simple brin (ADNsb) circulaire de ~ 2,7 kb protégées par une capside en double icosaèdre de 22 nm de diamètre et de 38 nm de long environ (**Figure 19** ; Zerbini *et al.*, 2017).

Historiquement, ces phytovirus ont été regroupés au sein de **4 genres viraux** (*Mastrevirus*, *Curtovirus*, *Topocuvirus* et *Begomovirus*) en fonction de leur organisation génomique, leur insecte vecteur et leur gamme d'hôtes (Fauquet & Stanley, 2003). Ainsi, le genre ***Mastrevirus*** transmis par des cicadelles infecte les plantes monocotylédones et dicotylédones, le genre ***Curtovirus*** transmis également par cicadelles infecte les plantes dicotylédones, le genre ***Begomovirus*** transmis par l'aleurode *Bemisia tabaci* infecte principalement les plantes dicotylédones et enfin le genre ***Topocuvirus*** transmis par membracide infecte les plantes dicotylédones (Rojas *et al.*, 2005). Les genres *Begomovirus* et *Mastrevirus* sont ceux qui comportent le plus grand nombre d'espèces décrites à ce jour (Varsani *et al.*, 2014).

Figure 19. Organisation du génome des différents genres de géminivirus. Les ORF (V1, V2, V3, C1, etc.) sont codés par couleur en fonction de la fonction de leurs produits protéiques (*rep*, le gène codant pour la protéine associée à la réplication, *ren*, le gène codant pour la protéine de stimulation de la réplication; *trap*, le gène codant pour la protéine activatrice de la transcription; *cp*, le gène codant pour la protéine de capside; *mp*, le gène codant pour la protéine de mouvement; *nsp*, le gène codant pour la protéine de navette nucléaire). Les acronymes LIR, SIR et CR correspondent respectivement à la grande région intergénique, la petite région intergénique courte et la région commune. La tige boucle comprend l'origine de réplication et est indiquée dans la LIR (Zerbini *et al.*, 2017).

Le genre *Begomovirus* dont le virus type est le bean golden mosaic virus (BGMV) comportait en 2015 409 espèces validées par l'*International Committee on Taxonomy of Viruses* (ICTV) selon le seuil de distinction taxonomique de 91 % d'identité nucléotidique (Brown *et al.*, 2015) contre quarante-quatre espèces pour le genre *Mastrevirus* (seuil de distinction taxonomique de 78 % d'identité nucléotidique ; Muhire *et al.*, 2013 ; Zerbini *et al.*, 2017) dont le virus type est le maize streak virus (MSV).

L'utilisation de la polymérase *phi*29 pour l'amplification (Inoue-Nagata *et al.*, 2004) avant clonage des génomes de géminivirus associée à la diminution du coût de séquençage a permis de caractériser un grand nombre de nouvelles espèces ainsi que cinq nouveaux genres (Zerbini *et al.*, 2017) tels que le genre ***Becurtovirus***, ***Eragrovirus*** et ***Turncurtovirus*** (Varsani *et al.*, 2014) ainsi que les genres ***Capulavirus*** et ***Grablovirus*** (Varsani *et al.*, 2017).

Outre ces genres, il existe également plusieurs autres espèces de géminivirus très divergentes qui n'ont pas encore été attribuées à un genre telles que l'apple geminivirus (Liang *et al*., 2015), le **Camellia chlorotic dwarf associated virus** (Zhang *et al.*, 2018), le **Citrus chlorotic dwarf associated virus** (Loconsole *et al*., 2012), le **grapevine geminivirus A** (Al Rwahnih *et al*., 2017), le **mulberry mosaic dwarf associated virus** (Ma *et al*., 2015), le **passion fruit chlorotic mottle virus** (Fontenele, *et al*., 2018), le **tomato associated geminivirus 1** (Fontenele *et al*., 2017), le **tomato apical leaf curl virus** (Vaghi Medina *et al*., 2018), le **Limeum africanum associated virus**, le **Polygala garcinii associated virus** et le **Juncus maritimus associated virus** (Claverie *et al*., 2018). L'ensemble de ces espèces est en attente de validation par l'ICTV en raison du manque d'information sur leurs vecteurs et leur capacité à générer des particules virales en doublet caractéristiques des géminivirus (Varsani *et al.*, 2017).

5.2. L'origine des géminivirus

Tous les genres de la famille des géminivirus possèdent un génome contenant des gènes orientés bidirectionnellement mais seuls deux gènes codant pour la **protéine de capside** (CP) et la **protéine associée à la réplication** (Rep), sont communs à tous les genres. Plusieurs scénarios ont été proposés sur l'origine de virus à ADNsb circulaire codant pour une protéine Rep (*circular Rep-encoding ssDNA virus* ou CRESS-DNA viruses). L'un d'entre eux suggère que les virus CRESS-DNA sont issus d'une **recombinaison** entre des **plasmides** et des **virus à ARNsb eucaryotes** (Kazlauskas *et al.*, 2019 ; Krupovic *et al.*, 2009). Ainsi, les géminivirus modernes auraient possiblement évolué à partir de réplicons d'ADN extra-chromosomiques présents chez les ancêtres procaryotes ou eucaryotes primitifs des plantes modernes (Koonin &

Ilyina, 1992 ; Rojas *et al.*, 2005). En effet, des études de phylogénie et de *clustering* ont démontré que les Reps des géminivirus partagent possiblement un ancêtre commun avec les Reps codés par des plasmides de phytoplasmes, bactéries phytopathogènes (Kazlauskas *et al.*, 2019 ; Koonin & Ilyina, 1992 ; Rojas *et al.*, 2005). Par ailleurs, Krupovic et ses collaborateurs (2009) ont mis en avant la forte ressemblance structurelle de type *jelly-roll* (i.e. repliement des protéines en huit brins β arrangés en deux feuillets de quatre brins) entre la CP des géminivirus et celle de certains virus à ARNsb. Ainsi, le probable ancêtre commun avec les plasmides et la large répartition et conservation de la Rep tout au long du processus évolutif, rendent compte de l'**origine ancienne** du ou des ancêtre(s) des géminivirus (Hull, 2014).

5.3. Organisation génomique et fonctionnelle des mastrévirus

Les mastrévirus sont des virus **monopartites** (ADN-A) dont le génome d'environ 2,7 kb présente quatre **cadres ouverts de lecture** (régions codantes) ou *Open Reading Frames* (ORFs) et deux **régions intergéniques** (IR). Deux membres des mastrévirus, le MSV et le wheat dwarf virus (WDV) ont été étudiés de manière approfondie afin de déterminer l'organisation génomique et fonctionnelle des mastrévirus.

Les mastrévirus codent pour quatre ORFs regroupées par paire, **V1** et **V2** codés par le **brin viral** et, **C1** et **C2** (ORFs chevauchants) portés par le **brin complémentaire**. Dans le sens viral, l'ORF V2 code pour la **protéine de mouvement** (MP) qui intervient dans les mécanismes de diffusion du virus au sein de la plante hôte, en association avec la **protéine de capside** (CP), unité de base dans la constitution de la particule virale des géminivirus, codée quant à elle par l'ORF V1 (Fondong, 2013). Outre l'encapsidation du génome du virus, la CP a été associée à plusieurs autres fonctions, notamment la reconnaissance et la transmission par son insecte vecteur (Boulton, 2002 ; Caciagli *et al.*, 2009) ainsi que la protection et le transport (importation et exportation nucléaires) de l'ADN viral au sein de la plante hôte et dans les mouvements systémiques du virus entre cellules. Par ailleurs, la MP se liant au complexe CP-ADN viral, est responsable plus précisément du mouvement de l'ADN viral du noyau vers la périphérie cellulaire et de cellule à cellule via les plasmodesmes (Liu *et al.*, 1999, 2001, 1997 ; Rojas *et al.*, 2005). En outre, la

MP chez le MSV serait impliquée dans la gravité des symptômes chez leurs hôtes (Van der Walt *et al.*, 2008). Dans le sens complémentaire, les ORFs C1 et C2 codent après épissage pour la **protéine associée à la réplication** (Rep) alors que la RepA est uniquement exprimée à partir de l'ORF C1 (Dekker *et al.*, 1991 ; Mullineaux *et al.*, 1990 ; Wright *et al.*, 1997). La Rep est responsable de l'initiation de la réplication en se liant et en coupant le brin viral au niveau de l'origine de réplication (v-ori) alors que la **RepA** se lie aux protéines apparentées au rétinoblastome de la plante hôte afin d'empêcher ces dernières d'inférer dans la réplication virale (Heyraud *et al.*, 1993; Liu *et al.*, 1999). Ainsi, Rep et RepA sont essentiels à la réplication (Liu *et al.*, 1999).

Le génome des mastrévirus porte deux régions intergéniques (*Intergenic Region*, IR) qui séparent les deux paires d'ORFs (**Figure 19**). La **grande région intergénique** (*Large Intergenic Region,* LIR) comprend l'origine de réplication (v-ori) incluant la tige boucle (5' TAAT(A/G)TTAC 3', site d'initiation de la réplication) et des itérons (courtes séquences répétées constituant un site de reconnaissance et de liaison de la Rep). La **petite région intergénique** (*Small Intergenic Region*, SIR) contient les signaux de terminaison et polyadénylation pour les gènes du brin viral et complémentaire (Zerbini *et al.*, 2017).

Par ailleurs, deux cas unique d'association entre des molécules satellites et des mastrévirus ont été décrits. Il s'agit de l'association entre des alphasatellites et betasatellites et le chickpea chlorotic dwarf virus au Pakistan (CpCDV ; Hamza *et al.*, 2018) et le wheat dwarf India virus en Inde (WDIV ; Kumar *et al.*, 2014).

5.4. La réplication et la transcription des géminivirus

Les géminivirus n'étant pas capable de s'auto-répliquer, ils sont dépendants de la machinerie cellulaire de l'hôte et notamment de l'utilisation de l'ADN polymérase cellulaire pour leur réplication. La réplication à lieu selon deux mécanismes connus sous le nom de ***Rolling circle replication*** (RCR) et ***Recombination-dependent replication*** (RDR) (**Figure 20** ; Jeske *et al.*, 2001).

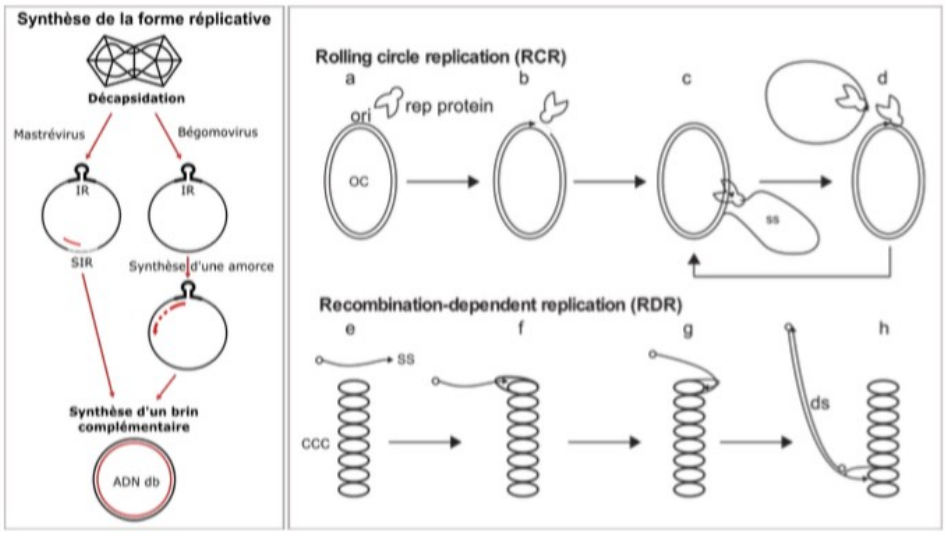


Figure 20. **Schématisation du processus de réplication des géminivirus.** Synthèse de la forme réplicative ADN db : Pour les mastrévirus après décapsidation de l'ADN viral (noir), la synthèse du brin complémentaire se fait à partir d'une amorce d'ADN (rouge) déjà présente et hybridée dans la région SIR (indiquée). Les ADN db circulaires serviront dans la réplication (Bernardi & Timchenko, 2008). Réplication en cercle roulant (RCR) : Etape a : accrochage de la protéine associée à la réplication (Rep) à l'origine de réplication (ori). Etape b : ouverture de l'ADN et liaison covalente de la Rep à l'extrémité 5'. Etape c : déplacement et réplication. Etape d : nouvelle ouverture de l'ADN, fermeture des ADN simple brin et relargage de la Rep. Réplication dépendante de la recombinaison (RDR) : Etape e : interaction entre un ADN simple brin incomplet et la forme super-enroulée de l'ADN viral (cccDNA) à des sites homologues. Etape f : recombinaison homologue. Etape g : élongation de l'ADN simple brin. Etape h : synthèse de l'ADN complémentaire et obtention d'un ADN double brin (Jeske *et al.*, 2001).

Après décapsidation, la réplication se produit dans le noyau de la cellule hôte via un intermédiaire d'ADN double brin (ADNdb). La synthèse de cet intermédiaire d'ADNdb est amorcée soit par une séquence d'oligonucléotides complémentaire à l'IR soit, dans le cas des mastrévirus par exemple, grâce à une amorce d'ADN déjà présente et hybridée dans la région SIR de l'ADN viral parent (Saunders *et al.*, 1991). Après la formation de la forme double brin, l'ADN viral s'associe avec les histones de la cellule hôte, formant des minichromosomes viraux prêts à être répliqués et transcrits (Pilartz & Jeske, 1992).

La RCR chez les géminivirus est analogue à la réplication des phages à ADN circulaire simple brin (Novick, 1998) et peut être décomposée en trois étapes à savoir l'initiation, l'élongation et la terminaison (Hanley-Bowdoin *et al.*, 2013). L'expression de la protéine Rep est cruciale à l'initiation de la RCR (Pilartz & Jeske, 1992 ; Saunders *et al.*, 1991). En effet, après reconnaissance et liaison au niveau des itérons présents dans la LIR de l'ADN viral, la protéine Rep initie la réplication en clivant le brin viral au niveau de l'origine de

réplication, plus précisément la tige-boucle contenant le nonanucléotide hautement conservé (5'-TAATATT↓AC-3) (Hanley-Bowdoin *et al.*, 2000). La protéine Rep se lie alors de façon covalente au niveau de l'extrémité 5' de l'ADN clivé. L'ADN viral, dans un état circulaire ouvert, devient alors une matrice pour la production en continu d'un nouveau brin d'ADN viral. Une fois la synthèse complète du nouvel ADN simple brin viral (ADNsb), celui-ci est alors circularisé et suivi du relargage de la Rep. Ce modèle de réplication en cercle roulant a été confirmé par microscopie électronique (Jeske *et al.*, 2001). L'électrophorèse sur gel bidimensionnel à haute résolution ainsi que les visualisations par microscopie électronique ont mis en évidence que les géminivirus étaient capables d'utiliser une voie de réplication dépendante de la recombinaison (RDR) analogue à celle des bactériophages T4. Ce mécanisme utilisant des intermédiaires réactionnels additionnels, repose sur l'interaction entre un ADN simple brin et la forme super-enroulée de l'ADN viral (cccDNA) induisant une recombination homologue, suivi de l'élongation de l'ADN simple brin et de l'obtention d'un ADN double brin (Mosig *et al.*, 2001 ; Preiss & Jeske, 2003).

Les génomes de géminivirus sont transcrits dans le noyau de manière **bidirectionnelle**, c'est à dire que les ARN messager (ARNm) résultent à la fois des ORFs du sens viral que du sens complémentaire. La transcription des mastrévirus aboutit à la production de plusieurs ARNm polyadénylés produisant la MP, la CP, la Rep et la RepA (Hanley-Bowdoin *et al.*, 2000 ; Wright *et al.*, 1997).

5.5. Diversité, gamme d'hôtes et transmission des mastrévirus

Le genre des mastrévirus comptent actuellement 40 espèces différentes (ICTV, https://talk.ictvonline.org/taxonomy/) sur la base d'un seuil de distinction taxonomique de 78 % d'identité nucléotidique sur l'ensemble du génome (Muhire *et al.*, 2013). La grande majorité des virus décrits à ce jour sont associés à des symptômes de striures et/ou de mosaïques.

5.5.1. Les mastrévirus infectant les plantes monocotylédones

La majorité des mastrévirus infectent les plantes monocotylédones de la famille des Poaceae et compte à ce jour 33 espèces différentes dont quinze initialement caractérisées sur **plantes cultivées** et dix-huit sur **plantes sauvages**. Quatorze espèces ont été caractérisées en Afrique et dans les îles du sud-ouest de l'Océan Indien (SWIO), dix en Australie, quatre en Eurasie, trois en Amérique, et enfin deux dans le Pacifique (Japon et Vanuatu ; Kraberger *et al.*, 2012, 2017 ; Muhire *et al.*, 2013).

5.5.1.1. Les African streak virus (AfSV)

La plupart des mastrévirus infectant les monocotylédones ont été identifiés en Afrique ou dans les îles environnantes et ont été collectivement appelés les **African streak virus** (AfSV). Il existe actuellement quatorze espèces d'AfSV reconnues, dont cinq infectent à la fois des espèces cultivées et non cultivées, cinq exclusivement identifées sur des cultures et quatre exclusivement caractérisées sur des plantes non cultivées (**Figure 21** ; Kraberger *et al.*, 2017).

Parmi les quatorze espèces identifiées en Afrique, le **MSV** est l'espèce qui a été la plus étudiée en raison de son impact majeur et dévastateur sur les cultures de maïs. Le MSV est responsable d'une des plus importantes virose du maïs en Afrique subsaharienne, avec des **pertes de rendement** pouvant atteindre les 100% dans certains cas (Thottappilly *et al.*, 1993). Ce virus a été observée pour la première fois en 1901 en Afrique du Sud par Claude Fuller mais ce n'est qu'en 1925 que Storey et ses collaborateurs ont mis en évidence l'infection virale et la transmission par cicadelle (Shepherd *et al.*, 2010). Le MSV infecte non seulement le maïs mais également d'autres cultures telles que l'orge, le blé, l'avoine, le seigle, la canne à sucre, le millet (Shepherd *et al.*, 2010), ainsi qu'un large panel d'espèces de poacées sauvages (39 espèces sauvages décrites comme hôtes à ce jour).

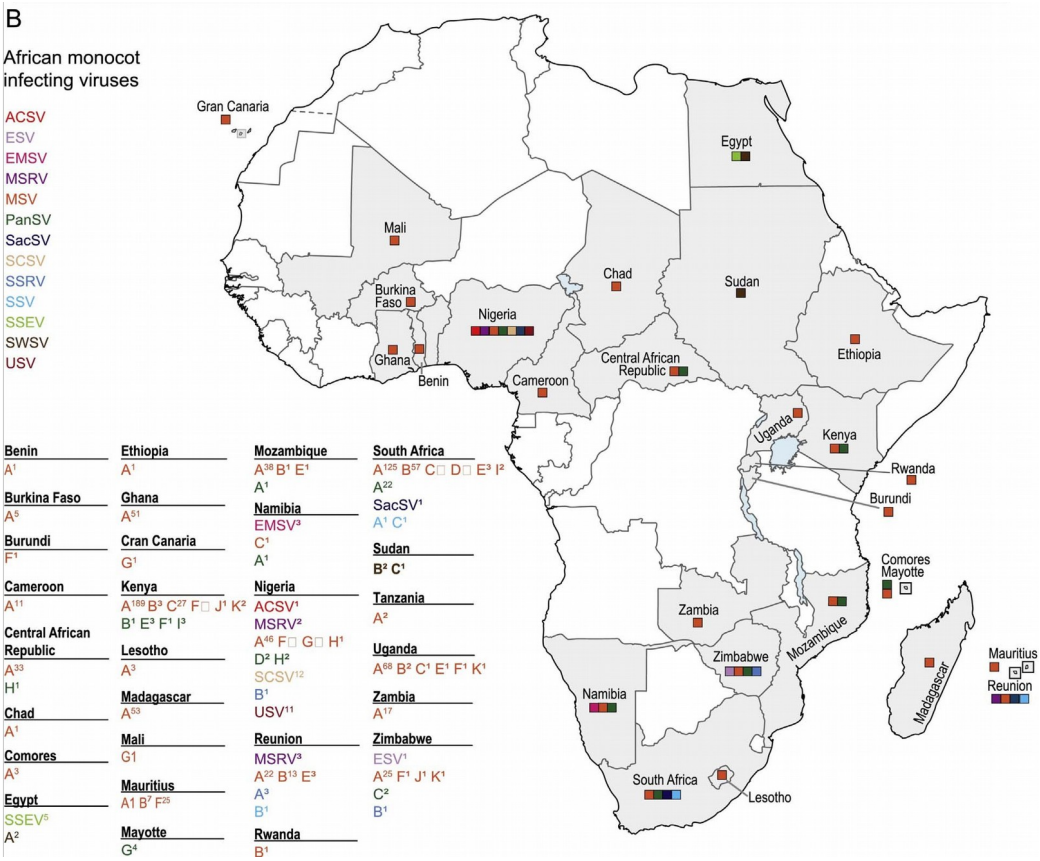


Figure 21. Représentation des espèces d'AfSV caractérisées dans les pays d'Afrique et les iles de l'océan Indien. Les différentes espèces d'AfSV sont représentées par un code couleur. Les lettres dans la liste des pays représentent les souches des espèces d'AFSV. Le nombre d'échantillons obtenu pour chaque espèce et souche dans les différents pays est indiqué par le numéro en indice à côté de l'acronyme de l'espèce et / ou de la lettre de souche.

Sur la base du seuil d'identité nucléotidique de 94% (Muhire *et al.*, 2013), **onze souches** de MSV ont été décrites allant du **MSV-A** au **MSV-K** mais seule la souche A est connue pour causer des symptômes sévères sur le maïs (Monjane *et al.*, 2011 ; Oluwafemi *et al.*, 2011 ; Shepherd *et al.*, 2010 ; Varsani *et al.*, 2008). La gamme d'hôtes du MSV-A décrite actuellement est constituée de quatre espèces cultivées et de 12 espèces sauvages. La souche A du MSV serait issue de la recombinaison entre les variants ancestraux du MSV-B et MSV-G/F (Varsani *et al.*, 2008). Les données actuelles supportent l'idée que cette recombinaison a vraisemblablement participé à l'émergence de cette souche de MSV sur maïs notamment en élargissant sa gamme de plantes hôtes sauvages et en favorisant sa dissémination à large échelle (Varsani *et al.*, 2008). Les autres souches (B à K) ont été identifiées à ce jour sur 41 espèces de poacées, incluant 3 espèces cultivées et 38 espèces sauvages.

D'autres espèces de mastrévirus ont été identifiées sur maïs avec le **maize streak Reunion virus** (**MSRV**) et le **maize streak dwarfing virus** (**MSDV**). Le MSRV a été initialement identifié sur des plants de maïs symptomatiques à La Réunion (Pande *et al.*, 2012), avant d'être par la suite décrit sur maïs en Chine et en Ethiopie (Chen *et al.*, 2015 ; Guadie *et al.*, 2019) mais aussi sur des poacées sauvages (*Setaria barbata* et *Rottboellia sp.*) à La Réunion et au Nigeria (Kraberger, *et al.*, 2017 ; Oluwafemi *et al.*, 2014). Le **MSDV** a quant à lui été identifié récemment sur des maïs symptomatiques en Ethiopie (Guadie *et al.*, 2019).

Six mastrévirus infectant la canne à sucre ont initialement été isolés en Afrique et dans les îles SWOI, à savoir le **Saccharum streak virus** (**SacSV** ; Lawry *et al.*, 2009), le **sugarcane chlorotic streak virus** (**SCSV** ; Yahaya *et al.*, 2017), le **sugarcane streak Egypt virus** (**SSEV** ; Bigarré *et al.*, 1999), le **sugarcane streak Reunion virus** (**SSRV** ; Shepherd *et al.*, 2008), le **sugarcane streak virus** (**SSV** ; Hughes *et al.*, 1992) et le **sugarcane white streak virus** (**SWSV** ; Candresse *et al.*, 2014). Pour ces trois dernières espèces virales, trois souches ont été décrites à ce jour. Le SWSV a été récemment caractérisé dans le Nouveau Monde, à la Barbade sur canne à sucre. Le SSRV a été également identifié à l'Ile de La Réunion, au Zimbabwe et au Nigeria sur trois espèces sauvages, *Eleusine coracana*, *Paspalum conjugatum* et *Setaria barbata.* Le SSV a également été décrit sur *Cenchrus echinatus* à l'Ile de La Réunion (Oluwafemi *et al.*, 2014 ; Shepherd *et al.*, 2008).

Cinq espèces d'AfSV à savoir, l'**Axonopus compressus streak virus** (**ACSV** ; Oluwafemi *et al.*, 2014), l'**Eragrostis minor streak virus** (**EMSV** ; (Martin *et al.*, 2011), **l'Eragrostis streak virus** (**ESV** ; Shepherd *et al.*, 2008)Shepherd *et al.*, 2008), le **Panicum streak virus** (PanSV ; Varsani *et al.*, 2008b) et l'**Urochloa streak virus** (**USV** ; Oluwafemi *et al.*, 2008) ont été identifiées initalement sur des plantes non cultivées. Actuellement, l'ACSV, l'EMSV et l'ESV n'ont pas été retrouvés sur d'autres espèces que respectivement l'*Axonopus compressus*, l'*Eragrostis minor* et l'*Eragrostis sp.* (Kraberger, Saumtally, *et al.*, 2017; Martin *et al.*, 2011; Oluwafemi *et al.*, 2014; Shepherd *et al.*, 2008). A l'heure actuelle, toutes ces espèces ont des hôtes

exclusivement sauvages, à l'exception du PanSV qui a aussi été isolé sur orge (*Hordeum vulgare)* et maïs (*Zea mays* ; Kraberger *et al.*, 2017). Ce dernier est la deuxième espèce la mieux caractérisée après le MSV avec **neuf souches** identifiées (**PanSV-A** à **PanSV-I**). Cette espèce présente un niveau de diversité similaire au MSV (Varsani *et al.*, 2008 ; Varsani *et al.*, 2009) et est également transmis par la cicadelle *Cicadulina mbila* (Briddon *et al.*, 1992). Le PanSV a été identifié à ce jour sur environ 17 espèces de poacées et a été décrit au sein de sept pays d'Afrique et à Mayotte.

5.5.1.2. Les mastrévirus non africains infectant les monocotylédones

Parmi les 19 espèces de mastrévirus infectant les monotylédones sur le reste du globe, six d'entre elles ont été caractérisées initialement sur plantes cultivées à savoir le **sugarcane striate virus** (**SCStV** ; Boukari *et al.*, 2017), le **maize striate mosaic virus** (MSMV ; Fontenele, *et al.*, 2018), le **barley dwarf virus** (**BDV**), le **oat dwarf virus** (**ODV** ; Schubert *et al.*, 2007), le wheat dwarf India virus (**WDIV** ; Kumar *et al.*, 2012) et le **wheat dwarf virus** (**WDV** ; Schubert *et al.*, 2014). Le WDV regroupe cinq souches différentes (**WDV-A** à **WDV-E**) et est le pathogène le plus dommageable et étudié en Eurasie. Il infecte d'autres cultures que le blé telles que l'épeautre, l'orge ou encore l'avoine, ainsi que neuf espèces sauvages (Parizipour *et al.*, 2017; Schubert *et al.*, 2007, 2014). Les 13 autres espèces de mastrévirus ont été caractérisées exclusivement sur plantes non cultivées dont la majorité uniquement sur leurs espèces de première description. Ces virus sont le **Bromus catharticus striate mosaic virus** (**BCSMV**), le **Digitaria ciliaris striate mosaic virus** (**DCSMV**), le **Digitaria didactyla striate mosaic virus** (**DDSMV**), le **rice latent virus 1** et **2** (**RLV-1** et **2** ; Kraberger *et al.*, 2017), le **Sporobolus striate mosaic virus 1** et **2** (**SSMV-1** et **2**), le **Paspalum dilatatum striate mosaic virus** (**PDSMV**), le **paspalum striate mosaic virus** (**PSMV**), le **Chloris striate mosaic virus** (**CSMV**), le **switchgrass mosaic associated virus 1** (**SgMaV-1** ; Agindotan *et al.*, 2015), le **Miscanthus streak virus** (**MiSV** ; Chatani *et al.*, 1991) et le **Digitaria streak virus** (**DSV** ; Dollet *et al.*, 1986).

5.5.1.3. Les mastrévirus infectant les plantes dicotylédones

Les mastrévirus infectant les plantes dicotylédones sont regroupés au sein de dix espèces virales, toutes identifiées sur des plantes cultivées et la moitié d'entre elles exclusivement en Australie. Ainsi, parmi ces dix espèces, sept ont été initialement caractérisées sur pois chiche alors que les trois autres ont été isolées respectivement sur haricot, patate douce et tabac. Les mastrévirus initialement caractérisées sur pois chiche sont le **chickpea chlorosis virus** (**CpCV** ; Hadfield *et al.*, 2012 ; Kraberger *et al.*, 2013 ; Thomas *et al.*, 2010), le **chickpea chlorosis Australia virus** (**CpCAV** ; Hadfield *et al.*, 2012), le **chickpea yellows virus** (**CpYV** ; Hadfield *et al.*, 2012), le **chickpea redleaf virus** (**CpRLV** ; Thomas *et al.*, 2010) exclusivement isolés en Australie, le **chickpea chlorotic dwarf Pakistan virus** (**CpCDPV**) et le **chickpea yellow dwarf virus** (**CpYDV**) exclusivement caractérisés au Pakistan (Kraberger *et al.*, 2015) et le **chickpea chlorotic dwarf virus** (**CpCDV**) localisé en Afrique du sud, en Afrique du Nord-Est, au Moyen Orient et dans le sous-continent Indien (Horn *et al.*, 1993 ; Kraberger *et al.*, 2013 ; Liu *et al.*, 1997 ; Nahid *et al.*, 2008). Enfin, le **tobacco yellow dwarf virus** (**TYDV**) décrit exclusivement en Australie (Ṭrebicki *et al.*, 2010). Ces pathogènes induisent généralement des symptômes comprenant un retard de croissance, la présence de chlorose et/ou de rougissement des feuilles et parfois même le brunissement du phloème. Le CpCDV est le pathogène le mieux caractérisé, identifié dans environ quinze pays. Ce virus infecte une large gamme d'hôtes, constituée de 21 espèces différentes appartenant aux Fabaceae, Solanaceae, Asteraceae, Malvaceae et Cucurbitaceae.

5.5.2. La transmission des mastrévirus

Les différents membres du genre *Mastrevirus* sont connus pour être transmis par des espèces de cicadelles issus de la famille des Cicadellidae. Le MSV est transmis par des cicadelles du genre ***Cicadulina*** regroupant 22 espèces parmi lesquelles 18 espèces sont présentes en Afrique et 14 en sont endémiques (Webb, 1987). Neuf des espèces africaines ont été identifiées comme étant capables de transmettre le MSV (Bosque-Pérez, 2000) mais

l'espèce la plus répandue en Afrique, ***Cicadulina mbila***, semble être le vecteur le plus efficace (**Figure 22** ; Storey, 1928 ; Reynaud 1988).

Figure 22. Vue latérale d'un adulte et d'une larve de *Cicadulina mbila* (Crédits Antoine Franck).

C. mbila a également été identifiée comme étant capable de transmettre le PanSV (Briddon *et al.*, 1992). Au sein des mastrévirus d'Eurasie, le WDV est transmis par les cicadelles *Psammotettix alienus* et *Psammotettix provincialis* (Ekzayez *et al.*, 2011 ; Kamali *et al.*, 2017 ; Wang *et al.*, 2014) et le ODV est également transmis par *P. alienus* (Schubert *et al.*, 2007). Parmi les mastrévirus infectant les dicotylédones, le CpCDV est transmis par les espèces de cicadelles *Orosius orientalis* en Inde (Horn *et al.*, 1993) et *Orosius albicinctus* en Syrie, en Australie et au Pakistan (Kumari *et al.*, 2004 ; Fletcher, 2009 ; Akhtar *et al.*, 2011). Par ailleurs, les espèces *Orosius orientalis* et *Anzygina zealandica* se sont révélées positives au TYDV en Australie, cependant aucune expérience de transmission n'a confirmé leur capacité à transmettre le TYDV (Ṭrebicki *et al.*, 2010).

6. Problématique et objectifs

Historiquement, les études menées en phytovirologie se sont principalement focalisées sur la diversité des virus pathogènes des plantes cultivées. En effet, les maladies phytovirales causent d'énormes pertes économiques en particulier dans les régions tropicales et sub-tropicales, où les populations sont fortement tributaires de l'agriculture. Ces dernières décennies en Afrique, un grand nombre d'espèces virales appartenant aux géminivirus a été associé à diverses épidémies notamment sur manioc avec les cassava mosaic geminiviruses (CMGs ; Legg & Fauquet, 2004) et sur maïs avec le MSV-A (Martin & Shepherd, 2009). Si le MSV-A est l'agent le plus dévastateur et le mieux étudié du genre *Mastrevirus*, d'autres espèces virales infectant le maïs (MSRV et MSDV) ou la canne à sucre (SacSV, SCSV, SSEV, SSRV, SSV, SWSV, SCStV) ont été caractérisées en Afrique (Pour revue, Shepherd *et al.*, 2010). Outre la description de toutes ces espèces de mastrévirus pathogènes sur cultures, les connaissances sur la diversité globale des mastrévirus restent néanmoins incomplètes. Les travaux de ces dernières années sur plantes non-cultivés ont permis à la fois d'identifier les espèces de mastrévirus préalablement décrites sur plantes cultivées mais aussi de caractériser de nouvelles espèces, indiquant une nouvelle fois l'importance d'étudier la diversité virale à l'échelle de l'écosystème plutôt qu'uniquement à partir d'un hôte cultivé. Le manque d'exhaustivité des études passées sur la diversité des populations virales s'explique certes par le fait que (i) la majorité des travaux de recherche s'est focalisée sur les maladies des plantes cultivées mais également (ii) par les difficultés à identifier les éventuels symptômes associés à l'infection de plantes sauvages par des virus pour lesquels une longue co-évolution avec l'hôte limite l'apparition de dommages visibles (Jones, 2009) et (iii) par les associations complexes avec la coexistence de plusieurs souches d'une même espèce (Péréfarres *et al.*, 2014), de plusieurs espèces d'un même genre (Legg & Thresh, 2000) ou de plusieurs espèces de familles différentes (Adams *et al.*, 2013), au sein d'une même plante. Le développement des approches métagénomiques s'est traduit par la caractérisation de la diversité des communautés virales et de leurs interactions à l'échelle d'écosystèmes entiers (Alexander *et al.*, 2014 ; Roossinck *et al.*, 2015 ; Stobbe & Roossinck,

2014), entraînant une remise en perspective de l'évolution virale et de l'émergence de certains agents pathogènes (Vayssier-Taussat *et al.*, 2014).

Aux vues des récentes caractérisations de nouveaux genres, de nouvelles espèces au sein des agro-écosystèmes et de la détection dans certains génomes de fragments recombinants encore non décrits chez les géminivirus, l'ensemble des données actuelles argumente l'hypothèse que ces biais de connaissances sur la diversité et la structuration des communautés virales concernent également les géminivirus (Kraberger *et al.*, 2017 ; Lefeuvre *et al.*, 2007). Les études récentes sur la caractérisation des géminivirus en Afrique et dans les îles du sud-ouest de l'océan Indien (SOOI) ont mis en évidence la présence d'une importante diversité de mastrévirus et d'une large gamme de plantes hôtes notamment de la famille des Poaceae. Ainsi, à La Réunion, près de la moitié des espèces d'AfSV décrites ont été identifiés principalement sur maïs (MSV, Peterschmitt *et al.*, 1996 ; MSRV, Pande *et al.*, 2012 et canne à sucre (SSV, Bigarré *et al.*, 1999 ; SSRV, Shepherd *et al.*, 2008 ; SCStV, Candresse *et al.*, 2014 ; SWSV, Boukari *et al.*, 2017).

L'objectif général de ce projet de thèse a été d'étudier et de comprendre l'écologie virale des communautés de mastrévirus de poacées en améliorant nos connaissances sur la diversité et la structure des communautés de mastrévirus.
Pour se faire, mon travail s'est focalisé dans un premier temps sur l'étude de la diversité des communautés de mastrévirus des poacées au sein d'un agro-écosystème réunionnais. Le premier chapitre de cette thèse a fait l'objet du développement d'une nouvelle méthodologie métagénomique axée sur les petits virus circulaires (i) facile à mettre en œuvre et (ii) permettant de traiter un grand nombre d'échantillons à moindre coût. Cette approche métagénomique a été appliquée à un ensemble d'espèces de poacées non cultivées collectées en novembre 2014 à l'échelle d'un agro-écosystème de taille restreinte de La Réunion. Cette méthode a non seulement permis la détection et la caractérisation de plusieurs mastrévirus connus à partir d'échantillons de plantes sans symptômes de striures caractéristiques mais également l'identification de trois nouvelles espèces de mastrévirus. De telles caractérisations ont ainsi démontré l'efficacité de la méthode pour mettre en

évidence une diversité virale jusque-là non caractérisée à partir de nombreux échantillons. Les principaux résultats obtenus dans le cadre de ce chapitre ont permis de (i) valider un protocole de métagénomique dédié aux petits virus à ADN circulaire et (ii) confirmer les biais de connaissances sur la diversité des mastrévirus à La Réunion et en général.

Au regard de ces premiers résultats et en se basant sur un échantillonnage élargi, les travaux présentés dans le deuxième chapitre de cette thèse a cherché à identifier de manière exhaustive les virus circulant dans l'agro-écosystème afin de déterminer leur distribution entre les différents hôtes et de déterminer les facteurs évolutifs ayant pu façonner la structure de cette communauté. Les résultats de la prospection de 2014 ont été intégrés à ceux des trois larges prospections étalées de 2016 à 2017. Ce travail a permis (i) d'élargir les connaissances sur les espèces de mastrévirus présentes et sur la gamme d'hôtes de ces virus mais aussi, (ii) d'établir les possibles réseaux d'échange de virus entre ces plantes hôtes. L'ensemble de nos résultats démontre la présence d'une structuration complexe au sein de cet agro-écosystème caractérisée par la présence (i) de virus généralistes (principalement le MSV) et de virus spécialistes, (ii) de plantes carrefours et de différentes co-infections qui soulignent l'importance de s'intéresser à la diversité globale des complexes viraux et au concept de pathobiome (Vayssier-Taussat *et al.*, 2014). Dans un second temps, du fait de cette structuration mixte et complexe et de la présence d'une forte diversité de mastrévirus au sein d'un agro-écosystème de taille réduite, notre étude s'est intéressée à déterminer l'importance de l'échange de gènes au sein de la communauté virale.

Chapitre 1

Chapitre 1

Mise en place d’une approche de métagénomique pour la caractérisation moléculaire de mastrévirus infectant les Poaceae à La Réunion

L’ensemble des données de la littérature soutiennent que la compréhension de l’évolution et l’adaptation des phytovirus pathogènes nécessite d’élargir l’étude des phytovirus à l’échelle des écosystèmes entiers (French & Holmes, 2020), plutôt que de se focaliser exclusivement sur l’étude des phytovirus pathogènes des cultures. L’intégration récente de l’écologie par la virologie végétale a nécessité le développement de nouveaux protocoles basés sur la métagénomique. En effet, de par l’abondance virale et le nombre important de plantes et d’insectes présents à l’échelle des écosystèmes, les techniques dites classiques employées en virologie (*e.g.* sérologie, PCR…) se sont avérées limitantes voir inadaptées (Hugenholtz & Tyson, 2008).

Contrairement à ces techniques classiques, les approches de métagénomique permettent l’analyse des communautés virales (virome) dans leur globalité à partir d’échantillons environnementaux (Roossinck *et al.*, 2010). Ces approches utilisant généralement les technologies de séquençage de nouvelle génération (NGS), ont considérablement élargi notre connaissance de la diversité et notre compréhension du rôle des virus dans l'environnement. Toutefois, la structure réelle des communautés virales reste encore largement inconnue. Une des raisons à cela est liée à la grande diversité des virus révélés par métagénomique et la difficulté d’attribuer ces virus au rang taxonomique de l’espèce. Ainsi, certains travaux fondateurs de la discipline (Bernardo *et al.*, 2018 ; Muthukumar *et al.*, 2009) ont dû se limiter à l’identification de la famille ou du genre viral plutôt qu’à celui de l’espèce, rendant difficile toute évaluation réelle des structures de communautés.

Afin de pallier à cette limitation, nous avons choisi dans ce travail de thèse de nous concentrer sur le genre des *Mastrevirus*, phytovirus transmis par cicadelles principalement aux plantes de la famille des Poaceae. Outre la description historique de virus de ce genre à La Réunion, un des intérêts

majeurs de l'étude des mastrévirus repose sur une spécificité génomique (virus à ADN circulaire simple brin de petite taille, particularité partagée, entre autre, avec tous les virus de la famille des *Geminiviridae*) qui autorise le développement d'un protocole de métagénomique dédié.

Le premier chapitre de cette thèse a été consacré à (i) l'élaboration d'une nouvelle approche de métagénomique adaptée aux géminivirus et (ii) l'évaluation des limites qu'impliquent une telle approche en terme de diagnostic. Cette activité a été réalisée à La Réunion, à l'échelle d'un agro-écosystème réunionnais de 10000m^2 sur le site expérimental du Cirad de Bassin Plat à Saint-Pierre. L'élaboration de notre approche de métagénomique a tout d'abord portée sur un nombre restreint d'échantillons (n = 144) issus d'une première prospection réalisée en novembre 2014 (**Chapitre 1 - Article 1**). Notre approche combine l'amplification en cercle roulant avec marquage aléatoire par amplification PCR (RCA-RA-PCR), le séquençage à haut débit (Illumina HiSeq) et l'assignation phylogénétique des séquences de mastrévirus obtenues. L'utilisation généralisée de protocoles basés sur la RCA avec la polymérase du phage *phi*29 a révolutionné l'étude des virus dont le génome est constitué de petites molécules d'ADN circulaire (Inoue-Nagata *et al.*, 2004 ; Varsani *et al.*, 2008) induisant notamment la découverte de nombreux mastrévirus infectant les Poaceae (Kraberger *et al.*, 2017 ; Varsani *et al.*, 2008a). Cette approche était donc particulièrement indiquée pour le développement d'un protocole de métagénomique dédié aux géminivirus.

L'étape d'amplification aléatoire utilisant la réaction en chaîne par polymérase (RA-PCR) avec des amorces comportant en 5' une étiquette spécifique (autrement appelée *tag* ou *MID* pour *multiplex identifier*) permet d'étiqueter chaque amplicon RCA avant leur multiplexage. Cette étiquette permet, après séquençage, de réattribuer chacune des séquences à un échantillon d'origine. L'un des points forts de notre approche réside d'ailleurs dans un degré de multiplexage élevé, pouvant atteindre 1200 échantillons par ligne Illumina, ce qui est supérieur a ceux employés dans les études précédentes de viromique végétale (Bernardo *et al.*, 2018 ; Roossinck *et al.*, 2010).

Par ailleurs, l'emploi des technologies de séquençage NGS aboutit à l'obtention rapide et massive de séquences à moindre coût (Barba *et al.*, 2013). Néanmoins, que ce soit du fait de la quantité élevée de données de séquences obtenues ou du fait de l'absence de séquences proches dans les bases de données de références, le classement des séquences obtenues par rapport à l'existant n'est pas une étape triviale. De ce fait, pour pallier à cette difficulté, notre approche inclus une procédure de classification des séquences qui, en complément de méthodes basées sur la recherche de similarité de type « BLAST », utilise une méthode de placement phylogénétique.
Cette méthode, si elle sera tout autant impactée par l'absence de référence connue que les méthodes alternatives d'assignation taxonomique couramment utilisées, présente l'avantage de placer l'étape de classification dans un cadre phylogénétique et statistique. Après identification des plantes infectées par les mastrévirus, l'étape de clonage et séquençage Sanger a permis d'obtenir les génomes complets de ces virus et de confirmer le diagnostic d'infection établi après le placement phylogénétique des *reads*.

Malgré les avancées majeures qu'ont induit ces approches de métagénomique, celles-ci ne sont pas exempts de limitations. En effet, l'approche RCA-RA-NGS peut induire (i) l'exclusion de certains virus pour des raisons méthodologiques (présence d'un GC *clamp* et d'une étape de cisaillement des acides nucléiques ; voir **Chapitre 1 - Article 1**), (ii) des biais d'extraction ou d'amplification (Gallet *et al.*, 2017 ; Lasken & Stockwell, 2007) ou encore (iii) la détection, même en de très faible quantité, de contaminations induites par des virus couramment étudiés dans le laboratoire, entre échantillons ou liées au séquençage lui-même (Degnan & Ochman, 2012 ; Kunin *et al.*, 2008 ; Rosseel *et al.*, 2014).

Du fait de ces limites, la détermination d'un seuil de détection approprié afin de statuer sur l'état sanitaire d'une plante constitue un enjeu majeur (Massart *et al.*, 2014). L'objectif de la dernière partie de ce chapitre a donc été d'évaluer le degré de confiance pouvant être accordé au diagnostic basé sur la métagénomique. Pour cela, cette évaluation a porté sur un nombre plus large d'échantillon que celui de l'étude pilote (campagne d'échantillonnage d'avril 2017, **Chapitre 2 - Article 3**) mais aussi sur l'analyse de témoins

positifs et négatifs obtenus après l'application du protocole de RCA-RA-NGS respectivement sur des extraits de plasmide pUC19 et d'ADN d'un plant de tomate sain. La présence de ces témoins nous a permis d'évaluer le niveau de contamination inter-échantillons et d'estimer des valeurs seuils de positivité. Une des sources connues de contamination est associée à des phénomènes dit d'*index hopping/switching*. Ces contaminations peuvent être liées à la présence d'amorces libres résiduelles dans la *librairie* de séquençage ou encore à la confusion de spot de séquençage, induisant lors du démultiplexage l'assignation de la séquence à un mauvais échantillon (Costello *et al.*, 2018 ; Sinha *et al.*, 2017). A l'aide d'une procédure de recherche de séquences identiques entre échantillons, nous avons pu démontrer l'existence de contaminations *inter-librairies* et leur prise en compte dans l'estimation des seuils de confiance (**Chapitre 1 - Article 2**). Enfin, l'utilisation de réplicats (à partir de l'étape de RA) s'est révélée déterminante pour l'évaluation du statut sanitaire des échantillons.

Bien que l'approche RCA-RA-NGS présente des limites, spécifiques ou inhérentes à la métagénomique, son potentiel a été validé par l'identification de deux espèces de mastrévirus préalablement décrites à La Réunion (MSV et MSRV) mais également de trois nouvelles espèces de mastrévirus jamais encore décrites et nommées Eleusine indica associated virus (EIAV), Melinis repens associated virus (MeRAV) et Sorghum arundinaceum associated virus (SAAV ; **Chapitre 1 - Article 1**). De telles identifications a partir d'un échantillonnage restreint (ici, 144 échantillons) sont très prometteuses et démontrent le potentiel de cette approche métagénomique dédiée à l'étude d'une famille virale. La preuve de concept de notre approche métagénomique RCA-RA-NGS (**Chapitre 1 - Article 1**) nous a ouvert la voie vers une nouvelle étude plus approfondie avec un effort d'échantillonnage plus important du même agro-écosystème afin de pouvoir établir plus finement les contours de la diversité des mastrévirus à La Réunion dans notre agrosystème mais aussi pour comprendre comment ceux-ci sont structuré entre leurs hôtes.

Article 1

Exploring the diversity of Poaceae-infecting mastreviruses on Reunion Island using a viral metagenomics-based approach

SCIENTIFIC REPORTS

nature research

OPEN

Exploring the diversity of Poaceae-infecting mastreviruses on Reunion Island using a viral metagenomics-based approach

Received: 7 April 2019
Accepted: 20 August 2019
Published online: 03 September 2019

Sohini Claverie[1,2], Alassane Ouattara[3,4], Murielle Hoareau[1], Denis Filloux[7,8], Arvind Varsani[5,6], Philippe Roumagnac[7,8], Darren P. Martin[9], Jean-Michel Lett[1] & Pierre Lefeuvre[1]

Mostly found in Africa and its surrounding islands, African streak viruses (AfSV) represent the largest group of known mastreviruses. Of the thirteen AfSV species that are known to infect either cultivated or wild Poaceae plant species, six have been identified on Reunion Island. To better characterize AfSV diversity on this island, we undertook a survey of a small agroecosystem using a new metagenomics-based approach involving rolling circle amplification with random PCR amplification tagging (RCA-RA-PCR), high-throughput sequencing (Illumina HiSeq) and the mastrevirus reads classification using phylogenetic placement. Mastreviruses that likely belong to three new species were discovered and full genome sequences of these were determined by Sanger sequencing. The geminivirus-focused metagenomics approach we applied in this study was useful in both the detection of known and novel mastreviruses. The results confirm that Reunion Island is indeed a hotspot of AfSV diversity and that many of the mastrevirus species have likely been introduced multiple times. Applying a similar approach in other natural and agricultural environments should yield sufficient detail on the composition and diversity of geminivirus communities to precipitate major advances in our understanding of the ecology and the evolutionary history of this important group of viruses.

Members of the *Mastrevirus* genus, one of the nine genera in the *Geminiviridae* family[1], cause diseases of economically important crops such as maize[2], wheat[3], sugarcane[4] and chickpea[5]. These viruses have ~2.5–2.7 kb circular single-stranded DNA genomes and are transmitted by leafhoppers (in the Cicadellidae family) to a range of either monocotyledonous or dicotyledonous host species.

Most of the known monocot-infecting mastreviruses have been identified either in Africa or surrounding islands. These mastreviruses have collectively been called the "African streak viruses" (AfSV). There are presently thirteen recognised AfSV species, nine of which have been found infecting both cultivated and non-cultivated host species, four of which have only been found infecting non-cultivated grasses[6].

Viruses that infect cultivated plants represent only a small fraction of ecosystem level viral diversity[7]. As with many other groups of viruses, only a small fraction of the mastreviruses that currently exist has been discovered[7–9]. It is very likely that large numbers of presently unknown viruses reside within the thousands of non-cultivated and largely unsampled plant species that are found within the unmanaged portions of terrestrial environments. The significance of non-cultivated plant species both in the functioning and maintenance of virus communities, and in the emergence of new viral pathogens, has remained largely unexplored. This is in part due to the fact that

[1]CIRAD, UMR PVBMT, F-97410, St Pierre, La Réunion, France. [2]Université de La Réunion, UMR PVBMT, Pôle de Protection des Plantes, 7 Chemin de l'IRAT, Saint-Pierre, 97410, France. [3]INERA, 01 BP 476, Ouagadougou 01, Burkina Faso. [4]Laboratoire Biosciences, Université Joseph KI-ZERBO, 03 BP 7021, Ouagadougou 03, Burkina Faso. [5]The Biodesign Center for Fundamental and Applied Microbiomics, Center for Evolution and Medicine, School of Life Sciences, Arizona State University, 1001 S. McAllister Ave, Tempe, AZ 85287-5001, USA. [6]Structural Biology Research Unit, Departement of Integrative Biomedical Sciences, University of Cape Town, Observatory, Cape Town, South Africa. [7]CIRAD, UMR BGPI, F-34398, Montpellier, France. [8]BGPI, Université de Montpellier, INRA, CIRAD, Montpellier SupAgro, F-34398, Montpellier, France. [9]Computational Biology Division, Departement of Integrative Biomedical Sciences, Institute of Infectious Diseases and Molecular Medicine, University of Cape Town, Observatory, South Africa. Correspondence and requests for materials should be addressed to P.L. (email: pierre.lefeuvre@cirad.fr)

many non-cultivated species display no easily identifiable symptoms when they are virus-infected[10,11], and partly because of the prohibitive cost and effort that is required to acquire sufficient viral genomic sequence data from large-enough numbers of plants to obtain a global view of viral diversity at the ecosystem scale.

The development and widespread use of rolling circle amplification (RCA) based protocols for the recovery of complete mastrevirus genome sequences has vastly increased the feasibility of identifying viruses in large numbers of non-cultivated plant samples[6]. RCA has facilitated both the discovery of many new mastreviruses, and the characterization of diverse strain groups within already known species. For example, although the "A-strain" of maize streak virus (MSV-A), which causes severe disease in maize, was characterized in the early 1980s and four other "grass-infecting" strains were characterized during the 1990s following extensive sampling of non-maize grass species, RCA-based appoaches enabled the discovery of six additional MSV strains in a single small study[12]. Moreover, characterization of these additional strains revealed that the last common ancestor of all known MSV-A isolates was likely a recombinant of two "grass adapted" MSV-variants: one belonging to the MSV-B strain and the other most closely related to the MSV-F and MSV-G strains[12]. Although there is no direct proof that this recombination event triggered the emergence of MSV-A as a maize pathogen, its discovery highlights once again the importance of considering virus diversity at the ecosystem scale rather than at the scale of a single pathosystem. Similar discoveries for viruses in other families are shifting the global plant disease paradigm and have highlighted the utility of studying the whole viral communities (virome) within the context of their biotic environments (pathobiomes) to achieve a better understanding of how, for example, crop pathogens emerge from largely hypovirulent natural viral communities[13,14].

Concomitant to, and in some cases driving our present perception of the roles that plant-viruses play in global ecosystems, has been the development and refinement of plant virus metagenomics techniques which enable the characterization of entire viral communities at ecosystem scales[9,15]. A growing number of methodologies have been developed that are based on the high throughput sequencing of total nucleic acids, nucleic acids from purified virus-like particles, or enriched virus-specific nucleic acids. By strategically tagging samples to preserve information on their host origins, one approach, called "eco-genomics"[11], has proven extremely fruitful in the discovery of hundreds of new plant-virus associations within natural ecosystems[16]. However, the main challenge facing the widespread application of such metagenomics protocols remains their cost and the technical challenges inherent of acquiring and analysing the massive data sets that are generated.

Here, by focusing solely on viruses with circular DNA genomes, the power of RCA has been harnessed to enrich for circular viral nucleic acids and, in so doing, avoid many of the cumbersome aspects of ecogenomics approaches. The metagenomics approach used here is easy to implement and involves denser pooling of samples than in previous studies, while still yielding enough viral genomic sequence data to enable the study of virus diversity at both the individual plant and ecosystem scales. In addition, along with a conventional similarity search procedure for read assignment, a statistical phylogenetic placement methodology was used to classify virus sequence reads. As a proof of concept, this new metagenomic methodology was applied to a set of non-cultivated plants sampled from a single farm on Reunion Island. In addition to finding viruses that belong to two of the six mastreviruses species that have previously been detected on Reunion Island, MSV and maize streak Reunion virus (MSRV)[6,17], this article further highlights the gaps that likely remain in our appreciation of mastrevirus diversity on this island by discovering mastreviruses that could represent three new mastrevirus species.

Methods

Sampling. Leaf samples of Poaceae plants (n = 144) were randomly collected regardless of whether they were symptomatic or asymptomatic on a 1000 m^2 fallow plot at the Bassin Plat CIRAD experimental facility (Latitude −21.3231; Longitude 55.4912) in Saint Pierre (Reunion Island) during November 2014 (see Supplementary Table S1 for details). Only two of the collected plants presented visible typical symptoms of streak disease. Samples were dried in an oven at 50 °C overnight and stored at room temperature before use.

Metagenomic approach. A metagenomic approach based on RCA-Random PCR amplification tagging (RCA-RA-PCR) followed by a high throughput sequencing was developed (Fig. 1). Total genomic DNA was extracted from dried leaf material using the DNeasy Plant DNA extraction kit (Qiagen, USA) according to the manufacturer's instructions and was then stored at −20 °C before use. RCA was carried out on each DNA extract using the Illustra Templiphi Kit (GE Healthcare, USA). A random amplification step using Polymerase Chain Reaction (RA-PCR) was then performed on diluted RCA products. This relied on the use of random primers each having at its 5′ extremity a single barcode of eight nucleotides, followed by six random nucleotides and the nucleotide motif TGGC (5′-BARCODE(8nt)-NNNNNN-TGGC-3′). A total of 160 barcodes previously defined[18] with an edit distance of three were used, meaning that a minimum of three single-nucleotide changes (insertions, deletions or substitutions) are required for one barcode to be confound with another. Each sample was subjected to two independent random PCR reactions using two distinct random primers. In a total volume of 25 µl, 1 µl of diluted RCA DNA (1:10) was mixed with 5 µl of 5X buffer, 1 µl of dNTP (2 mM), 2.5 µl of MgCl2 (1.25 mM), 10 µl of random primers (4.5 µM) and 0.25 µl of GoTaq polymerase (Promega, USA). After an initial denaturation step at 94 °C for 3 minutes, 35 PCR cycles (at 94 °C for 1 minute, 50 °C for 1 minute and 72 °C for 1 minute) were carried out before a final elongation step for 5 minutes at 72 °C.

To enable equimolar pooling of the RA-PCR products, hereafter called amplicons, quantification was carried out using the Quant-iT PicoGreen dsDNA Assay Kit (Thermo Fisher Scientific, USA). Along with standards (eight duplicates of a DNA standard ranging from 0 to 40 ng/µl), 2 µl of diluted amplicons (1:4) were added to 98 µl of Picogreen mix (diluted to 1/200 with Tris-EDTA 1 ×, supplied by the manufacturer) within a Fast Optical 96-well reaction plate (Applied Biosystems®). After vortexing and centrifugation at 600 rpm, plates were incubated in the dark at room temperature for five minutes. Fluorescence measurements were performed at 25 °C using a StepOnePlus Real-Time PCR system (Applied Biosystem, USA) at the excitation and emission

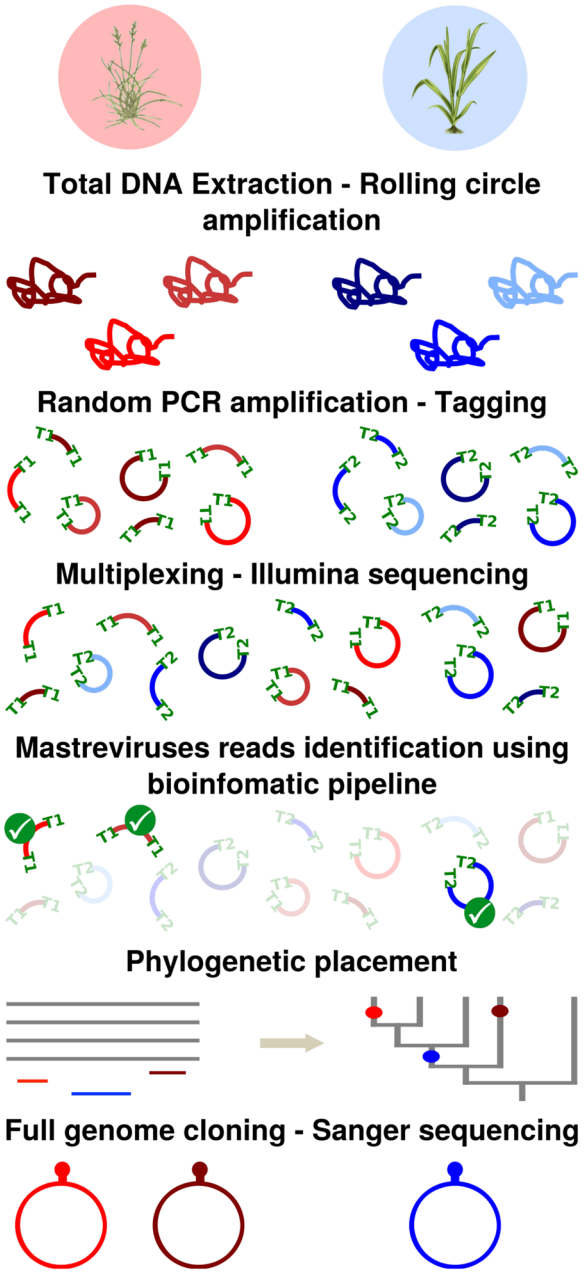


Figure 1. Schematic representation of the metagenomic approach. Total genomic DNA was extracted from dried leaf material, followed by a rolling circle amplification (RCA). A random amplification step using polymerase chain reaction (RA-PCR) combined with a tagging was performed. The use of distinct tags, here symbolized with the «T1» and «T2» labels, allows each sequence to be traced back to the original sample. During the multiplexing step, amplicons are equimolarly pooled before being submitted to Illumina library construction and sequencing. Mastrevirus sequences identified using similarity search (sequences with green ticks on the figure) were more precisely classified through phylogenetic placement analysis on mastreviruses reference alignment and tree (in grey). Viral classifications are then confirmed after cloning and Sanger sequencing using the RCA-RFLP protocol.

wavelengths of Picogreen (respectively 480 nm and 530 nm). DNA concentrations of amplicons were obtained based on the fluorescence curve of the standard. It is important to notice here that the amount of DNA obtained from each PCR reaction is mostly independent of the infection status of the sample and that all amplicons were used for downstream experiments.

The 288 amplicons (two PCR replicates for 144 samples) obtained in the study were combined in pools (with up to 160 amplicons per pool) with a maximum concentration ratio of 1.5 (*i.e.* no amplicons could have a concentration more than 1.5 times higher than any one of the others in the same pool). Amplicon pools were then purified using the Illustra GFX™ PCR DNA kit and gel purified (GE Healthcare, USA) according to the manufacturer's instructions. Once purified, the pools were quantified using the Qubit dsDNA BR Assay Kit for the

Qubit fluorometer (Thermo Fisher Scientific, USA) and checked on D5000 ScreenTape for 2200 TapeStation (Agilent Technologies, USA). Amplicon sizes ranged mostly from 300 bp to 5 kb. Amplicon pools were submitted to Illumina library construction and 2 × 250 bp paired-end sequencing on a Illumina HiSeq2500 sequencer at Genewiz (USA). A Covaris shearing step was performed to fragment amplicons to the desired size suitable for Illumina sequencing (between 200 and 700 bp) prior to library construction.

Metagenomic data analyses. A quality control step (sliding window of 30 bases with an average quality of 25 bases required) was first performed on raw Illumina reads using Trimmomatic[19] which was also used to remove Illumina adapter sequences (with a maximum mismatch of 2 bases, a palindromic match threshold in the range of 30 bases and a simple clip threshold of 10 bases). Sequences shorter than 80 nucleotides and unpaired reads were discarded. Trimmed reads were demultiplexed with exact matches using SABRE (https://github.com/najoshi/sabre). It is important to note here that, due to the Covaris shearing step, not all the amplicon pairs had an identifiable barcode on one or both reads. Pairs without any barcode were discarded from the analysis. After demultiplexing and primer sequence removal, overlapping paired-end reads were merged using PANDAseq[20] and non-overlapping paired-end reads were abutted. Merged and abutted sequences were then dereplicated using VSEARCH v1.9.5[21].

Taxonomic assignment of sequences. An initial fast taxonomic assignment of dereplicated sequences was performed using Kraken v2[22] with a custom database. This database was built with a maximum file size of 20 GB using the "max db size" option of "kraken2-build" from nucleotide sequences of archaeal, bacterial, plasmid, viral, fungal, plant, protozoan and human complete genomes within the NCBI Reference Sequence (RefSeq) and environmental sequencing project (env_nt) databases.

More precise classification of viral sequences, was achieved with similarity searches of both the viral RefSeq database using the algorithm in DIAMOND v0.9.19.120[23], and on a database containing only geminivirus and geminivirus satellite sequences using the "usearch_global" algorithm of VSEARCH. Both databases were obtained from GenBank in October 2017. To reduce the dataset size, reads with similarities to geminivirus sequences were clustered using SWARM v2.1.9[24] with the distance parameter set to 3.

Non-singleton clusters of mastrevirus sequences *i.e.* clusters gathering more than one read, were further classified with the phylogenetic placement approach implemented in pplacer (v1.1.alpha18-2-gcb55169)[25]. This phylogenetic placement method relies on a phylogenetic-based maximum-likelihood classification of sequences over a taxonomically informed reference alignment and tree which are combined into a so-called "reference package". To construct such a package, all mastrevirus complete genome sequences along with their taxonomic information were obtained from GenBank in April 2018. After linearization of the circular sequences at the virion strand origin of replication and alignment using MAFFT v7.310[26], the dataset was reduced using T-COFFEE v11[27] to the ten most informative sequences for each mastrevirus species. Note that because of the circular nature of mastrevirus genomes, the reference sequences were tandemly repeated to avoid biases in the alignment of metagenomic sequences that traversed the virion strand origin of replication. A maximum-likelihood phylogenetic tree was then constructed using FastTree v2.1.8[28]. The reference package was built using taxtastic v0.6.4 (https://github.com/fhcrc/taxtastic) and evaluated using rppr v1.1[25]. After alignment of the merged sequences to the reference package alignment using the "addfragments" option of MAFFT, the phylogenetic placement was performed using pplacer with default parameters and sequence classifications obtained using rppr and guppy v1.1[25]. These classifications were analysed using the BoSSA R package (https://cran.r-project.org/package=BoSSA). Groups of samples with similar viral infection profiles were obtained after the clustering of the Kantorovich–Rubinstein distance matrix directly obtained from the placement files using guppy.

Cloning and full genome sequencing. Representative samples from each viral infection profile group were selected for full genome cloning and sequencing. Full mastrevirus genomes were obtained using a RCA-RFLP based approach as previously described[29]. Briefly, 1 µl of RCA PCR was digested using *Bam*HI, *Eco*RI or *Nco*I to yield a ~2.7 kb fragment before purification using the Illustra GFX PCR DNA and Gel Band Purification Kit (GE Healthcare) according to the manufacturer's instructions. The resulting purified fragments were ligated to the pJET 1.2 cloning vector (Thermo Fisher Scientific) and used to transform competent *Escherichia coli* (JM109, Promega). Recombinant plasmids were purified using QIAprep Spin Miniprep Kit (Qiagen) and were Sanger sequenced by Macrogen Inc. (Netherlands) using primer walking. Full-length mastrevirus genomes were then assembled with Geneious v6.0.6 (http://www.geneious.com)[30].

Phylogenetic and recombination analyses. Full genome nucleotide sequences were subjected to a BLAST search of the NCBI nt database for preliminary species assignment. One sequence of each mastrevirus species and all the genomes previously characterized from Reunion Island were selected from GenBank. These representative mastrevirus sequences, the sequences from Reunion Island and the full genome sequences determined in this study were linearized at the virion strand origin of replication and aligned using MAFFT. Pairwise similarities between the sequences were determined using SDT v1.2[31]. Recombination events were detected within the full genome sequences from Reunion Island using the RDP[32], GENECONV[33], BOOTSCAN[34], MAXCHI[35], CHIMERA[36], SISCAN[37] and 3SEQ[38] methods included in the RDP4[39] program. Default settings were used and recombination events were considered significant when detected by at least three methods. Maximum-likelihood phylogenetic trees were constructed using FastTree v2.1.8[28] and were edited using the APE[40] R package.

Host identification. Sequencing of the *matK* and *rbcL* genes were performed on plant samples for which no confident genus identification could be achieved by visual inspection. PCRs were conducted before direct Sanger sequencing by Macrogen Inc. (Netherlands) as described previously[41]. After quality control, sequences

Plant species	Number of samples	Number of positive samples	Viral profile	Mastrevirus species				
				MSV	MSRV	EIAV	SAAV	MeRAV
Brachiaria umbellata	1	0						
Cenchrus echinatus	1	1	2	1 (1)	1 (1)			
Chloris gayana	15	0						
Cynodon dactylon	12	1	1	1				
Cyperus rotondus	1	0						
Dactyloctenium aegyptium	4	1	1	1 (1)				
Digitaria ciliaris	27	6	1,2	6 (2)	1			
Eleusine indica	5	1	3			1 (1)		
Melinis repens	21	2	5					2 (1)
Urochloa maxima	31	1	4				1	
Paspalum dilatatum	1	0						
Setaria pumila	8	0						
Sorghum arundinaceum	17	3	4				3 (2)	
Total	144	16	5	9 (4)	2 (1)	1 (1)	4 (2)	2 (1)

Table 1. Summary of sampled plants and viral assignments. The number in brackets refers to the number of full complete cloned genomes.

were classified to the genus level using the RDP[42] classifier against a database of *matK* and *rbcL* plant sequences obtained from GenBank.

Results and Discussion

Raw reads and the proportion of viral sequences. After quality control and barcode searches, 13 million sequence reads remained. The number of sequences obtained for the 144 Poaceae samples ranged from ~1000 to ~390000 with a mean of 108000. A first taxonomic assignment using Kraken revealed that 7% of the reads were detectably homologous with viral sequences, 3% with bacterial sequences and 28% with plant sequences. The remaining 62% of the reads were not detectably homologous to any sequences in the database. This high proportion of unclassified "dark matter" reads emphasises the large gap between the organisms which are known and those which remain to be discovered[43]. Among the reads likely to be of viral origin, 99% were most closely related to members of the genus *Mastrevirus*, 0.08% to geminivirus-associated satellites, 0.01% to members of the family *Genomoviridae* and the remaining 0.91% to members of other viral families. Importantly, no single sample could be associated with a high number of reads classified as geminivirus-associated satellites making these classifications mostly spurious.

The Poaceae hosts of mastreviruses. To minimize the impact of contamination and sequencing errors when determining whether to flag a plant sample as being infected by a mastrevirus, it was necessary to set a threshold number of mastrevirus-derived reads above which a plant would be considered as being infected. If this threshold was set to 100 reads, the number of mastrevirus-infected plants in the sample was 23 (15%) whereas if the threshold was set to 1000 the number of infected plants decreased to 16 (11%). This article focusses on the samples determined to be infected using the more conservative threshold (*i.e.* 1000).

The sixteen mastrevirus-infected plants represent eight Poaceae species: *Cenchrus echinatus* (n = 1/1); *Cynodon dactylon* (n = 1/12); *Dactyloctenium aegyptium* (n = 1/4); *Digitaria ciliaris* (n = 6/27); *Eleusine indica* (n = 1/5); *Melinis repens* (n = 2/21); *Urochloa maxima* (n = 1/31) and *Sorghum arundinaceum* (n = 3/17) (Table 1). It is important to note here that whereas the two symptomatic *D. ciliaris* plants had clear evidence of the type of foliar streaking that is characteristic of AfSV infections in grasses, all 14 other infected plants did not present with any discernible streak symptoms. Therefore, in concordance with previous plant virus metagenomic studies, most non-cultivated plants that are infected by viruses appear to show no obvious outward signs of infection[10,11]: a factor emphasizing the importance during plant viral metagenomics studies of sampling plants regardless of their apparent health status.

Of the eight mastrevirus-infected Poaceae species, five have already been described as mastrevirus hosts either on Reunion Island (*C. echinatus* and *D. ciliaris*) or elsewhere (*C. dactylon*, *E. indica* and *U. maxima*). The remaining three species, *D. aegyptium, M. repens and S. arundinaceum* have not previously been identified as mastrevirus hosts.

Molecular characterization of mastreviruses. The most common approach that is presently used to taxonomically assign viral Illumina sequencing reads is the use of pairwise similarity searches of annotated sequences within public reference databases. Generally, these similarity searches employ BLAST or BLAST-like algorithms. When close relatives of the viral sequence reads are in the database, it is possible to assign the reads to a specific taxonomic rank. However, besides the difficulty of interpreting E-values and converting these to genetic distance measurements so that sequences can be assigned to an appropriate taxonomic rank, BLAST and BLAST-like algorithms can yield high miss-assignment rates when the reference database is inappropriate or incomplete (*i.e.* does not contain sequences that are closely related to the reads[44]).

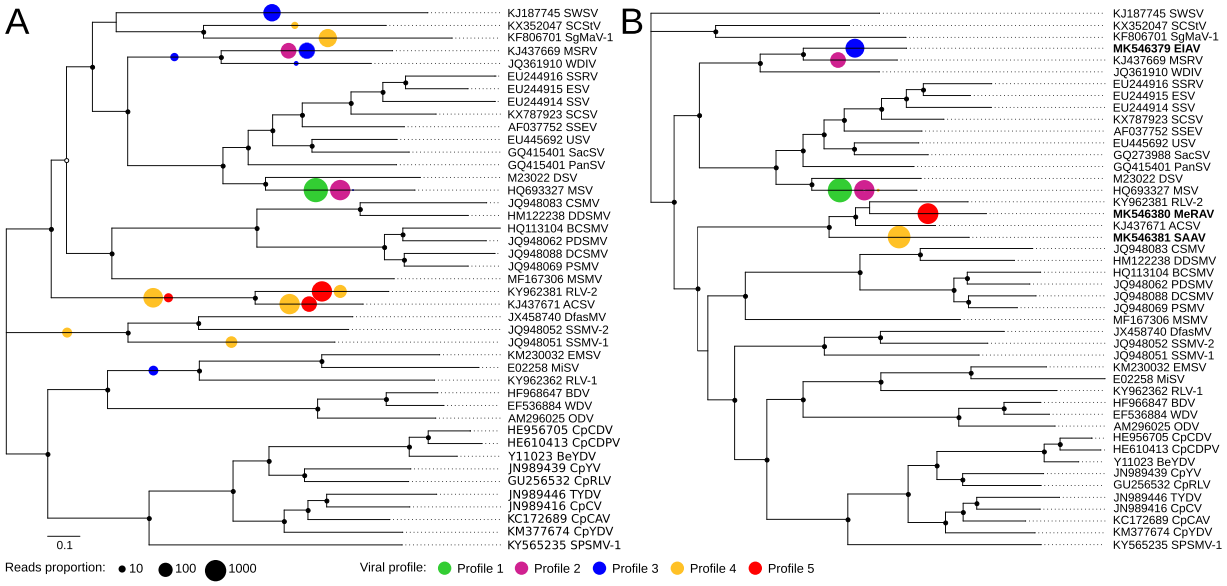


Figure 2. Phylogenetic placements of Illumina reads in a simplified Maximum-likelihood (ML) phylogenetic tree representing the breadth of known mastrevirus diversity. The ML phylogenetic tree was constructed from the complete genomes of 43 mastrevirus species (**A**) and three new mastreviruses cloned in this study (**B**). Open and closed circles on nodes indicate bootstrap support for the branches to their left of 70–89% and >=90% respectively. Phylogenetic placements are summarised with coloured circles on branches whose sizes are proportional to the number of sequencing reads it represent and colours are function of the infection profile. More detailed phylogenetic placement trees are available in Supplementary Figures S1 and S2 and complete names of mastrevirus species are available in Supplementary Table S2.

Phylogenetic placement is an alternative to similarity search-based methods of read assignment, which attempts to map query sequences to a fixed reference phylogenetic tree according to a model of evolution using maximum likelihood (ML), Bayesian or neighbour joining (NJ) methods[45]. Ideally a query sequence will be placed on a branch of the phylogenetic tree at the exact location where it would branch on the tree had it been used along with the reference sequences during tree construction. Therefore, if a query sequence is very similar to one or more of the reference sequences used to construct the tree then it should be placed on a branch within a clade close to the tips of the tree that contains its closest relatives. On the other hand, a divergent query sequence which has no close relatives among the reference sequences, should be assigned to a branch that is close to the base of the tree.

After phylogenetic assignment of the mastrevirus-derived sequence reads, the sixteen positive plant samples were grouped into five infection profiles, defined as group of samples which present with similar viral taxonomic assignments. Two infection profiles correspond with samples having sequences clearly assigned to known mastrevirus species (placed close to the tips, profiles 1 and 2) and three with sequences assignments indicative of infection with unknown mastreviruses (placement scattered on the tree and including several basal branches, profiles 3, 4 and 5; Fig. 2A and Supplementary Figure S1). The first infection profile (green spots) correspond to seven samples infected exclusively with MSV strains: *C. dactylon* (n = 1), *D. aegyptium* (n = 1) and *D. ciliaris* (n = 5). The second profile (purple spots) corresponds to two samples co-infected by MSV and MSRV strains: *C. echinatus* (n = 1) and *D. ciliaris* (n = 1). The seven other samples yielded assignments suggestive of novel mastrevirus species in that no species-level virus assignments were evident following the phylogenetic placement of reads from these samples: *E. indica* (n = 1; profile 3, blue spots); *U. maxima* (n = 1) and *S. arundinaceum* (n = 3; profile 4, yellow spots) and *M. repens* (n = 2; profile 5, red spots). Importantly, virus detection and identification was effective and was congruent between replicates for 12 out of the 16 positive samples. In the four remaining samples the the virus was detected and identifiedin only a single replicate.

To confirm the MSV-B assignments indicated by profile 1, three complete genomes were cloned and sequenced from *D. aegyptium* (n = 1) and *D. ciliaris* (n = 2) plants with profile 1 (Table 1). A complete genome obtained from *D. aegyptium* (n = 1, [Reunion-Bassin Plat-Dactyloctenium aegyptium-RE025-2014], accession no. MK546377) and another from *D. ciliaris* (n = 1, [Reunion-Bassin Plat-Digitaria ciliaris-RE019-2014], accession no. MK546376) share between 98.7 and 99.4% identity with previously determined MSV-B3 sequences. A third genome from another profile 1 *D. ciliaris* plant ([Reunion-Bassin Plat-Digitaria ciliaris-RE081-2014], accession no. MK546374) shares 94.1% identity with previously determined MSV-B1 sequences (Fig. 3).

Similarly, the phylogenetic assignments that were made for the profile 2 samples – which indicate the presence of mixed MSV-B3 and MSRV infections – were validated by cloning one full MSV-B3 genome ([Reunion-Bassin Plat-Cenchrus echinatus-RE001-2014], accession no. MK546375) and one full MSRV genome ([Reunion-Bassin Plat-Cenchrus echinatus-RE001-2014],accession no. MK546378) from a profile 2 *C. echinatus* plant (Table 1). The

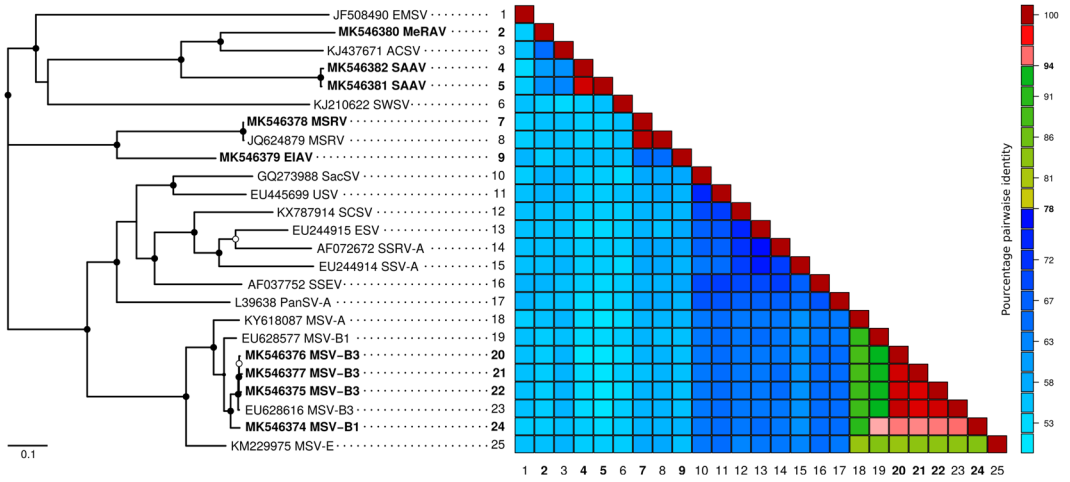


Figure 3. Maximum-likelihood phylogenetic tree (A) and pairwise sequence similarity matrix (B) of 16 known complete genomes of African streak viruses and the eight complete genomes determined in this study (indicated in bold font). The branches of the maximum likelihood tree are coloured according to geographical origins of the samples. Open and closed circles on nodes indicate bootstrap support for the branches to their left of 70–89% and >=90% respectively. Complete names of mastrevirus species are available in Supplementary Table S2.

cloned MSV-B3 genome shared 99.1% identity with previously determined MSV-B3 sequences and the cloned MSRV genome shared 99.8% identity with a previously determined MSRV-A sequence (Fig. 3).

MSV-B has previously been found infecting *C. echinatus* and *D. ciliaris*[12], but this is the first time it has been characterized in *D. aegyptium* and *C. dactylon*. MSRV has previously been isolated in diverse Poaceae species (*Rottboellia* sp., *Setaria barbata* and *Zea mays)* but has not previously been identified in *C. echinatus*. It must be noted here that the phylogenetic placement results for these mixed infections could not exclude the possibility that the apparent mixed infections in profile 2 plants could have been a consequence of these plants being infected by a recombinant of MSV and MSRV (rather than being attributable to a co-infection). However, our recovery of non-recombinant MSV-B and MSRV genomes from one of the profile 2 plants indicates that, even if MSV-MSRV recombinants occur in these plants, they are unlikely to represent a majority of the mastrevirus sequences within these plants.

Further, to confirm that the profiles 3, 4 and 5 samples did indeed contain novel mastreviruses, full length virus genomes were cloned and sequenced from at least one plant with profiles 3, 4 and 5 (Fig. 2B and Supplementary Figure S2): one from the *E. indica* (profile 3, blue spot, [Reunion-Bassin Plat-RE004-2014], accession no. MK546379), two from *S. arundinaceum* (profile 4, yellow spot, [Reunion-Bassin Plat-RE034-2014], [Reunion-Bassin Plat-RE084-2014], accession no. MK546381 and MK546382 respectively) and one from *M. repens* (profile 5, red spot, [Reunion-Bassin Plat-RE027-2014], accession no. MK546380; Table 1). When these newly determined sequences were included in the reference package and another round of phylogenetic placement performed (Fig. 2B and Supplementary Figure S2) all the reads of each of the seven profile 3, 4 and 5 samples were clearly assigned to clades containing the newly sequenced genomes. This indicated both that the cloned genomes do indeed represent all of the reads that were poorly assigned (placed on basal branches of the tree) during the first round of phylogenetic placement, and that there were no additional divergent mastrevirus sequences coinfecting the profiles 3, 4 and 5 plants.

The novel mastrevirus genome from *E. indica* ([Reunion-Bassin Plat-RE004-2014], accession no. MK546379) shares 65.6% nucleotide sequence identity with MSRV-A, that from *S. arundinaceum* ([Reunion-Bassin Plat-RE034-2014], [Reunion-Bassin Plat-RE084-2014], accession no. MK546381 and MK546382 respectively) shares 61.8–62.4% identity with *Axonopus compressus streak virus* (ACSV), and that from *M. repens* ([Reunion-Bassin Plat-RE027-2014], accession no. MK546380) shares 66.2% identity with ACSV (Fig. 3). Based on the current ICTV recommended pairwise similarity-based species demarcation threshold for mastreviruses (<78% identity[12]), these viruses therefore represent three novel mastrevirus species. The names *Eleusina indica associated virus* (EIAV), *Sorghum arundinaceum associated virus* (SAAV) and *Melinis repens associated virus* (MeRAV) were proposed for these putative species.

A maximum-likelihood (ML) phylogenetic tree constructed with these three sequences and a representative selection of other mastreviruses (Fig. 3) confirms the pairwise sequence analysis results that EIAV is most closely related to MSRV and that both SAAV and MeRAV are most closely related to ACSV. It also confirms that these three novel mastreviruses fall within the AfSV group. EIAV, SAAV and MeRAV have a typical mastrevirus genome architecture. However, whereas the EIAV and SAAV sequences contain a canonical geminivirus TAATATTAC nonanucleotide sequence at the probable origin of virion-strand DNA replication, the MeRAV genome has an unusual TAACATTGC nonanucleotide motif.

It is well known that intra- and inter-species recombination is very common in mastreviruses[42–44]. Because of the high degree of mastreviruses diversity on Reunion Island, all Reunion Island mastrevirus sequences were

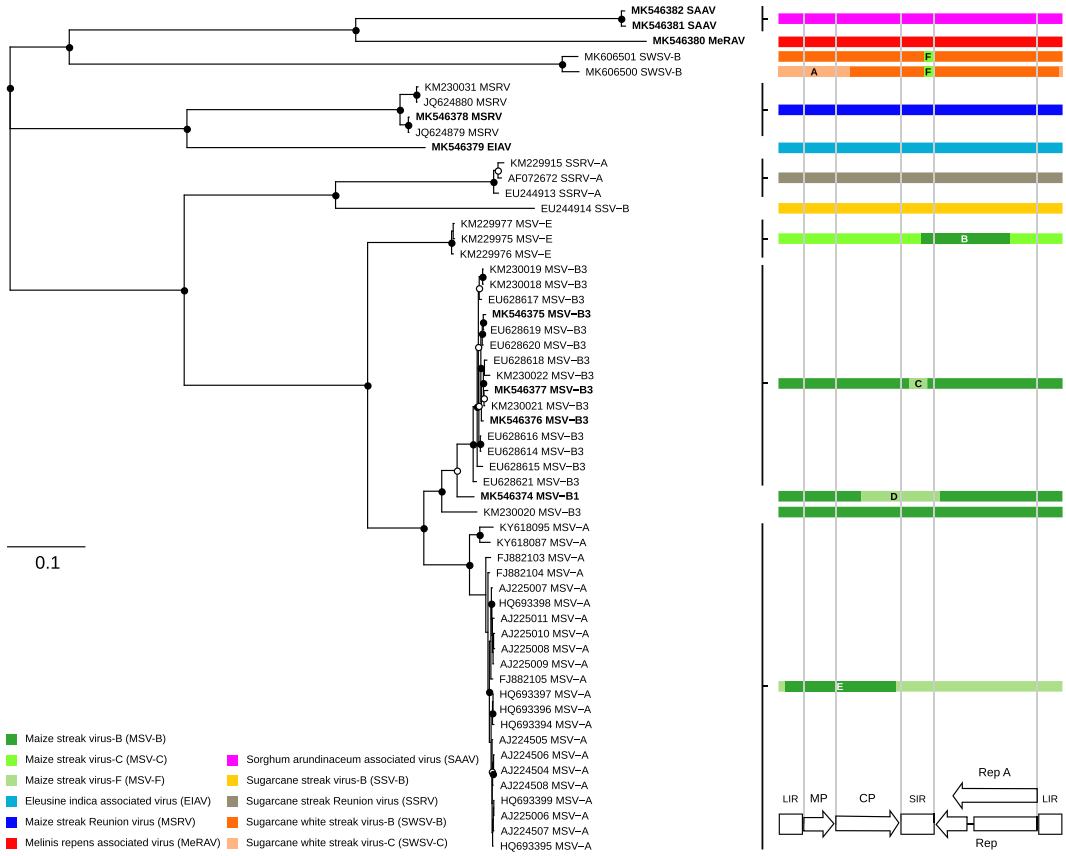


Figure 4. Phylogenetic relationships and recombination patterns among the AfSV species on Reunion Island. The maximum-likelihood phylogenetic tree contains 47 known complete genomes of monocot-infecting mastreviruses from Reunion Island and eight complete genomes determined in this study (indicated in bold font). The tree was rooted on chickpea chorosis virus (JN989413) as an outgroup (not shown). Open and closed circles on nodes indicate bootstrap support for the branches to their left of 70–89% and $>=90\%$ respectively. The schematic representation of recombination events detected by RDP4 using seven different methods: RDP, GENCONV, BOOTSCAN, MAXCHI, CHIMERA, SISCAN and 3SEQ. Arrows and blocks at the bottom correspond respectively to open reading frames (ORFs) and intergenic regions: movement protein (*MP*), coat protein (CP), replication-associated proteins (Rep and Rep A), long intergenic region (LIR) and small intergenic region (SIR)The colours of blocks represent the different AfSV species and strains. More details on each event (lettered A to F) are available in Supplementary Table S3.

therefore analysed for evidence of recombination (Fig. 4, Supplementary Table S3). No recombination events were identified in the newly characterized MeRAV, SAAV and EIAV genomes. However, six recombination events were detected in other viruses: five of which were intra-species events (A to E in Fig. 4, Supplementary Table S3) in the *Sugarcane white streak virus* strain B (SWSV-B), MSV-E, MSV-B3, MSV-B1 and MSV-A sequences; and one of which was an inter-species event (F) in the SWSV-B sequence (Fig. 4, Supplementary Table S3). The inter-species recombination event was detected in the short intergenic region (SIR) whereas the intra-species recombination events were identified in the long intergenic region (LIR), SIR, the virion sense open reading frames (ORFs), MP and CP, and the C-terminus portion of Rep and Rep A ORFs. The LIR and SIR have previously been determined to be recombination hotspots in AfSVs[12,46]. Recombination events in SIR regions in MSV-B3 genomes (event C) and in MP and CP regions in MSV-A (event E) have been previously reported[12]. However, it's the first description of recombination events in MSV-B1 (event D), SWSV-B (events A and F) and MSV-E (event B) in such genome regions (Fig. 4, Supplementary Table S3).

The inter-species recombination event involved a small genomic fragment transfer representing only 2.2% of the full genome size while the intra-species recombination events involved mainly large genomic fragment exchanges involving on average 24.7% of the full genome length. The sizes of these transferred genome fragments are on average larger than those previously detected in the genomes of mastreviruses infecting monocotyledonous plants[12,47,48].

Technical aspects with relevance for future studies of geminivirus diversity. From the perspective of future geminivirus diversity studies, it is noteworthy that the metagenomics approach used here permitted the pooling of 144 non-cultivated plants that were sampled at the scale of an individual farm. This degree of pooling, while higher than that used in previous plant metagenomics studies is still not at the upper limit of pooling that is possible using our approach. In fact, it should be possible to successfully pool up to 1,200 samples in a

full Illumina sequencing lane. Given such dense pooling, our approach should be economical for preliminarily testing the infection status of large number of plants such as that required for plant quarantine testing and epidemiological surveillance. From the perspective of basic viral ecology research, our phylogenetics based taxonomic assignment approach should enable deep insights into the known and unknown geminiviral populations that are present within any given population of plants no matter the sampling scale.

It is important to stress, however, that there are also potential pitfalls associated with the use of both RCA and RA. Firstly, the RCA reaction can generate chimeric products that could lead to the detection of artifactual recombinants[49]. Second, RCA is not unbiased in its amplification of circular molecules and it should not be used to infer the actual relative frequencies of genetic variants in a population[50]. Third, RCA products were shown to be only representative of 0.05–0.06% of the initial DNA templates for another family of plant infecting circular ssDNA viruses and may therefore result in the detection of only the most common viruses within co-infections where there is a large numerical imbalance between the titres of the coinfecting virus genomes[51]. Fourth, the random primers used for RA possess a DNA clamp (TGGC) on the 3′ extremity that is intended to promote PCR efficiency. However, this clamp will reduce the randomness of the hybridization site and will therefore introduce biases with respect to both the genomic regions that are amplified most efficiently, and the viral genomic sequences that will be amplified most frequently. These biases, probably reduced with the low 50 °C annealing temperature during the PCR step, would impact the ability to detect viral sequences and the assembly of full genome sequences from short reads which could in turn reduce the accuracy with which sequences can be taxonomically assigned. Fifth, our approach is unable to constrain amplicon sizes and therefore requires a shearing step prior to Illumina library construction, which can result in many sequenced amplicons lacking one or both of their tags; many of these sequences must therefore be disregarded during downstream analyses.

An improvement of our approach would have been the inclusion of positive and negative controls within the sequencing reactions that could have been used to rationally determine a suitable read frequency threshold to defend against sequencing contamination. Because of the absence of such controls, an extremely conservative frequency threshold was applied and results from many samples that likely contained detectable mastrevirus sequences were disregarded. Optimising the read frequency threshold[52] will be crucial for maximizing the sensitivity with which viruses can be detected within infected hosts without incurring an elevated false positive rate.

Conclusions

The geminivirus-focused metagenomics approach presented here has proven useful in both the large-scale detection of know mastreviruses and the discovery of novel mastreviruses. It is ideally suited to studies focused on circular DNA viruses (such as geminiviruses and genomoviruses). It is unclear whether the large diversity of mastreviruses on Reunion Island is a consequence of the importation of these viruses to the island or whether some of these viruses might have originated on the island. Numerous crop, medicinal and ornamental plants have all been imported to these islands over the past centuries and it is entirely plausible that many different viruses could have been ferried to the island within these plants. Certainly, Reunion Island is a hotspot of mastrevirus diversity and it is possible that additional sampling on the island within a more diverse set of biomes will reveal even greater mastrevirus diversity than that encountered in this study. In this regard, newer "geo-referenced" metagenomic approaches (called geo-metagenomics), would be ideally suited to examining the relationships between land use history, plant distributions and viral diversity[53].

References

1. Zerbini, F. M. *et al*. ICTV virus taxonomy profile: Geminiviridae. *Journal of General Virology* **98**, 131–133 (2017).
2. Martin, D. P. & Shepherd, D. N. The epidemiology, economic impact and control of maize streak disease. *Food Security* **1**, 305–315 (2009).
3. Lindsten, K. & Lindsten, B. Wheat dwarf — an old disease with new outbreaks in Sweden. *Journal of Plant Diseases and Protection* **106**, 325–332 (1999).
4. Bigarré, L. *et al*. Nucleotide sequence evidence for three distinct sugarcane streak mastreviruses. *Archives of Virology* **144**, 2331–2344 (1999).
5. Kraberger, S. *et al*. Molecular diversity of Chickpea chlorotic dwarf virus in Sudan: High rates of intra-species recombination - a driving force in the emergence of new strains. *Infection, Genetics and Evolution* **29**, 203–215 (2015).
6. Kraberger, S. *et al*. Molecular diversity, geographic distribution and host range of monocot-infecting mastreviruses in Africa and surrounding islands. *Virus Research* **238**, 171–178 (2017).
7. Alexander, H. M., Mauck, K. E., Whitfield, A. E., Garrett, K. A. & Malmstrom, C. M. Plant-virus interactions and the agro-ecological interface. *European Journal of Plant Pathology* **138**, 529–547 (2014).
8. Malmstrom, C. M., Melcher, U. & Bosque-Pérez, N. A. The expanding field of plant virus ecology: Historical foundations, knowledge gaps, and research directions. *Virus Research* **159**, 84–94 (2011).
9. Stobbe, A. H. & Roossinck, M. J. Plant virus metagenomics: what we know and why we need to know more. *Frontiers in Plant Science* **5**, 1–4 (2014).
10. Muthukumar, V. *et al*. Non-cultivated plants of the Tallgrass Prairie Preserve of northeastern Oklahoma frequently contain virus-like sequences in particulate fractions. *Virus Research* **141**, 169–173 (2009).
11. Roossinck, M. J. M. J. *et al*. Ecogenomics: Using massively parallel pyrosequencing to understand virus ecology. *Molecular Ecology* **19**, 81–88 (2010).
12. Varsani, A. *et al*. Recombination, decreased host specificity and increased mobility may have driven the emergence of maize streak virus as an agricultural pathogen. *Journal of General Virology* **89**, 2063–2074 (2008).
13. Duffy, S. & Holmes, E. C. Phylogenetic Evidence for Rapid Rates of Molecular Evolution in the Single-Stranded DNA Begomovirus Tomato Yellow Leaf Curl Virus. *Journal of Virology* **82**, 957–965 (2008).
14. Vayssier-Taussat, M. M. *et al*. Shifting the paradigm from pathogens to pathobiome: new concepts in the light of meta-omics. *Frontiers in Cellular and Infection Microbiology* **4**, 1–7 (2014).
15. Roossinck, M. J. & García-Arenal, F. Ecosystem simplification, biodiversity loss and plant virus emergence. *Current Opinion of Virology* **10**, 56–62 (2015).
16. Roossinck, M. J., Martin, D. P. & Roumagnac, P. Plant virus metagenomics: Advances in virus discovery. *Phytopathology,* **105**, 716–727 (2015).

17. Boukari, W. *et al*. Occurrence of a novel mastrevirus in sugarcane germplasm collections in Florida, Guadeloupe and Réunion. *Virology journal* **14**, 146 (2017).
18. Faircloth, B. C. & Glenn, T. C. Not all sequence tags are created equal: designing and validating sequence identification tags robust to indels. *PLoS One* **7**, e42543 (2012).
19. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: A flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114–2120 (2014).
20. Marsella, A., Bartram, A. K., Truszkowski, J., Brown, D. G. & Neufeld, J. D. PANDAseq: Paired-eND Assembler for Illumina sequences. *BMC Bioinformatics* **13**, 31 (2012).
21. Rognes, T., Flouri, T., Nichols, B., Quince, C. & Mahé, F. VSEARCH: a versatile open source tool for metagenomics. *PeerJ* **4**, e2584 (2016).
22. Wood, D. E. & Salzberg, S. L. Kraken: ultrafast metagenomic sequence classification using exact alignments. *Genome Biology* **15**, R46 (2014).
23. Tang, M. *et al*. Multiplex sequencing of pooled mitochondrial genomes - A crucial step toward biodiversity analysis using mito-metagenomics. *Nucleic Acids Research* **42**, 1–13 (2014).
24. Mahé, F., Rognes, T., Quince, C., de Vargas, C. & Dunthorn, M. Swarm v2: highly-scalable and high-resolution amplicon clustering. *PeerJ*, **3**, e1420 (2015).
25. Matsen, F. A., Kodner, R. B. & Armbrust, V. E. pplacer: linear time maximum-likelihood and Bayesian phylogenetic placement of sequences onto a fixed reference tree. *BMC Bioinformatics* **11**, 538 (2010).
26. Katoh, K. & Standley, D. M. MAFFT multiple sequence alignment software version 7: Improvements in performance and usability. *Molecular biology and evolution* **30**, 772–780 (2013).
27. Notredame, C., Higgins, D. G. & Heringa, J. T-coffee: A novel method for fast and accurate multiple sequence alignment. *Journal of molecular biology* **302**, 205–217 (2000).
28. Price, M. N., Dehal, P. S. & Arkin, A. P. FastTree 2 - Approximately maximum-likelihood trees for large alignments. *PLoS One* **5**, e9490 (2010).
29. Inoue-Nagata, A. K., Albuquerque, L. C., Rocha, W. B. & Nagata, T. A simple method for cloning the complete begomovirus genome using the bacteriophage φ29 DNA polymerase. *Journal of virological methods* **116**, 209–211 (2004).
30. Kearse, M. *et al*. Geneious basic: An integrated and extendable desktop software platform for the organization and aalysis of sequence data. *Bioinformatics* **28**, 1647–1649 (2012).
31. Muhire, B. M., Varsani, A. & Martin, D. P. SDT: A virus classification tool based on pairwise sequence alignment and identity calculation. *PLoS One*, **9** (2014).
32. Martin, D. & Rybicki, E. RDP: detection of recombination amongst aligned sequences. *Bioinformatics* **16**, 562–563 (2000).
33. Padidam, M., Sawyer, S. & Fauquet, C. M. Possible emergence of new geminiviruses by frequent recombination. *Virology* **265**, 218–225 (1999).
34. Martin, D. P., Posada, D., Crandall, K. A. & Williamson, C. A Modified Bootscan Algorithm for Automated Identification of Recombinant Sequences and Recombination Breakpoints. *AIDS Res. Hum. Retroviruses* **21**, 98–102 (2005).
35. Maynard Smith, J. Analysing the mosaic structure of genes. *Journal of Molecular Evolution* **34**, 126–129 (1992).
36. Posada, D. & Crandall, K. A. Intraspecific gene genealogies: Trees grafting into networks. *Trends in ecology & evolution* **16**, 37–45 (2001).
37. Gibbs, M. J., Armstrong, J. S. & Gibbs, A. J. Sister-scanning: A Monte Carlo procedure for assessing signals in recombinant sequences. *Bioinformatics* **16**, 573–582 (2000).
38. Boni, M. F., Posada, D. & Feldman, M. W. An exact nonparametric method for inferring mosaic structure in sequence triplets. *Genetics* **176**, 1035–1047 (2007).
39. Martin, D. P., Murrell, B., Golden, M., Khoosal, A. & Muhire, B. RDP4: Detection and analysis of recombination patterns in virus genomes. *Virus Evolution* **1**, 1–5 (2015).
40. Paradis, E. *et al* Package 'ape': analysis of phylogenetics and evolution, http://cran.r-project.org/web/packages/ape/ape.Pdf, 1–222, https://doi.org/10.1109/TMECH.2007.897281 (2011).
41. Charlery de la Masselière, M. *et al*. Changes in phytophagous insect host ranges following the invasion of their community: Long-term data for fruit flies. *Ecology and evolution* **7**, 5181–5190 (2017).
42. Wang, Q., Garrity, G. M., Tiedje, J. M. & Cole, J. R. Naïve Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. *Applied and environmental microbiology* **73**, 5261–5267 (2007).
43. Rosario, K. & Breitbart, M. Exploring the viral world through metagenomics. *Current opinion of virology* **1**, 289–297 (2011).
44. Berger, S. A., Krompass, D. & Stamatakis, A. Performance, accuracy, and web server for evolutionary placement of short sequence reads under maximum likelihood. *Systematic biology* **60**, 291–302 (2011).
45. Bazinet, A. L. & Cummings, M. P. A comparative evaluation of sequence classification programs. *BMC Bioinformatics*, **13** (2012).
46. Muhire, B., Varsani, A. & Martin, D. P. A genome-wide pairwise-identity-based proposal for the classification of viruses in the genus Mastrevirus (family Geminiviridae). *Archives of virology* **158**, 1411–1424 (2013).
47. Varsani, A. *et al*. Panicum streak virus diversity is similar to that observed for maize streak virus. *Archives of virology* **153**, 601–604 (2008).
48. Monjane, A. L. *et al*. Reconstructing the History of Maize Streak Virus Strain A Dispersal To Reveal Diversification Hot Spots and Its Origin in Southern Africa. *Journal of virology* **85**, 9623–9636 (2011).
49. Lasken, R. S. & Stockwell, T. B. Mechanism of chimera formation during the Multiple Displacement Amplification reaction. *BMC Biotechnology* **7**, 1–11 (2007).
50. Kim, M. S., Park, E. J., Roh, S. W. & Bae, J. W. Diversity and abundance of single-stranded DNA viruses in human feces. *Applied and environmental microbiology* **77**, 8062–8070 (2011).
51. Gallet, R., Fabre, F., Michalakis, Y. & Blanc, S. The number of target molecules of the amplification step limits accuracy and sensitivity in ultradeep-sequencing viral population studies. *Journal of virology* **91**, e00561–17 (2017).
52. Massart, S., Olmos, A., Jijakli, H. & Candresse, T. Current impact and future directions of high throughput sequencing in plant virus diagnostics. *Virus Research* **188**, 90–96 (2014).
53. Bernardo, P. *et al*. Geometagenomics illuminates the impact of agriculture on the distribution and prevalence of plant viruses at the ecosystem scale. *ISME Journal* **12**, 173–184 (2018).

Acknowledgements

The authors thank Martial Grondin, Jérémy Hascoat, Gérard Lebreton and Sarah Scussel for their excellent technical support and Penelope Hartnady for her suggestions and revisions. This work was supported by the European Union (ERDF, contract GURDT I2016-1731-0006632), the *Conseil Régional de La Réunion*, the Agropolis Fondation (Labex Agro – Montpellier, E-SPACE project number 1504-004) and CIRAD. SC is a recipient of a PhD fellowship from CIRAD and the Agropolis Fondation (E-SPACE). This work was conducted on the Plant Protection Platform (3 P, IBISA) and performed using the SouthGreen HPC server.

Author Contributions

S.C., J.M.L. and P.L. conceived and designed the experiments. S.C., A.O., M.H., J.M.L. and P.L. performed the experiments. S.C., D.F., A.V., P.R., D.P.M., J.M.L. and P.L. analysed the data. S.C., A.V., P.R., D.P.M., J.M.L. and P.L. wrote the paper. J.M.L. and P.L. secured funding for the project's execution.

Additional Information

Supplementary information accompanies this paper at https://doi.org/10.1038/s41598-019-49134-9.

Competing Interests: In the interests of transparency and to help readers to form their own judgements of potential bias, authors declare no competing interests in relation to the work described.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Article 2

Amplicon-based viromics: where is the limit?

Article 2

Amplicon-based viromics: where is the limit?

Sohini Claverie[1,2], Murielle Hoareau[1], Jean-Michel Lett[1], Pierre Lefeuvre[1]

[1]CIRAD, UMR PVBMT, F-97410 St Pierre, La Réunion, France
[2]Université de La Réunion, La Réunion, France

Keywords: metagenomics, amplicon, index-hopping, threshold, diagnostic

Abstract

Viral metagenomics (viromics) methods are gaining a lot of traction in the past years and were at the root of essential discoveries in the field of virology. The associated deep sequencing techniques are now used with very high multiplexing levels making it difficult to disentangle noise from signal. Here we determine the extent of inter-sample contamination after the sequencing of geminivirus populations from 1200 plant samples in a single Illumina lane. Our study points to some controls and filters that can be employed to assess the degree of confidence of health status diagnosis.

Introduction

Metagenomic methods, that usually uses next generation sequencing (NGS) technologies, were a real breakthrough for our understanding of life and allowed a complete redefinition of the diversity of micro-organisms (Koonin and Dolja, 2018; Lefeuvre *et al*., 2019). By leveraging the high multiplexing capacity of NGS methods, it is now possible to analyse several dozens of samples in a single sequencing run (Smith *et al*., 2010). Whereas high multiplexing was cardinal for eco-genomics studies, where sequences are tracked back to single samples (*i.e.* such as a given plant), high multiplexing is also associated to a higher risk of cross-sample contamination. Cross-contaminations arise when nucleotide sequences from a given sample are in the end associated to another one. Cross-contaminations can occur at distinct steps of the sequencing procedure, such as sample nucleic acid extraction, sequencing library construction, sequencing itself or bioinformatic post-treatment. Whereas there is the possibility to limit contamination during sample processing with extreme care or using an experimental design that would reduce cross-contaminations (Faircloth *et al*., 2012; Kircher *et al*., 2012; MacConaill *et al*., 2018), some of the inter sample contaminations are apparently “unavoidable”. Due to its ultra-deep sequencing nature, even the faintest contamination would be detected and may influence the conclusion of the experiment (Tosar *et al*., 2014). Moreover, several metagenomic procedures include one or more polymerase chain reaction (PCR) steps to amplify a target sequence. As a result, NGS-based approaches will be as

susceptible to contamination problems as are PCR-based assays. This is in fact frequently observed but rarely reported observation in the majority of laboratories that have investigated NGS approaches (Massart *et al.*, 2014). Beside the difficulties that viromics poses to diagnosis (Massart *et al*., 2017), where the correctness of results is paramount, deep sequencing challenge the definition of a positive sample (Massart *et al*., 2014): in most if not all plant samples that may be analysed, viral sequences will be discovered. Also, the most frequently used high throughput technology, using reversible terminated chemistry, is also sensitive to what is called index cross-talk. Cross-talk happens during a sequencing run, when a wrong index is associated to a sequencing cluster due to the presence of free indexed-primer in the library or to the confusion between cluster signals during the run (Carlsen *et al*., 2012; Mitra *et al*., 2015; Sinha *et al*., 2017). This confusion of spots has been evaluated and suggested to be more frequent (Sinha *et al*., 2017; Van Der Valk *et al*., 2019) when using the most recent Illumina exclusion amplification chemistry along with the patterned flow cells (used starting from Illumina Hiseq X, Illumina HiSeq4000 and Novaseq devices). Together this strongly highlight the necessity to run control along the samples during the metagenomic procedure and apply *ad hoc* bioinformatic filters to flag samples that are likely to be contaminated. In this paper, we analyse the results obtained after the Illumina sequencing of plant samples following the eco-genomic procedure described in Claverie *et al.* (2019). This procedure is designed for the sequencing of small circular DNA viruses (such as those belonging to the Geminiviridae, Genomoviridae or Circoviridae families) and was used to uncover the diversity of the phytovirus from the *Mastrevirus* genus in the context of an agro-ecosystem. The procedure, called RCA-RA-NGS involves a rolling circle amplification (RCA) step using the *phi*29 DNA polymerase, a PCR amplification using random primers (RA) for the tagging of sequences from each sample and the deep sequencing using Illumina (NGS) after pooling of the samples (multiplexing). In the original procedure, whereas the samples are treated in duplicate starting from the PCR step, no control where run along the samples. Here, we describe an extension of the procedure that include positive and negative controls. Using a set of bioinformatic filters, we evaluate the inter sample contamination to define threshold above which sample may be confidently considered as positive, in the context of our

problematic. The analysis of the network of shared sequences between samples also point to index-hopping (*i.e.* inter library sample contamination) as a potent source of contamination.

Material and methods

NGS dataset

The dataset analysed in this paper was obtained after performing the RCA-RA-NGS procedure (Claverie *et al.*, 2019) on a set of 800 plant samples collected in April 2017 in a 10000m^2 agro-ecosystem at the Bassin Plat CIRAD experimental facility (Latitude -21.3231; Longitude 55.4912) in Saint Pierre (Reunion). This agro-ecosystem is a mixed environment containing grassland, wasteland, fallow and agricultural plots. Leaf samples of all monocotyledonous plant species (cultivated and non-cultivated) were randomly collected regardless of their health status (with or without symptoms). Along with samples, negative (DNA extracts from a healthy tomato plant grown in the lab from seed) and positive controls (pUC19 plasmid as available as positive control within the Illustra TempliphiTM Kit) were treated along batch of 80 samples. A total of 2400 plant amplicons were sequenced in a single Illumina lane including 1200 plant samples (800 plant sample extracted using leaf soak extraction method described in Claverie *et al.*, in preparation; PhD manuscript chapter 2 and 400 samples of the 800 collected samples treated using Dneasy Plant Mini Kit, all treated in duplicate). Sequencing was performed using Illumina HiSeqXten with 2x150pb paired-end sequencing at Genewiz facilities (USA). After demultiplexing using SABRE (https://github.com/najoshi/sabre) and quality control using TRIMMOMATIC (parameters ILLUMINACLIP:adapter_file.fa: 2:30:10 SLIDINGWINDOW:30:30 MINLEN:100) (Bolger *et al.*, 2014), the sequence set comprises ~105 M pairs of sequences distributed in 15 sequencing libraries. Sequences were then trimmed at 100nt before merging overlapping reads using PANDAseq (Marsellla *et al.*, 2012). Viral sequences were classified after similarity searches against both the viral RefSeq database using the «blastx» algorithm implemented in DIAMOND v0.9.19.120 (Buchfink *et al.*, 2015), and a geminivirus and geminivirus satellite database using BLASTn. Both databases were obtained from GenBank in October 2017. In order to precisely determine the extent of inter-sample

contamination, we search for exact sequence match among samples and libraries. To this end, after a first round of clustering with exact match at the library scale using VSEARCH (Rognes *et al.*, 2016), unique sequences (*i.e.* singletons) were discarded to reduce the dataset size. All the remaining sequences (N=32 M) were clustered in a second step using exact match. Distribution of exact match between samples and libraries was then analysed using custom script in R and function from the igraph R library (Csardi, 2006).

Results and discussion

Control analyses

In each sequencing library, two different control were run along the samples. One of the control was a negative control: the RCA-RA-NGS procedure was performed on the DNA extracts of a healthy tomato plant (blue dots on **Figure 1**). The other was a positive control: the procedure was performed on pUC19 purified DNA (red dots on **Figure 1**). The number of geminivirus reads (supposedly absent from both type of control) and the number of pUC19 reads (supposedly absent from the plant control) were determined in every control. The distribution of pUC19 reads was high in all the plasmid control (mean of 141 K reads ranging from 32 to 299 K reads) but a mean of 17 pUC19 reads (ranging from 1 to 127 reads) were also detected in the plant control. Geminivirus reads were found in every control with a mean of 88 reads (and more than 100 reads in seven out of 30 control).

The number of reads point out to a minimum threshold that should be used for the determination of what is a positive sample. While this number of reads can be considered as high, it must be borne in mind that the metagenomic protocol involves a PCR procedure that would results in tens of thousands of amplicons, making it unlikely that a successful PCR would only results in less than a hundreds of sequences. Importantly, geminivirus reads always makes less than 1% of the total reads number of a given control. Both contamination of geminivirus reads and control reads are in accordance in term of number of reads and points to 1% of reads as a conservative number for contaminated reads and noise.

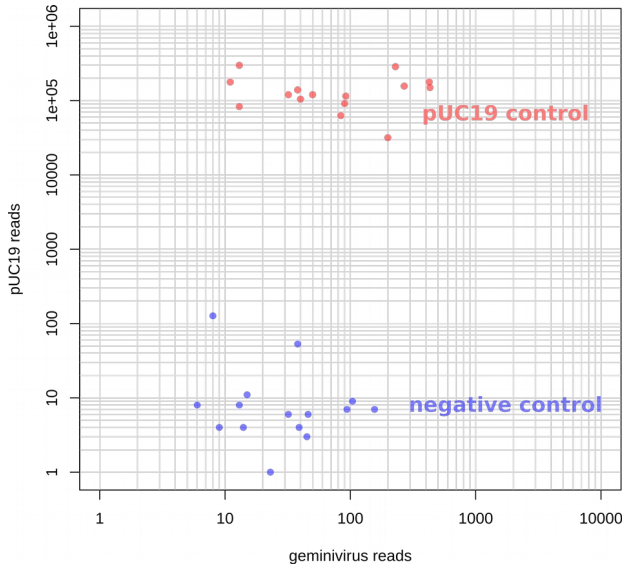


Figure 1. Plot of the number of pUC19 reads against the number of geminivirus reads for the positive control (plasmid control, red dots) and negative control (healthy plant control, blue dots).

Distribution reads per primer

As a reminder, the primers were used for a RA in order to tag amplicons obtained from the RCA step. Tags later allowed to tracked sequences back to every single sample after pooling. Each RCA product was randomly amplified in two distinct PCR reaction with two distinct tags. Each replicate was later included in distinct sequencing libraries. The primers are composed of a eight nucleotide unique tag on the 5’ end, followed with six random nucleotide and a four nucleotide clamp (5’-TAG(8nt)-NNNNNN-TGGC-3’). The specifics of the primers were discussed in Claverie *et al.* (2019). In order to determine if some of the primers used were ineffective, the number of reads per primers were plotted (**Figure 2**). The mean of the raw reads obtained per primers was 43,791 with a maximum of 283,556 and a minimum of 12 reads. The analyses of the number of raw reads per primers, revealed that the use of four of the primers resulted in ineffective amplifications (mean number of reads inferior to 1,000 reads). The samples amplified with these primers were thus removed from the analyses. Among the remaining 2,280 plants amplicon sets, 302 displayed a number of raw reads inferior to 10,000. All the samples for which one of the replicate present with less than 10,000 reads were removed from the analyses. It results to the loss of 223 additional samples. The loss of a

large fraction of the original sample set (~24%, (60+223)/1200) highlight one of the limitation of high multiplexing, the pooling of samples in equimolar ratios.

Virus amplification repeatability

The PCR amplicons associated with the remaining 917 samples (after the removal of ineffective primers and amplifications) were tested for repeatability. After identification of the geminivirus reads using similarity search algorithm, the repeatability of detection by comparing sample replicates was evaluated (**Figure 3**).

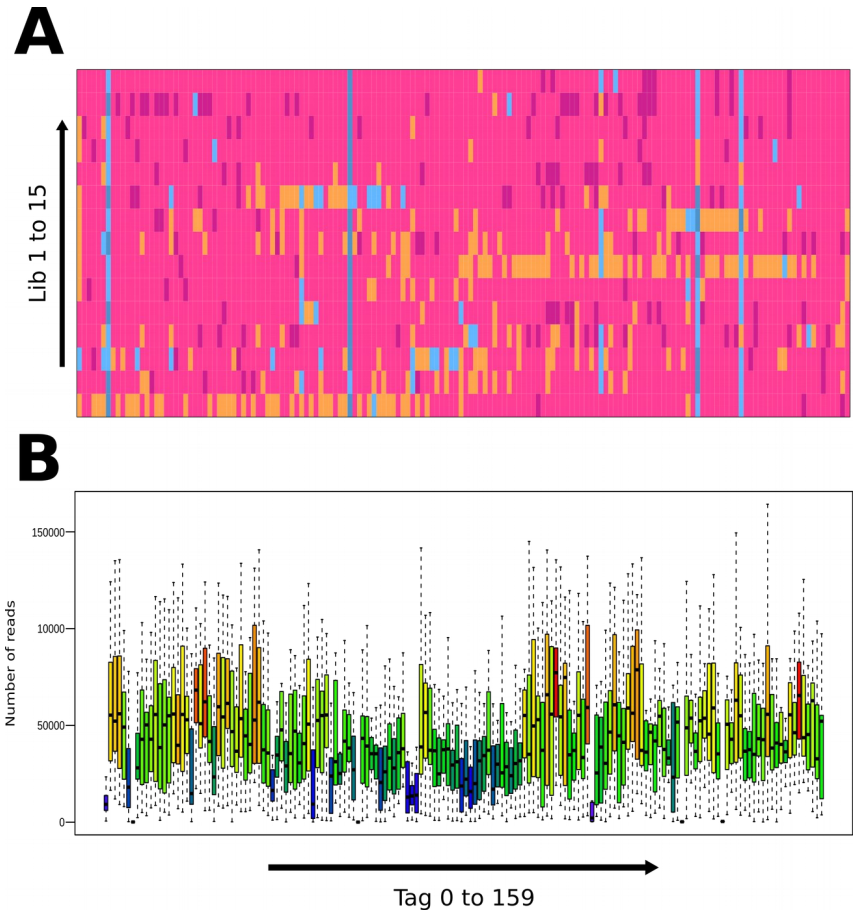


Figure 2. (A) Schematic representation of the number of raw reads obtained for each amplicons. Whereas libraries are presented as row, tags are presented as columns. The colour of the cell indicates the number of reads, relative to the maximum number of reads obtained. Lower number are coloured in dark blue colour whereas higher number are coloured in purple. **(B) Boxplots of reads number per tag among the 15 libraries.** The box represents the first and third quartile of the distribution whereas the thick black line present the median. The whiskers extend to 1.5 time the interquartile range. The colour of the boxplots indicates the number of reads, relative to the median number of reads obtained. Lower number are coloured in dark purple whereas higher number are coloured in red.

Both the number of geminivirus reads (**Figure 3A**) and the proportion of geminivirus reads for the sample (**Figure 3B**) were evaluated. It is apparent that most of the samples (N=783/917; 90 %) present with less than 100 geminivirus reads for both replicates. For all these samples, the maximum proportion of geminivirus reads for a replicate was 0.6%. These samples were considered as negative, *i.e.* not infected by a geminivirus. A total of 75 samples had more than 1000 geminivirus reads for both replicates. For these samples, the maximum proportion of viral reads range from 4.5 to 56 %. The remainning 59 samples did not present with a repeatable detection pattern. Whereas for the 29 samples that have more than a thousand reads in one of the replicates, it is suspected that one of the replicate did not perform well, for the 30 samples with mid-range reads number (from 100 to 1000 for at least one of the replicate) it is more difficult to determine the infection status. All these samples would require confirmation, either using the same procedure but with different primers or using other techniques, such as classic PCR (Pallás *et al.*, 2018), LAMP (Romero *et al*., 2019), geminivirus full genome cloning (Inoue-Nagata *et al*., 2004) or long-read sequencing (Filloux *et al*., 2018, Boykin *et al.*, 2019). It is important to note that the threshold used here are just indications to sort the samples between those that are confidently negative, those that are confidently positive and those that would requires further investigations.

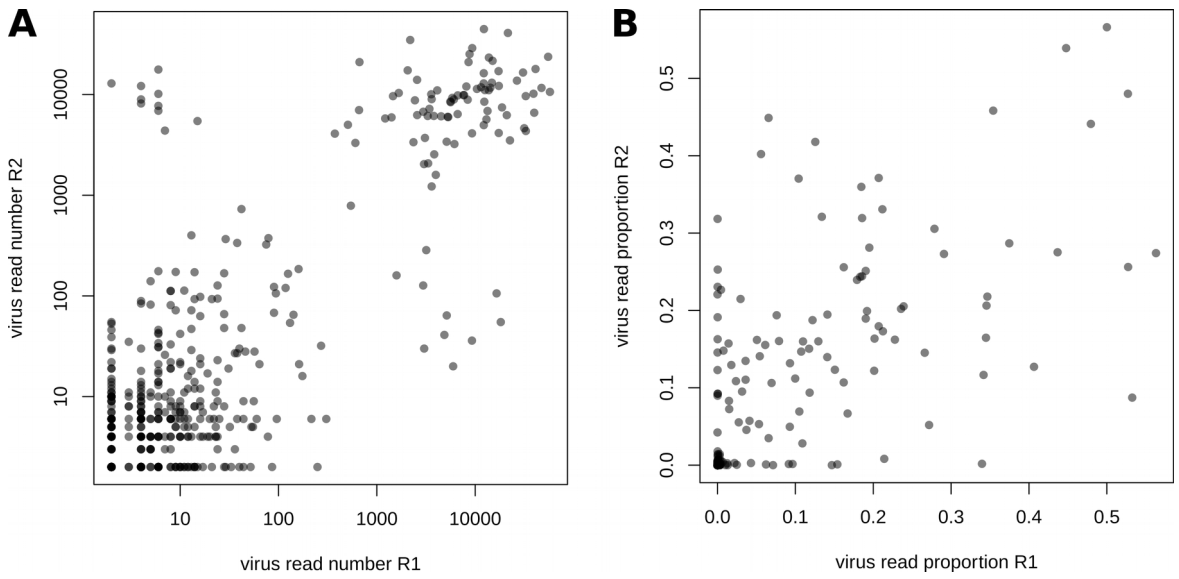


Figure 3. Plots of the number of geminivirus reads (A) and proportion of geminivirus reads (B) for both replicates obtained from a same sample.

Repeatability was also examined visually by comparing estimates of the numbers of geminivirus reads from the two replicates. Using a colour scale for the number of geminivirus reads (from dark blue to purple for low to high number of geminivirus reads), it is visible (**Figure 4** and **Supplementary Figures 1** to **3**) that globally, the repeatability is high with most of the samples with orange/purple colours in one plate presenting with orange/purple colours in the other plate. Some of the replicates are not reproducible due to the absence of amplification (very low total number of reads in one of the replicate). As a result, these samples were discarded from the analysis.

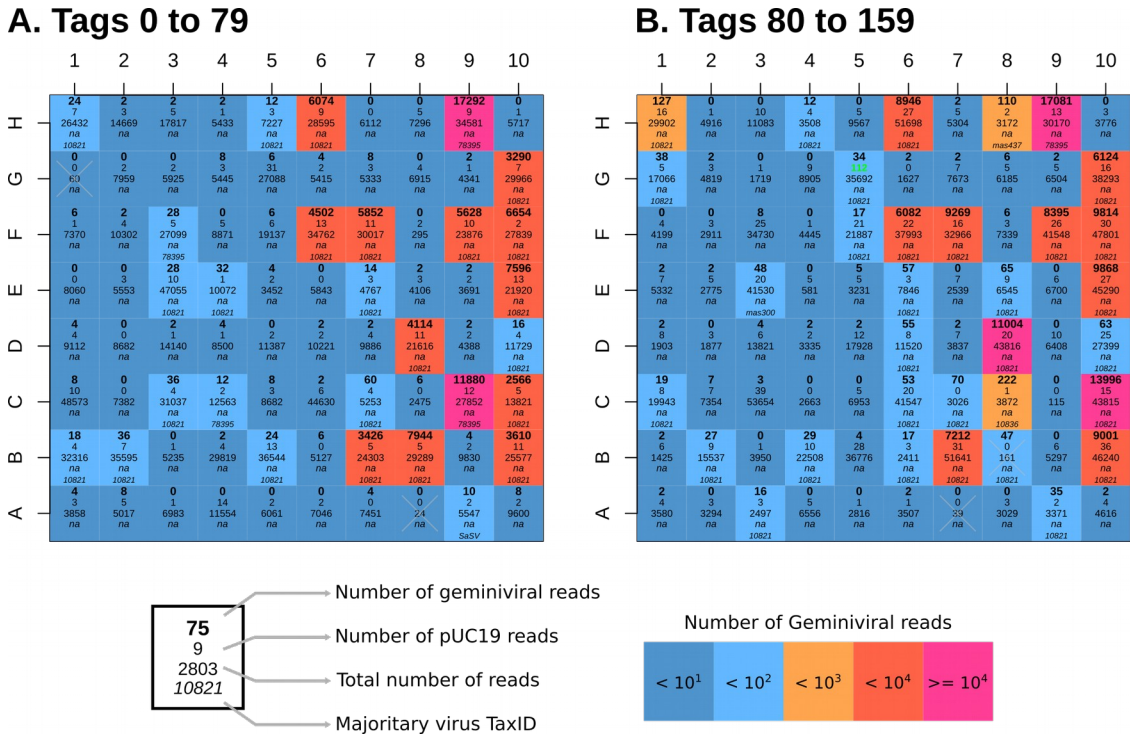


Figure 4. Schematic representation of the number of viral reads obtained for the replicates from 80 samples. Spatial representation of the reads correspond to that of the RCA and PCR plate. Replicates with tags 0 to 79 are presented on the left whereas the replicates associated with tags 80 to 159 are presented on the right. Cells are coloured according to the number of geminivirus reads (see the scale on the bottom right). Numbers and text of each cell is explained on the bottom left.

Inter library contamination

In order to determine inter sample contamination, we used a network approach where the sequenced amplicons were linked to one-another depending on the number of reads shared. Whereas the samples infected with similar viruses, and thus replicates of positive samples, are expected to share a high proportion of reads, in the absence of specific bias of cross sample

contamination, samples should randomly shared reads. The network analyses of shared reads revealed that replicates of the same positive sample (sample that have more than 1000 geminivirus reads for both replicates, a subset of the blue category in **Figure 5**) usually shared 37 % of their reads (standard deviation of 21.4 %).

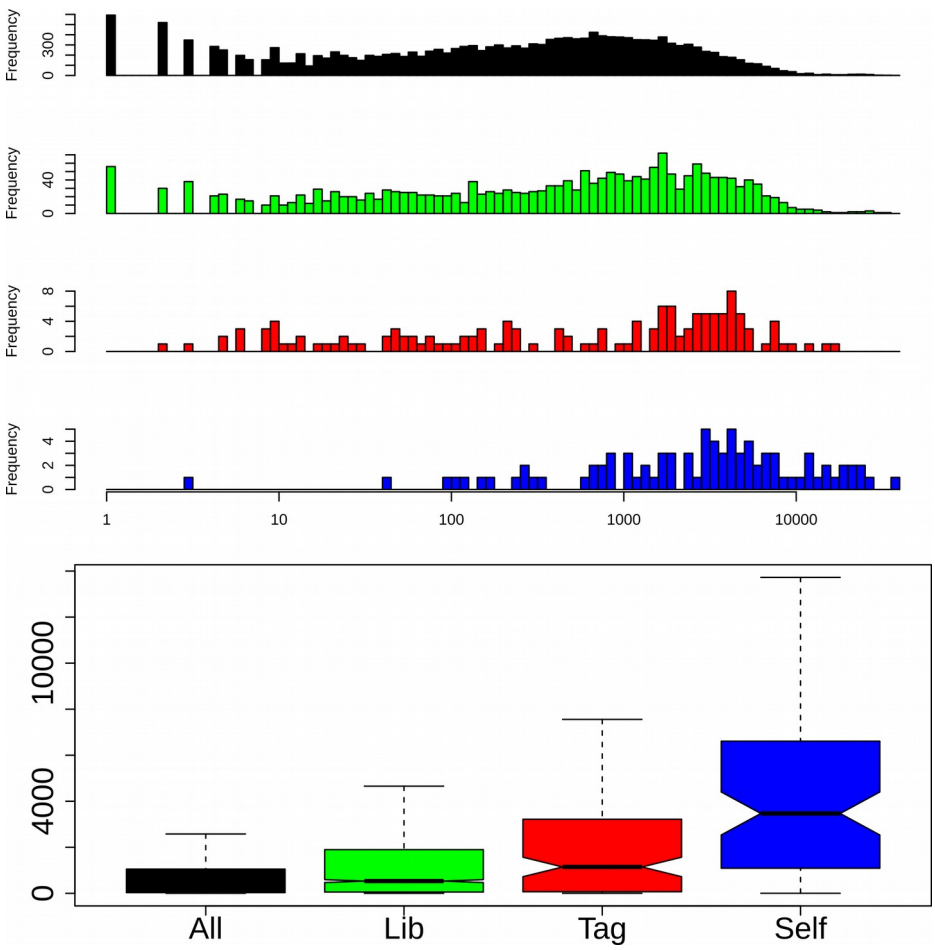


Figure 5. Number of reads shared between all the amplicon sets (black), amplicon sets from the same library (green), with a similar tag (red) or from a similar sample (blue). Boxplots at the bottom of the figure present the distribution of the number of reads. Boxes extend from the first to the third quartile whereas the thick black line represent the median. Notches represent the 95% confidence interval of the median. The absence of overlap of notches between two boxplots is a conservative test for significant difference of the median of the distributions.

Globally and within a single library, the number of shared reads was very variable. In fact, samples infected with the same viruses would naturally share more reads. Nevertheless, the analyses of the global number of shared reads revealed an increase in shared reads between samples with the same tags

(red category on **Figure 5**) in comparison to samples from the same library (green category on **Figure 5**). It must be noticed that this pattern of 'index-hopping' from one library to the other, mostly point to contamination associated with the sequencing procedure. In fact, if most of the cross sample contamination was associated with sample processing prior to sequencing, cross contamination would be more important within libraries than within tags. The analysis of shared reads, points to 16 samples (eight from the doubtful category and eight samples considered as positive) as having an abnormal proportion of reads shared with another sample. All these samples would require further scrutiny to determine whether index-hopping is associated to cross contamination or whether they simply share similar populations or viruses.

Concluding remarks

Whereas the use of deep sequencing in virology offers the possibility to uncover the most cryptic viruses, it is also associated to the difficulty to sort real signal from noise. Noise arise from inter-sample contamination during library preparation or sequencing itself. It points to the necessity to use a specific experimental design, that include replicates and controls, to estimate the confidence levels of the virus detection. While a high degree of multiplexing is achievable, a significant portion of the samples are in a grey area where it is difficult to confidently determine their health status. Besides the viral dark matter for which no reference sequence is available, causes for this uncertainty may be related to technical issues (failed molecular biology procedure or uneven pooling) or to the sample itself (samples collected at the very beginning of the infection, uneven distribution of the virus within the sample...). For such samples, either a second deep sequencing run or the use of another technique may be required. Nevertheless, beyond the field of medical diagnostic or quarantine (Martin *et al*., 2016), determination of health status from every single sample may not be essential and one may simply be interested in reducing the false positive detection rate. In such case, and if no specific systematic bias is suspected (*e.g.* enrichment of a specific type of plant or specific viruses within the doubtful sample), all the samples falling within the grey area may be considered as negative or discarded from the

analyses. Viromic has fulfilled most of its promises so far and is driving a complete paradigm shift in the diversity and ecology of viruses. It is now the reference technique for a wide range of studies and applications in virology. Although the limitations of the NGS technologies and approaches need to be carefully evaluated and considered, their use for diagnostic purposes is under way.

Acknowledgements

The authors thank Martial Grondin, Jérémy Hascoat, Gérard Lebreton and Sarah Scussel for their excellent technical support. This work was supported by the European Union (ERDF, contract GURDT I2016-1731-0006632), the *Conseil Régional de La Réunion*, the Agropolis Fondation (Labex Agro – Montpellier, E-SPACE project number 1504-004) and CIRAD. SC is a recipient of a PhD fellowship from CIRAD and the Agropolis Fondation (E-SPACE). This work was conducted on the Plant Protection Platform (3P, IBISA).

Authors information

Affiliations

CIRAD, UMR PVBMT, Pôle de Protection des Plantes, 7 chemin de l'Irat, 97410 Saint Pierre, La Réunion, France
Sohini Claverie, Murielle Hoarau, Jean-Michel Lett & Pierre Lefeuvre

Author contributions

S.C., J.M.L and P.L. conceived and designed the experiments. S.C., M.H., J.M.L and P.L. performed the experiments. S.C., J.M.L and P.L. analysed the data. S.C, J.M.L and P.L. wrote the paper. J.M.L. and P.L. secured funding for the project's execution.

Corresponding author

Correspondence to Pierre Lefeuvre

Competing interests

In the interests of transparency and to help readers to form their own judgements of potential bias, authors declare no competing interests in relation to the work described.

References

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

Boykin, L.M., Sseruwagi, P., Alicai, T., Ateka, E., Mohammed, I.U., Stanton, J.-A.L., Kayuki, C., Mark, D., Fute, T., Erasto, J., Bachwenkizi, H., Muga, B., Mumo, N., Mwangi, J., Abidrabo, P., Okao-Okuja, G., Omuut, G., Akol, J., Apio, H.B., Osingada, F., Kehoe, M.A., Eccles, D., Savill, A., Lamb, S., Kinene, T., Rawle, C.B., Muralidhar, A., Mayall, K., Tairo, F. & Ndunguru, J.*et al* (2019). Tree Lab: Portable Genomics for Early Detection of Plant Viruses and Pests in Sub-Saharan Africa. *Genes* (Basel) *10*.

Buchfink B., Xie, C. & Huson, D.H. (2015). "Fast and sensitive protein alignment using DIAMOND". *Nature Methods*, 12, 59-60.

Carlsen, T., Aas, A.B., Lindner, D., Vrålstad, T., Schumacher, T. & Kauserud, H. (2012). Don’t make a mista(g)ke: is tag switching an overlooked source of error in amplicon pyrosequencing studies? *Fungal Ecology*, 5, 747–749.

Claverie, S., Ouattara, A., Hoareau, M., Filloux, D., Varsani, A., Roumagnac, P., Martin, D.P., Lett, J.-M., & Lefeuvre, P. (2019). Exploring the diversity of Poaceae-infecting mastreviruses on Reunion Island using a viral metagenomics-based approach. *Scientific reports*, 9(1), 1-11.

Csardi, G., & Nepusz, T. (2006). The igraph software package for complex network research. *InterJournal, Complex Systems*, 1695(5), 1–9.

Faircloth, B,C., & Glenn, T.C. (2012). Not all sequence tags are created equal: designing and validating sequence identification tags robust to indels. *PloS one*, 7(8).

Filloux, D., Fernandez, E., Loire, E., Claude, L., Galzi, S., Candresse, T., Winter, S., Jeeva, M.L., Makeshkumar, T., Martin, D.P. *et al*& Roumagnac, P. (2018). Nanopore-based detection and characterization of yam viruses. *Scientific reports*, 8.

Inoue-Nagata, A.K., Albuquerque, L.C., Rocha, W.B. & Nagata, T. (2004). A simple method for cloning the complete begomovirus genome using the bacteriophage phi29 DNA polymerase. Journal of virological methods, 116, 209–211.

Kircher, M., Sawyer, S. & Meyer, M. (2012). Double indexing overcomes inaccuracies in multiplex sequencing on the Illumina platform. *Nucleic Acids Research*, 40, e3–e3.

Koonin, E.V. & Dolja, V.V. (2018). Metaviromics: a tectonic shift in understanding virus evolution. *Virus Research*. 246, A1–A3.

Lefeuvre, P., Martin, D.P., Elena, S.F., Shepherd, D.N., Roumagnac, P. & Varsani, A. (2019). Evolution and ecology of plant viruses. Nature reviews microbiology. 17, 632–644.

MacConaill, L.E., Burns, R.T., Nag, A., Coleman, H.A., Slevin, M.K., Giorda, K., Light, M., Lai, K., Jarosz, M., McNeill, M.S., *et al*Ducar, M.D., Meyerson, M. & Thorner, A.R. (2018). Unique, dual-indexed sequencing adapters with UMIs effectively eliminate index cross-talk and significantly improve sensitivity of massively parallel sequencing. *BMC Genomics,* 19.

Marsella, A., Bartram, A. K., Truszkowski, J., Brown, D. G. & Neufeld, J.D. (2012). PANDAseq: Paired-eND Assembler for Illumina sequences. *BMC Bioinformatics*, 13, 31.

Martin, R.R., Constable, F. & Tzanetakis, I.E. (2016). Quarantine Regulations and the Impact of Modern Detection Methods. Annual review of phytopathology, 54, 189–205.

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

Massart, S., Candresse, T., Gil, J., Lacomme, C., Predajna, L., Ravnikar, M., Reynard, J.-S., Rumbou, A., Saldarelli, P., Škorić, D., Vainio, E.J., Valkonen, J.P., Vanderschuren, H. Varveri, C. & Wetzel, T.*et al*. (2017). A Framework for the Evaluation of Biosecurity, Commercial, Regulatory, and Scientific Impacts of Plant Viruses and Viroids Identified by NGS Technologies. *Frontiers in microbiology*, 8.

Mitra, A., Skrzypczak, M., Ginalski, K. & Rowicka, M. (2015). Strategies for achieving high sequencing accuracy for low diversity samples and avoiding sample bleeding using Illumina platform. *PloS one*, 10(4), e0120520.

Pallás, V., Sánchez-Navarro, J.A. & James, D. (2018). Recent Advances on the Multiplex Molecular Detection of Plant Viruses and Viroids. Front Microbiol *9*.

Rognes, T., Flouri, T., Nichols, B., Quince, C. & Mahé, F. (2016). VSEARCH: a versatile open source tool for metagenomics. *PeerJ,* 4, e2584.

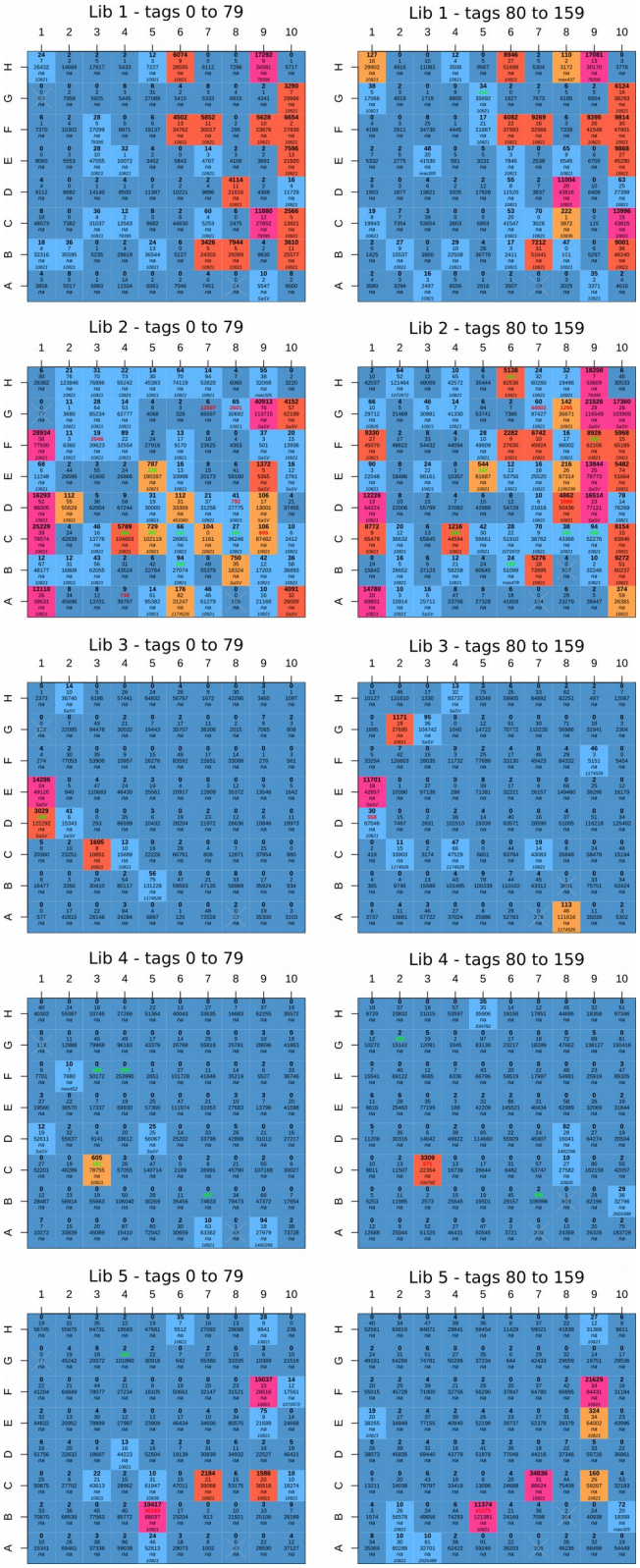
Romero, J.L.R., Carver, G.D., Arce Johnson, P., Perry, K.L. & Thompson, J.R. (2019). A rapid, sensitive and inexpensive method for detection of grapevine red blotch virus without tissue extraction using loop-mediated isothermal amplification. *Archive of virology*, 164, 1453–1457.

Sinha, R., Stanley, G., Gulati, G.S., Ezran, C., Travaglini, K.J., Wei, E., Chan, C.K.F., Nabhan, A.N., Su, T. & Morganti, R.M. (2017). Index switching causes “spreading-of-signal” among multiplexed samples in Illumina HiSeq 4000 DNA sequencing. *BioRxiv*, 125724.

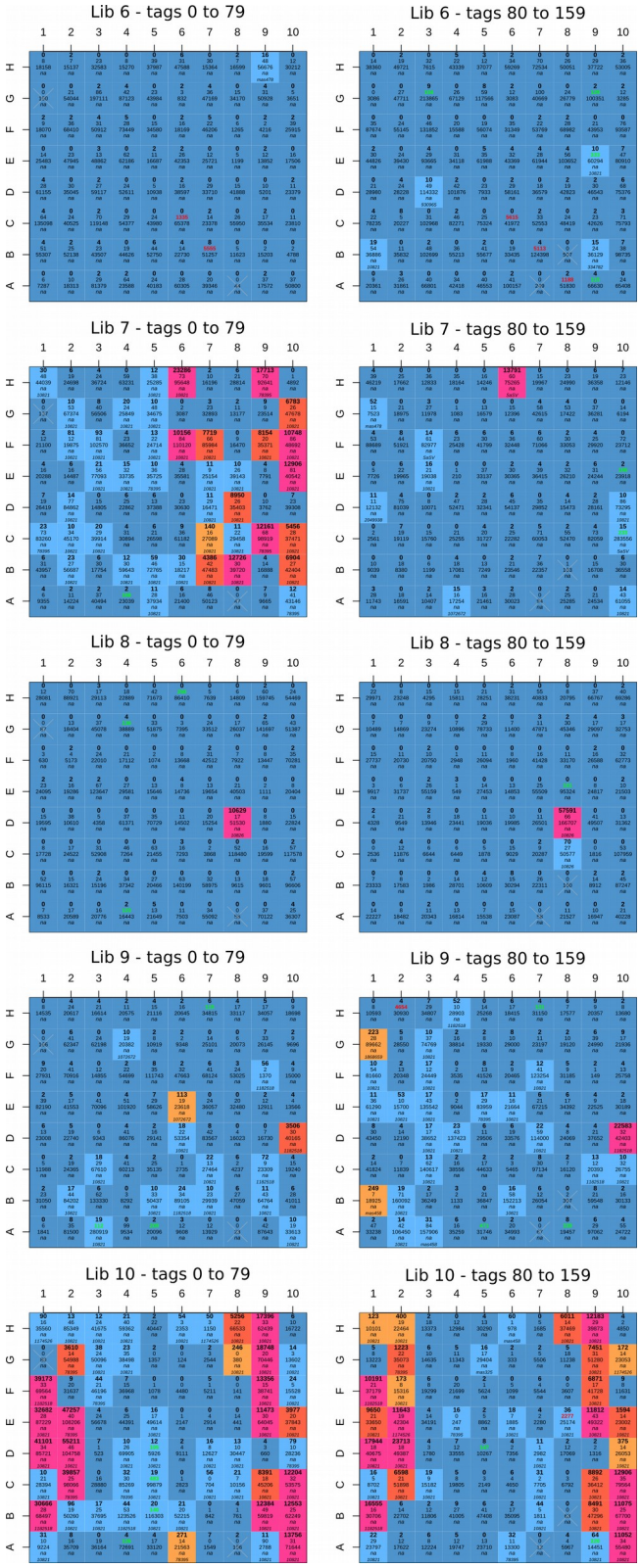
Smith, A.M., Heisler, L.E., St. Onge, R.P., Farias-Hesson, E., Wallace, I.M., Bodeau, J., Harris, A.N., Perry, K.M., Giaever, G. & Pourmand, N. (2010). Highly-multiplexed barcode sequencing: an efficient method for parallel analysis of pooled samples. *Nucleic Acids Research*, 38, e142–e142.

Tosar, J.P., Rovira, C., Naya, H., & Cayota, A. (2014). Mining of public sequencing databases supports a non-dietary origin for putative foreign miRNAs: underestimated effects of contamination in NGS. *Rna*, *20*(6), 754-757.

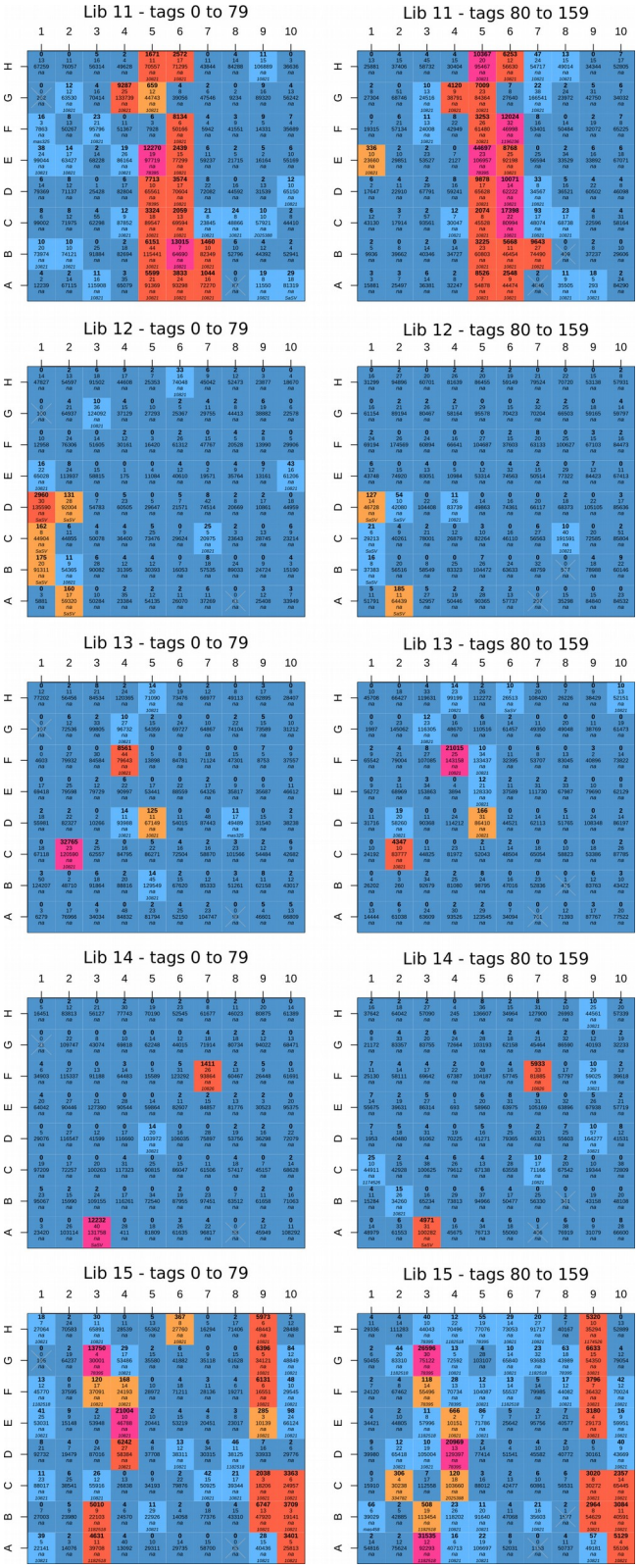
Van Der Valk, T., Vezzi, F., Ormestad, M., Dalén, L. & Guschanski, K. (2019). Index hopping on the Illumina HiseqX platform and its consequences for ancient DNA studies. Molecular ecology resources.



Supplementary Figure 1. Panel represents all the replicates' pair for all the samples from librairies 1 to 5. Each plate is a schematic representation of the number of viral reads obtained for the replicates from 80 samples. Replicates with tags 0 to 79 are presented on the left whereas the replicates associated with tags 80 to 159 are presented on the right. Cells are coloured according to **Figure 5**. Numbers and text of each cell is explained on **Figure 5**.



Supplementary Figure 2. Panel represents all the replicates' pair for all the samples from librairies 6 to 10. Each plate is a schematic representation of the number of viral reads obtained for the replicates from 80 samples. Replicates with tags 0 to 79 are presented on the left whereas the replicates associated with tags 80 to 159 are presented on the right. Cells are coloured according to **Figure 5**. Numbers and text of each cell is explained on **Figure 5**.



Supplementary Figure 3. Panel represents all the replicates' pair for all the samples from libraries 11 to 15. Each plate is a schematic representation of the number of viral reads obtained for the replicates from 80 samples. Replicates with tags 0 to 79 are presented on the left whereas the replicates associated with tags 80 to 159 are presented on the right. Cells are coloured according to **Figure 5**. Numbers and text of each cell is explained on **Figure 5**.

Chapitre 2

Chapitre 2

Étude de la structure et de la dynamique des communautés de mastrévirus à l'échelle d'un agro-écosystème

Les agro-écosystèmes sont des environnements complexes comportant des plantes cultivées, sauvages et des adventices de cultures (Burdon & Thrall, 2008). Ces environnements constituent de véritables interfaces entre les systèmes cultivés et le monde sauvage, offrant de nouvelles opportunités pour les virus associés aux plantes sauvages et cultivées d'interagir et de se déplacer entre les différents types d'hôtes (Shates *et al.*, 2019). Très peu d'études ont été menées à l'échelle des agro-écosystèmes (Bernardo *et al.*, 2018) alors qu'il est devenu évident que c'est de la promiscuité entre plantes cultivées et non cultivées que peuvent émerger de nouveaux variants viraux (Alexander *et al.*, 2014).

Au vu de la coexistence de plusieurs espèces de mastrévirus à l'échelle d'une parcelle agricole (**Chapitre 1**), nous avons consacré ce deuxième chapitre de thèse à une étude plus approfondie (i) de la diversité des mastrévirus, (ii) de la structure de la communauté qu'ils forment et (iii) de la dynamique de ces virus à l'échelle de celle-ci. Pour cela, notre étude pilote (**Chapitre 1 – Article 1**) a été élargi via un échantillonnage plus important au sein du même agro-écosystème. Ainsi, trois autres campagnes d'échantillonnage ont été menées entre novembre 2016 et 2017 avec la collecte aléatoire et sans *a priori* de plantes cultivées, d'adventices des cultures et de plantes sauvages (**Chapitre 2 - Article 3**). Globalement, 8 % des échantillons ont été identifiés comme étant infectés par des mastrévirus. Le taux d'infection des plantes cultivées (3%) s'est révélé être nettement inférieur à ceux des adventices (8%) et des plantes sauvages (14%). Au total sur les 30 espèces de Poaceae échantillonnées, 60 % d'entre elles ce sont avérés être des hôtes de mastrévirus, incluant trois espèces de plantes cultivées, trois espèces de plantes sauvages et douze espèces d'adventices. Par ailleurs, les espèces de Poaceae les plus infectées (taux d'infection supérieur à 10 % et avec une taille d'échantillon supérieure à 20) sont *Brachiaria umbellata* (40 %), *Digitaria*

ciliaris (28 %), *Cenchrus echinatus* (19 %), *Eleusine indica* (15 %) et *Sorghum arundinaceum* (15 %). Parmi ces cinq espèces, quatre correspondent à des adventices, deux à des plantes pérennes et deux autres ont une saisonnalité mixte (annuelle/pérenne). Cette large gamme d'hôtes chez les plantes non cultivées est en accord avec de précédentes études de métagénomique virale menées au sein d'environnements naturels comme au Costa Rica (Roossinck *et al.*, 2010), en Oklahoma (Muthukumar *et al.*, 2009), en France et en Afrique du Sud (Bernardo *et al.*, 2018). Cependant, contrairement à nos travaux, ces études avaient révélées des taux d'infection supérieurs pour les plantes cultivées par rapport aux plantes sauvages.

Dans un second temps, les *reads* obtenus à partir de chacun des échantillons infectés ont permis la confirmation de la présence de mastrévirus précédemment caractérisés à La Réunion tels que le MSV-B majoritairement, le MSV-A, le MSRV, le SSRV et le SWSV ainsi que les trois nouvelles espèces identifiées lors de l'étude pilote : le EIAV, le MeRAV et le SAAV (**Chapitre 1 - Article 1**). De plus, la souche F du MSV ainsi que le PanSV ont également été détectés pour la première fois à La Réunion alors qu'ils avaient déjà été caractérisés respectivement à Maurice et à Mayotte. Le clonage et le séquençage de génomes complets a permis la validation de la présence de ces virus à l'exception du SWSV et du MSV-A (**Chapitre 2 - Article 3**). Si pour le SWSV, un génome partiel a été obtenu, aucun génome partiel n'a pu être cloné pour le MSV-A.

En complément à ces mastrévirus, notre étude a permis la caractérisation d'une nouvelle espèce d'alphasatellite identifiée en novembre 2016 (n = 1) puis de nouveau en avril 2017 (n = 4) en association avec un MSV-B1 cloné. Cet alphasatellite a été nommée Sorghum mastrevirus-associated alphasatellite (SMasA) (**Chapitre 2 - Article 4**). Cette découverte ouvre de nouvelles questions sur le rôle des alphasatellites dans l'écologie de ce virus, mais aussi sur leur association possible avec d'autres espèces de mastrévirus. Le rôle des alphasatellites est encore très peu connu d'une manière générale. Seule une étude récente a permis de démontrer expérimentalement que l'association entre le mastrévirus wheat dwarf Indian virus (WDIV) et le guar leaf curl alphasatellite (GLCuA) provoquait l'expression de symptômes plus sévères de maladie (Kumar *et al.*, 2014). La diminution de l'accumulation de

petits ARN interférents (ARNsi) chez les plantes co-infectées suggére le rôle potentiel de l'alphasatellite dans la suppression du *RNA silencing*. Dans notre cas, aucun symptôme visible n'a été identifié sur les plantes co-infectées. Des expérimentations d'agro-inoculation du MSV-B1 en présence et en absence de l'alphasatellite permettrait d'évaluer l'éventuel avantage liée à la présence de l'alphasatellite pour le MSV-B1.

Hormis la co-infection MSV-B1/SMasA, vingt autres plantes co-infectées par des mastrévirus ont été identifiés. Toutes ces co-infections impliquent le MSV-B (B1 ou B3). Elles ont été confirmées par clonage et séquençage Sanger pour six échantillons de quatre espèces de Poaceae, à savoir *C. echinatus*, *C. gayana*, *D. ciliaris* et *B. umbellata.* Par ailleurs, cinq espèces peuvent constituer des hôtes potentiels pour plus de trois espèces virales à savoir *C. echinatus* (n = 6), *C. gayana* (n = 3), *B. umbellata* (n = 3), *S. arundinaceum* (n=3) et *U. maxima* (n=3). Leur capacité à héberger plusieurs virus font de ces plantes de véritables carrefours viraux où l'échange génétique par recombinaison pourrait être favorisé. L'analyse de détection de recombinaison menée sur les génomes nouvellement caractérisés a montré la présence de deux recombinaisons convergentes, d'origines distinctes, localisées dans la même région génomique et impliquant les MSV-A et MSV-B comme parents vraisemblables, chez les deux clades d'isolats réunionnais de MSV-B1 caractérisés dans notre étude. Le rôle adaptatif potentiel de cette recombinaison reste à être évalué. Le fait que certaines de ces plantes carrefours sont des adventices des cultures (n = 4) pourrait également favoriser les échanges viraux entre les plantes cultivées et non-cultivées, et l'émergence de nouveaux variants recombinants sur les plantes cultivées. Enfin, ces espèces étant pérennes (n = 3) ou mixtes (n = 2), elles pourraient jouer un rôle essentiel dans le maintien des populations virales dans le temps.

Notre effort d'échantillonnage d'espèces de Poaceae et de détection des mastrévirus par une approche de métagénomique nous a permis d'appréhender la structure des communautés de mastrévirus au sein de cet agro-écosystème. Une structure complexe composée de virus généralistes et spécialistes a été mise à jour. Les virus spécialistes sont définis comme ayant évolué pour infecter une ou quelques espèces hôtes, alors que les virus

généralistes ont la capacité d'infecter de nombreux hôtes d'espèces, de genres ou de familles différents (Elena *et al.*, 2009). L'analyse des gammes d'hôtes des virus caractérisés dans cette étude suggère de classer les MSV-B et SSRV comme virus généralistes et les autres espèces comme virus spécialistes. En effet, le MSV-B et le SSRV ont été détectés respectivement dans 16 et cinq espèces de plantes hôtes. À l'exception du SWSV et du MSRV qui n'ont été identifiés que chez une seule espèce de Poaceae, les six autres espèces de mastrévirus (EIAV, MeRAV, MSV-A, MSV-F, PanSV, SAAV) ont été détectés chez au moins deux espèces de plantes.

Les données théoriques et les évidences empiriques suggèrent que les organismes évoluant dans des environnements homogènes auront tendance à être plus spécialisés que ceux évoluant dans des environnements hétérogènes (Elena & Lenski, 2003 ; McLeish *et al.*, 2018 ; Nikolin *et al.*, 2012 ; Wilson & Yoshimura, 1994). Dans notre étude, si nous avons identifié plus de virus spécialistes que de virus généralistes, aucune modularité n'a été détectée. Un réseau modulaire serait caractérisé par la présence de groupes de virus infectant des groupes d'hôtes bien distincts et avec peu de relation inter-groupes. Cette absence de modularité pourrait être associée à des variations environnementales marquées. Parmi celles-ci, l'impact des changements de structure du réseau d'interactions suite aux introductions de nouveaux virus ou de nouveaux hôtes pourrait être essentiel dans le contexte d'îles océaniques, milieux connus comme très sensibles aux invasions biologiques (Bellard *et al.*, 2017 ; Charlery de la Masselière *et al.*, 2017). Notre analyse sur le réseau d'interactions entre les Poaceae et les mastrévirus suggère plutôt un fort niveau d'emboîtement (*nestedness*), avec notamment la gamme d'hôte du MSV-B englobant celles des autres virus. L'existence d'une gamme d'hôte plus large offrirait au MSV plus d'opportunités de transmission et de survie (Woolhouse *et al.*, 2001). L'absence de modularité significative dans notre réseau suggère une instabilité de l'écosystème étudié en terme de structure et temporalité. C'est notamment le cas pour le fonctionnement des agro-écosystèmes qui sont fortement perturbés par les activités humaines et par les méthodes culturales intensives ainsi que par les changements réguliers de l'agro-biodiversité et la biodiversité en général.

La structure des communautés de phytovirus transmis par insectes vecteurs pourrait en réalité être principalement le reflet des préférences alimentaires de ces vecteurs. En effet, la combinaison entre les préférences d'hôtes, la distribution d'hôtes et le schéma de dispersion des vecteurs pourra conditionner les possibilités de transmission interspécifique des virus (Woolhouse *et al.*, 1997). De nombreux insectes piqueurs suceurs sont des généralistes, mais certains se spécialisent sur une seule espèce hôte. Ainsi, l'identification précise des espèces d'insectes vecteurs pourrait nous permettre de savoir si l'adoption des différentes stratégies par les mastrévirus de cette étude est corrélée ou non à la gamme d'hôtes de leurs vecteurs. Ainsi, parallèlement à la dernière campagne d'échantillonnage réalisée en Novembre 2017, et même si ce type d'analyse ne permet pas d'affirmer le statut de vecteur, un échantillonnage d'hémiptères a été réalisé afin de déterminer la diversité des mastrévirus associés à ces hémiptères. Un total de 400 hémiptères représentant 7 familles (**Annexe I**) et 32 espèces différentes a été collecté à partir de 15 genres de Poaceae. L'extraction d'ADN total des hémiptères a été réalisée sur la base du protocole d'extraction individuelle d'insectes décrit par Delatte *et al.*, (2005), qui a été optimisée et adaptée aux cicadelles virulifères obtenues en laboratoire (*Cicadulina mbila*/MSV-A). Celui-ci comporte une extraction enzymatique par la protéinase K à 37°C durant 12 heures dans un tampon d'extraction contenant du Tris HCl 10mM, du KCl 50mM, du Tween 20 0.45 %, ainsi que du Nonidet P-40 0.45 %. Les ADNs totaux ainsi obtenus ont été utilisés selon notre approche de RCA-RA-NGS. Seuls deux des 400 échantillons se sont révélés être positifs pour des mastrévirus à savoir une cicadelle de l'espèce *Empoasca sp.* prélevée sur *Sorghum arundinaceum* et un psylle prélevé sur *Panicum sp.*. Les *reads* ont été assignés au MeRSV et MSV pour le psylle mais uniquement au MeRSV pour la cicadelle. Étant donné le faible nombre d'hémiptères positifs (2/400), nous avons choisi de répéter la manipulation sur des pools de cinq insectes de la même espèce quand cela était possible afin d'augmenter nos possibilités de détection de virus. Néanmoins, aucun de ces pools s'est avéré positif alors que les témoins *C. mbila* virulifères obtenus en conditions contrôlées suite à une phase d'alimentation infectieuse sur une plante de maïs agroinoculée avec le clone agroinfectieux du MSV-A, se sont tous révélés positifs. Deux hypothèses sont proposées pour expliquer nos résultats : (i) une très faible

fréquence d'insectes virulifères *in natura*, et/ou (ii) des charges virales inférieures au seuil de détection de notre approche chez les insectes échantillonnés en comparaison des insectes témoins infectés en laboratoire sur lesquels nous avons optimisé notre protocole d'extraction.

Globalement les travaux menés dans ce chapitre ont permis de mettre à jour la structure de la communauté de mastrévirus rencontrée dans un agro-écosystème de La Réunion. Cette structure est composée de virus généralistes et spécialistes cohabitant ensemble. Déterminer les facteurs influençant la topologie et la dynamique du réseau d'interaction plante hôte-virus-vecteur représente la prochaine étape de ce travail. En particulier, déterminer l'évolution dans le temps de cette structure de communauté de mastrévirus et sa robustesse aux changements biotiques et abiotiques sera essentiel à notre compréhension de l'évolution des communautés virales en général.

Article 3

Metagenomics revealed the structure of Mastrevirus-host network within agro-ecosystems

Article 3

Metagenomics revealed the structure of Mastrevirus-host network within agro-ecosystems

Sohini Claverie[1,2], Murielle Hoareau[1], Denis Filloux[3] Arvind Varsani[4,5], Philippe Roumagnac[3,6], Darren P. Martin[7], Jean-Michel Lett[1], Pierre Lefeuvre[1]

[1]CIRAD, UMR PVBMT, F-97410 St Pierre, La Réunion, France
[2]Université de La Réunion, La Réunion, France
[3]CIRAD, UMR BGPI, F-34398 Montpellier, France
[4]The Biodesign Center for Fundamental and Applied Microbiomics, Center for Evolution and Medicine, School of Life Sciences, Arizona State University, Tempe, AZ, USA
[5]Structural Biology Research Unit, Departement of Integrative Biomedical Sciences, University of Cape Town, Rondebosch, Cape Town, South Africa
[6]BGPI, INRA, CIRAD, SupAgro, Université de Montpellier, Montpellier, France
[7]Computational Biology Division, Departement of Integrative Biomedical Sciences, Institute of Infectious Diseases and Molecular Medicine, University of Cape Town, Observatory, South Africa

Keywords: ecology, viral metagenomics, mastreviruses, ssDNA viruses, community, Poaceae

Abstract

The *Mastrevirus* genus (family *Geminiviridae*) contains circular single-stranded DNA viruses transmitted by leafhopper vectors to a wide range of plants, including monocots and dicots. The most studied mastreviruses are those circulating in Africa and infecting monocots. This group of mastrevirus, referred to the African streak viruses (AfSV), comprises 14 species infecting cultivated and wild Poaceae species. Three decades of mastreviruses research in Reunion demonstrate the presence of nine of these species. Importantly, successive introductions, typical of islands' biogeography, are suspected. Understanding how viral communities are structured and evolve remains essential to understand virus ecology and emergence. Therefore, to determine mastrevirus diversity and host ranges in Reunion, we undertook an extensive survey in a single sampling site of one acre including crop fields, orchards and non-cultivated areas. After four sampling campaigns, 2886 samples of 30 cultivated and non-cultivated Poaceae species were collected, most of whom had no visible typical symptoms of streak disease. Using a metagenomic procedure devised for viruses with small circular DNA genomes, we determined the full set of mastreviruses infecting our samples. The identified Mastrevirus species were confirmed by cloning and Sanger sequencing. Our results provide an exhaustive view of the mastrevirus-host association network within an agro-ecosystem. The topology of this network suggests the co-existence of viruses ranging from generalist to specialist and that certain hosts may act as hub of viral diversity.

Introduction

Historically, viral diversity studies have mainly focused on pathogenic viruses because of their recrudescence and their major impact on human health (Ebola in humans; Coltart *et al*., 2017), animal (avian influenza in animals; Alexander, 2007) or on plants with agronomic (maize streak disease; Shepherd *et al*., 2010) or ornamental interest. Multiple factors were associated to the emergence of these viruses and the role of host-jumps from non-cultivated plants of non virulent viruses has been pointed out as the main origin for several important crop associated diseases (Fargette *et al*., 2006). It is thought that most interactions of native viruses with native plants are long co-evolved and symptomless and that their interactions with new hosts introduced for agriculture frequently results in the expression of disease (Elena *et al*., 2014). There has therefore been a real awareness of the necessity to obtain a more holistic vision of the diversity and dynamics of plant viruses.

In order to fill this gap of knowledge, metagenomics (Stobbe & Roossinck, 2014) and metatranscriptomics (Shi *et al*., 2018) methods have been leveraged. Beside revealing the existence of a large diversity of viruses, that dwarf that discovered through conventional methods (Shi *et al*., 2016), it also revealed that viruses are likely essential components of the ecosystems (French & Holmes, 2020). It lead to the emergence of the plant viral ecology field that focuses primarily on the study of (i) virus-vector-plant interactions in ecosystems, and (ii) the influence of ecosystem properties on the distribution and evolution of these interactions (Islam *et al*., 2017; Malmstrom *et al*., 2011). Whereas the seminal study on plant virus ecology revealed the abundance and the high prevalence of viruses in nature (Muthukumar *et al*., 2009; Roossinck *et al*., 2010), following work attempts to properly measure the viral diversity and dynamics at the scale of the agro-ecosystem, the true scale where emergence of disease begins. An agro-ecosystem is an environment which contains cultivated plants, weeds (non-cultivated plants associated with crops) and wild (non-cultivated plants non associated with crops) ranging from alien to native origins. Importantly, agro-ecosystems are characterised by interphases between long preserved environment and more

recent culture (Alexander *et al*., 2014), resulting in modified interactions within the otherwise long-evolved plant-virus networks (Bernardo *et al*., 2018; Malmstrom *et al*., 2005).

Despite the inherent difficulties of metagenomics studies to identify viruses at the genus level, studies on agro-interfaces in multiple environments demonstrate the impact of agriculture on virus diversity and prevalence (Bernardo *et al*., 2018). Noteworthy, they were able to uncovered new virus from the *Geminiviridae* family that would represent a diversity breadth as large as that already known. In a recent study and for a related viral genus, the *Mastrevirus* genus, Claverie *et al.* (2019) also demonstrated the existence of diverse and unknown viruses cohabiting within agro-ecosystems. Building on these results and in order to gain insights in the diversity and structure of plant viruses, we expanded this pilot study with a more thorough sequential sampling of the same agro-ecosystem in Reunion.

Members of the genus *Mastrevirus* (*Geminiviridae* family) are single-stranded circular DNA viruses of about 2.7kb, transmitted by several species of leafhoppers of the genus *Cicadulina* and infecting both monocotyledonous and dicotyledonous plants. Until now, a total of 36 different species have been characterized including 15 species that were initially identified on cultivated plants and 21 species on non-cultivated plants. Most of the mastreviruses infecting monocotyledons have been identified in Africa or in the surrounding Islands of the southwest Indian Ocean (SOOI) and have been collectively called African streak viruses (AfSV; Kraberger *et al*., 2017).

Despite the important effort of full genome sequencing, the knowledge of host mastrevirus association is profoundly biased by sampling, both in term of collected hosts and geography. Beside the fact that the majority of sequences were obtained from cultivated plants, it remains difficult to determine if the currently known host ranges are the effective realisation of virus infection capacity or merely the imprints of host availability in region were the distinct viruses were collected. In Reunion, more than half of AfSVs have been identified (9 species out of 17) and it was demonstrated that several species were existing in sympatry at the field scale (Claverie *et al*., 2019; Kraberger *et*

al., 2017). Therefore, an eco-genomic approach (analysis of viral diversity from ecosystems at the plant level) devised for geminivirus (Claverie *et al*., 2019) was used to study the diversity and prevalence of a single phytovirus genus and uncover the structure of the host virus associations.

Materials and Methods

Sampling

Samples were collected in a ~10000m^2 agro-ecosystem at the Bassin Plat CIRAD experimental facility (Latitude -21.3231; Longitude 55.4912) in Saint Pierre (Reunion). This agro-ecosystem is a mixed environment containing grassland, wasteland, fallow and agricultural plots. Leaf samples of all monocotyledonous Poaceae species (cultivated and non-cultivated) were randomly collected regardless of their health status (with or without symptoms) during four sampling campaigns on November 2014 (N=144), November 2016 (N=1,196), April 2017 (N=746) and November 2017 (N=800) (see **Table 1** for details). A total of 2,886 leaf samples including 30 plant species from 24 genera was collected. Of the 2,886 samples, 115 samples belonging to 8 different species, presented visible typical symptoms of streak disease. Samples were dried in an oven at 50°C overnight and stored at room temperature before use. Plant life-traits histories and origins, such as their crops status, life cycle and invasiveness were determined by local botanical experts.

Host identification

For samples with no confident genus or species identification obtained after visual inspection, a sequencing of the *matK* and *rbcL* genes was performed as described previously (Charlery de la Masselière *et al*., 2017). Polymerase chain reaction (PCR) amplification were conducted before direct Sanger sequencing by Macrogen Europe (Netherlands). After a quality control, sequences were classified using the RDP classifier (Wang *et al*., 2007) against a database of *matK* and *rbcL* plant sequences obtained from GenBank.

Table 1. Summary of sampled Poaceae species for each campaign. The number in brackets refers to the number of samples showing typical streak symptoms.

Tribe	Species	Crop status	Seasonality	Origin	Nov 2014	Nov 2016	Apr 2017	Nov 2017	Total
Andropogoneae	*Bothriochloa insculpta*	Wild	Annual	Indigenous	0	0	50 (1)	0	50 (1)
	Chrysopogon zizanioides	Wild	Perennial	Indigenous	0	10	0	0	10
	Saccharum sp.	Cultivated	Perennial	Alien	0	77	20	24	121
	Sorghum arundinaceum	Weed	Annual/Perennial	Cryptogenic	17	102	97	73	289
	Zea mays	Cultivated	Annual	Alien	1	107 (14)	17 (9)	40 (8)	165 (31)
Bromeae	*Bromus catharticus*	Wild	Annual	Indigenous	0	4	0	0	4
Cynodonteae	*Chloris gayana*	Weed	Perennial	Cryptogenic	15	132	75	66	288
	Chloris virgata	Wild	Annual/Perennial	Cryptogenic	0	24	22	22	68
	Cynodon dactylon	Weed	Perennial	Indigenous	12	134 (1)	27	53	226 (1)
	Dactyloctenium aegyptium	Weed	Annual	Indigenous	4	8	32	1	45
	Eleusine indica	Weed	Annual	Cryptogenic	5	22	4	54	85
Cypereae	*Cyperus rotundus*	Weed	Perennial	Cryptogenic	1	48	14	30	93
Eragrostideae	*Eragrostis minor*	Wild	Annual	Indigenous	0	50	65	69	184
Paniceae	*Brachiaria brizantha*	Wild	Perennial	Cryptogenic	0	0	0	2	2
	Brachiaria decumbens	Wild	Perennial	Cryptogenic	0	10	0	0	10
	Brachiaria umbellata	Wild	Perennial	Indigenous	1	15	51 (19)	62 (16)	129 (35)
	Cenchrus echinatus	Weed	Perennial	Indigenous	1	19 (5)	34 (4)	38 (2)	92 (11)
	Digitaria ciliaris	Weed	Annual	Cryptogenic	27	99 (20)	5 (4)	18	149 (24)
	Digitaria debilis	Wild	Annual	Cryptogenic	0	0	1 (1)	0	1 (1)
	Echinochloa colona	Wild	Annual	Cryptogenic	0	0	3	0	3
	Melinis repens	Weed	Annual/Perennial	Cryptogenic	21	76	95	61	253
	Pennisetum clandestinum	Weed	Perennial	Alien	0	7	0	0	7
	Setaria pumila	Weed	Annual	Alien	6	14	1	0	21
	Urochloa maximum	Weed	Annual/Perennial	Alien	32	142	115	80 (11)	369 (11)
	Urochloa miliaceum	Cultivated	Annual	Cryptogenic	0	0	0	32	32
	Urochloa sp.	Wild	Perennial	Cryptogenic	0	0	0	30	30
Paspaleae	*Paspalum dilatatum*	Weed	Perennial	Alien	1	6	17	18	42
	Paspalum urvillei	Weed	Perennial	Alien	0	0	0	1	1
Poeae	*Avena sativa*	Cultivated	Annual	Alien	0	84	0	10	94
Zoysieae	*Sporobolus africanus*	Weed	Perennial	Alien	0	6	1	16	23
		Total			144	1196 (40)	746 (38)	800 (37)	2886 (115)

Total genomic DNA extraction

Two distinct DNA extraction procedures were used. In the first, DNeasy Plant Mini Kit (Qiagen) were used according to the manufacturer's instructions and DNA extracts were stored at -20°C before use. In a second procedure, total genomic DNA was extracted using a «leaf soak» extraction protocol based on Roberts *et al*. (2000). Extracts were prepared by adding TPS buffer containing 100mM of Tris HCl, 1M of KCl and 10mM of EDTA (pH 8.4) and an incubation at 95°C during 10 minutes. Leaf soak extracts were readily used without storage. Whereas 400 samples from the April 2017 campaign (half of the samples from this campaign) were treated with both procedure, the samples from the 2014 campaign were only treated using the DNeasy Plant Mini Kit whereas all the remaining samples (November 2016, November 2017 campaigns and half of the samples from the April 2017 campaign) were only treated with the leaf soak procedure.

Eco-genomic approach

An eco-genomic approach based on rolling circle amplification (RCA) and random PCR amplification (RCA-RA-PCR) followed by high throughput sequencing was used as described in Claverie *et al.*, 2019. Briefly, RCA was achieved on each DNA extract using the Illustra TempliphiTM Kit (GE Healthcare) followed by a random amplification coupled with a tagging step. Along with samples, negative (DNA extracts from a healthy tomato plant grown in the lab from seed) and positive (pUC19 plasmid as available as positive control within the Illustra TempliphiTM Kit) were treated along batch of 80 samples. After quantification of amplicons, as described in Claverie *et al.* (2019), up to 160 amplicons were then pooled together before library construction with a maximum concentration ratio of 1.5 (*i.e.* no amplicons could have a concentration more than 1.5 higher than any another one in the same pool). Pools were then purified using the Illustra® GFX™ PCR DNA and gel band purification kit (GE Healthcare) according to the manufacturer's instructions before quantification and quality control using the Qubit dsDNA BR Assay Kit for the Qubit fluorometer (Thermo Fisher Scientific) and D5000 ScreenTape for 4200 TapeStation (Agilent Technologies). Library construction (including a Covaris shearing step) and Illumina sequencing (HiSeq2500 and 2x250pb PE sequencing for the November 2014 samples and HiSeqXten with

2x150pb PE sequencing for the other samples) were performed at Genewiz (USA). After a quality control and Illumina adapters removal following the pipeline described in Claverie *et al.*, 2019, viral sequences were classified with similarity searches against both the viral RefSeq database using the «blastx» algorithm implemented in DIAMOND v0.9.19.120 (Tang *et al*., 2014), and a geminivirus and geminivirus satellite database using BLASTn. Both databases were obtained from GenBank in October 2017. To reduce the dataset size, reads with similarities to geminivirus sequences were clustered using SWARM v2.1.9 (Mahé *et al*., 2015) with the distance parameter set to three.

Mastrevirus taxonomic assignment and analyses

If metagenomics data are associated with usually extremely large amount of data, it also translate in cross-sample and cross-library contamination that are potentially problematic to detect. Because of the difficulty to confidently define the health status from samples of any kind using this technique, we use replicates, employ a series of control and applied bio-informatic filters (described in Claverie *et al.*, in preparation; PhD manuscript chapter 1) to sort the sample in four categories: the confidently negative samples, the confidently positive samples, doubtful samples (*i.e.* samples probably negative but for which confirmation would be required) and failed samples (*i.e.* samples for which the number of reads was insufficient to be assigned to one of the three previous categories).

Shortly, besides analysing the samples duplicate (*i.e.* each sample was run in duplicate with distinct tags and within different sequencing libraries), the analysis of positive and negative controls from each library were used to determine minimum thresholds for sample classification. Among these, total number of reads, number of singleton reads, the number and size of cluster of reads, repeatability between replicates and cross-sample and library contamination indices were used (Claverie *et al.*, in preparation; PhD manuscript chapter 1). Sequences were further classified using the phylogenetic placement approach implemented in pplacer (Matsen *et al*., 2010) as described in Claverie *et al.* (2019). Placement was first performed against a database of all mastrevirus complete genome sequences available

in GenBank in April 2018. For sequences classified as maize streak virus from the first placement round, another round of placement was performed against a database which includes representatives of all the MSV strains complete genome sequences as obtained from GenBank in August 2019. Classifications were analysed using the BoSSA R package (https://cran.r-project.org/package =BoSSA). Groups of samples with similar viral assignment profiles were obtained after the clustering of the Kantorovich–Rubinstein distance matrix directly obtained from the placement files using guppy v1.1 (Matsen *et al.*, 2010).

Structure of plant–virus association and host range measurements

Contingency matrices with either the number of reads or the number of samples with assignation to each viral species were obtained after phylogenetic placement and positive samples filtering. Bipartite networks were generated using the R bipartite package (Dormann *et al.*, 2009). The structural pattern of the plant-virus interaction matrix was analyzed statistically to determine its degree of nestedness and modularity.

Nestedness is observed when a specialist virus interacts with a subset of the host range of generalist viruses (Weitz *et al.*, 2013). Modularity accounts for interactions of one or a few viruses exclusively with certain groups of plants (Weitz *et al.*, 2013). Quantitative nestedness was calculated using two different algorithms, weighted NODF (Almeida-Neto & Ulrich, 2011) and Wine (Galeano *et al.*, 2009). The quantitative modularity measurements were performed using two algorithms, one developed by Beckett (2016) and the other by Dormann and Strauss (2014). The statistical analyses of nestedness and modularity were obtained after 10,000 permutations of the interaction matrix using the null model function of the Vegan R Package (Oksanen *et al.*, 2019). For nestedness, permutations of the matrix that preserved the sum of positive plants per plant species («r0_both») or the individual values per plant species («r0_samp») were performed. For modularity, permutations that preserved the sum of positive plant per virus («c0_both») and the individual values of infection per viruses («c0_samp») were performed.

The plant species and virus species Shannon alpha diversities and turnover beta diversities were calculated using the number equivalent of Shannon entropy and turnover (Jost, 2007) using the Vegetarian R package (Charney & Record, 2009). The standard errors of each index (number equivalent of Shannon entropy and turnover) were estimated after 10,000 bootstrap iterations.

Cloning and full genomes sequencing

At least one sample from each viral group as defined after phylogenetic placement was selected for full genome cloning and sequencing. Full mastrevirus genomes were obtained using a RCA-RFLP as previously described in Inoue-Nagata *et al*. (2004). Briefly, 1µL of RCA amplicon was digested using several enzymes (*Acc*I, *Bam*HI, *Eco*RI, *Kpn*I, *Nco*I, *Nde*I, *Pst*I, *Sac*I, *Sal*I and *Sph*I) to yield a ~2.7kb fragment. Concomitantly to the restriction procedure, 1µL of RCA amplicon was amplified using back to back primers designed from reads alignment (**Supplementary Table 1**). Before the ligation to pJET 1.2 cloning vector (Thermo Fisher Scientific, USA), all fragments from restriction or PCR were purified using the Illustra GFX PCR DNA and Gel Band Purification Kit (GE Healthcare, USA) according to the manufacturer's instructions. The ligated products were then cloned into *Escherichia coli* (JM109, Promega) and selected plasmids were purified using QIAprep Spin Miniprep Kit (Qiagen) and were completely sequenced by Macrogen Europe (Netherlands) using primer walking. Full-length mastrevirus genomes were assembled with Geneious v6.0.6 (Kearse *et al*., 2012; http://www.geneious.com). Full nucleotide sequences were subjected to a BLAST search on the NCBI nt database for preliminary species assignment. Pairwise similarity comparisons of full nucleotide sequences of our cloned genomes to the closest mastrevirus species were obtained using SDT v1.2 (Muhire *et al*., 2014).

Phylogenetic and recombination analyses

One cloned sequence of each mastrevirus species per sample and all the genomes previously characterized from Reunion were selected and were linearized at the virion strand origin of replication and aligned using MAFFT (Katoh & Standley, 2013). Maximum-likelihood phylogenetic trees were constructed using FastTree v2.1.8 (Price *et al*., 2010) and were edited using

the APE R package (Paradis *et al*., 2004). Recombination events were detected within the full genome sequences from Reunion using the RDP (Martin and Rybicki, 2000), GENECONV (Padidam *et al*., 1999), BOOTSCAN (D.P. Martin *et al*., 2005a), MAXCHI (Smith, 1992), CHIMERA (Martin *et al*., 2005b), SISCAN (Gibbs *et al*., 2000) and 3SEQ (Boni *et al*., 2007) methods included in the RDP4 program (Martin *et al*., 2015). Default settings were used and recombination events were considered significant when detected by at least four methods.

Results and discussion

Global Poaceae diversity within an agro-ecosystem in Reunion

In this study, to examine the mastrevirus distribution across an agro-ecosystem, an eco-genomic approach was conducted on 2886 plant samples collectively representing 30 Poaceae species. Whereas sampling was randomly carried out, it must be noticed that the number of samples from each plant species reflects their abundance in the environment. The majority of samples are of non-cultivated plant species (2474/2886, 86%) and had no visible streak symptoms (2771/2886, 96%). In all, 69% of samples stem from weeds (1983/2886, 15 species; **Table 1**), 17% from wild plants (491/2886, eleven species) and 14% from crops (412/2886, four species). Most plants present a perennial (37%, 1074/2886, 14 species) or annual/perennial life cycle (*i.e.* plants that can behave as annual or perennial depending on growing conditions, 34%, 979/2886, four species). The remaining plants (29%, 833/2886, twelve species) were classified as annuals (**Table 1**). Nine plant species were alien plants (29%, 843/2886), eight were indigenous (26%, 740/2886) and 13 were cryptogenic (*i.e.* of unknown origin, 45%, 1303/2886, 13 species; **Table 1**).

In order to compare the plant species richness from one campaign to the other, the equivalent index of Shannon diversity and turnover from one sampling to the other were calculated (**Tables 2** and **3**). Beside November 2014, when a restricted number of sample was collected, the effective number of species and the composition of the samples are mostly equivalent from one campaign to the other. The low turnover index values, ranging between 0.11

and 0.25, indicate that the sets of Poaceae species sampled overlap significantly. It must be noticed however that most of the cultivated samples (65%, 268/412) were collected in the November 2016 campaign.

Table 2. Shannon equivalent number of alpha diversity of sampled and infected Poaceae species for the global survey and for each campaign based on reads assignments. The number in squared brackets represents the standard error.

Survey	Shannon equivalent numbers		
	Poaceae species alpha diversity		Virus species alpha diversity
	Sampling	Infection	
All	18.1 [± 0.3]	10.2 [± 0.7]	3.5 [± 0.3]
November 2014	8.5 [± 0.5]	6.1 [± 1.0]	3.5 [± 0.6]
November 2016	15.1 [± 0.3]	6.7 [± 0.9]	2.2 [± 0.3]
April 2017	13.2 [± 0.3]	6.2 [± 0.7]	2.8 [± 0.3]
November 2017	17.0 [± 0.3]	4.6 [± 0.6]	4.1 [± 0.5]

Global mastrevirus prevalence and host range

Firstly, 17% (503/2886) of the samples were discarded from the analysis as having a too low number of raw sequencing reads for at least one replicates (*i.e.* sample for which the infectious status could not confidently be determined, see the «failed» category in **Figure 1** and **Supplementary Table 2**). Based on the analysis of the replicates and controls, and after identification of probable inter-sample contaminations (Claverie *et al.*, in preparation; PhD manuscript chapter 1), the infectious status of the remaining 2383 samples were then determined after filtering the contingency matrix for the detection of false positives. A total of 194 samples (8.1%) were determined as confidently infected by a mastrevirus, whereas for 102 other samples (4.3%) the infectious status was most likely negative but requires confirmation. If these doubtful samples were to be considered as infected by a mastrevirus, it would increase the mastrevirus prevalence to 12% (**Figure 1** and **Supplementary Table 2**). In the following analyses and in order to keep a conservative approach, these 102 samples were considered as negative. Importantly, the doubtful samples were well distributed among the distinct plant species, suggesting the absence of a systematic bias in the detection of infection. These spurious detections also highlight the limitation of current

metagenomics methods when high multiplexing is used. This underlines the necessity to use a series of filters to determine infected samples or to complement the diagnostic with a complementary detection method (Massart *et al*., 2017).

Table 3. Turnover of beta diversity of sampled and infected Poaceae species and beta diversity of virus species between the different campaigns. The number in squared brackets represents the standard error.

Survey comparison	Turnover of Poaceae species beta diversity		Turnover of virus species beta diversity
	Sampling	Infection	
Nov 2014 – Nov 2016	0.13 [± 0.02]	0.25 [± 0.08]	0.13 [± 0.07]
Nov 2014 – Apr 2017	0.20 [± 0.02]	0.34 [± 0.09]	0.16 [± 0.06]
Nov 2014 – Nov 2017	0.25 [± 0.02]	0.58 [± 0.13]	0.30 [± 0.09]
Nov 2016 – Apr 2017	0.14 [± 0.01]	0.42 [± 0.08]	0.07 [± 0.03]
Nov 2016 – Nov 2017	0.11 [± 0.01]	0.62 [± 0.11]	0.15 [± 0.06]
Apr 2017 – Nov 2017	0.17 [± 0.01]	0.39 [± 0.05]	0.25 [± 0.06]

Among the 30 Poaceae species that were collected, 60% (18/30) were identified as mastrevirus hosts. These hosts' species include three crops (ten samples), twelve weeds (132 samples) and three wild species (52 samples; **Figure 1** and **Supplementary Table 2**). The rates of mastrevirus infection were significantly different (with a maximum p-value of $1x10^{-3}$ for all the pairwise comparisons) between the cultivated plants (3% 10/363), weeds (8 %, 132/1636) and wild plants (14 %, 52/384). The most infected Poaceae species (infection rate higher than 10% and sampling size superior to 20) were *Brachiaria umbellata* (40%), *Digitaria ciliaris* (28%), *Cenchrus echinatus* (19%), *Eleusine indica* (15%) and *Sorghum arundinaceum* (15%).

In order to find out whether the prevalence of mastreviruses is equivalent from one campaign to the other, the overall infection rates as well as the infection rates of the three plant types (*i.e.* crops, weeds and wild plants) were analysed. The November 2014 campaign was excluded due to the lower sampling effort carried out.

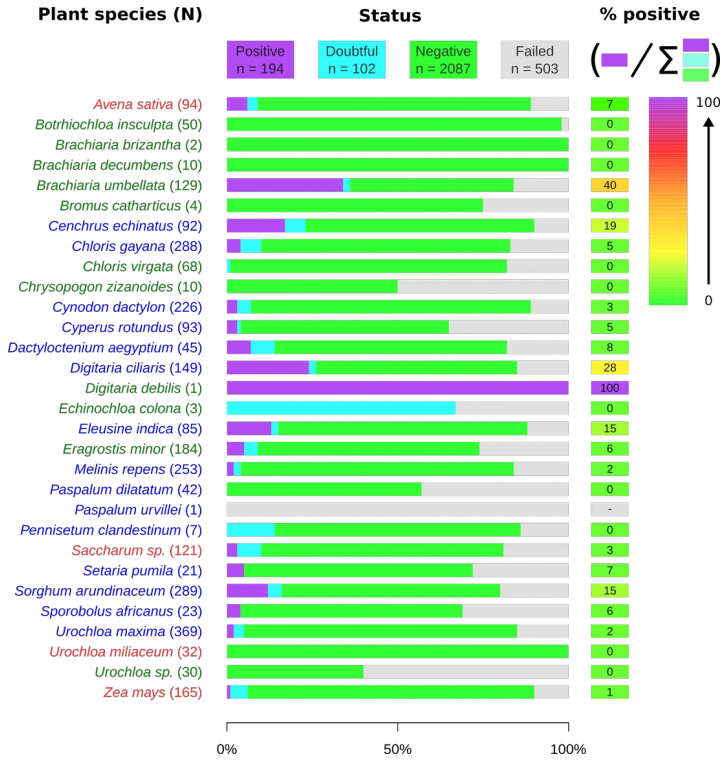


Figure 1. Proportions of Poaceae species for each assignment status and infection rates for each Poaceae species. Colours of Poaceae species names correspond to the crop status such as cultivated (in red), wild (in green) and weed (in blue). The number in brackets refers to the number of collected samples. Histogram colours refer to the infectious status such as positive (in purple), doubtful (in light blue), negative (in green) and failed (in grey). The number of infected samples for ecah Poaceae species is indicated in the cell on the right. Cells are coloured according to the infection rate (see the scale on the top right). More details in **Supplementary Table 2**.

A significant difference (maximum pairwise comparison p-value = $2.1x10^{-6}$) between the infection rate of April 2017 and those of November 2016 and 2017 was revealed. Indeed, the infection rate in April 2017 was higher (14%, 88/627) than those of November 2016 and 2017 (5%, 55/1023 and 6%, 35/630 respectively; **Supplementary Table 3**). The differences being apparently associated (maximum pairwise comparison p-value = $1.2x10^{-5}$) with the higher infection rate in weeds in April 2017 (14%, 59/429) than in November 2016 (6%, 40/691) and in November 2017 (4%, 17/414; **Supplementary Table 3**). For the samples with a sampling size superior to 15 collected samples, when infection rates are compared between sampling campaigns, there are no significantly differences between survey excepted for *C. echinatus* and *S. arundinaceum*. The infection rates of these two weeds are significantly higher in April 2017 (33%, 11/33 and 34%, 29/84 respectively) than in November 2016 (5%, 1/19 and 2%, 2/74 respectively) and November 2017 (8%, 3/30 and 0%, 0/60 respectively; **Supplementary Table 3**).

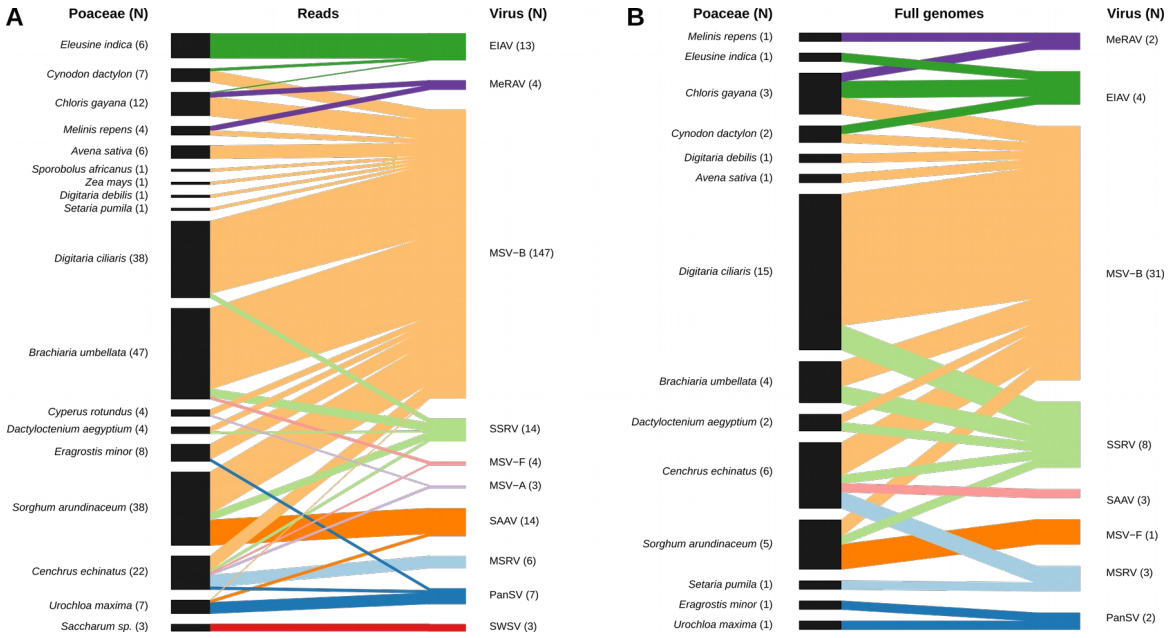


Figure 2. Tanglegrams representing the association between plant species (left side of the diagram) and viral species and strains (right side of the diagram) with in the size of the boxes and links proportional to the number of assigned reads in (**A**) and proportional to the number o samples with infection confirmed through Sanger sequencing in (**B**). Numbers in bracket indicate the number of samples.

Mastrevirus diversity

After the taxonomic classification of viral sequences, a total of eight species of mastreviruses have been identified (**Figure 2A**). Only one mastrevirus, maize streak virus strain B (MSV-B), was found in all campaigns, and only MSV-A was detected based on read assignments in a single campaign (**Supplementary Table 4**). Five mastreviruses were found in three campaigns (eleusine indica associated virus (EIAV), panicum streak virus (PanSV), maize streak Reunion virus (MSRV), sugarcane streak Reunion virus (SSRV) and sorghum arundinaceum associated virus (SAAV)). While three mastreviruses and strains were found only in two campaigns (melinis repens associated virus (MeRAV), strain F of MSV (MSV-F) and sugarcane white streak virus (SWSV)) (**Supplementary Table 4**). MSV-A and -B, MSRV strain A (MSRV-A), SSRV strain A (SSRV-A) and SWSV have been previously identified in Reunion and elsewhere. EIAV, MeRAV and SAAV have only been characterised in Reunion so far (Claverie *et al.*, 2019). PanSV and MSV-F represent the first descriptions in Reunion (**Figure 2A**). Only MSV-A and SWSV were not confirmed through full genome sequencing (**Figure 2B**).

The most prevalent species was MSV with three strains identified (MSV-A in three samples, MSV-B in 147 samples and MSV-F in four samples; **Figure 2A** and **Supplementary Table 4**). It is noteworthy that we described for the first time the F strain of MSV in Reunion. It was identified in four samples of non-cultivated plant and the detection was confirmed with the cloning of one full genome. MSV-F has previously been detected in Burundi, Nigeria, Uganda, Zimbabwe and Mauritius and exclusively on non-cultivated plants. MSV-B was detected in 147 samples and 31 full genomes were obtained. Among the 31 genomes, 15 were from strain B1 (minimum identify of 94% shared with MSV-B1 from South Africa, EU628592) and 16 were from strain B3 (minimum identity of 99% shared with MSV-B3 from Reunion, EU628616). The genome sequences of MSV-B3 obtained in this study as well as the MSV-B3 previously characterized in Reunion, Mauritius and Kenya are part of a single clade. MSV-B1 isolates were distributed in two clades with largely overlapping host ranges (**Figure 3**). Indeed, isolates from the two clades were identified on four common hosts. Solely *S. arundinaceum* and *C. echinatus* were specific to clade 1 and clade 2 respectively.

MSV-A was detected in three samples but the detection was not confirmed through full genome sequencing. The absence of a genome sequence therefore makes it difficult to determine whether MSV-A is actually present in the samples or if the reads assigned to MSV-A are traces of the introgression of MSV-A material within another viral species through recombination. Indeed, the detection of a recombinant region (events C and D, **Figure 3**) in the two clades of Reunionese isolates of MSV-B1 that originates from MSV-A goes this way. Nevertheless, if these plants were actually infected with MSV-A, its low prevalence (a maximum of three samples for MSV-A, in comparison to 147 samples for MSV-B) is unexpected. MSV-A has been extensively described in Africa and the SWIO Islands from maize but also from sugarcane and non-cultivated plants (Storey & McClean, 1930; Varsani *et al.*, 2008). This pathogen is the most devastating of the mastreviruses and has a wide geographical distribution in Africa. It has in fact been repeatedly characterized in the Indian Ocean Islands (Madagascar, Comoros, Mauritius and Reunion). Whereas the descriptions were only performed on maize, other studies in Africa demonstrate the ability of this virus strain to naturally infect a large

range of weeds and non-cultivated plants. Accordingly, it has been proposed that the success of the A strain in comparison to other MSV strains was in part attributable to a larger host range (Varsani *et al*., 2008).

The first observations of MSD in Reunion were in the 1970s (Baudin, 1976; Delpuech *et al*., 1986) and the first MSV-A clone was sequenced in 1990s by Peterschmitt *et al*. (1996). However, because of the high impact of severe disease associated with MSV-A on maize crops, maize varieties resistant to MSV-A were selected in Reunion from a tropical composite population resistant to streak disease Reunion in the 1990s (Pernet *et al*., 1999a; Pernet *et al*., 1999b). Rapid diffusion of these resistant varieties has led to a rapid decrease in the prevalence of MSD to a virtual absence of symptom description on maize crops in Reunion. The absence of MSV-A detection in the agro-ecosystems and the scope of our study could be related to the diffusion of resistant maize cultivars limiting the spread and the maintenance of MSV-A in the environment.

MSRV was detected in six plants exclusively from *C. echinatus*, a non-cultivated species. The detection was confirmed with the cloning of two full-genomes from *C. echinatus.* Unexpectedly, one full-genome of MSRV was cloned from *Setaria pumila* that the assignment of the reads suggested positive to MSV-B (**Figure 2** and **Supplementary Table 4**). These complete genomes share 100% identity with each other and most closely related to the isolates already obtained in Reunion. The MSRV was identified on both maize and non-cultivated plants in Nigeria (*Setaria barbata* and *Rottboellia sp.*) and Reunion (*C. echinatus* and *S. pumila*) and only on maize in Ethiopia and China (Chen *et al*., 2015; Claverie *et al*., 2019; Guadie *et al*., 2019; Hadfield *et al*., 2012; Oluwafemi *et al*., 2014).

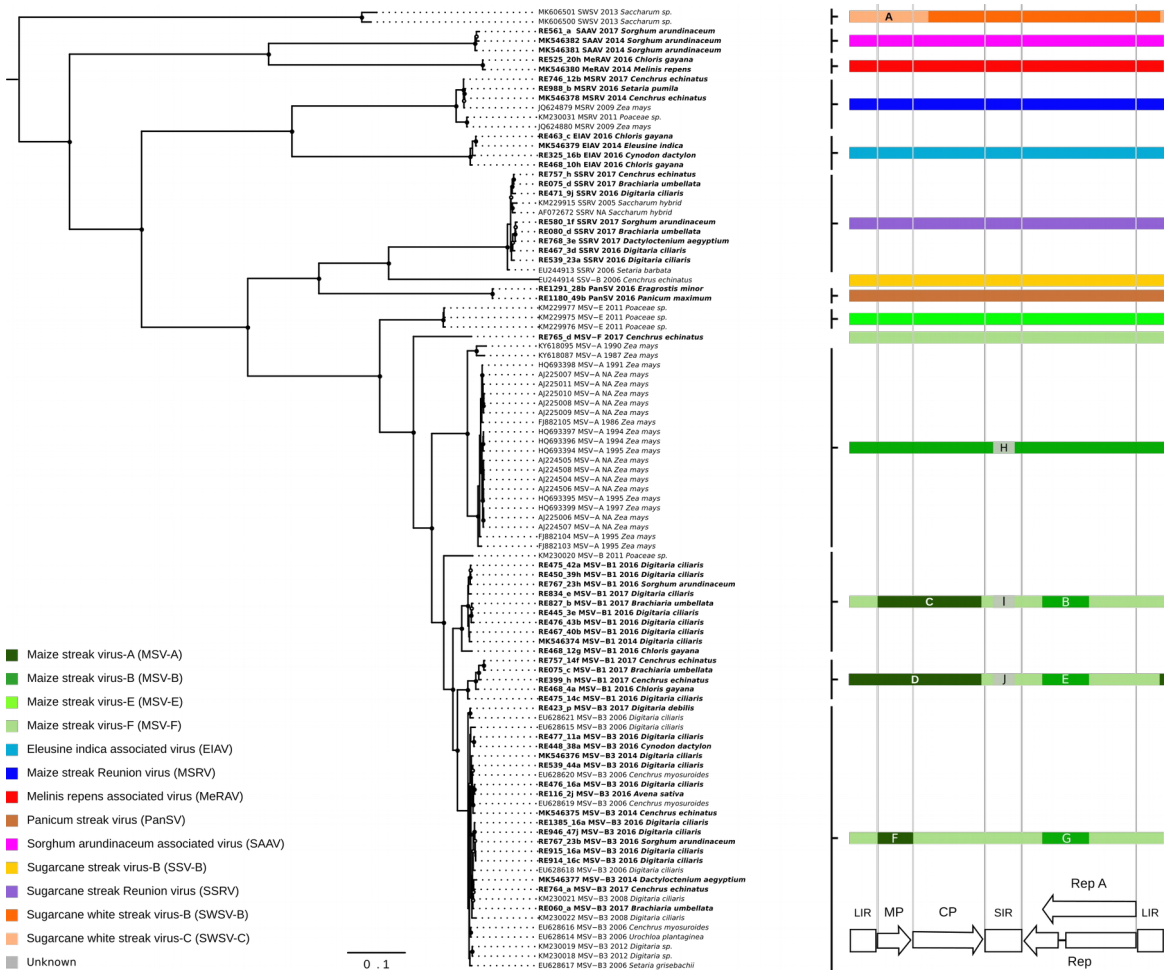


Figure 3. **Phylogenetic relationships and recombination patterns among the AfSV species in Reunion.** The maximum-likelihood phylogenetic tree contains 47 known complete genomes of monocot-infecting mastreviruses from Reunion and 54 complete genomes determined in this study (indicated in bold font). The tree was rooted on chickpea chlorosis virus (JN989413) as an outgroup (not shown). Open and closed circles on nodes indicate bootstrap support for the branches to their left of 70-89% and >=90% respectively. The schematic representation of recombination events detected using RDP4. Arrows and blocks at the bottom correspond respectively to open reading frames (ORFs) and intergenic regions: movement protein (MP), coat protein (CP), replication-associated proteins (Rep and Rep A), long intergenic region (LIR) and small intergenic region (SIR). The colours of blocks represent the different AfSV species and strains.

SSRV was detected in 14 samples exclusively from five non-cultivated species (*B. umbellata*, *C. echinatus*, *Digitaria aegyptium*, *D. ciliaris* and *S. arundinaceum*) and the detection was confirmed with the cloning of eight full genome sequences, sharing between 99% identity with isolates already obtained in Reunion (**Figure 2** and **Supplementary Table 4**). SSRV has been previously described on both sugarcane and non-cultivated plants in Reunion (*S. barbata*) and only on non-cultivated plants in Nigeria (*Eleusine coracana*) and Zimbabwe (*Paspalum conjugatum*; (Bigarré *et al.*, 1999; Kraberger *et al.*, 2017; Oluwafemi *et al.*, 2008; Shepherd *et al.*, 2008; Varsani *et al.*, 2008).

SWSV was only detected in three samples of sugarcane and the detection was only confirmed with the cloning of a partial genome. Interestingly, SWSV is a recently discovered mastrevirus that has been described in Reunion, Egypt, Sudan and Barbados (Boukari *et al*., 2017; Candresse *et al*., 2014; Claverie *et al*., 2019). It is thought that its worldwide diffusion was achieved through sugarcane transfer after quarantine control. The absence of clear symptoms and the use of conventional virus detection methods to determine plant health status apparently played a major role in the diffusion.

The three species recently characterized in Reunion EIAV, MeRAV and SAAV (Claverie *et al*., 2019), initially described from the samples collected during the November 2014 survey, were detected in samples from other campaign but not for all of them (EIAV in November 2014, 2016 and 2017; MeRAV in November 2014 and 2016; SAAV in November 2014 and 2016 and April 2017; **Supplementary Table 4**). Based on the read assignments, thirteen, four and fourteen samples were identified as infected by EIAV, MeRAV and SAAV respectively. MeRAV and EIAV were identified on other Poaceae species than the one of their first identification, unlike SAAV which was only detected on *S. arundinaceum* (n = 14). EIAV was identified in three non-cultivated species (*Chloris gayana*, *Cynodon dactylon* and *E. indica*) and MeRAV was detected in two non-cultivated species (*C. gayana* and *Melinis repens*; **Figure 2A** and **Supplementary Table 4**). A total of eight complete genomes of EIAV (n = 4), MeRAV (n = 2) and SAAV (n = 3) were obtained (**Figure 2B** and **Supplementary Table 4**). These genomes share 61%, 69% and 66% identity with axonopus compressus streak virus (KJ437671) respectively. One variant of EIAV that display a 36 nucleotides insertion in the *mp* gene (positions 416 to 451 relative to MK546379) was identified in one *C. dactylon* plant.

The last species identified based on read assignments was PanSV which represents a first in Reunion. Indeed, in the SWIO Islands, PanSV has only been characterized in Mayotte (PanSV-G). In total, seven samples, all of *Urochloa maxima* (n=5), *C. echinatus* (n=1) and *Eragrostis minor* (n=1) were found to be infected by PanSV. The presence of the virus was confirmed by sequencing two full genomes from two samples (**Figure 2B** and **Supplementary Table 4**). Considering the 94% strain demarcation level for mastreviruses (Muhire *et*

al., 2013), the Reunionese isolates of PanSV correspond to a new strain provisionally named PanSV-J (92% identity with PanSV-C from Zimbabwe, EU224264). This description expand its known distribution range and challenge the belief that PanSV has a more restricted geographical distribution than MSV (Kraberger *et al*., 2017).

Ecology of mastreviruses

First, it must be borne in mind that while we study the distribution of vector-transmitted viruses within a set of plant species in an agro-ecological system, the distribution of the virus between these plants may potentially be best explained with the ecology of their vectors. Indeed, plant viruses generally have a limited number of efficient vectors, whose intrinsic characteristics (dynamics, preferences, etc.) will have a strong influence on the ecology of the virus (Elena *et al*., 2014). The host range of a vector-transmitted virus is necessarily limited by the host range of the vector's preferences (Power, 2008). Therefore, any expansion of the host range of the vector would result in an expansion of the host range of viruses transmitted by that vector (Harrison & Robinson 1999). Mastreviruses are transmitted by leafhopper (family Cicadellidae). MSV is transmitted by eight leafhopper species of the genus *Cicadulina,* which includes 22 recognized species (Webb, 1987). *C. mbila* Naudé and *C. storeyi* China, described in Reunion (Bonfils *et al*., 1994), represent the most important vectors of MSV in Africa (Shepherd *et al*., 2010). Due to the limited knowledge available on the presence of *C. mbila, C. storeyi* or other leafhoppers and on their potential differential abilities to transmit the distinct mastrevirus species, our analysis was restricted to the plant virus interaction and transmission will remain a black box.

The overall host range of mastreviruses uncovered during the study was represented as a bipartite network (also called «tanglegrams», **Figures 2A and 2B**). In this representation, virus species are linked to the plant species they infect. While in **Figure 2A** links width and boxes sizes are function of the number of reads assigned to each virus per plant, on **Figure 2B**, they are proportional to the number of samples whose infection was confirmed through full genome Sanger sequencing. Out of the 18 Poaceae species infected by mastreviruses, 16 (*i.e.* almost the totality of infected species) were host of

MSV-B (**Figure 2A**). With the exception of SWSV and MSRV which have been identified in only one Poaceae species, the other seven mastreviruses (EIAV, MeRAV, MSV-A, MSV-F, PanSV, SAAV, SSRV) have been detected in at least two plant species (**Figure 2A**). The presence of mastreviruses infecting a restricted or broader host range would echo the notion of generalist and specialist viruses. Specialist viruses are defined as having evolved to infect one or very few host species, whereas generalist viruses have succeeded in infecting many hosts of different species and even higher taxonomic units (Elena *et al*., 2009). Whereas there is no definitive threshold in the number of plant species that a generalist would infect, the inspection of the relative host range in our mastreviruses community would lead to the classification of MSV-B (16 plant species) and SSRV (five plant species) as generalists and the other species as specialists. Although generalist viruses have more opportunity for transmission and survival due to their larger host range, theoretical (Wilson & Yoshimura, 1994) and empirical evidence (Elena & Lenski, 2003) suggests that organisms evolving in homogeneous environments tend to be more specialized than those evolving in heterogeneous environments (Nikolin *et al*., 2012). The fitness of generalists in homogenous environments is expected to be below that of specialist. However, in some cases of shared environments, generalists can be equally fit as specialists (Remold, 2012). Hosts availability will determine how viruses use the resources. Thus, the heterogeneity of the environment in time and space may favour generalists and explain their persistence (Fraile *et al*., 2017).

Table 4. Statistical significance of nestedness and modularity in an infectivity matrix between 10 virus species and strains and 18 plant species.

Analysis	Algorithm	r0_samp	r0_both	c0_samp	c0_both
Nestedness	Weighted NODF	<0.0001	<0.0001	-	-
	Wine	<0.0001	<0.0001	-	-
Modularity	Beckett	-	-	0.1	0.1
	Dormann & Strauss	-	-	0.1	0.1

In order to statistically assess the characteristics of the host-virus associations, analyses of nestedness and modularity of the structure of bipartite networks were performed (**Table 4**). Nestedness occurs when some viruses interact with subsets of the partners of some other species. Conversely, modularity is observed when virus species interact exclusively with certain type of hosts, resulting in groups of infection with no or little interconnectivity (*i.e.* no shared hosts between groups of viruses). These two type of structure are not mutually exclusive. Indeed, it is possible to observe nestedness within modules. Nestedness was found highly significant (p-value < 0.001; **Table 4**) whatever the algorithm or permutation method used. Modularity was maximal with seven modules: five modules were composed of a single species (cases of SWSV, PanSV, EIAV, SAAV and MeRAV), a module gathered MSV-B and MSV-F and another gathered MSRV, SSRV and MSV-A; **Figure 4B**). However, regardless of the algorithm or permutation methods used, no significant difference was observed between the modularities of randomly simulated matrices and the original interaction matrix, suggesting the absence of a true modularity (**Table 4**). These results while rejecting the presence of infection modules confirm the nestedness of the interaction network with most of the host range being nested in that of MSV-B. Four host species were particularly important for the nested structure (**Figure 4A**) with *C. echinatus* (infected by seven virus species; MSV-B, SSRV, PanSV, MSRV, MSV-F and MSV-A), *B. umbellata* (three; MSV-B, SSRV and MSV-F), *C. gayana* (three; MSV-B, EIAV and MeRAV) and *S. arundinaceum* (three; MSV-B, SSRV and SAAV).

Several hypotheses have been proposed to explain a nested structure of an interaction network. It is first hypothesized that it may result from an adaptation process called 'gene for gene' (Agrawal & Lively, 2003). The gene-for-gene co-evolution scheme accounts for mutation conferring to a virus the ability to infect a new plant while retaining its ability to infect its original hosts (Weitz *et al.*, 2013). The 'inverse gene for gene' evolutionary scheme is another possible explanation for the nestedness of an interaction network.

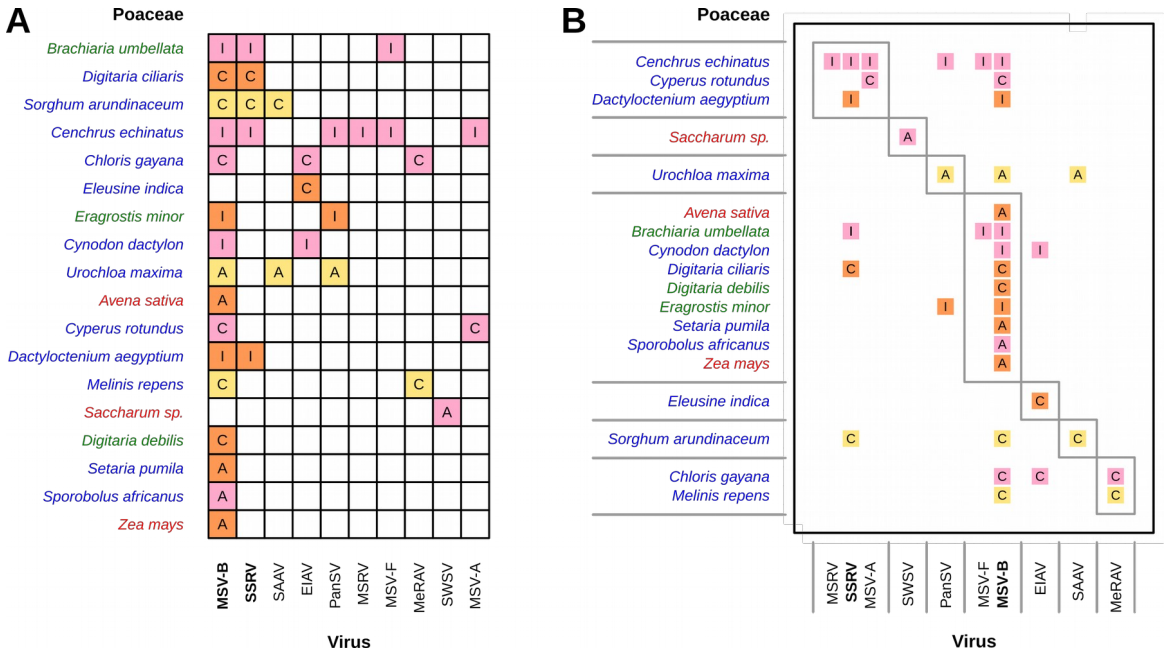


Figure 4. Interaction network representing the association between plant species and viral species and strains with the representation of the nestedness in (A) and of the modularity in (B). Colours of Poaceae species names correspond to the crop status such as cultivated (in red), wild (in green) and weed (in blue). Squares colours refer to seasonality status such as perennial (in pink), annual (in orange) and mix status (in yellow). The various origins of Poaceae species are presented with a letter with alien (A), indigenous (I) or cryptogenic (C). Generalist virus species are indicated in bold font.

This pattern is observed when infection requires recognition of the host by the pathogen and the host becomes resistant by losing the receptor targeted by the pathogen (Morris & Moury, 2019). Plant-virus species interaction matrixes are in general nested, due to some generalists, and modular due to some viruses specialised in infecting plants with similar characteristics. In our study, beside the absence of modularity in the interaction matrixes, there was no correlation between the type of plants and the virus it hosts (**Figure 4**). Indeed, viruses with multiple hosts infect plants with different life-traits histories or from distinct taxonomic groups. The absence of infection pattern may be associated with the restricted number of partners analysed, making it difficult to uncover general statistical association. However, it could also be linked to an instability of the analysed community. Islands are highly susceptible to biological invasions and extinctions and frequent disruption of the interaction networks are observed at multiple scale and for multiple biological interaction types (Bellard *et al.*, 2017; Charlery de la Masselière *et al.*, 2017; Sugiura, 2010; Traveset *et al.*, 2015). It is possible that newcomers,

both for plants and virus, disturb the structure of interactions and that no equilibrium is reached, if it ever could. In fact, beside several exotic weeds and wild plants, oats (*Avena sativa*) and millet (*Urochloa miliaceum*) are not traditional crops in Reunion. Also, while there is few historic information on the presence of mastreviruses in Reunion, the SWSV, at least, was probably recently introduced. Whereas this virus represent currently an outlier in the community (present on a single plant without any coinfection), the co-occurrence of multiple virus with distinct host-ranges in a settings that contains different hosts than that of the region from where they evolved may lead to new host-virus network structure.

In particular, co-infections may results in genetic exchange and emergence of variants more adapt to the available hosts (Lefeuvre & Moriones, 2015). The majority of Poaceae species (12/18) are hosts for more than two different virus species (**Figures 2A**). More importantly, twenty co-infections were determined. Viral coinfection were confirmed via Sanger cloning and sequencing in six samples from four Poaceae species, namely *C. echinatus*, *C. gayana*, *D. ciliaris* and *B. umbellata* (**Table 5**). Besides, Sanger sequencing revealed MSV-B1/MSV-B3 co-infection in two samples, one from *D. ciliaris* and *S. arundinaceum*. All the coinfections have been identified with or between MSV-B (B1 or B3 ; **Table 5**). Plant species that were detected with multiple types of coinfection are *C. echinatus* (n = 4), *B. umbellata* (n = 2), *D. ciliaris* (n = 2) and *S. arundinaceum* (n = 2). Five plant species were of particularly interest since they were infected with more than three viral species: *C. echinatus* (n = 6), *C. gayana* (n = 3), *B. umbellata* (n = 3), *S. arundinaceum* (n=3) and *U. maxima* (n=3; **Figures 2A** and **4**). These species correspond to four weeds and one wild species (*B. umbellata*) but also to four perennial and one mixed life-traits histories (*U. maxima* ; **Table 1**). A longer lifespan with alternating cycles of dormancy and growth may results in a higher probability to encounter viruses and thus to higher co-infection frequencies. Taken together, not only can these species be hosts for different mastreviruses and have a higher prevalence of mastreviruses, but they may also contain several mastrevirus species or strains simultaneously. These plants can behave as viral hubs where genetic exchange through recombination can be promoted.

Table 5. Coinfection identifcation in Poaceae species. The number of co-infected samples were identified by Sanger cloning and sequencing. Informations in brackets correspond to incomplete coinfection cloning.

Poaceae species	Coinfection type	Number of coinfected sample	Number of cloned sample
Brachiaria umbellata	MSV-B/MSV-F	2	-
	MSV-B/SSRV	2	1
Cenchrus echinatus	MSV-B/MSRV	1	1
	MSV-B/MSV-A	1	-
	MSV-B/SSRV	2	-
	MSV-B/SSRV/MSV-A	1	1 (MSV-B/SSRV)
Chloris gayana	MSV-B/EIAV	1	1
Cynodon dactylon	MSV-B/EIAV	1	-
Cyperus rotundus	MSV-B/MSV-A	1	-
Dactyloctenium aegyptium	MSV-B/SSRV	1	-
*Digitaria ciliaris**	MSV-B/SSRV	2	2
*Sorghum arundinaceum**	MSV-B/SAAV	4	-
Urochloa maxima	MSV-B/PanSV	1	-

*one clone of MSV-B1/MSV-B3 coinfection obtained

It is well known that intra- and interspecies recombination is common in mastreviruses (Varsani *et al*., 2008). Its cardinal example remains the MSV-A strain, which is the probable descendant of a recombinant virus between the ancestral variants of MSV-B and MSV-G/F (Varsani *et al*., 2008). Due to their diversity and the presence of coinfections, all mastrevirus sequences obtained during the surveys were analysed for evidence of recombination. The analyses of 54 sequences obtained from this project, 47 additional sequences previously obtained from Reunion along with 961 AfSV from elsewhere did not lead to the identification of recombinants specific to Reunion. However, ten recombination events have been identified, some of which being previously described (Varsani *et al*., 2008). Seven events were intra-species (events A to G in SWSV, MSV-B3 and MSV-B1; **Figure 3**) and three involved unknown parents (Events H to J in in MSV-A and MSV-B1; **Figure 3**). The unknown recombination events (Events H to J) were detected in the short intergenic region (SIR) whereas the intra-species recombination events were identified in the long intergenic region (LIR), SIR, the virion sense open reading frames (ORFs), MP and CP, and the C-terminus portion of Rep and Rep A ORFs. Both

the LIR and SIR have been previously determined to be recombination hotspots in AfSVs (Muhire *et al*., 2013; Varsani *et al*., 2008).

Conclusion

Our study reveals the breadth and abundance of mastreviruses infecting wild plants, weeds and cultivated plants in a small agro-ecological landscape. Ten viruses and strains were uncovered on 18 plant species. The virus-host association network is strongly structured and most notably comport generalists viruses along with specialists. Globally, a greater infection rate was revealed from non-cultivated hosts than from crops, contradicting previous discoveries. It must be noticed however that within the studied agro-ecosystems, whereas non-cultivated plants are abundantly present at any time in the year, crops are localised and periodically removed. The absence of crops and the use of crops that are not frequently grown in Reunion (such as oats and millet) may make it less likely for the viruses circulating in the field to be adapted to these hosts. Also, striking is the presence in most of the infected plants of MSV-B. This great generalist was found in almost all the infected plant species (N=16, excluding *Saccharum sp.* and *E. indica*). The domination of the agro-ecosystem with a generalist is unexpected following the theory and the adage: «*Jake of all-trade, master of none*» (Elena *et al*., 2009). In the long term, specialists are supposed to evolve and their presence would likely translate in a modular structure of the infection network. Nevertheless, in disrupted ecosystems, that experience spatial and temporal variation in abiotic and biotic conditions (such as hosts availability), generalists can be maintained and nestedness occurs (Valverde *et al*., 2020).
Whereas temporality in biotic and abiotic conditions may strongly imprints interaction networks, the structure of the network itself may influence virus evolution. Some plants were demonstrated to behave as viral hubs that for viruses prone for recombination such as mastreviruses would lead to the emergence of new variants with potentially altered fitness, virulence or host ranges. Whether these would in turn modify the network structure or make it more robust remains an open question. Future periodic surveys of the same sampling site would certainly revealed if the currently observed patterns are maintained through time or if disruption of the systems are frequent.

Acknowledgements

The authors thank Martial Grondin, Jérémy Hascoat, Gérard Lebreton and Sarah Scussel for their excellent technical support. This work was supported by the European Union (ERDF, contract GURDT I2016-1731-0006632), the *Conseil Régional de La Réunion*, the Agropolis Fondation (Labex Agro – Montpellier, E-SPACE project number 1504-004) and CIRAD. SC is a recipient of a PhD fellowship from CIRAD and the Agropolis Fondation (E-SPACE). This work was conducted on the Plant Protection Platform (3P, IBISA).

Authors information

Affiliations

CIRAD, UMR PVBMT, Pôle de Protection des Plantes, 7 chemin de l'Irat, 97410 Saint Pierre, La Réunion, France
Sohini Claverie, Murielle Hoarau, Jean-Michel Lett & Pierre Lefeuvre

CIRAD, UMR BGPI, TA A-54/K, Campus International de Baillarguet, 34398 Montpellier Cedex 5, France
Denis Filloux & Philippe Roumagnac

The Biodesign Center for Fundamental and Applied Microbiomics, Center for Evolution and Medicine, School of Life Sciences, Arizona State University, 1001 S. McAllister Ave, Tempe, AZ 85287-5001, USA
Arvind Varsani

Computational Biology Division, Departement of Integrative Biomedical Sciences, Institute of Infectious Diseases and Molecular Medicine, University of Cape Town, Observatory 7925, South Africa
Darren P. Martin

Author contributions

S.C., J.M.L and P.L. conceived and designed the experiments. S.C., M.H., J.M.L and P.L. performed the experiments. S.C., D.F.., A.V., P.R., D.P.M., J.M.L and P.L.

analysed the data. S.C, A.V., P.R., D.P.M., J.M.L and P.L. wrote the paper. J.M.L. and P.L. secured funding for the project's execution.

Corresponding author

Correspondence to Pierre Lefeuvre

Competing interests

In the interests of transparency and to help readers to form their own judgements of potential bias, authors declare no competing interests in relation to the work described.

References

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

Alexander, D. J. (2007). An overview of the epidemiology of avian influenza. *Vaccine*, *25*(30 SPEC. ISS.), 5637–5644.

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

Almeida-Neto, M., & Ulrich, W. (2011). A straightforward computational approach for measuring nestedness using quantitative matrices. *Environmental Modelling and Software*, *26*(2), 173–178.

Baudin, P. (1976). Etude d'une souche du virus de la mosaique de la canne a sucre. *Agronomie Tropicale*, *32*, 180–204.

Beckett, S. J. (2016). Improved community detection in weighted bipartite networks. *Royal Society Open Science*, *3*(1).

Bellard, C., Rysman, J., Leroy, B., Claud, C., & Mace, G. M. (2017). A global picture of biological invasion threat on islands. *Nature Ecology & Evolution*, *1*(12), 1862-1869.

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

Bigarré, L., Salah, M., Granier, M., Frutos, R., Thouvenel, J. C., & Peterschmitt, M. (1999). Nucleotide sequence evidence for three distinct sugarcane streak mastreviruses. *Archives of Virology*, *144*(12), 2331–2344.

Bonfils, J., Quilici, S., & Reynaud, B. (1994). Les Hémiptères Auchénorhynques de l ’ ile de la Réunion. *Bulletin de La Société Entomologique de France*, *99*(3), 227–240.

Boni, M. F., Posada, D., & Feldman, M. W. (2007). An exact nonparametric method for inferring mosaic structure in sequence triplets. *Genetics*, *176*(2), 1035–1047.

Boukari, W., Alcalá-briseño, R. I., Kraberger, S., Fernandez, E., Filloux, D., Daugrois, J., Comstock, J. C., Lett, J., Martin, D. P., Varsani, A., Roumagnac, P., Polston, J. E., & Rott, P. C. (2017). *Occurrence of a novel mastrevirus in sugarcane germplasm collections in Florida , Guadeloupe and Réunion*. 1–8.

Candresse, T., Filloux, D., Muhire, B., Julian, C., Galzi, S., Fort, G., Bernardo, P., Daugrois, J. H., Fernandez, E., Martin, D. P., Varsani, A., & Roumagnac, P. (2014). Appearances can be deceptive: Revealing a hidden viral infection with deep sequencing in a plant quarantine context. *PLoS ONE*, *9*(7).

Charlery de la Masselière, M., Ravigné, V., Facon, B., Lefeuvre, P., Massol, F., Quilici, S., & Duyck, P. F. (2017). Changes in phytophagous insect host ranges following the invasion of their community: Long-term data for fruit flies. *Ecology and Evolution*, *7*(14), 5181–5190.

Charney, N., & Record, S. (2009). Jost Diversity Measures for Community Data. Package 'vegetarian.' *R Package*, *2*, 3-3.

Chen, S., Huang, Q., Wu, L., & Qian, Y. (2015). Identification and characterization of a maize-associated mastrevirus in China by deep sequencing small RNA populations. *Virology Journal*, *12*(1), 1–9.

Claverie, S., Ouattara, A., Hoareau, M., Filloux, D., Varsani, A., Roumagnac, P., Martin, D. P., Lett, J. M., & Lefeuvre, P. (2019). Exploring the diversity of Poaceae-infecting mastreviruses on Reunion Island using a viral metagenomics-based approach. *Scientific Reports*, *9*(1), 1–11.

Coltart, C. E. M., Lindsey, B., Ghinai, I., Johnson, A. M., & Heymann, D. L. (2017). The Ebola outbreak, 2013–2016: Old lessons for new epidemics. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *372*(1721), 2013–2016.

Delpuech, I., Bonfils, J., Leclant, F., Delpuech, I., Bonfils, J., & Leclant, F. (1986). Contribution à l ' étude des virus du maïs transmis par homoptères auchénorrhynques à l ' île de la Réunion

Dormann, C. F., Fründ, J., Blüthgen, N., & Gruber, B. (2009). Indices, graphs and null models: Analyzing bipartite ecological networks. *The Open Ecology Journal*, *2*, 7–24.

Dormann, C. F., & Strauss, R. (2014). A method for detecting modules in quantitative bipartite networks. *Methods in Ecology and Evolution*, *5*(1), 90–98.

Elena, S. F., Agudelo-Romero, P., & Lalic, J. (2009). The Evolution of Viruses in Multi-Host Fitness Landscapes. *The Open Virology Journal*, *3*(1), 1–6.

Elena, S. F., Fraile, A., & García-Arenal, F. (2014). Evolution and emergence of plant viruses. In *Advances in Virus Research, 8*, 161-191.

Elena, S. F., & Lenski, R. E. (2003). Evolution experiments with microorganisms: The dynamics and genetic bases of adaptation. *Nature Reviews Genetics*, *4*(6), 457–469.

Fargette, D., Konaté, G., Fauquet, C., Muller, E., Peterschmitt, M., & Thresh, J. M. (2006). Molecular Ecology and Emergence of Tropical Plant Viruses. *Annual Review of Phytopathology*, *44*(1), 235–260.

Fraile, A., McLeish, M. J., 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 Research*, *241*, 68–76.

French, R. K., & Holmes, E. C. (2020). An Ecosystems Perspective on Virus Evolution and Emergence. *Trends in Microbiology*, *28*(3), 165–175.

Galeano, J., Pastor, J. M., & Iriondo, J. M. (2009). Weighted-Interaction Nestedness Estimator (WINE): A new estimator to calculate over frequency matrices. *Environmental Modelling and Software*, *24*(11), 1342–1346.

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

Guadie, D., Tesfaye, K., Knierim, D., Winter, S., & Abraham, A. (2019). Molecular analysis of maize (Zea mays L.)-infecting mastreviruses in Ethiopia reveals marked diversity of virus genomes and a novel species. *Virus Genes*, *0*(0), 0.

Hadfield, J., Thomas, J. E., Schwinghamer, M. W., Kraberger, S., Stainton, D., Dayaram, A., Parry, J. N., Pande, D., Martin, D. P., & Varsani, A. (2012). Molecular characterisation of dicot-infecting mastreviruses from Australia. *Virus Research*, *166*(1–2), 13–22.

Inoue-Nagata, A. K., Albuquerque, L. C., Rocha, W. B., & Nagata, T. (2004). A simple method for cloning the complete begomovirus genome using the bacteriophage φ29 DNA polymerase. *Journal of Virological Methods*, *116*(2), 209–211.

Islam, W., Zhang, J., Adnan, M., Noman, A., Zaynab, M., & Wu, Z. (2017). Plant virus ecology: A glimpse of recent accomplishments. *Applied Ecology and Environmental Research*, *15*(1), 691–705.

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

Katoh, K., & Standley, D. M. (2013). MAFFT multiple sequence alignment software version 7: Improvements in performance and usability. *Molecular Biology and Evolution*, *30*(4), 772–780.

Kearse, M., Moir, R., Wilson, A., Stones-Havas, S., Cheung, M., Sturrock, S., Buxton, S., Cooper, A., Markowitz, S., Duran, C., Thierer, T., Ashton, B., Meintjes, P., & Drummond, A. (2012). Geneious basic: An integrated and extendable desktop software platform for the organization and aalysis of sequence data. *Bioinformatics*, *28*(12), 1647–1649.

Kraberger, S., Saumtally, S., Pande, D., Khoodoo, M. H. R., Dhayan, S., Dookun-Saumtally, A., Shepherd, D. N., Hartnady, P., Atkinson, R., Lakay, F. M., Hanson, B., Redhi, D., Monjane, A. L., Windram, O. P., Walters, M.,

Oluwafemi, S., Michel-Lett, J., Lefeuvre, P., Martin, D. P., & Varsani, A. (2017). Molecular diversity, geographic distribution and host range of monocot-infecting mastreviruses in Africa and surrounding islands. *Virus Research*, *238*(June), 171–178.

Lefeuvre, P., & Moriones, E. (2015). Recombination as a motor of host switches and virus emergence: Geminiviruses as case studies. In *Current Opinion in Virology, 10*, 14-19.

Mahé, F., Rognes, T., Quince, C., de Vargas, C., & Dunthorn, M. (2015). Swarm v2: highly-scalable and high-resolution amplicon clustering. *PeerJ*, *3*, e1420.

Malmstrom, C. M., Hughes, C. C., Newton, L. A., & Stoner, C. J. (2005). Virus infection in remnant native bunchgrasses from invaded California grasslands. *New Phytologist*, *168*(1), 217–230.

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

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

Martin, D.P., Posada, D., Crandall, K. A., & Williamson, C. (2005). A Modified Bootscan Algorithm for Automated Identification of Recombinant Sequences and Recombination Breakpoints. *AIDS Research and Human Retroviruses*, *21*(1), 98–102.

Martin, Darren P., Murrell, B., Golden, M., Khoosal, A., & Muhire, B. (2015). RDP4: Detection and analysis of recombination patterns in virus genomes. *Virus Evolution*, *1*(1), 1–5.

Martin, Darren P., Van Walt, E. Der, Posada, D., & Rybicki, E. P. (2005). The evolutionary value of recombination is constrained by genome modularity. *PLoS Genetics*, *1*(4), 0475–0479.

Massart, S., Candresse, T., Gil, J., Lacomme, C., Predajna, L., Ravnikar, M., Reynard, J. S., Rumbou, A., Saldarelli, P., Škoric, D., Vainio, E. J., Valkonen, J. P. T., Vanderschuren, H., Varveri, C., & Wetzel, T. (2017). A framework for the evaluation of biosecurity, commercial, regulatory, and scientific impacts of plant viruses and viroids identified by NGS technologies. *Frontiers in Microbiology*, *8*(1), 45.

Matsen, F. A., Kodner, R. B., & Armbrust, V. E. (2010). pplacer: linear time maximum-likelihood and Bayesian phylogenetic placement of sequences onto a fixed reference tree. *BMC Bioinformatics*, *11*, 538.

Maynard Smith, J. (1992). Analysing the Mosaic Structure of Genes. *Journal of Molecular Evolution*, *34*, 126–129.

Morris, C. E., & Moury, B. (2019). Revisiting the Concept of Host Range of Plant Pathogens. *Annual Review of Phytopathology*, *57*(1), 63–90.

Muhire, B. M., Varsani, A., & Martin, D. P. (2014). SDT: A virus classification tool based on pairwise sequence alignment and identity calculation. *PLoS ONE*, *9*(9).

Muhire, B., Martin, D. P., Brown, J. K., Navas-Castillo, J., Moriones, E., Zerbini, F. M., Rivera-Bustamante, R., Malathi, V. G., Briddon, R. W., & Varsani, A. (2013). A genome-wide pairwise-identity-based proposal for the classification of viruses in the genus Mastrevirus (family Geminiviridae). *Archives of Virology*, *158*(6), 1411–1424.

Muthukumar, V., Melcher, U., Pierce, M., Wiley, G. B., Roe, B. A., Palmer, M. W., Thapa, V., Ali, A., & Ding, T. (2009). Non-cultivated plants of the Tallgrass Prairie Preserve of northeastern Oklahoma frequently contain virus-like sequences in particulate fractions. *Virus Research*, *141*(2), 169–173.

Nikolin, V. M., Osterrieder, K., von Messling, V., Hofer, H., Anderson, D., Dubovi, E., Brunner, E., & East, M. L. (2012). Antagonistic Pleiotropy and Fitness Trade-Offs Reveal Specialist and Generalist Traits in Strains of Canine Distemper Virus. *PLoS ONE*, *7*(12), 1–9.

Oksanen, J., Blanchet, F. G., Friendly, M., Kindt, R., Legendre, P., Mcglinn, D., Minchin, P. R., O'hara, R. B., Simpson, G. L., Solymos, P., Henry, M., Stevens, H., Szoecs, E., & Maintainer, H. W. (2019). Package "vegan" Title Community Ecology Package. *Community Ecology Package*, *2*(9), 1–297.

Oluwafemi, S., Kraberger, S., Shepherd, D. N., Martin, D. P., & Varsani, A. (2014). A high degree of African streak virus diversity within Nigerian maize fields includes a new mastrevirus from Axonopus compressus. *Archives of Virology*, *159*(10), 2765–2770.

Oluwafemi, S., Varsani, A., Monjane, A. L., Shepherd, D. N., Owor, B. E., Rybicki, E. P., & Martin, D. P. (2008). A new African streak virus species from Nigeria. *Archives of Virology*, *153*(7), 1407–1410.

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

Paradis, E., Blomberg, S., Bolker, B., Brown, J., Claude, J., Cuong, H. S., Desper, R., Didier, G., Durand, B., Dutheil, J., Ewing, R., Gascuel, O., Guillerme, T., Heibl, C., Ives, A., Jones, B., Krah, F., Lawson, D., Lefort, V., … Vienne, D. de. (2004). APE: analyses of phylogenetics and evolution in R language. *Bioinformatics*, *20*(2), 289-290.

Pernet, A., Hoisington, D., Dintinger, J., Jewell, D., Jiang, C., Khairallah, M., Letourmy, P., Marchand, J. L., Glaszmann, J. C., & González De León, D. (1999a). Genetic mapping of maize streak virus resistance from the Mascarene source. II. Resistance in line CIRAD390 and stability across germplasm. *Theoretical and Applied Genetics*, *99*(3–4), 540–553.

Pernet, A., Hoisington, D., Franco, J., Isnard, M., Jewell, D., Jiang, C., Marchand, J. L., Reynaud, B., Glaszmann, J. C., & González de León, D. (1999b). Genetic mapping of maize streak virus resistance from the Mascarene source. I. Resistance in line D211 and stability against different virus clones. *Theoretical and Applied Genetics*, *99*(3–4), 524–539.

Peterschmitt, M., Granier, M., Frutos, R., & Reynaud, B. (1996). Infectivity and complete nucleotide sequence of the genome of a genetically distinct strain of maize streak virus from Reunion Island. *Archives of Virology*, *141*(9), 1637–1650.

Power, A. G. (2008). Community Ecology of Plant Viruses. In *Plant Virus Evolution*, 15-26.

Price, M. N., Dehal, P. S., & Arkin, A. P. (2010). FastTree 2 - Approximately maximum-likelihood trees for large alignments. *PLoS O*, *5*(3), e9490.

Remold, S. (2012). Understanding specialism when the jack of all trades can be the master of all. *Proceedings of the Royal Society B: Biological Sciences*, *279*(1749), 4861–4869.

Roberts, C. A., Dietzgen, R. G., Heelan, L. A., & MaclLean, D. J. (2000). Real-time RT-PCR fluorescent detection of tomato spotted wilt virus. *Journal of Virological Methods*, *88*(1), 1–8.

Roossinck, M. J., Saha, P., Wiley, G. B., Quan, J., White, J. D., Lai, H., Chavarría, F., Shen, G., & Roe, B. A. (2010). Ecogenomics: Using massively parallel pyrosequencing to understand virus ecology. *Molecular Ecology*, *19*(SUPPL. 1), 81–88.

Shepherd, D. N., Martin, D. P., Van Der Walt, E., Dent, K., Varsani, A., & Rybicki, E. P. (2010). Maize streak virus: An old and complex “emerging” pathogen. *Molecular Plant Pathology*, *11*(1), 1–12.

Shepherd, D. N., Varsani, A., Windram, O. P., Lefeuvre, P., Monjane, A. L., Owor, B. E., & Martin, D. P. (2008). Novel sugarcane streak and sugarcane streak Reunion mastreviruses from southern Africa and la Réunion. *Archives of Virology*, *153*(3), 605–609.

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

Shi, M., Zhang, Y. Z., & Holmes, E. C. (2018). Meta-transcriptomics and the evolutionary biology of RNA viruses. *Virus Research*, *243*(October 2017), 83–90.

Stobbe, A. H., & Roossinck, M. J. (2014). Plant virus metagenomics: what we know and why we need to know more. *Frontiers in Plant Science*, *5*(April), 1–4.

Storey, H. H., & McClean, A. P. D. (1930). *The transmission of streak disease between maize, sugar cane and wild grasses, 4*.

Sugiura, S. (2010). Species interactions-area relationships: Biological invasions and network structure in relation to island area. *Proceedings of the Royal Society B: Biological Sciences*, *277*(1689), 1807–1815.

Tang, M., Tan, M., Meng, G., Yang, S., Su, X., Liu, S., Song, W., Li, Y., Wu, Q., Zhang, A., & Zhou, X. (2014). Multiplex sequencing of pooled mitochondrial genomes - A crucial step toward biodiversity analysis using mito-metagenomics. *Nucleic Acids Research*, *42*(22), 1–13.

Traveset, A., Olesen, J. M., Nogales, M., Vargas, P., Jaramillo, P., Antolín, E., Trigo, M. M., & Heleno, R. (2015). Bird-flower visitation networks in the Galápagos unveil a widespread interaction release. *Nature Communications*, *6*(6), 1-6.

Valverde, S., Vidiella, B., Montañez, R., Fraile, A., Sacristán, S., & García-Arenal, F. (2020). Coexistence of nestedness and modularity in host–pathogen infection networks. *Nature Ecology and Evolution*.

Varsani, A., Shepherd, D. N., Monjane, A. L., Owor, B. E., Erdmann, J. B., Rybicki, E. P., Peterschmitt, M., Briddon, R. W., Markham, P. G., Oluwafemi, S., Windram, O. P., Lefeuvre, P., Lett, J. M., & Martin, D. P. (2008). Recombination, decreased host specificity and increased mobility may have driven the emergence of maize streak virus as an agricultural pathogen. *Journal of General Virology*, *89*(9), 2063–2074.

Wang, Q., Garrity, G. M., Tiedje, J. M., & Cole, J. R. (2007). Naïve Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. *Applied and Environmental Microbiology*, *73*(16), 5261–5267.

Webb, M. D. (1987). Species recognition in Cicadulina leafhoppers (Hemiptera: Cicadellidae), vectors of pathogens of Gramineae. *Bulletin of Entomological Research*, *77*(4), 683–712.

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

Wilson, D. S., & Yoshimura, J. (1994). On the coexistence of specialists and generalists. *The American Naturalist*, *144*(4), 692–707.

Supplementary Table 1. Back to back primers design for cloning.

Primer name	Sequence
alpha_b2b_forward	5’-ATGATAGAAATATCATCTGGGTGTGCGGCACT-3’
alpha_b2b_reverse	5'-CAGGGCCCTGGTTAATTCTCTCATTAAAACATATC-3’
EISV_b2b_forward	5'-GGGGTTGACGGTGAAGTCTCGTTG-3’
EISV_b2b_reverse	5'-AAGTACGGGAAGAAGAAAAGAATTCCTGGAGGA-3'
EISV-like_b2b_forward	5'-CCATAATCTCGTTCTTGGTCTGTTTAGTGGAATTC-3'
EISV-like_b2b_reverse	5'-CAGACATAATCTCTAGCTCCACCAGTAAACAGGAC-3'
SASV_b2b_forward	5'-GGAATATACCTCGGGAGTAACCTGGTAGATGTTCT-3'
SASV_b2b_reverse	5'-CTTCTACACCCTAACGTAGATGCTGAGGGAGAC-3'
sat_b2b_forward	5'-GATCTCCTCTGATTGACGTGGAAGATCGAAGCAGA-3'
sat_b2b_reverse	5'-CTTAACTACGCGGTCATTGAGGAAATTAAGGATC-3'
mastre_b2b_forward	5'-ATTATTCTGGTAGTAGTTATGGACCCCTAAGC-3'
mastre_b2b_reverse	5'-GTGTACCTCCGTTTAACCACAAAGCGATGACAGAC-3'
MSRV_b2b_forward	5'-GAAGCTCAAGCATGGTTCCGGTTAAGCAAC-3'
MSRV_b2b_reverse	5'-CGTTGGTAATTGTCTGGATCTGGAGACTG-3'
MSRV_b2b_forward1	5'-GTTCACACTCGAGTCAAATGGAAGAAGGAATGATG-3'
MSRV_b2b_reverse1	5'-GTGTACCTCCGTTTAACCACAAAGCGATGACAGAC-3'
MSV_b2b_forward	5'-CCATCCCTTCAGATCCAGACACTCCAGCATG-3'
MSV_b2b_reverse	5'-CCTATCGGCCTTGCTTCCAGCCTTCTTC-3'
PanSV_b2b_forward	5'-CTACAAGGTTGCCCTGGACTACCACTTC-3'
PanSV_b2b_reverse	5'-GTCAGAGTCTCGTTGGTGTGGCGTTC-3'
SSRV_b2b_forward	5'-GTGGTGTTTGTGATATCCTCGGGTCCTATTCAC-3'
SSRV_b2b_reverse	5'-CCGAGGGAACGGTGATCATTGACGTG-3'
SSRV_b2b_forward1	5'-GCGGACAGGAAGCAAACACATACAATACTC-3'
SSRV_b2b_reverse1	5'-CTATCTGCGGACTATGTAATGTAACGCCTCC-3'

Supplementary Table 2. Number of Poaceae species for each assignment status.

Species	Status				Total
	Positive	Doubtful	Negative	Failed	
Avena sativa	6	3	75	10	94
Bothriochloa insculpta	0	0	49	1	50
Brachiaria brizantha	0	0	2	0	2
Brachiaria decumbens	0	0	10	0	10
Brachiaria umbellata	43	3	62	21	129
Bromus catharticus	0	0	3	1	4
Cenchrus echinatus	16	5	62	9	92
Chloris gayana	11	18	209	50	288
Chloris virgata	0	1	55	12	68
Chrysopogon zizanoides	0	0	5	5	10
Cynodon dactylon	6	8	186	26	226
Cyperus rotundus	3	1	57	32	93
Dactyloctenium aegyptium	3	3	31	8	45
Digitaria ciliaris	36	3	88	22	149
Digitaria debilis	1	0	0	0	1
Echinochloa colona	0	2	0	1	3
Eleusine indica	11	1	62	11	85
Eragrostis minor	8	8	120	48	184
Melinis repens	4	4	203	42	253
Paspalum dilatatum	0	0	24	18	42
Paspalum urvillei	0	0	0	1	1
Pennisetum clandestinum	0	1	5	1	7
Saccharum sp.	3	9	86	23	121
Setaria pumila	1	0	14	6	21
Sorghum arundinaceum	34	13	185	57	289
Sporobolus africanus	1	0	15	7	23
Urochloa maxima	6	10	296	57	369
Urochloa miliaceum	0	0	32	0	32
Urochloa sp.	0	0	12	18	30
Zea mays	1	9	139	16	165
Total	194	102	2087	503	2886

Supplementary Table 3. Positive number of Poaceae species for each survey.

Species	Positive				All
	Nov 2014	Nov 2016	Apr 2017	Nov 2017	
Avena sativa	0	6	0	0	6
Bothriochloa insculpta	0	0	0	0	0
Brachiaria brizantha	0	0	0	0	0
Brachiaria decumbens	0	0	0	0	0
Brachiaria umbellata	0	1	27	15	43
Bromus catharticus	0	0	0	0	0
Cenchrus echinatus	1	1	11	3	16
Chloris gayana	0	6	5	0	11
Chloris virgata	0	0	0	0	0
Chrysopogon zizanoides	0	0	0	0	0
Cynodon dactylon	1	3	2	0	6
Cyperus rotundus	0	1	2	0	3
Dactyloctenium aegyptium	1	0	2	0	3
Digitaria ciliaris	6	24	5	1	36
Digitaria debilis	0	0	1	0	1
Echinochloa colona	0	0	0	0	0
Eleusine indica	1	0	0	10	11
Eragrostis minor	0	7	0	1	8
Melinis repens	2	1	1	0	4
Paspalum dilatatum	0	0	0	0	0
Paspalum urvillei	0	0	0	0	0
Pennisetum clandestinum	0	0	0	0	0
Saccharum sp.	0	1	0	2	3
Setaria pumila	0	1	0	0	1
Sorghum arundinaceum	3	2	29	0	34
Sporobolus africanus	0	0	1	0	1
Urochloa maxima	1	1	1	3	6
Urochloa miliaceum	0	0	0	0	0
Urochloa sp.	0	0	0	0	0
Zea mays	0	0	1	0	1
Total	16	55	88	35	194

Supplementary Table 4. Summary of taxonomic classification of viral sequences per Poaceae species for each survey. The number in brackets corresponds to the number of Poaceae samples for which partial genomes were obtained.

Survey	Virus species / strain	Poaceae Species	Number of infected Poaceae sample	Number of Poaceae sample with full cloned genomes
Nov 2014	EIAV	*Eleusine indica*	1	1
	MeRAV	*Melinis repens*	2	1
	MSRV	*Cenchrus echinatus*	1	1
	MSV-B	*Cenchrus echinatus*	1	1
		Cynodon dactylon	1	-
		Dactyloctenium aegyptium	1	1
		Digitaria ciliaris	6	2
	SAAV	*Sorghum arundinaceum*	3	2
		Urochloa maxima	1	-
Nov 2016	EIAV	*Chloris gayana*	1	2
		Cynodon dactylon	1	1
	MeRAV	*Chloris gayana*	2	1
	MSV-B	*Avena sativa*	6	1
		Brachiaria umbellata	1	-
		Cenchrus echinatus	1	-
		Chloris gayana	4	2
		Cynodon dactylon	3	1
		Cyperus rotundus	1	-
		Digitaria ciliaris	23	12
		Eragrostis minor	6	-
		Melinis repens	1	-
		Setaria pumila	1	-
		Sorghum arundinaceum	1	1
		Urochloa maxima	1	-
	MSRV	*Setaria pumila*	-	1
	PanSV	*Eragrostis minor*	1	1
		Urochloa maxima	1	1
	SAAV	*Sorghum arundinaceum*	1	-
	SSRV	*Digitaria ciliaris*	3	2
	SWSV	*Saccharum sp.*	1	(1)
Apr 2017	MSV-A	*Cenchrus echinatus*	2	-
		Cyperus rotundus	1	-
	MSV-B	*Brachiaria umbellata*	26	3

		Cenchrus echinatus	5	3
		Chloris gayana	5	-
		Cynodon dactylon	2	-
		Cyperus rotundus	2	-
		Dactyloctenium aegyptium	2	-
		Digitaria ciliaris	5	1
		Digitaria debilis	1	1
		Melinis repens	1	-
		Sorghum arundinaceum	21	-
		Sporobolus africanus	1	-
		Zea mays	1	-
	MSV-F	*Cenchrus echinatus*	1	1
	MSRV	*Cenchrus echinatus*	5	1
	PanSV	*Urochloa maxima*	1	-
	SAAV	*Sorghum arundinaceum*	9	1
	SSRV	*Brachiaria umbellata*	3	2
		Cenchrus echinatus	3	1
		Dactyloctenium aegyptium	1	1
		Sorghum arundinaceum	1	1
Nov 2017	EIAV	*Eleusine indica*	10	-
	MSV-B	*Brachiaria umbellata*	13	-
		Cenchrus echinatus	2	-
		Digitaria ciliaris	1	-
		Eragrostis minor	1	-
	MSV-F	*Brachiaria umbellata*	3	-
	PanSV	*Cenchrus echinatus*	1	-
	SSRV	*Brachiaria umbellata*	1	-
	SWSV	*Saccharum sp.*	2	-

Article 4

Sorghum mastrevirus-associated alphasatellites: new geminialphasatellites associated with an African streak mastrevirus infecting wild Poaceae plants on Reunion

Article 4

Sorghum mastrevirus-associated alphasatellites: new geminialphasatellites associated with an African streak mastrevirus infecting wild Poaceae plants on Reunion

Sohini Claverie[1,2], Arvind Varsani[3,4], Murielle Hoareau[1], Denis Filloux[5,6], Philippe Roumagnac[5,6], Darren P. Martin[7], Pierre Lefeuvre[1], Jean-Michel Lett[1]

[1]CIRAD, UMR PVBMT, F-97410 Saint-Pierre, La Réunion, France
[2]Université de La Réunion, UMR PVBMT, F-97410 Saint-Pierre, La Réunion, France
[3]The Biodesign Center for Fundamental and Applied Microbiomics, Center for Evolution and Medicine, School of Life Sciences, Arizona State University, Tempe, AZ, USA
[4]Structural Biology Research Unit, Department of Clinical Laboratory Sciences, University of Cape Town, Rondebosch, Cape Town, South Africa
[5]CIRAD, UMR BGPI, F-34398 Montpellier, France
[6]BGPI, Université de Montpellier, CIRAD, INRA, Montpellier SupAgro, F-34398 Montpellier, France
[7]Computational Biology Division, Department of Integrative Biomedical Sciences, Institute of Infectious Diseases and Molecular Medicine, University of Cape Town, Observatory, South Africa

Abstract

Nine complete nucleotide sequences of geminialphasatellites recovered from the wild Poaceae *Sorghum arundinaceum* collected in Reunion are described and analyzed. While the helper geminivirus was identified as an isolate of maize streak virus (genus *Mastrevirus*; family *Geminiviridae*), the geminialphasatellite genomes were most closely related to, and shared ~63% identity with, clecrusatellites. Even though the geminialphasatellite molecules lack an adenine rich-region, they have the typical size of geminialphasatellites, encode a replication-associated protein in the virion sense and have probable stem-loop structures at their probable virion strand origins of replication. According to the proposed geminialphasatellites species and genus demarcation thresholds (88% and 70% nucleotide identity, respectively), the genomes identified here represent a new species within a new genus for which we propose the name Sorghum mastrevirus-associated alphasatellite.

Annotated sequence record

Members of the genus *Mastrevirus* (family *Geminiviridae*) are vectored by leafhoppers (Hemiptera: Cicadellidae) to a wide range of monocotyledonous and dicotyledonous plants. The circular single stranded DNA (ssDNA) genomes of mastreviruses are ~2.5-2.7kb in length and have at least four large open reading frames coding for a movement protein (MP) and a capsid protein (CP) on the virion sense genome strand and a replication associated protein (Rep) and RepA on the complementary sense genome strand. While circular ssDNA satellite molecules are often associated with members of the geminivirus genus *Begomovirus*, only a few recent reports have described their association with mastreviruses (Hamza *et al*., 2018; Kumar *et al*., 2014; Rosario *et al*., 2013) and there have not been any reports of satellite molecules associated with the best characterized groups of mastreviruses such as the African streak viruses (AfSVs).

Geminivirus satellite molecules are ~ 0.7-1.3kb in size and they fall into two families, *Alphasatellitidae* and *Tolecusatellitidae*. Geminivirus-associated alphasatellites (family *Alphasatellitidae,* subfamily *Geminialphasatellitinae*) are divided into four genera *Ageyesisatellite*, *Clecrusatellite*, *Colecusatellite* and *Gosmusatellite* (Rob W. Briddon *et al*., 2018). These satellite molecules generally contain two conserved regions, an adenine rich (A-rich) region and a Rep encoding gene. As a consequence of encoding a Rep, they are able to initiate their own replication. This is unlike tolecusatellites (formerly betasatellites), which depend on the Rep of their helper virus. Nonetheless, members of both *Alphasatellitidae* and *Tolecusatellitidae* are dependent on their helper virus for encapsidation, movement within the plant and transmission by insect vectors (R. W. Briddon & Stanley, 2006).

A total of five samples of the wild Poaceae *Sorghum arundinaceum* were randomly collected in December 2016 (n=1) and April 2017 (n=4) within an agro-ecosystem at the Bassin Plat CIRAD experimental facility (Latitude -21.3231; Longitude 55.4912) in Saint Pierre (Reunion). None of these plants presented visible symptoms of streak disease. Total DNA was extracted from dried leaf material using the DNeasy Plant DNA extraction kit (Qiagen). Viral and satellite DNA molecules were amplified by rolling circle amplification (RCA) using the Illustra TempliPhi Amplification Kit (GE Healthcare). RCA amplified products were digested with *Acc*I, *Bam*HI and *Sal*I endonucleases to obtain unit length genomic molecules (**Supplementary Table 1**). The purified fragments were cloned in the pJET 1.2 vector (Thermo Fisher Scientific) and Sanger sequenced by primer-walking at Macrogen Europe. Sequences were assembled with Geneious v6.0.6 (Kearse *et al*., 2012) and were subjected to a BLAST search for preliminary identification.

Among the assembled genomic DNA sequences we identified maize streak virus (MSV) like sequences as well as sequences with detectable similarity to geminialphasatellites. Based on the sequences of these geminialphasatellite-like molecules, two pairs of abutting primers were designed for a more thorough search and recovery of additional geminialphasatellite sequences (**Supplementary Table 1**). After PCR amplification and purification, the amplicons were cloned and Sanger sequenced as described above. Pairwise

similarities between the complete geminialphasatellite sequences were determined using SDT v1.2 with pairwise MUSCLE alignments and deletion of gaps (Muhire *et al*., 2013). Finally, a Maximum-Likelihood phylogenetic tree (with 1000 bootstrap replicates and GTR+I+G4 substitution model) was constructed from the alignment of geminialphasatellite sequences using MEGA6 (Tamura *et al*., 2013).

The complete genome of an isolate of MSV (MN901976) was obtained from one sample (**Supplementary Table 1**), which shared 99.8% nucleotide identity with an isolate of the B1 strain of MSV (MK546374) that had previously been isolated from *Digitaria ciliaris* in Reunion. Besides this one MSV sequence we also identified nine complete geminialphasatellite-like molecules (MN901967 to MN901975; **Supplementary Table 1**). These nine sequences presented a minimum nucleotide identity of 98% to one another and were most closely related (63% pairwise identity) to clecrusatellites: cucurbit yellow mosaic alphasatellite from Pakistan (CYMA-[PK-51SA-12]; KT948075), whitefly associated Guatemala alphasatellite from Guatemala (WasGA-[GT-GtTo2-12]; KT099170) and whitefly associated Puerto Rico alphasatellite from Puerto Rico (WasPRA-[PR-PR3/6-10]; KT099173) (**Figure 1**). At between 1531 and 1532 nt in length the nine new sequences have a size that is typical of geminialphasatellites. They also have other features typical of known geminialphasatellites such as a stem-loop structure with a TAGTATT$^{\downarrow}$AC sequence at their presumed virion strand origin of replication (indicated by $^{\downarrow}$), and a Rep encoding gene (**Supplementary Figure 1**). However, these molecules lacked a A-rich region that is usually found downstream of the *rep* genes of geminialphasatellites. Additionally, a polyadenylation motif, AAATAA (nucleotides 1124-1129), is located at the 3’ terminus of the Rep ORFs (**Supplementary Figure 2**).

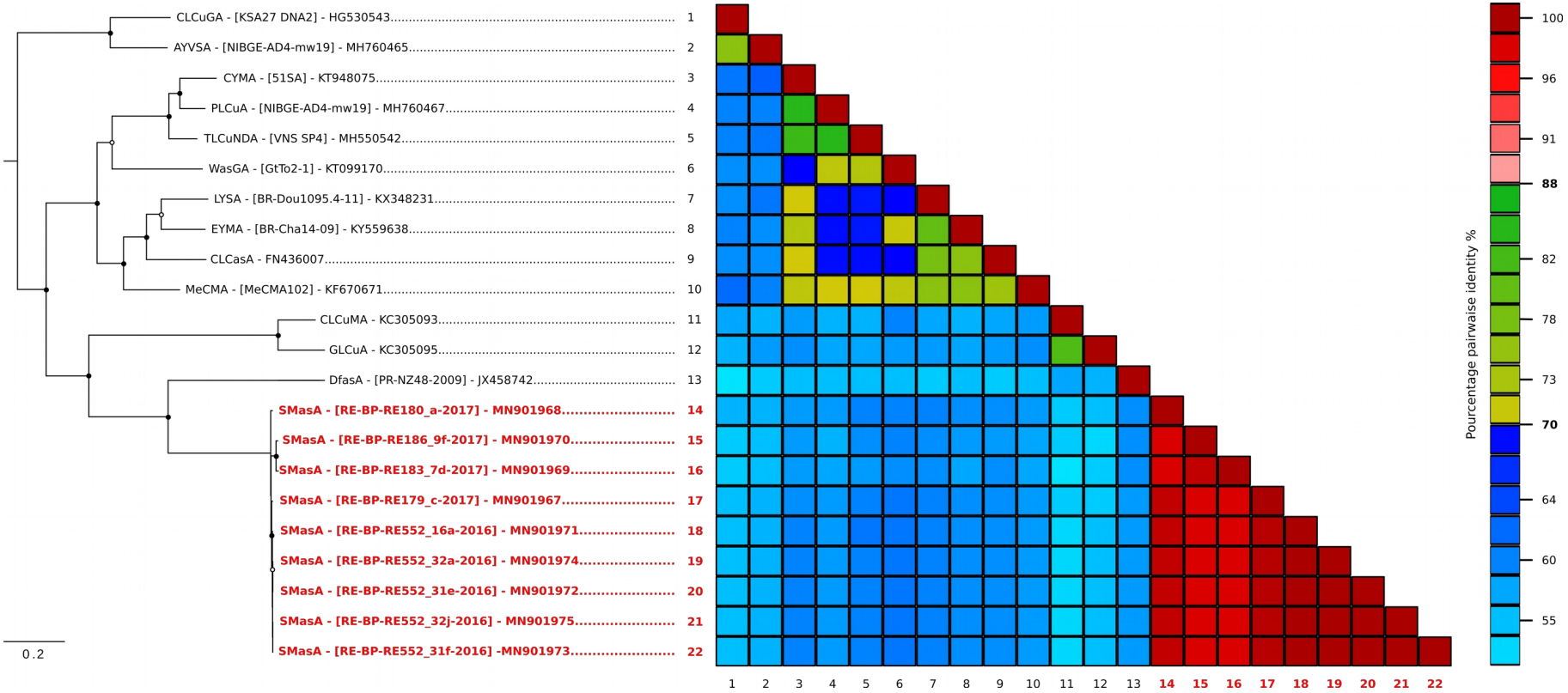


Figure 1 Maximum-likelihood (ML) phylogenetic tree showing relationships between complete genome sequences of Sorghum mastrevirus associated alphasatellite identified in this study (highlighted in bold and red) and representative geminialphasatellite sequences from the four genera *Ageyesisatellite*, *Clecrusatellite*, *Colecusatellite* and *Gosmusatellite* described in the subfamily *Geminialphasatellitinae*. The tree was rooted with the representative DNA-R nucleotide sequence of the nanovirus milk vetch dwarf virus (AB027511). Numbers at nodes indicate bootstrap scores (1000 replicates). Only bootstrap values greater than 70% are displayed. Acronyms and accession numbers of nucleotide sequences are available in **Supplementary Table 1** and **Supplementary Table 2**.

The predicted Rep protein shares ~59% amino acid sequence identity with that of the melon chlorotic mosaic alphasatellite (MeCMA-[MeCMA140]; KF670681). As with all other known geminialphasatellite encoded Rep proteins, the predicted Rep of the new geminialphasatellite-like molecules has a tyrosine residue located at amino acid 82 (**Supplementary Figure 2**), which, by analogy with other known Reps, is likely to participate in the initiation of rolling circle replication (Koonin & Ilyina, 1992). A consensus NTP binding motif (GGEGKT/S) such as that which is present in the Reps of both nanoviruses and geminiviruses (Laufs *et al*., 1995) was also identified in the new geminialphasatellite-like sequences (amino acids 177-182 of the Rep protein; **Supplementary Figure 2**). The Maximum-Likelihood phylogenetic tree shows that the complete nucleotide sequences of sorghum mastrevirus-associated alphasatellites cluster together and form a distinct cluster in the sub-family *Geminialphasatellitinae* (**Figure 1**).

To our knowledge, this is the first report of a natural association between a geminialphasatellite and an AfSV. Based on the proposed species and genus demarcation thresholds of geminialphasatellites (88% and 70% nucleotide identity, respectively [4]), the mastrevirus-associated alphasatellite sequences would be classified into a new geminialphasatellite species and would likely also be assigned to a new genus. We have tentatively named these geminialphasatellites Sorghum mastrevirus-associated alphasatellite (SMasA). Since SMasA was identified from wild plants without streak symptoms, biological tests should be performed to characterize any contribution of SMasA to the biological traits of MSV.

GenBank accession number

MN901967 – MN901975

Acknowledgements

This study was funded by the European Union (ERDF, contract GURDT I2016-1731-0006632), the Agropolis Fondation (Labex Agro – Montpellier, E-SPACE project number 1504-004), the *Conseil Régional de la Réunion* and CIRAD. SC

is a recipient of a PhD fellowship from CIRAD and the Agropolis Fondation (E-SPACE). This work was conducted on the Plant Protection Platform (3P, IBISA).

Authors information

Affiliations

CIRAD, UMR PVBMT, Pôle de Protection des Plantes, 7 chemin de l'Irat, 97410 Saint Pierre, La Réunion, France
Sohini Claverie, Murielle Hoarau, Jean-Michel Lett & Pierre Lefeuvre

CIRAD, UMR BGPI, TA A-54/K, Campus International de Baillarguet, 34398 Montpellier Cedex 5, France
Denis Filloux & Philippe Roumagnac

The Biodesign Center for Fundamental and Applied Microbiomics, Center for Evolution and Medicine, School of Life Sciences, Arizona State University, 1001 S. McAllister Ave, Tempe, AZ 85287-5001, USA
Arvind Varsani

Computational Biology Division, Departement of Integrative Biomedical Sciences, Institute of Infectious Diseases and Molecular Medicine, University of Cape Town, Observatory 7925, South Africa
Darren P. Martin

Author contributions

S.C., J.M.L and P.L. conceived and designed the experiments. S.C., M.H., J.M.L and P.L. performed the experiments. S.C., D.F.., A.V., P.R., D.P.M., J.M.L and P.L. analysed the data. S.C, A.V., P.R., D.P.M., J.M.L and P.L. wrote the paper. J.M.L. and P.L. secured funding for the project's execution.

Corresponding author

Correspondence to Jean-Michel Lett

Competing interests

In the interests of transparency and to help readers to form their own judgements of potential bias, authors declare no competing interests in relation to the work described.

References

Briddon, R. W., & Stanley, J. (2006). Subviral agents associated with plant single-stranded DNA viruses. *Virology*, *344*(1), 198–210. https://doi.org/10.1016/j.virol.2005.09.042

Briddon, Rob W., Martin, D. P., Roumagnac, P., Navas-Castillo, J., Fiallo-Olivé, E., Moriones, E., Lett, J. M., Zerbini, F. M., & Varsani, A. (2018). Alphasatellitidae: a new family with two subfamilies for the classification of geminivirus- and nanovirus-associated alphasatellites. *Archives of Virology*, *163*(9), 2587–2600. https://doi.org/10.1007/s00705-018-3854-2

Hamza, M., Tahir, M. N., Mustafa, R., Kamal, H., Khan, M. Z., Mansoor, S., Briddon, R. W., & Amin, I. (2018). Identification of a dicot infecting mastrevirus along with alpha- and betasatellite associated with leaf curl disease of spinach (Spinacia oleracea) in Pakistan. *Virus Research*, *256*(August), 174–182. https://doi.org/10.1016/j.virusres.2018.08.017

Kearse, M., Moir, R., Wilson, A., Stones-Havas, S., Cheung, M., Sturrock, S., Buxton, S., Cooper, A., Markowitz, S., Duran, C., Thierer, T., Ashton, B., Meintjes, P., & Drummond, A. (2012). Geneious basic: An integrated and extendable desktop software platform for the organization and aalysis of sequence data. *Bioinformatics*, *28*(12), 1647–1649. https://doi.org/10.1093/bioinformatics/bts199

Koonin, E. V., & Ilyina, T. V. (1992). Geminivirus replication proteins are related to prokaryotic plasmid rolling circle DNA replication initiator proteins. *Journal of General Virology*, *73*(10), 2763–2766. https://doi.org/10.1099/0022-1317-73-10-2763

Kumar, J., Kumar, J., Singh, S. P., & Tuli, R. (2014). Association of Satellites with a Mastrevirus in Natural Infection: Complexity of Wheat Dwarf India Virus Disease. *Journal of Virology*, *88*(12), 7093–7104. https://doi.org/10.1128/jvi.02911-13

Laufs, J., Schumacher, S., Geisler, N., Jupin, I., & Gronenborn, B. (1995). Identification of the nicking tyrosine of geminivirus Rep protein. *FEBS Letters*, *377*(2), 258–262. https://doi.org/10.1016/0014-5793(95)01355-5

Muhire, B., Martin, D. P., Brown, J. K., Navas-Castillo, J., Moriones, E., Zerbini, F. M., Rivera-Bustamante, R., Malathi, V. G., Briddon, R. W., & Varsani, A. (2013). A genome-wide pairwise-identity-based proposal for the classification of viruses in the genus Mastrevirus (family Geminiviridae). *Archives of Virology*, *158*(6), 1411–1424. https://doi.org/10.1007/s00705-012-1601-7

Rosario, K., Padilla-Rodriguez, M., Kraberger, S., Stainton, D., Martin, D. P., Breitbart, M., & Varsani, A. (2013). Discovery of a novel mastrevirus and alphasatellite-like circular DNA in dragonflies (Epiprocta) from Puerto Rico. *Virus Research*, *171*(1), 231–237. https://doi.org/10.1016/j.virusres.2012.10.017

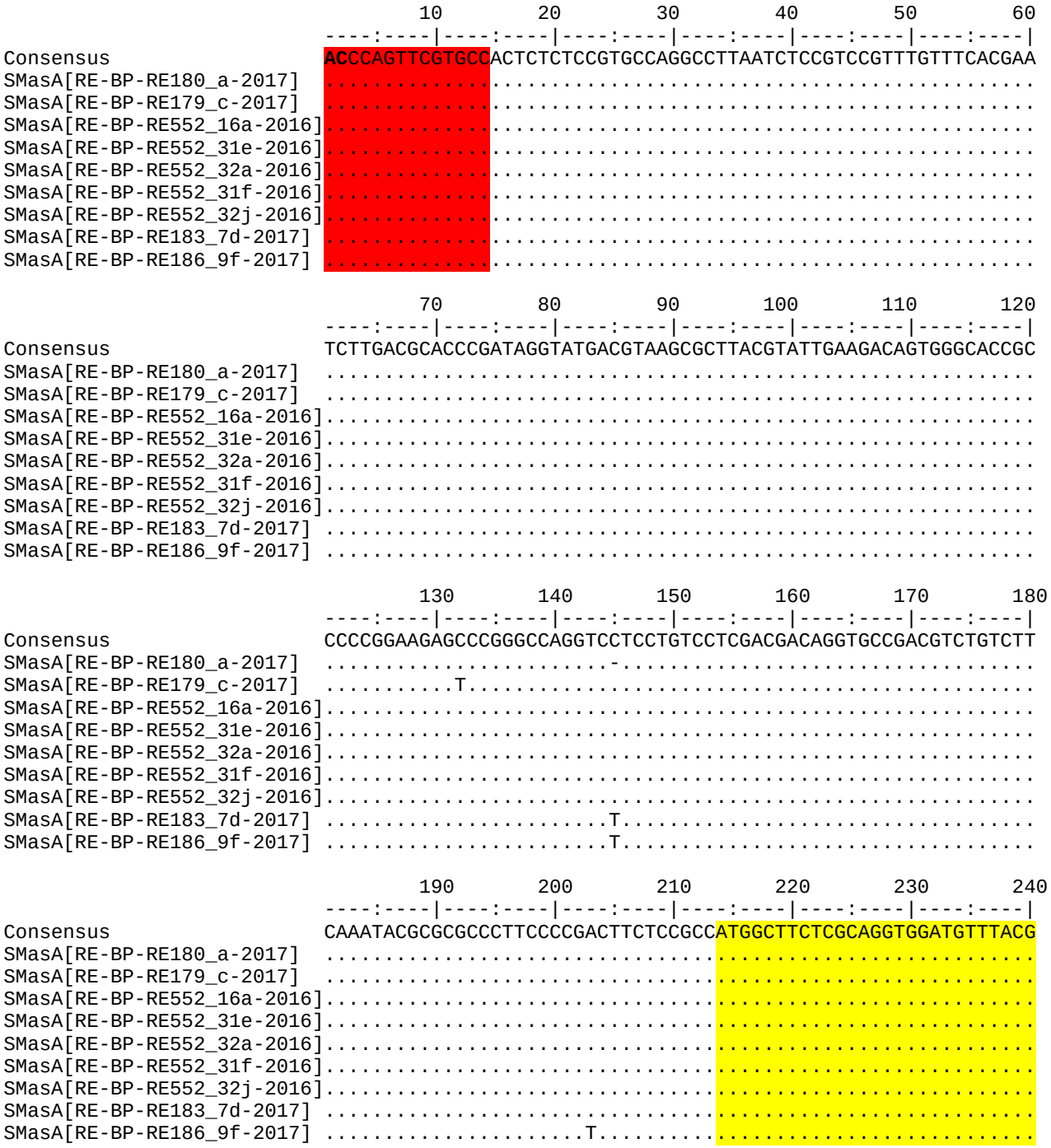
Tamura, K., Stecher, G., Peterson, D., Filipski, A., & Kumar, S. (2013). MEGA6: Molecular evolutionary genetics analysis version 6.0. *Molecular Biology and Evolution*, *30*(12), 2725–2729. https://doi.org/10.1093/molbev/mst197

Supplementary Table 1. Isolate names, GenBank accession numbers and primers or restricted enzymes used for the molecular characterisation of geminialphasatellites and their associated mastrevirus.

Satellite and associated mastrevirus isolates	Accession numbers	Primers pair or restriction enzyme used for cloning
SMasA[RE-BP-RE179_c-2017]	MN901967	alpha_F 5’-ATGATAGAAATATCATCTGGGTGTGCGGCACT-3’ alpha_R 5’-CAGGGCCCTGGTTAATTCTCTCATTAAAACATATC-3'
SMasA[RE-BP-RE180_a-2017]	MN901968	alpha_F 5’-ATGATAGAAATATCATCTGGGTGTGCGGCACT-3’ alpha_R 5’-CAGGGCCCTGGTTAATTCTCTCATTAAAACATATC-3'
SMasA[RE-BP-RE183_7d-2017]	MN901969	alpha_F 5’-ATGATAGAAATATCATCTGGGTGTGCGGCACT-3’ alpha_R 5’-CAGGGCCCTGGTTAATTCTCTCATTAAAACATATC-3'
SMasA[RE-BP-RE186_9f-2017]	MN901970	alpha_F 5’-ATGATAGAAATATCATCTGGGTGTGCGGCACT-3’ alpha_R 5’-CAGGGCCCTGGTTAATTCTCTCATTAAAACATATC-3'
SMasA[RE-BP-RE552_16a-2016]	MN901971	sat_F 5’-GATCTCCTCTGATTGACGTGGAAGATCGAAGCAGA-3’ sat_R 5’-CTTAACTACGCGGTCATTGAGGAAATTAAGGATC-3’
SMasA[RE-BP-RE552_31e-2016]	MN901972	*Acc*I
SMasA[RE-BP-RE552_31f-2016]	MN901973	*Acc*I
SMasA[RE-BP-RE552_32a-2016]	MN901974	*Sal*I
SMasA[RE-BP-RE552_32j-2016]	MN901975	*Sal*I
MSV-B1[RE-BP-RE179_e-2017]	MN901976	*Bam*HI

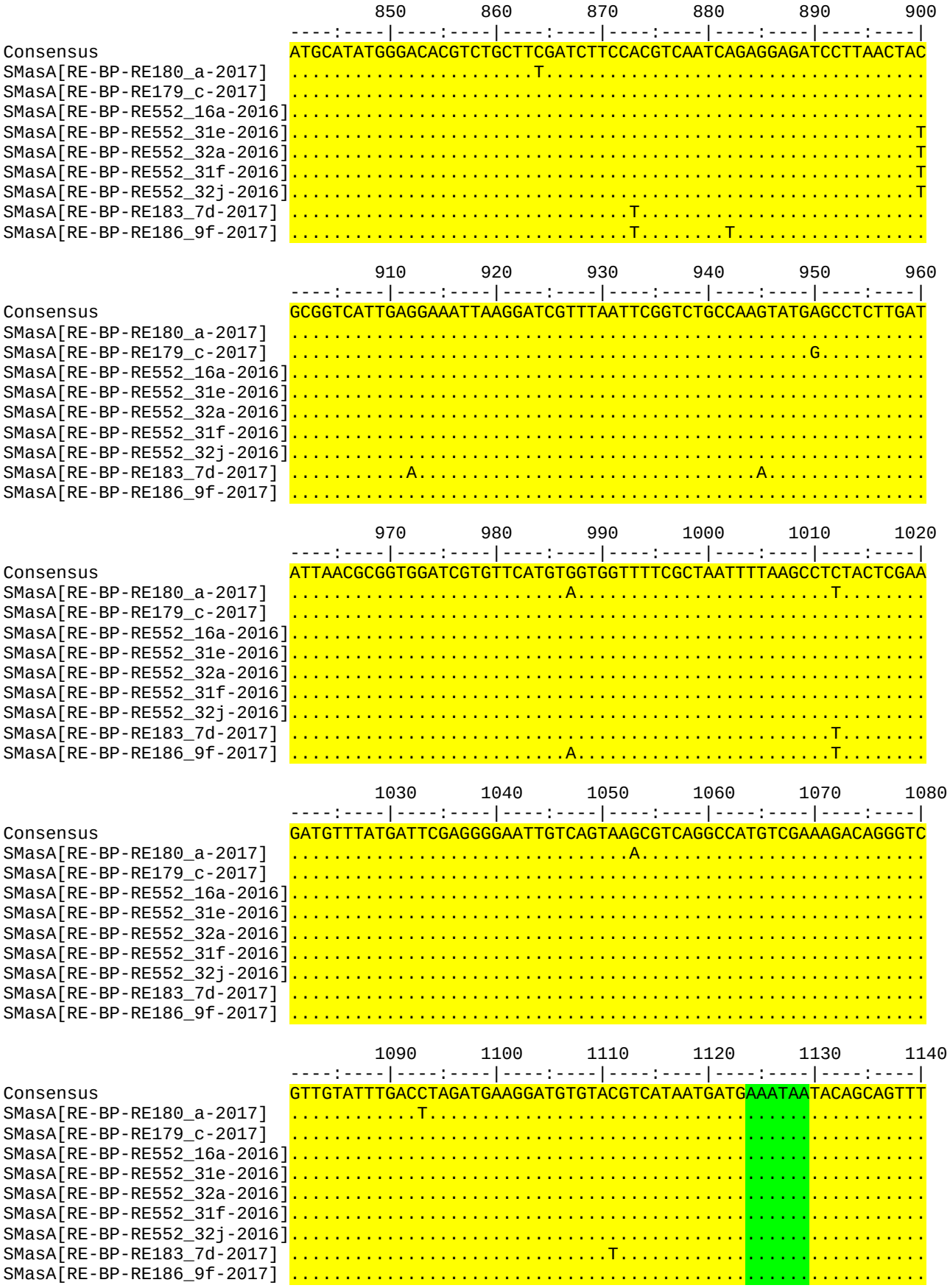
Supplementary Table 2. Geminialphasatellites, acronyms and accession numbers used in this study.

Genus	Geminialphasatellites species	Acronym	Accession number
Ageyesisatellite	*Ageratum yellow vein Singapore alphasatellite*	AYVSGA	AJ416153
	Cotton leaf curl Saudi Arabia alphasatellite	CLCuSAA	HG530543
Clecrusatellite	*Cleome leaf crumple alphasatellite*	ClLCrA	FN436007
	Whitefly associated Guatemala alphasatellite 2	WfaGA 2	KT099170
	Melon chlorotic mosaic alphasatellite	MeCMA	HM163578
	Whitefly associated Puerto Rico alphasatellite 1	WfaPRA 1	KT099173
	Cucurbit yellow mosaic alphasatellite	CYMA	KT948075
Colecusatellidae	*Cotton leaf curl Multan alphasatellite*	CLCuMuA	AJ132344
	Cassava mosaic Madagascar alphasatellite	CMMGA	HE984148
	Cotton leaf curl Gezira alphasatellite	CLCuGeA	EU589450
	Cotton leaf curl Egypt alphasatellite	CLCuEA	AJ512960
	Tomato leaf curl Buea alphasatellite	ToLCuBuA	FN675299
Gosmusatellite	*Gossypium mustelinum symptomless alphasatellite*	GMusSLA	EU384656
Unassigned	*Dragonfly associated alphasatellite*	DfaA	JX458742
	Whitefly associated Guatemala alphasatellite 1	WfaGA 1	KT099172

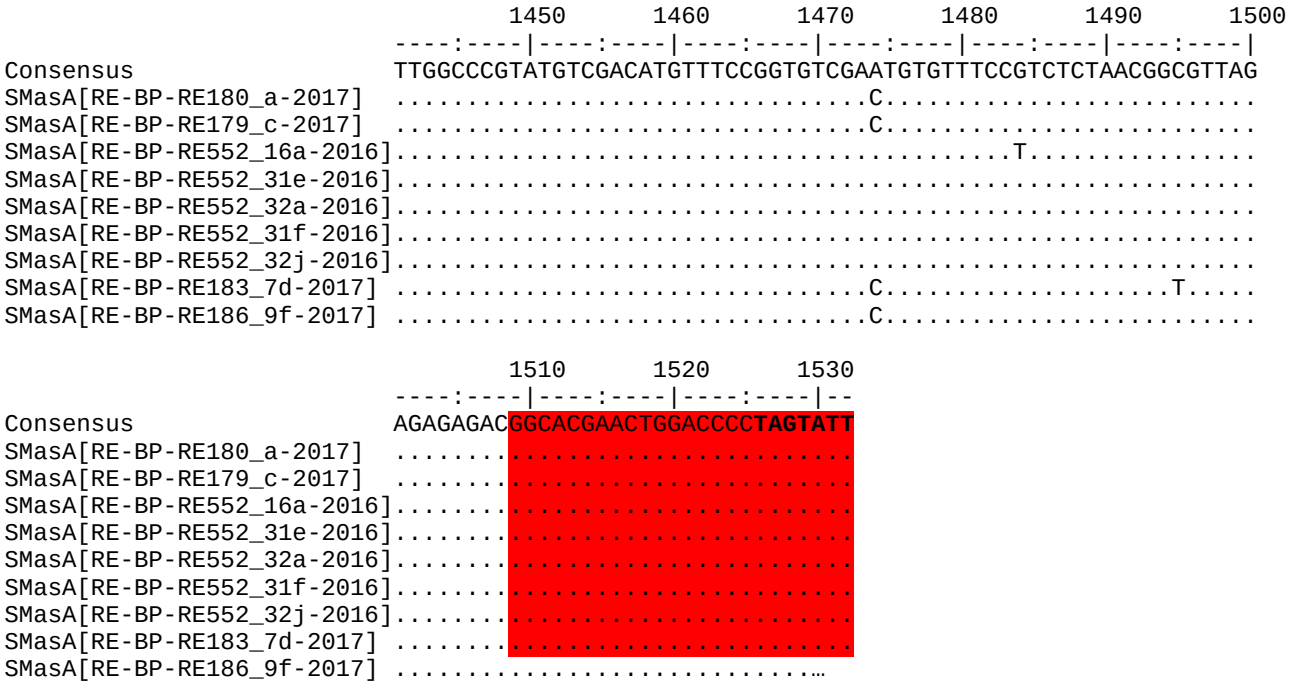

10 20 30 40 50 60
Consensus ACCCAGTTCGTGCCACTCTCTCCGTGCCAGGCCTTAATCTCCGTCCGTTTGTTTCACGAA
SMasA[RE-BP-RE180_a-2017]
SMasA[RE-BP-RE179_c-2017]
SMasA[RE-BP-RE552_16a-2016]
SMasA[RE-BP-RE552_31e-2016]
SMasA[RE-BP-RE552_32a-2016]
SMasA[RE-BP-RE552_31f-2016]
SMasA[RE-BP-RE552_32j-2016]
SMasA[RE-BP-RE183_7d-2017]
SMasA[RE-BP-RE186_9f-2017]
70 80 90 100 110 120
Consensus TCTTGACGCACCCGATAGGTATGACGTAAGCGCTTACGTATTGAAGACAGTGGGCACCGC
SMasA[RE-BP-RE180_a-2017]
SMasA[RE-BP-RE179_c-2017]
SMasA[RE-BP-RE552_16a-2016]
SMasA[RE-BP-RE552_31e-2016]
SMasA[RE-BP-RE552_32a-2016]
SMasA[RE-BP-RE552_31f-2016]
SMasA[RE-BP-RE552_32j-2016]
SMasA[RE-BP-RE183_7d-2017]
SMasA[RE-BP-RE186_9f-2017]
130 140 150 160 170 180
Consensus CCCCGGAAGAGCCCGGGCCAGGTCCTCCTGTCCTCGACGACAGGTGCCGACGTCTGTCTT
SMasA[RE-BP-RE180_a-2017]
SMasA[RE-BP-RE179_c-2017]
SMasA[RE-BP-RE552_16a-2016]
SMasA[RE-BP-RE552_31e-2016]
SMasA[RE-BP-RE552_32a-2016]
SMasA[RE-BP-RE552_31f-2016]
SMasA[RE-BP-RE552_32j-2016]
SMasA[RE-BP-RE183_7d-2017]
SMasA[RE-BP-RE186_9f-2017]
190 200 210 220 230 240
Consensus CAAATACGCGCGCCCTTCCCCGACTTCTCCGCCATGGCTTCTCGCAGGTGGATGTTTACG
SMasA[RE-BP-RE180_a-2017]
SMasA[RE-BP-RE179_c-2017]
SMasA[RE-BP-RE552_16a-2016]
SMasA[RE-BP-RE552_31e-2016]
SMasA[RE-BP-RE552_32a-2016]
SMasA[RE-BP-RE552_31f-2016]
SMasA[RE-BP-RE552_32j-2016]
SMasA[RE-BP-RE183_7d-2017]
SMasA[RE-BP-RE186_9f-2017]

```
                                 250       260       270       280       290       300
                                 ----:----|----:----|----:----|----:----|----:----|----:----|
Consensus                        CTTTTTGAGGATTTTCCCTCTCCGCCATTCGCAGACCTTCCTGAGTCTGCCGAGTATTTA
SMasA[RE-BP-RE180_a-2017]        ............................T............................
SMasA[RE-BP-RE179_c-2017]        ........................................................................
SMasA[RE-BP-RE552_16a-2016]      ........................................................................
SMasA[RE-BP-RE552_31e-2016]      ........................................................................
SMasA[RE-BP-RE552_32a-2016]      ........................................................................
SMasA[RE-BP-RE552_31f-2016]      ........................................................................
SMasA[RE-BP-RE552_32j-2016]      ........................................................................
SMasA[RE-BP-RE183_7d-2017]       ........................................................................
SMasA[RE-BP-RE186_9f-2017]       ..........C.....A.......................................................

                                 310       320       330       340       350       360
                                 ----:----|----:----|----:----|----:----|----:----|----:----|
Consensus                        ATCTGTCAGAAGGAGAAGGCCCCCACCACTGGGAAGATTCATCTCCAGGGTTTTATTGTA
SMasA[RE-BP-RE180_a-2017]        ........................................................................
SMasA[RE-BP-RE179_c-2017]        ........................................................................
SMasA[RE-BP-RE552_16a-2016]      ........................................................................
SMasA[RE-BP-RE552_31e-2016]      ........................................................................
SMasA[RE-BP-RE552_32a-2016]      ........................................................................
SMasA[RE-BP-RE552_31f-2016]      ........................................................................
SMasA[RE-BP-RE552_32j-2016]      ........................................................................
SMasA[RE-BP-RE183_7d-2017]       ........................................................................
SMasA[RE-BP-RE186_9f-2017]       ........................................................................

                                 370       380       390       400       410       420
                                 ----:----|----:----|----:----|----:----|----:----|----:----|
Consensus                        TTGAAGTCTCCCAGGCGGATTACTTTTCTTCGGAAATTTCTCGGTAAATCTGCGCATCTT
SMasA[RE-BP-RE180_a-2017]        ............................C.....................T......
SMasA[RE-BP-RE179_c-2017]        ............A............C.............................
SMasA[RE-BP-RE552_16a-2016]      ........................................................................
SMasA[RE-BP-RE552_31e-2016]      ........................................................................
SMasA[RE-BP-RE552_32a-2016]      ........................................................................
SMasA[RE-BP-RE552_31f-2016]      ........................................................................
SMasA[RE-BP-RE552_32j-2016]      ........................................................................
SMasA[RE-BP-RE183_7d-2017]       ............A............C.............................
SMasA[RE-BP-RE186_9f-2017]       ............................C.............................

                                 430       440       450       460       470       480
                                 ----:----|----:----|----:----|----:----|----:----|----:----|
Consensus                        GAACATGCTCGCTCTAAGTCCTCAAGTTGCAGAGATTATTGCCGAAAGGATGCTACAAGA
SMasA[RE-BP-RE180_a-2017]        ........................................................................
SMasA[RE-BP-RE179_c-2017]        ........................................................................
SMasA[RE-BP-RE552_16a-2016]      ........................................................................
SMasA[RE-BP-RE552_31e-2016]      ..............................................................T.......
SMasA[RE-BP-RE552_32a-2016]      ..............................................................T.......
SMasA[RE-BP-RE552_31f-2016]      ........................................................................
SMasA[RE-BP-RE552_32j-2016]      ........................................................................
SMasA[RE-BP-RE183_7d-2017]       ........................................................................
SMasA[RE-BP-RE186_9f-2017]       ........................................................................

                                 490       500       510       520       530       540
                                 ----:----|----:----|----:----|----:----|----:----|----:----|
Consensus                        ACCGATGGTCCGTGGGAGTATGGTGTCTTCGCGGAGCAGGGCAGCAAGGCGAGGAAAGCT
SMasA[RE-BP-RE180_a-2017]        ........................................................................
SMasA[RE-BP-RE179_c-2017]        ........................................................................
SMasA[RE-BP-RE552_16a-2016]      ........................................................................
SMasA[RE-BP-RE552_31e-2016]      ........................................................................
SMasA[RE-BP-RE552_32a-2016]      ........................................................................
SMasA[RE-BP-RE552_31f-2016]      ........................................................................
SMasA[RE-BP-RE552_32j-2016]      ........................................................................
SMasA[RE-BP-RE183_7d-2017]       ........................................................................
SMasA[RE-BP-RE186_9f-2017]       ........................................................................
```

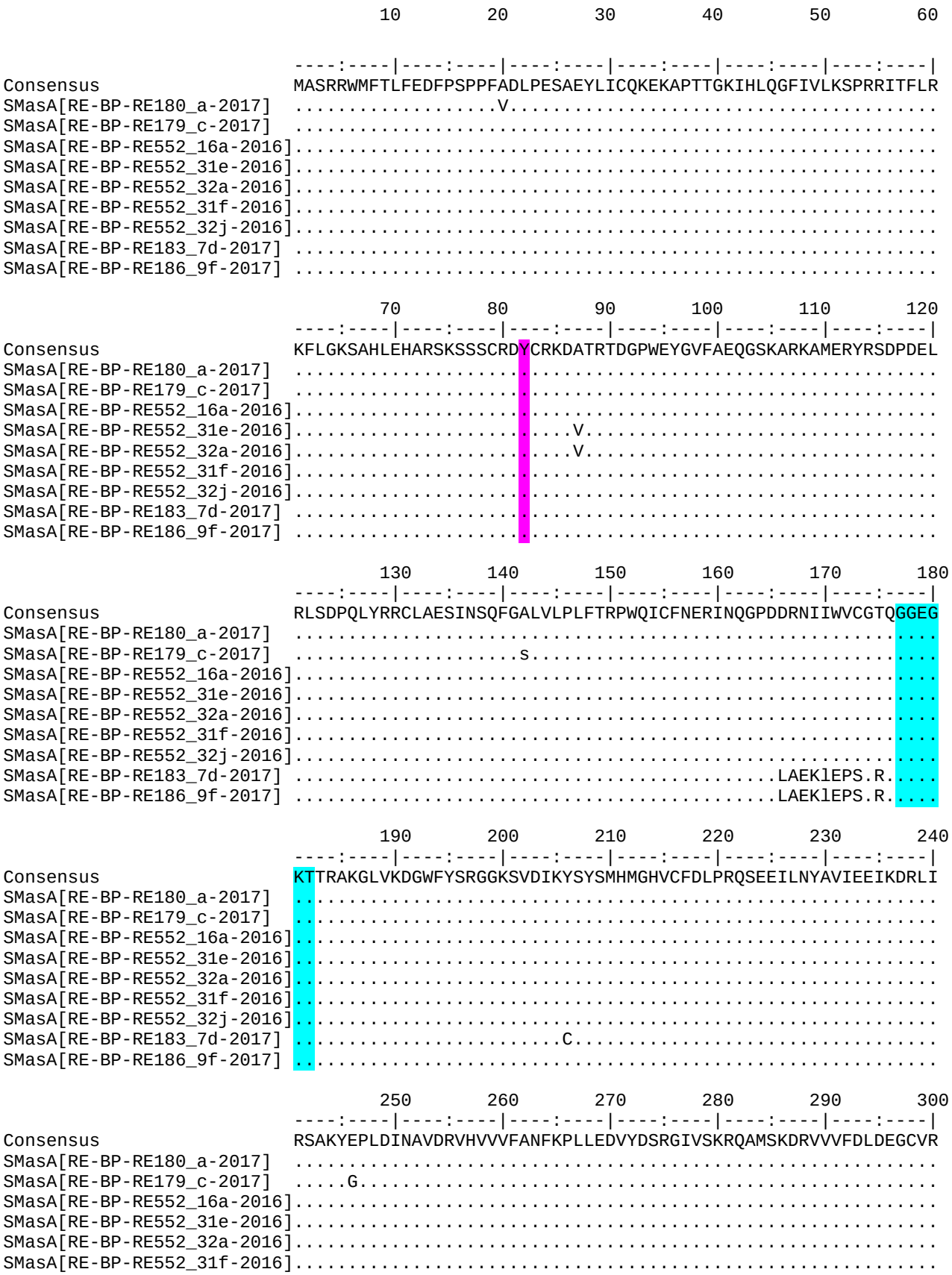
```
                                 550       560       570       580       590       600
                                 ----:----|----:----|----:----|----:----|----:----|----:----|
Consensus                        ATGGAACGCTATCGTAGTGACCCGGATGAACTACGCCTATCTGACCCTCAATTGTATCGT
SMasA[RE-BP-RE180_a-2017]        ..................................C.....................
SMasA[RE-BP-RE179_c-2017]        ..................................C.....................
SMasA[RE-BP-RE552_16a-2016]      ............................................................
SMasA[RE-BP-RE552_31e-2016]      ............................................................
SMasA[RE-BP-RE552_32a-2016]      ............................................................
SMasA[RE-BP-RE552_31f-2016]      ............................................................
SMasA[RE-BP-RE552_32j-2016]      ............................................................
SMasA[RE-BP-RE183_7d-2017]       ..................................C.....................
SMasA[RE-BP-RE186_9f-2017]       ..................................C.....................

                                 610       620       630       640       650       660
                                 ----:----|----:----|----:----|----:----|----:----|----:----|
Consensus                        CGATGCCTGGCGGAATCGATTAATTCGCAGTTCGGTGCTTTGGTATTGCCTCTATTTACT
SMasA[RE-BP-RE180_a-2017]        ............................................................
SMasA[RE-BP-RE179_c-2017]        ..................................T.....................
SMasA[RE-BP-RE552_16a-2016]      ............................................................
SMasA[RE-BP-RE552_31e-2016]      ............................................................
SMasA[RE-BP-RE552_32a-2016]      ............................................................
SMasA[RE-BP-RE552_31f-2016]      ............................................................
SMasA[RE-BP-RE552_32j-2016]      ............................................................
SMasA[RE-BP-RE183_7d-2017]       ............................................................
SMasA[RE-BP-RE186_9f-2017]       ............................................................

                                 670       680       690       700       710       720
                                 ----:----|----:----|----:----|----:----|----:----|----:----|
Consensus                        CGACCGTGGCAGATATGTTTTAATGAGAGAATTAACCAGGGCCCTGATGATAGAAATATC
SMasA[RE-BP-RE180_a-2017]        ............................................................
SMasA[RE-BP-RE179_c-2017]        ............................................................
SMasA[RE-BP-RE552_16a-2016]      ............................................................
SMasA[RE-BP-RE552_31e-2016]      ............................................................
SMasA[RE-BP-RE552_32a-2016]      ............................................................
SMasA[RE-BP-RE552_31f-2016]      ............................................................
SMasA[RE-BP-RE552_32j-2016]      ............................................................
SMasA[RE-BP-RE183_7d-2017]       ..............................................CT.GCTG.A.AA
SMasA[RE-BP-RE186_9f-2017]       ..............................................CT.GCTG.A.AA

                                 730       740       750       760       770       780
                                 ----:----|----:----|----:----|----:----|----:----|----:----|
Consensus                        ATCTGGGTGTGCGGCACTCAGGGCGGTGAAGGCAAAACGACGAGGGCGAAGGGCCTCGTG
SMasA[RE-BP-RE180_a-2017]        ............................................................
SMasA[RE-BP-RE179_c-2017]        ............................................................
SMasA[RE-BP-RE552_16a-2016]      ............................................................
SMasA[RE-BP-RE552_31e-2016]      ............................................................
SMasA[RE-BP-RE552_32a-2016]      ............................................................
SMasA[RE-BP-RE552_31f-2016]      ............................................................
SMasA[RE-BP-RE552_32j-2016]      ............................................................
SMasA[RE-BP-RE183_7d-2017]       C..GA.CCA.C...A.GA..........................................
SMasA[RE-BP-RE186_9f-2017]       C..GA.CCA.C...A.GA..........................................

                                 790       800       810       820       830       840
                                 ----:----|----:----|----:----|----:----|----:----|----:----|
Consensus                        AAGGATGGATGGTTTTACTCCCGCGGAGGGAAATCTGTCGATATCAAATACTCGTACTCT
SMasA[RE-BP-RE180_a-2017]        ............C...............................................
SMasA[RE-BP-RE179_c-2017]        ............................................................
SMasA[RE-BP-RE552_16a-2016]      ............................................................
SMasA[RE-BP-RE552_31e-2016]      ............................................................
SMasA[RE-BP-RE552_32a-2016]      ............................................................
SMasA[RE-BP-RE183_7d-2017]       ............................................................
SMasA[RE-BP-RE552_32j-2016]      ............................................................
SMasA[RE-BP-RE183_7d-2017]       ............C...............................G.............
SMasA[RE-BP-RE186_9f-2017]       ............C...............................................
```


850 860 870 880 890 900
Consensus ATGCATATGGGACACGTCTGCTTCGATCTTCCACGTCAATCAGAGGAGATCCTTAACTAC
SMasA[RE-BP-RE180_a-2017] .....T.....
SMasA[RE-BP-RE179_c-2017] .....
SMasA[RE-BP-RE552_16a-2016] .....
SMasA[RE-BP-RE552_31e-2016] .....T
SMasA[RE-BP-RE552_32a-2016] .....T
SMasA[RE-BP-RE552_31f-2016] .....T
SMasA[RE-BP-RE552_32j-2016] .....T
SMasA[RE-BP-RE183_7d-2017] .....T.....
SMasA[RE-BP-RE186_9f-2017] .....T.....T.....
910 920 930 940 950 960
Consensus GCGGTCATTGAGGAAATTAAGGATCGTTTAATTCGGTCTGCCAAGTATGAGCCTCTTGAT
SMasA[RE-BP-RE180_a-2017] .....
SMasA[RE-BP-RE179_c-2017] .....G.....
SMasA[RE-BP-RE552_16a-2016] .....
SMasA[RE-BP-RE552_31e-2016] .....
SMasA[RE-BP-RE552_32a-2016] .....
SMasA[RE-BP-RE552_31f-2016] .....
SMasA[RE-BP-RE552_32j-2016] .....
SMasA[RE-BP-RE183_7d-2017] ..........A.....A.....
SMasA[RE-BP-RE186_9f-2017] .....
970 980 990 1000 1010 1020
Consensus ATTAACGCGGTGGATCGTGTTCATGTGGTGGTTTTCGCTAATTTTAAGCCTCTACTCGAA
SMasA[RE-BP-RE180_a-2017] .....A.....T.....
SMasA[RE-BP-RE179_c-2017] .....
SMasA[RE-BP-RE552_16a-2016] .....
SMasA[RE-BP-RE552_31e-2016] .....
SMasA[RE-BP-RE552_32a-2016] .....
SMasA[RE-BP-RE552_31f-2016] .....
SMasA[RE-BP-RE552_32j-2016] .....
SMasA[RE-BP-RE183_7d-2017] .....T.....
SMasA[RE-BP-RE186_9f-2017] .....A.....T.....
1030 1040 1050 1060 1070 1080
Consensus GATGTTTATGATTCGAGGGGAATTGTCAGTAAGCGTCAGGCCATGTCGAAAGACAGGGTC
SMasA[RE-BP-RE180_a-2017] .....A.....
SMasA[RE-BP-RE179_c-2017] .....
SMasA[RE-BP-RE552_16a-2016] .....
SMasA[RE-BP-RE552_31e-2016] .....
SMasA[RE-BP-RE552_32a-2016] .....
SMasA[RE-BP-RE552_31f-2016] .....
SMasA[RE-BP-RE552_32j-2016] .....
SMasA[RE-BP-RE183_7d-2017] .....
SMasA[RE-BP-RE186_9f-2017] .....
1090 1100 1110 1120 1130 1140
Consensus GTTGTATTTGACCTAGATGAAGGATGTGTACGTCATAATGATGAAATAATACAGCAGTTT
SMasA[RE-BP-RE180_a-2017] ....T.....
SMasA[RE-BP-RE179_c-2017] .....
SMasA[RE-BP-RE552_16a-2016] .....
SMasA[RE-BP-RE552_31e-2016] .....
SMasA[RE-BP-RE552_32a-2016] .....
SMasA[RE-BP-RE552_31f-2016] .....
SMasA[RE-BP-RE552_32j-2016] .....
SMasA[RE-BP-RE183_7d-2017] .....T.....
SMasA[RE-BP-RE186_9f-2017] .....

```
                                 1150      1160      1170      1180      1190      1200
                                 ----:----|----:----|----:----|----:----|----:----|----:----|
Consensus                        TGAGTATGTTTCTATTGATCGTGGAGCACAGCAAGGCGAGGGATCGAGCGCTGTGCGGAA
SMasA[RE-BP-RE180_a-2017]        ...........A.A.....................................................
SMasA[RE-BP-RE179_c-2017]        ...................................................................
SMasA[RE-BP-RE552_16a-2016]      ...................................................................
SMasA[RE-BP-RE552_31e-2016]      ...................................................................
SMasA[RE-BP-RE552_32a-2016]      ...................................................................
SMasA[RE-BP-RE552_31f-2016]      ...................................................................
SMasA[RE-BP-RE552_32j-2016]      ...................................................................
SMasA[RE-BP-RE183_7d-2017]       ..............................G....................................
SMasA[RE-BP-RE186_9f-2017]       ...........A.......................................................

                                 1210      1220      1230      1240      1250      1260
                                 ----:----|----:----|----:----|----:----|----:----|----:----|
Consensus                        AGGTACGGAGTAGGGCGGAAGGGCGGAATTAGGGAAGATGGGCCATATCCGTATTCATGT
SMasA[RE-BP-RE180_a-2017]        ...........................................................G.......
SMasA[RE-BP-RE179_c-2017]        ...................................................................
SMasA[RE-BP-RE552_16a-2016]      ...................................................................
SMasA[RE-BP-RE552_31e-2016]      ...................................................................
SMasA[RE-BP-RE552_32a-2016]      ...................................................................
SMasA[RE-BP-RE552_31f-2016]      ...................................................................
SMasA[RE-BP-RE552_32j-2016]      ...................................................................
SMasA[RE-BP-RE183_7d-2017]       ...................................................................
SMasA[RE-BP-RE186_9f-2017]       ...........................................................G.......

                                 1270      1280      1290      1300      1310      1320
                                 ----:----|----:----|----:----|----:----|----:----|----:----|
Consensus                        TATCTTTAGATTACAGAGACCTACCAGGTCGCATAAAAAGCCCAACCAGGGCGCACAGAA
SMasA[RE-BP-RE180_a-2017]        ...................................................................
SMasA[RE-BP-RE179_c-2017]        ...................................................................
SMasA[RE-BP-RE552_16a-2016]      ...................................................................
SMasA[RE-BP-RE552_31e-2016]      ...................................................................
SMasA[RE-BP-RE552_32a-2016]      ...................................................................
SMasA[RE-BP-RE552_31f-2016]      ...................................................................
SMasA[RE-BP-RE552_32j-2016]      ...................................................................
SMasA[RE-BP-RE183_7d-2017]       ...............................................................T..
SMasA[RE-BP-RE186_9f-2017]       ............T......................................................

                                 1330      1340      1350      1360      1370      1380
                                 ----:----|----:----|----:----|----:----|----:----|----:----|
Consensus                        GGCCCACCAGGCCGTAACTCGAGGCCCACCAGGCCGCATAGAAGATATGGGCCTGAATTC
SMasA[RE-BP-RE180_a-2017]        ...................................................................
SMasA[RE-BP-RE179_c-2017]        ...................................................................
SMasA[RE-BP-RE552_16a-2016]      ...................................................................
SMasA[RE-BP-RE552_31e-2016]      ...................................................................
SMasA[RE-BP-RE552_32a-2016]      ...................................................................
SMasA[RE-BP-RE552_31f-2016]      ...................................................................
SMasA[RE-BP-RE552_32j-2016]      ...................................................................
SMasA[RE-BP-RE183_7d-2017]       ...................................................................
SMasA[RE-BP-RE186_9f-2017]       ...................................................................

                                 1390      1400      1410      1420      1430      1440
                                 ----:----|----:----|----:----|----:----|----:----|----:----|
Consensus                        CGGATCCGTCCCTTACTTTTTTGCAAAGGCCTGGCCCATTTGGTGCTTTTTCCCGCGCTT
SMasA[RE-BP-RE180_a-2017]        ...................................................................
SMasA[RE-BP-RE179_c-2017]        ...................................................................
SMasA[RE-BP-RE552_16a-2016]      ...................................................................
SMasA[RE-BP-RE552_31e-2016]      ...................................................................
SMasA[RE-BP-RE552_32a-2016]      ...................................................................
SMasA[RE-BP-RE552_31f-2016]      ...................................................................
SMasA[RE-BP-RE552_32j-2016]      ...................................................................
SMasA[RE-BP-RE183_7d-2017]       ..............T....................................................
SMasA[RE-BP-RE186_9f-2017]       ...................................................................
```



Supplementary Figure 1 Nucleotide sequence alignment of sorghum mastrevirus associated alphasatellite. The sequences were linearised at the nick site (↓) of the conserved nonanucleotide sequence TAGTATT↓AC (bold font) located at the stem-loop structure (highlighted in red). The polyadenylation motif AAATAA (nucleotides 1124-1129, highlighted in green) is located at the 3’ of the Rep gene (highlighted in yellow).


10 20 30 40 50 60
Consensus MASRRWMFTLFEDFPSPPFADLPESAEYLICQKEKAPTTGKIHLQGFIVLKSPRRITFLR
SMasA[RE-BP-RE180_a-2017] .....V.....
SMasA[RE-BP-RE179_c-2017]
SMasA[RE-BP-RE552_16a-2016]
SMasA[RE-BP-RE552_31e-2016]
SMasA[RE-BP-RE552_32a-2016]
SMasA[RE-BP-RE552_31f-2016]
SMasA[RE-BP-RE552_32j-2016]
SMasA[RE-BP-RE183_7d-2017]
SMasA[RE-BP-RE186_9f-2017]
70 80 90 100 110 120
Consensus KFLGKSAHLEHARSKSSSCRDYCRKDATRTDGPWEYGVFAEQGSKARKAMERYRSDPDEL
SMasA[RE-BP-RE180_a-2017]
SMasA[RE-BP-RE179_c-2017]
SMasA[RE-BP-RE552_16a-2016]
SMasA[RE-BP-RE552_31e-2016] ....V.....
SMasA[RE-BP-RE552_32a-2016] ....V.....
SMasA[RE-BP-RE552_31f-2016]
SMasA[RE-BP-RE552_32j-2016]
SMasA[RE-BP-RE183_7d-2017]
SMasA[RE-BP-RE186_9f-2017]
130 140 150 160 170 180
Consensus RLSDPQLYRRCLAESINSQFGALVLPLFTRPWQICFNERINQGPDDRNIIWVCGTQGGEG
SMasA[RE-BP-RE180_a-2017]
SMasA[RE-BP-RE179_c-2017] .....s.....
SMasA[RE-BP-RE552_16a-2016]
SMasA[RE-BP-RE552_31e-2016]
SMasA[RE-BP-RE552_32a-2016]
SMasA[RE-BP-RE552_31f-2016]
SMasA[RE-BP-RE552_32j-2016]
SMasA[RE-BP-RE183_7d-2017] .....LAEKlEPS.R.....
SMasA[RE-BP-RE186_9f-2017] .....LAEKlEPS.R.....
190 200 210 220 230 240
Consensus KTTRAKGLVKDGWFYSRGGKSVDIKYSYSMHMGHVCFDLPRQSEEILNYAVIEEIKDRLI
SMasA[RE-BP-RE180_a-2017]
SMasA[RE-BP-RE179_c-2017]
SMasA[RE-BP-RE552_16a-2016]
SMasA[RE-BP-RE552_31e-2016]
SMasA[RE-BP-RE552_32a-2016]
SMasA[RE-BP-RE552_31f-2016]
SMasA[RE-BP-RE552_32j-2016]
SMasA[RE-BP-RE183_7d-2017] .....C.....
SMasA[RE-BP-RE186_9f-2017]
250 260 270 280 290 300
Consensus RSAKYEPLDINAVDRVHVVVFANFKPLLEDVYDSRGIVSKRQAMSKDRVVVFDLDEGCVR
SMasA[RE-BP-RE180_a-2017]
SMasA[RE-BP-RE179_c-2017] .....G.....
SMasA[RE-BP-RE552_16a-2016]
SMasA[RE-BP-RE552_31e-2016]
SMasA[RE-BP-RE552_32a-2016]
SMasA[RE-BP-RE552_31f-2016]

```
SMasA[RE-BP-RE552_32j-2016].....................................................
SMasA[RE-BP-RE183_7d-2017] .....................................................C
SMasA[RE-BP-RE186_9f-2017] .....................................................



                          ----:----
Consensus                 HNDEIIQQF
SMasA[RE-BP-RE180_a-2017]  .........
SMasA[RE-BP-RE179_c-2017]  .........
SMasA[RE-BP-RE552_16a-2016].........
SMasA[RE-BP-RE552_31e-2016].........
SMasA[RE-BP-RE552_32a-2016].........
SMasA[RE-BP-RE552_31f-2016].........
SMasA[RE-BP-RE552_32j-2016].........
SMasA[RE-BP-RE183_7d-2017] .........
SMasA[RE-BP-RE186_9f-2017] .........
```

Supplementary Figure 2 Amino acid sequence of the Rep gene alignment of sorghum mastrevirus associated alphasatellite. The predicted Rep amino acid sequence contains the tyrosine residue (amino acid 82, highlighted in purple) and the NTP-binding motif GGEGKT (amino acids 177-182, highlighted in blue).

Discussion générale

Discussion générale

Ces dernières décennies, nous avons assisté à une véritable prise de conscience du manque d'informations fondamentales sur la diversité et la dynamique des virus de plantes à l'échelle des écosystèmes. En particulier, suite aux travaux de métagénomique et à la découverte d'une diversité virale insoupçonnée notamment dans les plantes sauvages (Roossinck *et al.*, 2015 ; Stobbe & Roossinck, 2014), il est devenu évident qu'il fallait appréhender la diversité virale dans sa globalité par une approche écologique intégrative (Wren *et al.*, 2006 ; Roossinck *et al.*, 2010). L'échelle des agro-écosystèmes (Alexander *et al.*, 2014 ; Shates *et al.*, 2019), s'est rapidement imposée pour ce type d'étude, du fait de la diversité de la flore associée (Burdon & Thrall, 2008), mais surtout parce que ces systèmes sont composés d'interfaces entre milieux cultivés et sauvages (Bernardo *et al.*, 2018). Ces interfaces seraient des zones de perturbations des communautés de plantes hôtes, de vecteurs et de virus pouvant aboutir à l'émergence de nouveaux variants viraux présentant de nouvelles capacités biologiques (Shates *et al.*, 2019).

Toutefois, en raison de la difficulté d'identifier l'ensemble des virus associé à un écosystème quel qu'il soit et de la complexité des interactions que ceux-ci nouent entre eux, avec leurs hôtes et leurs vecteurs, ce travail de thèse s'est focalisé sur l'étude de la diversité et l'écologie d'un genre unique de virus sur une famille unique de plante ; les mastrévirus infectant les Poaceae. Dans un premier temps, nous avons développé une approche de métagénomique virale spécifique aux molécules d'ADN circulaire de petite taille, tels que les génomes de mastrévirus, et permettant le multiplexage dense d'échantillons (**Chapitre 1**). Cette approche a été utilisée pour l'étude de la diversité et de la prévalence des mastrévirus à l'échelle des plantes individuelles d'un agro-écosystème de La Réunion. Dans un second temps, nous avons tenté d'élucider la structure et la gamme d'hôtes de ces virus et de décrire les caractéristiques de leur structuration en communautés virales (**Chapitre 2**).

Les approches métagénomiques : une solution adaptée pour l'étude des mastrévirus à l'échelle d'un agro-écosystème

Plusieurs études ont utilisées des méthodes de métagénomique afin de caractériser des virus de la famille des géminivirus ou associées à celle-ci (Bernardo *et al.*, 2018; Candresse *et al.*, 2014 ; Kraberger *et al.*, 2013 ; Kreuze *et al.*, 2009 ; Ng *et al.*, 2012 ; Ng *et al.*, 2011). Toutefois, encore peu d'études ont été menées à l'échelle des agro-écosystèmes entiers (Bernardo *et al.*, 2018 ; Muthukumar *et al.*, 2009). L'approche de métagénomique développée dans le cadre de cette thèse, en combinant l'amplification en cercle roulant (RCA) avec marquage aléatoire par PCR (RA), le séquençage à haut débit Illumina (NGS) et la classification des *reads* par placement phylogénétique (**Chapitre 1 – Article 1**), a permis de répondre à ce challenge. Notre approche a permis de travailler avec un niveau très élevé de multiplexage des échantillons, allant jusqu'à 1 200 échantillons par ligne de séquençage Illumina, bien supérieur à celui utilisé dans les précédentes études de métagénomique végétale (384 et 634 échantillons multiplexés respectivement par Roossinck *et al.* (2010) et Bernardo *et al.* (2018)). Compte tenu de cette mise en commun des échantillons, notre approche s'est révélée économique pour tester au préalable le statut d'infection d'un grand nombre de plantes dans le cas par exemple d'une surveillance épidémiologique des géminivirus. Cependant, il est important de souligner que le multiplexage se pratique au dépend de la sensibilité.

En effet, bien que la sensibilité de notre approche a semble-t-il été suffisante pour l'étude de la diversité virale, elle pourrait devenir cruciale pour la mise en œuvre d'un outil de diagnostic basée sur la métagénomique. Ce qui constitue la plus grande force de la métagénomique, à savoir la quantité massive de données obtenues sans à priori, constitue également sa plus grande faiblesse. Au-delà des limitations spécifiques à la méthode RCA-RA-NGS (discutée dans le **Chapitre 1**, à savoir la présence d'un *clamp* sur les amorces d'amplification aléatoire et la fragmentation des amplicons avant séquençage) ou plus génériques à la biologie moléculaire (biais d'extraction et d'amplification (Gallet *et al.*, 2017), produits chimériques (Lasken & Stockwell,

2007), il peut être difficile de déterminer quels échantillons sont positifs sans une analyse approfondie du nombre et de la nature des *reads* viraux associés à chaque échantillon. En effet, les études de métagénomique sont également connues pour la présence non négligeable de contaminations (Degnan & Ochman, 2012 ; Kunin *et al.*, 2008 ; Rosseel *et al.*, 2014). Si ces contaminations peuvent être dues à des virus couramment étudiés dans le laboratoire d'étude, ou aux contaminations inter-échantillons, elles peuvent aussi être associées à la nature de la technologie de séquençage employée. La mauvaise interprétation des signaux d'émission de fluorescence (Dohm *et al.*, 2008) ou encore les phénomènes dit d'*index hopping/switching* liés à la confusion de spots de séquençage ou à la présence d'amorces libres dans la librairie de séquençage, induisant lors du démultiplexage une mauvaise assignation à l'échantillon d'origine (Costello *et al.*, 2018 ; Sinha *et al.*, 2017).

Ainsi, les potentielles contaminations et erreurs d'identification de virus sur la base de peu de *reads* soulèvent clairement des questions sur le seuil de détection à considérer. Une possible stratégie seraient tout d'abord d'appliquer les mêmes précautions mises en œuvre pour les méthodes de diagnostics classiques (ELISA et (RT)-PCR) aux diagnostics utilisant les technologies NGS, à savoir l'inclusion de témoins négatifs et positifs aux différentes étapes du diagnostic ou encore la réalisation de réplicats (Kwok & Higuchi, 1989 ; Wright *et al.*, 1993). D'autre stratégies, tels que la recherche de taux de singletons élevés, de signatures de contaminations inter-échantillons ou inter-librairies peuvent être appliquées afin de limiter les faux-positifs (**Chapitre 2**). Les résultats d'études métagénomiques peuvent alors nécessiter confirmation, par détection classique (*e.g.* PCR, ELISA), séquençage Sanger, répétition de l'échantillon par la même méthode ou l'utilisation d'une technique de viromique alternative. Pour cette dernière solution, les techniques de séquençage dites de troisième génération (TGS, séquençage sans amplification de la cible au préalable, production de longues lectures) feraient de bons candidats. Malgré des taux d'erreurs brutes variant entre 10 et 30 %, des algorithmes de correction ont été développés afin d'améliorer la précision par nucléotide (après assemblage, aux alentours de 99,99 % pour Pacific Biosciences et supérieur à 99,95 % pour le séquençage par Nanopore; Lee *et al.*, 2016). De plus, contrairement aux séquençage Illumina, le nombre

de génomes viraux détectables dans l'échantillon par ces méthodes de TGS serait directement proportionnel à la charge virale (Kiselev *et al.*, 2020).

Une plateforme de TGS telle que le MinION a déjà prouvé son efficacité pour la détection et/ou le séquençage de divers virus chez l'Homme (notamment Ebola ; Hayden, 2015 ; Quick *et al.*, 2016), les animaux (le cowpox virus par exemple ; Kilianski *et al.*, 2015) et les plantes (le yam mild mosaic virus par exemple ; Filloux *et al.*, 2018). Ces nouvelles techniques de séquençage pourraient dans le cas de notre étude, améliorer le diagnostic réalisé, notamment en confirmant les échantillons identifiés comme étant positifs ainsi que clarifier les échantillons dont le statut infectieux semble douteux. De plus, la technologie MinION offre l'avantage de s'affranchir des grands centres de séquençage ou prestataires de service extérieurs, pour réaliser les études supplémentaires de confirmation au sein du laboratoire.

Enfin, un nombre important de séquences d'ADN ne présentant pas suffisamment de similarité avec les séquences nucléotidiques actuellement archivées dans les bases de données publiques est souvent mis à jour par métagénomique (Rosario & Breitbart, 2011). Ces données qualifiées de « matière noire » peuvent constituer jusqu'à 60 à 99 % des séquences récupérées. Ainsi, les échantillons qu'on qualifierait de négatif, seraient négatifs sur la base de nos connaissances actuelles mais pourraient possiblement héberger des virus encore non caractérisés. Cette «limitation» associée à la métagénomique a le bénéfice de souligner les limites de nos connaissances et de réaffirmer que la virologie représente un champ d'étude en plein bouleversement (Koonin & Dolja, 2018).

La Réunion : un véritable hot spot de diversité des African streak virus

Le terme 'African streak virus' (AfSV) regroupe les mastrévirus infectant des plantes monocotylédones et identifiés en Afrique et dans les îles du Sud-Ouest de l'océan Indien (SOOI). Parmi les 14 espèces connues avant le début de mes travaux de thèse, cinq avaient été précédemment identifiés à La Réunion (Kraberger *et al.*, 2017 ; Varsani *et al.*, 2008). Dans un premier temps, mes

travaux ont permis de détecter certains des virus déjà identifiés localement, avec le MSV (souches A et B), le MSRV, le SSRV et le SWSV mais aussi certains virus décrits dans l'océan Indien mais pas à La Réunion, à savoir le MSV-F et le PanSV. De plus, trois nouvelles espèces de mastrevirus appelées Eleusine indica associated virus (EIAV), Melinis repens associated virus (MeRAV) et Sorghum arundinaceum associated virus (SAAV ; **Chapitres 1 et 2**) ainsi qu'une nouvelle espèce d'alphasatellite associée au MSV nommée temporairement Sorghum mastrevirus-associated alphasatellite ont été caractérisées (**Chapitre 2 – Article 4**). Si nos résultats suggèrent que La Réunion représente un hot-spot de diversité des mastrévirus, ils mettent surtout en avant les limites des études précédentes, basées sur un faible nombre d'échantillons et un nombre restreint d'espèces de plantes cultivées (comme le maïs et la canne à sucre). En effet, si l'échantillonnage massif de plantes non cultivés ont permis de mettre en évidence une diversité d'espèces virales inconnue (PanSV, EIAV, MeRAV et SAAV), nos travaux représentent la première démonstration de la coexistence de cette diversité virale dans une même zone géographique de taille restreinte et au sein d'un agro-écosystème formé de plantes sauvages et cultivées.

Association des mastrévirus avec des ADN satellites

Alors que des molécules satellites d'ADNsb circulaires sont souvent associées à des membres du genre *Begomovirus* (*Geminiviridae*), seuls quelques études récentes ont décrit leur association avec des mastrévirus du Nouveau Monde (Rosario *et al.*, 2013) et du sous-continent Indien (Hamza *et al.*, 2018 ; Kumar *et al.*, 2014). Au cours de nos investigations de la diversité génétique des mastrévirus nous avons caractérisé une nouvelle espèce de géminialphasatellite sur une espèce non cultivée de Poaceae *Sorghum arundinaceum* en association avec le MSV, que nous avons provisoirement nommée Sorghum mastrevirus-associated alphasatellite (SMasA ; **Chapitre 2 – Article 4**). Cette caractérisation représente la première description d'un géminialphasatellite avec un AfSV en Afrique et souligne que les associations complexes entre des géminivirus et des molécules d'ADN satellites ne sont pas l'apanage des bégomovirus.

Les molécules d'ADN satellites des géminivirus appartiennent à deux familles les *Alphasatellitidae*, comprenant deux sous-familles les géminialphasatellites et les nanoalphasatellites, et les *Tolecusatellitidae* (anciennement betasatellites ; Hamza *et al.*, 2018). Alors que les tolécusatellites dépendent du virus assistant à la fois pour l'encapsidation, le mouvement et la réplication, les géminialphasatellites codent pour leur propre protéine associée à la réplication (Rep), leur permettant de se répliquer de façon autonome (Briddon & Stanley, 2006 ; Fiallo-Olivé *et al.*, 2012). De nombreuses études ont montré que les tolécusatellites jouent un rôle direct dans le développement et la gravité des symptômes de nombreuses maladies associées à des bégomovirus alors que les géminialphasatellites semblaient seulement impliqués dans la modulation des symptômes associés aux bégomovirus ou aux complexes bégomovirus-tolécusatellites (Briddon & Stanley, 2006 ; Patil & Fauquet, 2010). De récents travaux ont démontré expérimentalement que l'association entre le bégomovirus Euphorbia yellow mosaic virus (EuYMV) et le géminialphasatellite Euphorbia yellow mosaic alphasatellite (EuYMA) était responsable de symptômes plus sévères, mais que le géminialphasatellite impactait négativement la capacité de transmission du virus par son vecteur *B. tabaci* et potentiellement interférait dans sa capacité de dissémination au champ (Mar *et al.*, 2017).

De manière équivalente, Kumar *et al.* (2014) ont montré expérimentalement que le mastrévirus wheat dwarf India virus (WDIV) en association avec les géminialphasatellites cotton leaf curl Multan alphasatellite (CLCuMA) ou Guar leaf curl alphasatellite (GLCuA) induisaient des infections avec des symptômes plus sévères. Par ailleurs, ces auteurs ont démontré que ces géminialphasatellites étaient des suppresseurs du *RNA silencing* (ou ANR interférence - mécanisme de défense naturelle des plantes vis à vis des virus), et que le maintien de ces alphasatellites sur le terrain représentait un avantage sélectif pour le WDIV pour surmonter le mécanisme de défense de la plante hôte via *RNA silencing* (Kumar *et al.*, 2014). Dans le cadre de notre étude, malgré la caractérisation du SMasA à partir de plantes sauvages ne présentant pas de symptômes apparents de striure, la réalisation de tests biologiques en conditions contrôlées permettrait d'évaluer la contribution potentielle de ce géminialphasatellite dans les caractéristiques biologiques du MSV.

Quel historique d'introduction des mastrévirus à La Réunion ?

Grace à la disponibilité de nombreuses séquences génomiques et le développement de modèles d'inférences dédiés (Kühnert *et al.*, 2011 ; Picard *et al.*, 2017), la reconstruction de la diffusion de virus dans l'espace et le temps a permis de nombreuses avancées dans la compréhension des épidémies tels que l'identification des sources (Pande *et al.*, 2017) ou des routes privilégiés (Talbi *et al.*, 2010). Dans le cas des géminivirus, ces méthodes d'inférence ont été utilisées pour reconstruire la diffusion mondiale du TYLCV (Mabvakure *et al.*, 2016), pan-Africaine du MSV (Monjane *et al.*, 2011) ou encore les routes d'invasion des bégomovirus du manioc à Madagascar (De Bruyn *et al.*, 2016). Reconstruire les historiques d'introduction de mastrévirus à La Réunion permettrait de mieux comprendre comment et à quelle vitesse cet assemblage de virus s'est constitué. Toutefois, parmi les AfSV de La Réunion, seuls le MSV et le PanSV se prêteraient à de telles analyses, du fait du nombre et de la diversité d'origine géographique des séquences disponibles dans les bases de données. En se basant simplement sur des analyses phylogénétiques, dans le cas du MSV-B, cinq clades majeurs ont été identifiés (**Annexe II**):

- le clade 1 constitué des isolats de MSV-B3 caractérisés à La Réunion, à Maurice et au Kenya,
- le clade 2 constitué exclusivement d'isolats MSV-B1 de La Réunion,
- le clade 3 constitué exclusivement d'isolats de MSV-B1 de La Réunion partageant quatre espèces communes d'hôtes avec le clade 2,
- le clade 4 comprenant les isolats MSV-B1/2 isolés au Rwanda, Kenya, Uganda et à La Réunion,
- le clade 5 constitué des isolats MSV-B1 isolés en Afrique du Sud et au Mozambique.

Sur la base de la phylogénie, deux introductions distinctes ont été suggérées entre l'Afrique de l'Est et les îles de l'océan Indien (*cf.* flèches rouges sur l'**Annexe II**). Cette analyse suggère aussi que les isolats de MSV-B caractérisés dans cette étude, à la fois le MSV-B3 de La Réunion, mais aussi les souches recombinantes de MSV-B1, auraient pu diverger localement. Ces hypothèses ainsi que la diversité décrite pour le MSV-B sont cohérents avec

les travaux de Varsani *et al.*, (2008), qui suggéraient une forte différentiation génétique des populations locales de MSV-B. L'étude de la phylogéographie du PanSV a également permis d'identifier une structuration géographique des souches (Kraberger *et al.*, 2017 ; Varsani *et al.*, 2009), en fonction de certaines régions d'Afrique Centrale et Australe et plus précisément de certains pays comme le Kenya, Mayotte et La Réunion (**Annexe III**). Ce schéma de structuration géographique de la diversité génétique et de la dispersion vraisemblable du PanSV et de la souche B du MSV, principalement inféodés aux plantes non cultivées, est sensiblement différent de celui de la souche A du MSV adaptée au maïs. Le MSV-A s'est diffusé largement et rapidement sur le continent africain et dans les îles environnantes à partir de son centre d'origine l'Afrique Australe au cours des 150 dernières années (Harkins *et al.*, 2009 ; Varsani *et al.*, 2008, 2009). Ces différences de dispersion pourraient s'expliquer par les activités humaines notamment la diffusion de matériel végétal infecté et la propagation de vecteurs virulents et infectieux.

Globalement, l'amélioration de notre compréhension des processus de migration, des zones d'origines majeures et de la fréquence des introductions passées de mastrévirus (ou de plantes hôtes) à La Réunion nous renseignerait à la fois sur les virus susceptibles d'être introduits mais surtout nous permettrait d'analyser de manière rétrospective comment et à quelle vitesse la composition et la structure de la communauté de mastrévirus ont évolué jusqu'à maintenant.

Les stratégies d'adaptation aux hôtes : généralistes versus spécialistes

Sur les 30 espèces de plantes échantillonnées, 18 se sont révélées être des hôtes des mastrévirus avec majoritairement des espèces non cultivées (15/18). Notre étude a permis (i) d'élargir la gamme d'hôtes connues des mastrévirus à La Réunion, avec un total de 22 espèces de plantes, et (ii) de mettre en évidence des différences au niveau de l'étendue de la gamme d'hôtes de ces virus (**Chapitre 2**). Parmi les espèces de mastrévirus identifiées, seules deux (SWSV et MSRV) n'ont été retrouvées que sur un seul et unique hôte. A l'opposé, deux espèces se sont distinguées par des gammes

d’hôtes très larges à savoir le MSV-B et le SSRV avec 16 et cinq espèces d’hôtes respectives. Ces différences d’étendues de la gamme d’hôtes rendent compte de deux stratégies d’adaptation opposées, la spécialisation et le généralisme. Un virus spécialiste ne peut infecter, se multiplier efficacement et être transmis que chez des hôtes d'une ou de quelques espèces taxonomiquement apparentées alors qu’un virus généraliste aura une gamme d’hôtes plus large composée de différentes espèces, souvent de taxons non apparentés (McLeish *et al.*, 2018). Il est important de noter ici que la définition de virus généraliste ou spécialiste est somme toute relative. Pour certains groupes de virus, un virus généraliste représente un virus capable d’infecter des plantes de familles différentes. Dans un tel contexte, l’ensemble des mastrévirus seraient à considérer comme spécialiste si leurs gammes d’hôtes étaient comparées à celle du cucumber mosaic virus qui représente un des exemples emblématiques les mieux décrits de virus généraliste avec une gamme d’hôtes supérieure à 1000 espèces comprenant des plantes monocotylédones et dicotylédones (Elena *et al.*, 2009 ; Jacquemond, 2012 ; Scholthof *et al.*, 2011).

La sélection naturelle favoriserait la spécialisation plutôt que le généralisme (Futuyma & Moreno, 1988) et cela même si les virus généralistes, du fait d’une gamme d’hôtes plus large, auraient plus d’opportunité de transmission et de survie (Woolhouse *et al.*, 2001). Cette hypothèse repose sur le concept de compromis dans l'adaptation à différents hôtes (ou *adaptative trade-off*) qui indique qu’un virus ne peut pas maximiser sa valeur adaptative pour toutes les plantes disponibles au sein d’un environnement. En particulier, du fait de la compacité des génomes viraux et en conséquence de la forte épistasie (*i.e.* interaction entre deux ou plusieurs gènes), une mutation bénéfique chez un hôte peut s’avérer délétère chez un autre hôte (*i.e.* pléiotropie antagoniste, Moury & Simon, 2011 ; Remold, 2012 ; Whitlock, 1996) . Malgré cela, de nombreux virus émergents se sont souvent avérés être des virus généralistes tels que le tomato yellow leaf curl virus (Famille *Geminiviridae* ; Ioannou *et al.*, 1987), le cucumber mosaic virus (Famille Bromoviridae ; Edwardson & Christie, 1991) ou encore le chickpea chlorotic dwarf virus (Famille Geminiviridae ; Kraberger *et al.*, 2015). Ces constats rendent compte de l’influence d’autres facteurs que la sélection naturelle,

avec en particulier l'hétérogénéité environnementale et la variabilité de disponibilité d'hôte dans l'espace et le temps (McLeish *et al.*, 2018).

Notre analyse sur le réseau d'interaction entre les espèces de Poaceae et les mastrévirus caractérisés dans un agro-écosystème réunionnais suggère un fort niveau d'« emboîtement » (*nestedness*), avec notamment la gamme d'hôte du MSV-B englobant celles des autres virus. Il est théoriquement attendu qu'un réseau d'interaction hôte-pathogène puisse être à la fois emboîté et modulaire si une diversité d'habitat coexiste de manière stable (Valverde *et al.*, 2020). Dans notre cas, la présence de virus spécialistes (huit espèces virales identifiées comme telles) suggère la présence de manière stable d'espèces hôtes qui s'avèrent être principalement des espèces pérennes. En effet, les données actuelles de la littérature suggèrent que la spécialisation est observée chez les espèces virales qui bénéficient de ressources constantes et abondantes (Futuyma & Moreno, 1988). Néanmoins, l'absence de modularité significative dans notre réseau suggère une instabilité de l'écosystème en terme de structure et temporalité sur le pas de temps d'adaptation des mastrévirus. Cette instabilité pourrait être une explication à un des paradoxes de notre étude. En effet, nos travaux ont montré que les plantes cultivées présentent les taux d'infection les plus faibles, alors que d'autres études ont montré que les plantes cultivées présentent généralement des prévalences virales supérieures aux plantes non cultivées (Bernardo *et al.*, 2018). Si on ne peut écarter l'hypothèse d'une spécificité du modèle des mastrévirus des Poaceae, il est important de noter que dans notre agro-écosystème, exception faite de la canne à sucre, les Poaceae cultivées sont majoritairement des plantes saisonnières et exploitées sur de petites surfaces. Aussi, certaines cultures comme l'avoine et le millet ne sont pas des cultures traditionnelles à La Réunion. L'irrégularité de la présence de ces plantes dans l'agro-écosystème pourrait expliquer une plus faible prévalence virale et l'absence de virus spécialiste de ces cultures.

L'intrigante absence de l'emblématique MSV-A

Le mastrévirus le plus emblématique, qui a donné son nom au genre taxonomique, est le MSV, dont la souche A est régulièrement responsable

d'épidémies de striure sur la culture de maïs en Afrique (Shepherd *et al.*, 2010). De manière intéressante, malgré la détection de *reads* de MSV-A dans trois plantes non cultivées, tous nos efforts de clonage de génomes complets ont échoué. Cette difficulté à isoler et séquencer le MSV-A malgré la détection de *reads* est possiblement due à la présence des MSV-B recombinants (**Chapitres 1** et **2**) dont des portions génomiques recombinantes correspondent au MSV-A. La recombinaison a depuis longtemps été identifiée comme un moteur majeur de l'évolution des géminivirus, aboutissant à de nouvelles souches (Kraberger *et al.*, 2013) ; de nouvelles espèces (Tiendrébéogo *et al.*, 2012) mais également de nouveaux genres (Hernández-Zepeda *et al.*, 2013). Tous les MSV-B caractérisés dans cette étude (**Chapitres 1** et **2**), que ce soit les deux clades de MSV-B1 ou le clade de MSV-B3, ont été identifiés comme étant de possibles recombinants avec des ancêtres du MSV-A et MSV-F. Ces recombinants ont été isolés dans des espèces de plantes (i) pérennes, (ii) présentant les plus fortes prévalences de mastrévirus et (iii) se comportant comme des plantes « carrefours » c'est à dire des lieux de co-infections (**Chapitre 2**). La caractérisation de recombinants majoritairement à partir des plantes adventices n'est pas étonnante sachant qu'elles sont considérées comme des *mixing vessels* (ou conteneurs de mélanges de la diversité génétique virale) capables de générer de nouveaux variants viraux recombinants en raison (i) de leur proximité et de leur adaptation au milieu agricole et (ii) de leurs fréquentes infections multiples (Roshan *et al.*, 2019 ; Silva *et al.*, 2012). Les isolats MSV-B1 et MSV-B3 caractérisés présentent de fortes similitudes sur les zones de recombinaison (*i.e.* la MP, CP et les parties C-terminales de la Rep et RepA), ce qui pourrait rendre compte d'une convergence adaptative. En effet, la MP et la CP sont impliquées (i) dans le mouvement à courte et à longue distance du virus dans la plante et (ii) potentiellement dans la gravité des symptômes provoquée par le MSV-A (Van Der Walt *et al.*, 2008). L'introgression d'une MP-CP de type MSV-A dans un fond génétique de type MSV-B pourrait représenter un évènement d'adaptation.

Malgré nos travaux, l'intrigante question concernant la disparition du MSV-A à La Réunion (ou *a minima* de la forte baisse de prévalence) reste en suspens. En effet, cette souche a historiquement été caractérisée à La Réunion

(Peterschmitt *et al.*, 1996) et utilisée de façon récurrente dans les années 1990s pour la sélection en plein champ de maïs résistant (Pernet *et al.*, 1999a ; Pernet *et al.*, 1999b). Ce programme d'amélioration variétale pour la résistance du maïs vis-à-vis des principales viroses du maïs en Afrique notamment le MSV a abouti à la sélection de variétés multi-résistantes. La vulgarisation et l'utilisation de ces variétés résistantes à La Réunion a conduit à une diminution rapide de la prévalence de la MSD jusqu'à une quasi absence de description de symptômes de cette maladie sur les cultures de maïs. L'éventuelle « disparition » du MSV-A pourrait s'expliquer à la fois par l'utilisation de cultivars de maïs résistants mais aussi par la présence d'une souche de MSV-B recombinante et compétitrice. Cette hypothèse de compétition inter-souches devra toutefois être confirmée, à la fois par la validation de l'absence de génomes complets de MSV-A dans nos échantillons par les nouveaux outils de séquençage de troisième génération (MinION), ainsi que par l'évaluation de leurs traits biologiques (abondance virale, virulence, efficacité de transmission) en infections simples et mixtes. Ces dernières expérimentations de compétition permettront de comparer les valeurs adaptatives des isolats réunionnais historiques et récents de MSV-A et MSV-B et notamment des isolats recombinants de MSV-B nouvellement caractérisés.

Le destin des virus intimement lié à celui de leur insecte vecteur

La gamme d'hôte d'un phytovirus transmis par insecte vecteur ne pourra jamais être plus large que celle de ce dernier (Power & Flecker, 2003). Par conséquent, les vecteurs et leurs préférences alimentaires déterminent *in fine* l'étendue maximale du réseau d'interaction hôte-virus. La combinaison entre les préférences d'hôtes, la distribution des hôtes et le schéma de dispersion des vecteurs conditionne les possibilités de transmission interspécifique (Woolhouse *et al.*, 1997). Alors que de nombreux insectes piqueurs suceurs sont des généralistes, d'autres se sont spécialisés au cours de leur évolution sur une ou quelques espèce(s) de plantes hôtes (Kettle, 1995). En conséquence, l'identification des vecteurs de mastrévirus et de leur gamme d'hôte devrait permettre de déterminer si l'adoption d'une stratégie

généraliste ou spécialiste pour un virus est corrélée ou non à la gamme d'hôte de leur insecte vecteur.

Ainsi, vu l'importance du vecteur dans le réseau des interactions hôte-virus potentielles et en complément de l'étude de la diversité des mastrevirus à l'échelle de plantes individuelles, il nous parait essentiel de réaliser un travail similaire pour les insectes potentiellement vecteurs. Des études de métagénomiques, appelées VEM pour *vector enabled metagenomics* (ou métagénomique par l'intermédiaire du vecteur) ont déjà prouvé leur efficacité pour identifier la présence de mastrévirus dans un agro-écosystème à partir des insectes (Fontenele *et al.*, 2018 ; Rosario *et al.*, 2013). Malgré nos tentatives infructueuses (**Chapitre 2**), il nous semble nécessaire de poursuivre nos investigations. En effet, malgré la collecte et le traitement de plusieurs centaines de cicadelles et autres insectes piqueurs suceurs, nous avons détecté la présence de séquences nucléotidiques de mastrévirus dans deux insectes à savoir une cicadelle de l'espèce *Empoasca sp.* prélevée sur *Sorghum arundinaceum* et un psylle prélevé sur *Panicum sp.*. Aucun mastrevirus n'a été identifié dans les 27 individus évalués de *C. mbila*, décrit à ce jour comme le plus efficace vecteur du MSV en Afrique (Webb, 1987). L'absence de détection de mastrévirus au champ par l'intermédiaire des insectes piqueurs suceurs, alors même que les témoins *C. mbila* obtenus en conditions contrôlées se sont révélés positifs, pose la question de la fréquence des insectes virulifères *in natura* et *in fine* de la probabilité pour un insecte de croiser la route d'une plante virosée dans un agro-écosystème. Répondre à cette question essentielle nécessitera de poursuivre nos investigations.

Pour aller plus loin...

Globalement, notre étude renforce l'idée que la structure de communauté virale est dynamique et résulte d'un processus complexe dont la compréhension requiert de meilleures connaissances (i) de la transmission et des préférences alimentaires des populations de vecteurs (Goldbach & Peters, 1994; Harrison & Robinson, 1999; Power, 2008), (ii) de la disponibilité en hôte, (iii) des interactions virus-hôte, (iv) de la synergie virale et de l'éventuelle

implication des ADN satellites (Power, 2008 ; Zhou, 2013) ou encore (v) des processus évolutifs et des pressions de sélection (Seal *et al.*, 2006).
A ce titre, l'étude quantitative de la réponse de la communauté virale au changement serait primordiale et permettrait de comprendre comment cette structure de communauté a été et sera modifiée par l'arrivée de nouveaux virus et/ou plantes hôtes et/ou le changement de fréquence des virus et de leurs hôtes sous l'effet des activités humaines. Ceci est particulièrement vrai pour les îles océaniques comme La Réunion qui sont très impactées par les invasions biologiques et les activités humaines. Les données actuelles suggèrent l'existence d'un gradient altitudinal avec des régions de faible altitude très sujettes aux invasions biologiques et des régions de plus hautes altitudes mieux conservées (Tassin *et al.*, 2004 ; Whittaker, 2007). La réalisation d'une étude comparative de différents habitats (sites agricoles et sauvages) le long d'un tel gradient devrait permettre de mieux comprendre l'impact du changement et de la perturbation des écosystèmes sur la structure des communautés virales. En effet, comprendre comment l'hétérogénéité environnementale affecte la structure des réseaux hôtes-pathogènes est une condition préalable à la prévision de la dynamique et de l'émergence des maladies (Valverde *et al.*, 2020).

Ces études de terrain seraient idéalement complémentées par des études aux laboratoires afin de déterminer les *fitness* relatives de ces virus, confronter les gammes d'hôtes potentielles à celles observées en milieu naturel, mais aussi comprendre si les évolutions éventuellement constatées, telle que la recombinaison MSV-A/MSV-B, sont adaptatives ou non. La découverte d'une molécule satellite associée avec le MSV-B soulève d'ailleurs la question du rôle que ces molécules pourraient avoir sur l'étendue de la gamme d'hôte des virus associés. Les géminialphasatellites sont suspectés d'intervenir dans le *RNA silencing* et leur présence pourrait participer aux rouages du réseau d'interaction hôtes-virus. Des études plus approfondies de leur fonction *in vivo* (*fitness*, virulence, transmissibilité) pourraient être envisagées.

Enfin, les changements récents de la réglementation concernant la réduction des pesticides et en particulier l'interdiction d'utilisation du glyphosate (en principe en 2022 pour les agriculteurs), largement utilisé comme herbicide sur

les cultures de canne à sucre contre de nombreuses adventices dont les Poaceae, entraînera une plus grande promiscuité entre les Poaceae cultivées et non cultivées, modifiant les dynamiques des populations virales et favorisant éventuellement de nouvelles rencontres entre des mastrévirus spécialistes et inféodés à la canne à sucre et d'autres spécialistes et généralistes. Si les virus ont une influence sur la structure des communautés virales, les modifications de cette structure pourrait en retour impacter leur évolution. L'impact potentiellement majeur de la recombinaison, largement documentée chez les géminivirus, pourrait être ici essentielle à l'évolution de la communauté virale.

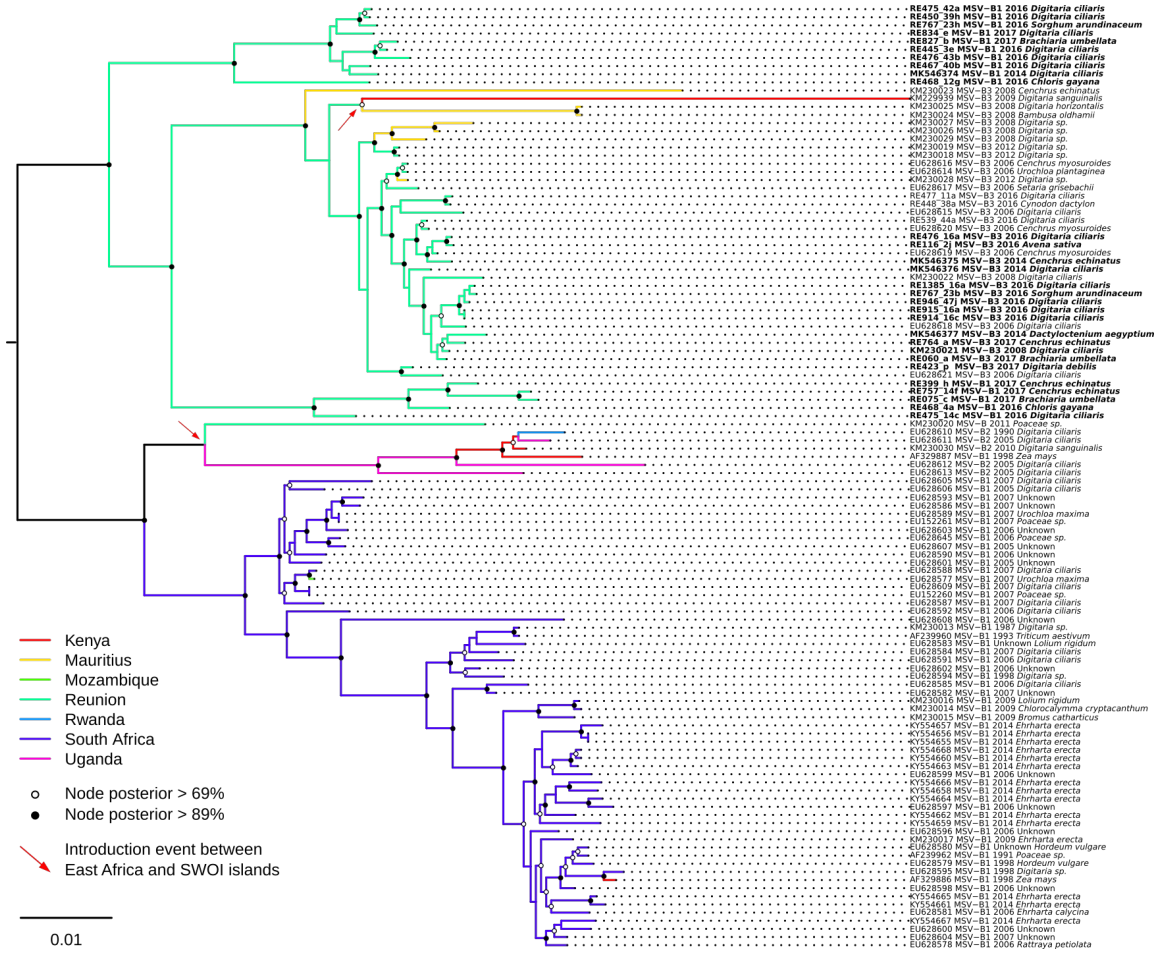
Annexes

Annexe I

Supplementary Table 1. Summary of taxonomic classification of collected insects species.

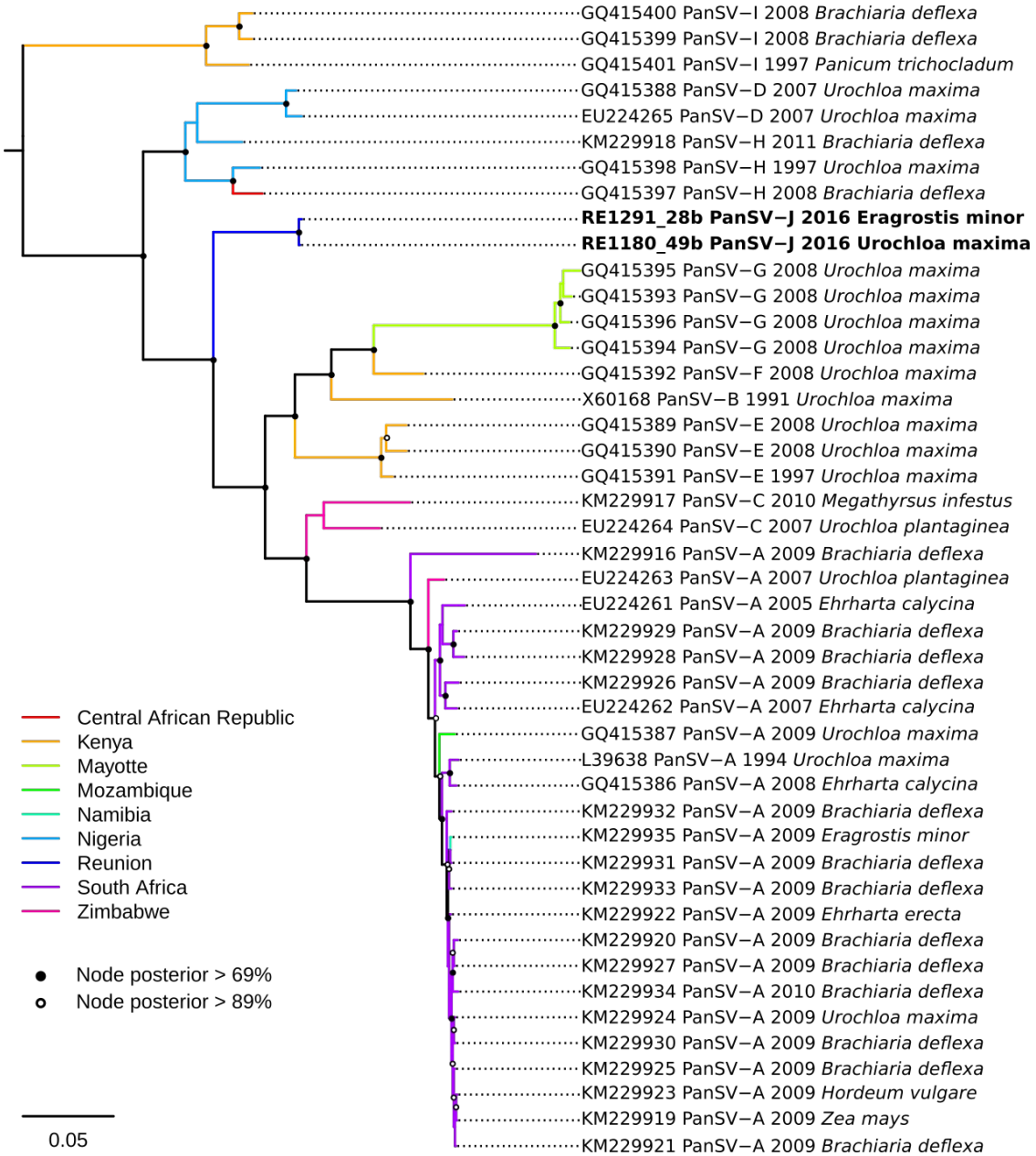
Family	Species	Number of collected samples	Total
Aphididae	*Puceron ND1*	5	28
	Sipha flava	23	
Cicadellidae	*Austroagallia caboverdensis*	28	232
	Balclutha rosea	4	
	Balclutha rufofasciata	9	
	Balclutha saltuella	8	
	Balclutha sp	2	
	Cicadelle ND2	12	
	Cicadelle ND4	4	
	Cicadulina mbila	27	
	Empoasca sp	67	
	Exitianus capicola	31	
	Exitianus frontalis	3	
	Exitianus sp	11	
	Macropsis sp	4	
	Penthimiola bella	2	
	Recilia mica	20	
Delphacidae	*Peregrinus maidis*	43	43
Membracidae	*Membracide ND1*	1	3
	Membracide ND2	2	
Miridae	*Punaise ND10*	2	48
	Punaise ND11	1	
	Punaise ND2	13	
	Punaise ND3	1	
	Punaise ND4	2	
	Punaise ND5	3	
	Punaise ND6	7	
	Punaise ND8	1	
	Trigonotylus tenuis	18	
Psyllidae	*Psylle ND1*	22	22
Tropiduchidae	*Fulgore ND2*	20	24
	Kallitaxila murcia	4	
Total		**400**	

Annexe II



Supplementary Figure 1. Maximum Likelihood phylogenetic tree of MSV-B. The maximum-likelihood phylogenetic tree contains 85 known complete genomes of MSV-B and 31 complete genomes determined in this study (indicated in bold font). The tree was rooted on MSV-A (KY618086) as an outgroup (not shown). Open and closed circles on nodes indicate bootstrap support for the branches to their left of 70-89% and >=90% respectively. Branches are coloured according to the location state of their descendant nodes. Probable introduction events between East Africa and the SWIO islands are indicated with red arrows.

Annexe III



Supplementary Figure 2. Maximum Likelihood phylogenetic tree of PanSV. The maximum-likelihood phylogenetic tree contains 43 known complete genomes of PanSV and 2 complete genomes determined in this study (indicated in bold font). The tree was rooted on MSV-A (KY618086) as an outgroup (not shown). Open and closed circles on nodes indicate bootstrap support for the branches to their left of 70-89% and >=90% respectively. Branches are coloured according to the location state of their descendant nodes.

Références

REFERENCES

Acosta-Leal, R., Duffy, S., Xiong, Z., Hammond, R. W., & Elena, S. F. (2011). Advances in Plant Virus Evolution: Translating Evolutionary Insights into Better Disease Management. *Phytopathology*, *101*(10), 1136–1148.

Adams, I. P., Miano, D. W., Kinyua, Z. M., Wangai, A., Kimani, E., Phiri, N., Reeder, R., Harju, V., Glover, R., Hany, U., Souza-Richards, R., Deb Nath, P., Nixon, T., Fox, A., Barnes, A., Smith, J., Skelton, A., Thwaites, R., Mumford, R., & Boonham, N. (2013). Use of next-generation sequencing for the identification and characterization of maize chlorotic mottle virus and sugarcane mosaic virus causing maize lethal necrosis in kenya. *Plant Pathology*, *62*(4), 741–749.

Agindotan, B. O., Domier, L. L., & Bradley, C. A. (2015). Detection and characterization of the first North American mastrevirus in switchgrass. *Archives of Virology*, *160*(5), 1313–1317.

Agnew, P., C. Koella, J., & Michalakis, Y. (2000). Host life history responses to parasitism. *Microbes and Infection*, *2*(8), 891–896.

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

Aiewsakun, P., & Katzourakis, A. (2016). Time-Dependent Rate Phenomenon in Viruses. *Journal of Virology*, *90*(16), 7184–7195.

Akhtar, K. P., Ahmad, M., Shah, T. M., & Atta, B. M. (2011). Transmission of chickpea chlorotic dwarf virus in chickpea by the leafhopper Orosius albicinctus (Distant) in Pakistan -short communication. *Plant Protection Science*, *47*(1), 1–4.

Al Rwahnih, M., Alabi, O. J., Westrick, N. M., Golino, D., & Rowhani, A. (2017). Description of a novel monopartite geminivirus and its defective subviral genome in grapevine. *Phytopathology*, *107*(2), 240–251.

Alexander, D. J. (2007). An overview of the epidemiology of avian influenza. *Vaccine*, *25*(30), 5637–5644.

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

Ali, A., Li, H., Schneider, W. L., Sherman, D. J., Gray, S., Smith, D., & Roossinck, M. J. (2006). Analysis of Genetic Bottlenecks during Horizontal Transmission of Cucumber Mosaic Virus. *Journal of Virology*, *80*(17), 8345–8350.

Almeida-Neto, M., & Ulrich, W. (2011). A straightforward computational approach for measuring nestedness using quantitative matrices. *Environmental Modelling and Software*, *26*(2), 173–178.

Alnasir, J., & Shanahan, H. P. (2015). Investigation into the annotation of protocol sequencing steps in the Sequence Read Archive. *GigaScience*, *4*(1), 1–11.

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

Astier, S., Abouy, J., Maury, Y., Robaglia, C., & Lecoq, H. (2007). Principles of Plant Virology, Genome, Pathogenicity, Virus. *Ecology. Paris: Science Publisher*.

Atanasova, N. S., Roine, E., Oren, A., Bamford, D. H., & Oksanen, H. M. (2012). Global network of specific virus-host interactions in hypersaline environments. *Environmental Microbiology*, *14*(2), 426–440.

Baltimore, D. (1971). Expression of animal virus genomes. *Bacteriological Reviews*, *35*(3), 235–241.

Bao, X., & Roossinck, M. J. (2013). A life history view of mutualistic viral symbioses: Quantity or quality for cooperation? *Current Opinion in Microbiology*, *16*(4), 514–518.

Barba, M., Czosnek, H., & Hadidi, A. (2013). Historical perspective, development and applications of next-generation sequencing in plant virology. *Viruses*, *6*(1), 106–136.

Baudin, P. (1976). Etude d’une souche du virus de la mosaique de la canne a sucre. *Agronomie Tropicale*, *32*, 180–204.

Bazinet, A. L., & Cummings, M. P. (2012). A comparative evaluation of sequence classification programs. *BMC Bioinformatics*, *13*(1), 92.

Beckett, S. J. (2016). Improved community detection in weighted bipartite networks. *Royal Society Open Science*, *3*(1), 140536.

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

Bellard, C., Rysman, J., Leroy, B., Claud, C., & Mace, G. M. (2017). A global picture of biological invasion threat on islands. *Nature Ecology & Evolution*, *1*(12)1862-1869.

Benitez-Alfonso, Y., Faulkner, C., Ritzenthaler, C., & Maule, A. J. (2010). Plasmodesmata: Gateways to local and systemic virus infection. *Molecular Plant-Microbe Interactions*, *23*(11), 1403–1412.

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

Betancourt, M., Fereres, A., Fraile, A., & Garcia-Arenal, F. (2008). Estimation of the Effective Number of Founders That Initiate an Infection after Aphid Transmission of a Multipartite Plant Virus. *Journal of Virology*, *82*(24), 12416–12421.

Bigarré, L., Salah, M., Granier, M., Frutos, R., Thouvenel, J. C., & Peterschmitt, M. (1999). Nucleotide sequence evidence for three distinct sugarcane streak mastreviruses. *Archives of Virology*, *144*(12), 2331–2344.

Blanc, S., & Michalakis, Y. (2016). Manipulation of hosts and vectors by plant viruses and impact of the environment. *Current Opinion in Insect Science*, *16*, 36–43.

Bonfils, J., Quilici, S., & Reynaud, B. (1994). Les Hémiptères Auchénorhynques de l ’ ile de la Réunion. *Bulletin de La Société Entomologique de France*, *99*(3), 227–240.

Boni, M. F., Posada, D., & Feldman, M. W. (2007). An exact nonparametric method for inferring mosaic structure in sequence triplets. *Genetics*, *176*(2), 1035–1047.

Borer, E. T., Hosseini, P. R., Seabloom, E. W., & Dobson, A. P. (2007). Pathogen-induced reversal of native dominance in a grassland community. *Proceedings of the National Academy of Sciences of the United States of America*, *104*(13), 5473–5478.

Bos, L. (1999). *Plant viruses, unique and intriguing pathogens: a textbook of plant virology*. Backhuys Publishers.

Bosque-Pérez, N. A. (2000). Eight decades of maize streak virus research. *Virus Research*, *71*(1–2), 107–121.

Boukari, W., Alcalá-briseño, R. I., Kraberger, S., Fernandez, E., Filloux, D., Daugrois, J., Comstock, J. C., Lett, J., Martin, D. P., Varsani, A., Roumagnac, P., Polston, J. E., & Rott, P. C. (2017). Occurrence of a novel mastrevirus in sugarcane germplasm collections in Florida, Guadeloupe and Réunion. *Virology journal, 14*(1), 146

Boulton, M. I. (2002). Functions and interactions of mastrevirus gene products. *Physiological and Molecular Plant Pathology*, *60*(5), 243–255.

Bousalem, M., Douzery, E. J. P., & Fargette, D. (2000). High genetic diversity, distant phylogenetic relationships and intraspecies recombination events among natural populations of Yam mosaic virus: A contribution to understanding potyvirus evolution. *Journal of General Virology*, *81*(1), 243–255.

Bragard, C., Caciagli, P., Lemaire, O., Lopez-Moya, J. J., MacFarlane, S., Peters, D., Susi, P., & Torrance, L. (2013). Status and Prospects of Plant Virus Control Through Interference with Vector Transmission. *Annual Review of Phytopathology*, *51*(1), 177–201.

Branton, D., Deamer, D. W., Marziali, A., Bayley, H., Benner, S. A., Butler, T., Ventra, M. Di, Garaj, S., Hibbs, A., Jovanovich, S. B., Krstic, P. S., Lindsay, S., Sean, X., Riehn, R., Soni, G. V, Tabard-cossa, V., & Wanunu, M. (2009). The potential and challenges of nanopore sequencing. *Nature Biotechnology*, *26*(10), 1146–1153.

Breitbart, M., Hewson, I., Felts, B., Mahaffy, J. M., Nulton, J., Salamon, P., & Rohwer, F. (2003). Metagenomic Analyses of an Uncultured Viral Community from Human Feces Metagenomic Analyses of an Uncultured Viral Community from Human Feces Downloaded from http://jb.asm.org/ on December 8 , 2013 by National Institute of Technology and Evaluation. *Journal of Bacteriology*, *185*(20), 6220–6223.

Breitbart, M., & Rohwer, F. (2005). Here a virus, there a virus, everywhere the same virus? *Trends in Microbiology*, *13*(6), 278–284.

Briddon, R. W., Lunness, P., Chamberlin, L. C. L., Pinner, M. S., Brundish, H., & Markham, P. G. (1992). The nucleotide sequence of an infectious insect-transmissible clone of the geminivirus Panicum streak virus. *Journal of General Virology*, *73*(5), 1041–1047.

Briddon, R. W., & Stanley, J. (2006). Subviral agents associated with plant single-stranded DNA viruses. *Virology*, *344*(1), 198–210.

Briddon, Rob W., Martin, D. P., Roumagnac, P., Navas-Castillo, J., Fiallo-Olivé, E., Moriones, E., Lett, J. M., Zerbini, F. M., & Varsani, A. (2018). Alphasatellitidae: a new family with two subfamilies for the classification of geminivirus- and nanovirus-associated alphasatellites. *Archives of Virology*, *163*(9), 2587–2600.

Briddon, Rob W., Patil, B. L., Bagewadi, B., Nawaz-Ul-Rehman, M. S., & Fauquet, C. M. (2010). Distinct evolutionary histories of the DNA-A and DNA-B components of bipartite begomoviruses. *BMC Evolutionary Biology*, *10*(1),97.

Brown, J. K., Zerbini, F. M., Navas-Castillo, J., Moriones, E., Ramos-Sobrinho, R., Silva, J. C. F., Fiallo-Olivé, E., Briddon, R. W., Hernández-Zepeda, C., Idris, A., Malathi, V. G., Martin, D. P., Rivera-Bustamante, R., Ueda, S., & Varsani, A. (2015). Revision of Begomovirus taxonomy based on pairwise sequence comparisons. *Archives of Virology*, *160*(6), 1593–1619.

Burdon, J. J., & Thrall, P. H. (2008). Pathogen evolution across the agro-ecological interface: implications for disease management. *Evolutionary Applications*, *1*(1), 57–65.

Caciagli, P., Medina Piles, V., Marian, D., Vecchiati, M., Masenga, V., Mason, G., Falcioni, T., & Noris, E. (2009). Virion Stability Is Important for the Circulative Transmission of Tomato Yellow Leaf Curl Sardinia Virus by Bemisia tabaci, but Virion Access to Salivary Glands Does Not Guarantee Transmissibility. *Journal of Virology*, *83*(11), 5784–5795.

Candresse, T., Filloux, D., Muhire, B., Julian, C., Galzi, S., Fort, G., Bernardo, P., Daugrois, J. H., Fernandez, E., Martin, D. P., Varsani, A., & Roumagnac, P. (2014). Appearances can be deceptive: Revealing a hidden viral infection with deep sequencing in a plant quarantine context. *PLoS ONE*, *9*(7).

Casteel, C. L., Yang, C., Nanduri, A. C., De Jong, H. N., Whitham, S. A., & Jander, G. (2014). The NIa-Pro protein of Turnip mosaic virus improves growth and reproduction of the aphid vector, Myzus persicae (green peach aphid). *Plant Journal*, *77*(4), 653–663.

Chaisson, M. J. P., Huddleston, J., Dennis, M. Y., Sudmant, P. H., Malig, M., Hormozdiari, F., Antonacci, F., Surti, U., Sandstrom, R., Boitano, M., Landolin, J. M., Stamatoyannopoulos, J. A., Hunkapiller, M. W., Korlach, J., & Eichler, E. E. (2015). Resolving the complexity of the human genome using single molecule sequencing. *National Institutes of Health*, *517*(7536), 608–611.

Charlery de la Masselière, M., Ravigné, V., Facon, B., Lefeuvre, P., Massol, F., Quilici, S., & Duyck, P. F. (2017). Changes in phytophagous insect host ranges following the invasion of their community: Long-term data for fruit flies. *Ecology and Evolution*, *7*(14), 5181–5190.

Charlesworth, B. (2009). Fundamental concepts in genetics: Effective population size and patterns of molecular evolution and variation. *Nature Reviews Genetics*, *10*(3), 195–205.

Charney, N., & Record, S. (2009). Jost Diversity Measures for Community Data. Package 'vegetarian.' *R Package 2.3-3*, *1*, Available at:cran.r-project.org/web/packages/veget.

Chatani, M., Matsumoto, Y., Mizuta, H., Ikegami, M., Boulton, M. I., & Davies, J. W. (1991). The nucleotide sequence and genome structure of the geminivirus miscanthus streak virus. *Journal of General Virology*, *72*(10), 2325–2331.

Check Hayden, E. (2015). Pint-sized DNA sequencer impresses first users. *Nature*, *521*(7550), 15–16.

Chen, S., Huang, Q., Wu, L., & Qian, Y. (2015). Identification and characterization of a maize-associated mastrevirus in China by deep sequencing small RNA populations. *Virology Journal*, *12*(1), 1–9.

Chin, C. S., Alexander, D. H., Marks, P., Klammer, A. A., Drake, J., Heiner, C., Clum, A., Copeland, A., Huddleston, J., Eichler, E. E., Turner, S. W., & Korlach, J. (2013). Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data. *Nature Methods*, *10*(6), 563–569.

Chiu, C. Y. (2013). Viral pathogen discovery. *Current Opinion in Microbiology*, *16*(4), 468–478. https://doi.org/10.1016/j.mib.2013.05.001

Chu, Y., Jeon, J., Yea, S., Kim, Y., Yun, S., Lee, Y., & Kim, K. (2002). Double-Stranded RNA Mycovirus from. *Society*, *68*(5), 2529–2534.

Claverie, S., Bernardo, P., Kraberger, S., Hartnady, P., Lefeuvre, P., Lett, J. M., Galzi, S., Filloux, D., Harkins, G. W., Varsani, A., Martin, D. P., & Roumagnac, P. (2018). From Spatial Metagenomics to Molecular Characterization of Plant Viruses: A Geminivirus Case Study. In *Advances in Virus Research, 101*, 55-83.

Claverie, S., Ouattara, A., Hoareau, M., Filloux, D., Varsani, A., Roumagnac, P., Martin, D. P., Lett, J. M., & Lefeuvre, P. (2019). Exploring the diversity of Poaceae-infecting mastreviruses on Reunion Island using a viral metagenomics-based approach. *Scientific Reports*, *9*(1), 1–11.

Coltart, C. E. M., Lindsey, B., Ghinai, I., Johnson, A. M., & Heymann, D. L. (2017). The Ebola outbreak, 2013–2016: Old lessons for new epidemics. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *372*(1721), 2013–2016.

Commins, J., Toft, C., & Fares, M. A. (2009). Computational biology methods and their application to the comparative genomics of endocellular symbiotic bacteria of insects. *Biological Procedures Online*, *11*(1), 52–78.

Cooper, I., & Jones, R. A. C. (2006). Wild Plants and Viruses: Under-Investigated Ecosystems. *Advances in Virus Research*, *67*(1), 1–47.

Costello, M., Fleharty, M., Abreu, J., Farjoun, Y., Ferriera, S., Holmes, L., Granger, B., Green, L., Howd, T., Mason, T., Vicente, G., Dasilva, M., Brodeur, W., DeSmet, T., Dodge, S., Lennon, N. J., & Gabriel, S. (2018). Characterization and remediation of sample index swaps by non-redundant dual indexing on massively parallel sequencing platforms. *BMC Genomics*, *19*(1), 1–10.

Crameri, A., Raillard, S. A., Bermudez, E., & Stemmer, W. P. C. (1998). DNA shuffling of a family of genes from diverse species accelerates directed evolution. *Nature*, *391*(6664), 288–291.

Dáder, B., Then, C., Berthelot, E., Ducousso, M., Ng, J. C. K., & Drucker, M. (2017). Insect transmission of plant viruses: Multilayered interactions optimize viral propagation. *Insect Science*, *24*(6), 929–946.

Dayaram, A., Galatowitsch, M. L., Argüello-Astorga, G. R., van Bysterveldt, K., Kraberger, S., Stainton, D., Harding, J. S., Roumagnac, P., Martin, D. P., Lefeuvre, P., & Varsani, A. (2016). Diverse circular replication-associated protein encoding viruses circulating in invertebrates within a lake ecosystem. *Infection, Genetics and Evolution*, *39*, 304–316.

De Bruyn, A., Harimalala, M., Zinga, I., Mabvakure, B. M., Hoareau, M., Ravigné, V., Walters, M., Reynaud, B., Varsani, A., Harkins, G. W., Martin, D. P., Lett, J. M., & Lefeuvre, P. (2016). Divergent evolutionary and epidemiological dynamics of cassava mosaic geminiviruses in Madagascar. *BMC Evolutionary Biology*, *16*(1).

Degnan, P. H., & Ochman, H. (2012). Illumina-based analysis of microbial community diversity. *ISME Journal*, *6*(1), 183–194.

Dekker, E. L., Woolston, C. J., Xue, Y., Cox, B., & Mullineaux, P. M. (1991). Transcript mapping reveals different expression strategies for the bicistronic RNAs of the geminivirus wheat dwarf virus. *Nucleic Acids Research*, *19*(15), 4075–4081.

Delatte, H., Reynaud, B., Granier, M., Thornary, L., Lett, J. M., Goldbach, R., & Peterschmitt, M. (2005). A new silverleaf-inducing biotype Ms of Bemisia tabaci (Hemiptera: Aleyrodidae) indigenous to the islands of the south-west Indian Ocean. *Bulletin of Entomological Research*, *95*(01), 29–35.

Delpuech, I., Bonfils, J., Leclant, F., Delpuech, I., Bonfils, J., & Leclant, F. (1986). Contribution à l'étude des virus du maïs transmis par homoptères auchénorrhynques à l ' île de la Réunion. Communication CIRAD.

Delwart, E. L. (2007). Viral metagenomics. *Reviews in Medical Virology*, *17(2)*, 115-131.

Dietrich, C., & Maiss, E. (2003). Fluorescent labelling reveals spatial separation of potyvirus populations in mixed infected Nicotiana benthamiana plants. *Journal of General Virology*, *84*(10), 2871–2876.

Dinsdale, E. A., Pantos, O., Smriga, S., Edwards, R. A., Angly, F., Wegley, L., Hatay, M., Hall, D., Brown, E., Haynes, M., Krause, L., Sala, E., Sandin, S. A., Thurber, R. V., Willis, B. L., Azam, F., Knowlton, N., & Rohwer, F. (2008). Microbial ecology of four coral atolls in the Northern Line Islands. *PLoS ONE*, *3*(2).

Dodds, J. A. (1984). Plant viral double stranded RNA. *Annual review of phytopathology*, *22*(1), 151-168.

Dohm, J. C., Lottaz, C., Borodina, T., & Himmelbauer, H. (2008). Substantial biases in ultra-short read data sets from high-throughput DNA sequencing. *Nucleic Acids Research*, *36*(16).

Dolja, V. V., & Koonin, E. V. (2011). Common origins and host-dependent diversity of plant and animal viromes. *Current Opinion in Virology*, *1*(5), 322–331.

Dollet, M., Accotto, G. P., & Lisa, V. (1986). A geminivirus, serologically related to maize streak virus, from Digitaria sanguinalis from Vanuatu. *Journal of General Virology*, *67*(5), 933–937.

Domingo-Calap, P., Cuevas, J. M., & Sanjuán, R. (2009). The fitness effects of random mutations in single-stranded DNA and RNA bacteriophages. *PLoS Genetics*, *5*(11), 1–7.

Dormann, C. F., Fründ, J., Blüthgen, N., & Gruber, B. (2009). Indices, graphs and null models: Analyzing bipartite ecological networks. *The Open Ecology Journal*, *2*, 7–24.

Dormann, C. F., & Strauss, R. (2014). A method for detecting modules in quantitative bipartite networks. *Methods in Ecology and Evolution*, *5*(1), 90–98.

Duffy, S., & Holmes, E. C. (2008). Phylogenetic Evidence for Rapid Rates of Molecular Evolution in the Single-Stranded DNA Begomovirus Tomato Yellow Leaf Curl Virus. *Journal of Virology*, *82*(2), 957–965.

Duffy, Siobain, Turner, P. E., & Burch, C. L. (2006). Pleiotropic costs of niche expansion in the RNA bacteriophage Φ6. *Genetics*, *172*(2), 751–757.

Edwardson, J. R., & Christie, R. G. (1991). Cucumoviruses. *CRC handbook of viruses infecting legumes*, 293-319.

Eid, J., Fehr, A., Gray, J., Luong, K., Lyle, J., Otto, G., Peluso, P., Rank, D., Baybayan, P., Bettman, B., Bibillo, A., Bjornson, K., Chaudhuri, B., Christians, F., Cicero, R., Clark, S., Dalal, R., DeWinter, A., Dixon, J., … Turner, S. (2009). Real-time DNA sequencing from single polymerase molecules. *Science*, *323*(5910), 133–138.

Ekzayez, A. M., Kumari, S. G., & Ismail, I. (2011). First report of Wheat dwarf virus and its vector (Psammotettix provincialis) affecting wheat and barley crops in Syria. *Plant Disease*, *95*(1), 76.

Elena, S. F., Agudelo-Romero, P., & Lalic, J. (2009). The Evolution of Viruses in Multi-Host Fitness Landscapes. *The Open Virology Journal*, *3*(1), 1–6.

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

Elena, S. F., & Lenski, R. E. (2003). Evolution experiments with microorganisms: The dynamics and genetic bases of adaptation. *Nature Reviews Genetics*, *4*(6), 457–469.

Escriu, F. (2017). Diversity of Plant Virus Populations: A Valuable Tool for Epidemiological Studies. *Genetic Diversity*, 3-18.

Faillace, C. A., Lorusso, N. S., & Duffy, S. (2017). Overlooking the smallest matter: viruses impact biological invasions. *Ecology Letters*, *20*(4), 524–538.

Fargette, D., Konaté, G., Fauquet, C., Muller, E., Peterschmitt, M., & Thresh, J. M. (2006). Molecular Ecology and Emergence of Tropical Plant Viruses. *Annual Review of Phytopathology*, *44*(1), 235–260.

Fauquet, C. M., & Stanley, J. (2003). Geminivirus classification and nomenclature: Progress and problems. *Annals of Applied Biology*, *142*(2), 165–189.

Fereres, A., & Moreno, A. (2009). Behavioural aspects influencing plant virus transmission by homopteran insects. *Virus Research*, *141*(2), 158–168.

Fiallo-Olivé, E., Martínez-Zubiaur, Y., Moriones, E., & Navas-Castillo, J. (2012). A novel class of DNA satellites associated with New World begomoviruses. *Virology*, *426*(1), 1–6.

Filloux, D., Fernandez, E., Loire, E., Claude, L., Galzi, S., Candresse, T., Winter, S., Jeeva, M. L., Makeshkumar, T., Martin, D. P., & Roumagnac, P. (2018). Nanopore-based detection and characterization of yam viruses. *Scientific Reports*, *8*(1), 1–11.

Fletcher, M. (2009). Identification keys and checklists for the leafhoppers, planthoppers and their relatives occurring in Australia and neighbouring areas (Hemiptera: Auchenorrhyncha).

Folimonova, S. Y., Robertson, C. J., Shilts, T., Folimonov, A. S., Hilf, M. E., Garnsey, S. M., & Dawson, W. O. (2010). Infection with Strains of Citrus Tristeza Virus Does Not Exclude Superinfection by Other Strains of the Virus. *Journal of Virology*, *84*(3), 1314–1325.

Fondong, V. N. (2013). Geminivirus protein structure and function. *Molecular Plant Pathology*, *14*(6), 635–649.

Fontenele, R. S., Abreu, R. A., Lamas, N. S., Alves-Freitas, D. M. T., Vidal, A. H., Poppiel, R. R., Melo, F. L., Lacorte, C., Martin, D. P., Campos, M. A., Varsani, A., & Ribeiro, S. G. (2018). Passion fruit chlorotic mottle virus: Molecular characterization of a new divergent geminivirus in Brazil. *Viruses*, *10*(4).

Fontenele, R. S., Alves-Freitas, D. M. T., Silva, P. I. T., Foresti, J., Silva, P. R., Godinho, M. T., Varsani, A., & Ribeiro, S. G. (2018). Discovery of the first maize-infecting mastrevirus in the Americas using a vector-enabled metagenomics approach. *Archives of Virology*, *163*(1), 263–267.

Fontenele, R. S., Lamas, N. S., Lacorte, C., Lacerda, A. L. M., Varsani, A., & Ribeiro, S. G. (2017). A novel geminivirus identified in tomato and cleome plants sampled in Brazil. *Virus Research*, *240*(5), 175–179.

Fontes, E. P. B., Gladfelter, H. J., Schaffer, R. L., Petty, I. T. D., & Hanley-Bowdoin, L. (1994). Geminivirus replication origins have a modular organization. *Plant Cell*, *6*(3), 405–416.

Fraile, A, Escriu, F., Aranda, M. A., Malpica, J. M., Gibbs, A. J., & García-Arenal, F. (1997). A century of tobamovirus evolution in an Australian population of Nicotiana glauca. *Journal of Virology*, *71*(11), 8316–8320.

Fraile, Aurora, & García-Arenal, F. (2010). The coevolution of plants and viruses: resistance and pathogenicity. *Advances in Virus Research*, *76*, 1–32.

Fraile, Aurora, McLeish, M. J., 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 Research*, *241*, 68–76.

French, R. K., & Holmes, E. C. (2020). An Ecosystems Perspective on Virus Evolution and Emergence. *Trends in Microbiology*, *28*(3), 165–175.

French, R., & Stenger, D. C. (2003). E VOLUTION OF W HEAT S TREAK M OSAIC V IRUS : Dynamics of Population Growth Within Plants May Explain Limited Variation . *Annual Review of Phytopathology*, *41*(1), 199–214.

Futuyma, D. J., & Moreno, G. (1988). The evolution of ecological specialization. *Annual Review of Ecology and Systematics, 19*(20), 207–233.

Gadhave, K. R., Dutta, B., Coolong, T., & Srinivasan, R. (2019). A non-persistent aphid-transmitted Potyvirus differentially alters the vector and non-vector biology through host plant quality manipulation. *Scientific Reports*, *9*(1), 1–12.

Galeano, J., Pastor, J. M., & Iriondo, J. M. (2009). Weighted-Interaction Nestedness Estimator (WINE): A new estimator to calculate over frequency matrices. *Environmental Modelling and Software*, *24*(11), 1342–1346.

Gallet, R., Fabre, F., Michalakis, Y., & Blanc, S. (2017). The Number of Target Molecules of the Amplification Step Limits Accuracy and Sensitivity in Ultradeep-Sequencing Viral Population Studies. *Journal of Virology*, *91*(16), 561-17.

García-Arenal, F., Fraile, A., & Malpica, J. M. (2001). V Ariability and G Enetic S Tructure of P Lant V Irus P Opulations . *Annual Review of Phytopathology*, *39*(1), 157–186.

García-Arenal, F., Fraile, A., & Malpica, J. M. (2003). Variation and evolution of plant virus populations. *International Microbiology*, *6*(4), 225–232.

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

Garcia-Diaz, M., & Bebenek, K. (2007). Multiple functions of DNA polymerases. *Critical Reviews in Plant Sciences*, *26*(2), 105–122.

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

Gnanasekaran, P., & Chakraborty, S. (2018). Biology of viral satellites and their role in pathogenesis. *Current Opinion in Virology*, *33*, 96–105.

Goldbach, R., & Peters, D. (1994). Possible causes of the emergence of tospovirus diseases. In *Seminars in Virology,5*(2), 113-120.

Goodwin, S., McPherson, J. D., & McCombie, W. R. (2016). Coming of age: Ten years of next-generation sequencing technologies. In *Nature Reviews Genetics, 17*(6).

Grubman, M. J., & Baxt, B. (2004). *Foot-and-Mouth Disease*. *17*(2), 465–493.

Gu, H., Zhang, C., & Ghabrial, S. A. (2007). Novel naturally occurring Bean pod mottle virus reassortants with mixed heterologous RNA1 genomes. *Phytopathology*, *97*(1), 79–86.

Guadie, D., Tesfaye, K., Knierim, D., Winter, S., & Abraham, A. (2019). Molecular analysis of maize (Zea mays L.)-infecting mastreviruses in Ethiopia reveals marked diversity of virus genomes and a novel species. *Virus Genes*, *55*(3), 339-345.

Gutierrez, C., & Gutierrez, C. (1999). Review Geminivirus DNA replication. *Genome*, *56*, 313–329.

Hadfield, J., Thomas, J. E., Schwinghamer, M. W., Kraberger, S., Stainton, D., Dayaram, A., Parry, J. N., Pande, D., Martin, D. P., & Varsani, A. (2012). Molecular characterisation of dicot-infecting mastreviruses from Australia. *Virus Research*, *166*(1–2), 13–22.

Hall, R. J., Wang, J., Todd, A. K., Bissielo, A. B., Yen, S., Strydom, H., Moore, N. E., Ren, X., Huang, Q. S., Carter, P. E., & Peacey, M. (2014). Evaluation of rapid and simple techniques for the enrichment of viruses prior to metagenomic virus discovery. *Journal of Virological Methods*, *195*, 194–204.

Hamelin, F. M., Hilker, F. M., Sun, T. A., Jeger, M. J., Hajimorad, M. R., Allen, L. J. S., & Prendeville, H. R. (2017). The evolution of parasitic and mutualistic plant–virus symbioses through transmission-virulence trade-offs. *Virus Research*, *241*, 77–87.

Hamza, M., Tahir, M. N., Mustafa, R., Kamal, H., Khan, M. Z., Mansoor, S., Briddon, R. W., & Amin, I. (2018). Identification of a dicot infecting mastrevirus along with alpha- and betasatellite associated with leaf curl disease of spinach (Spinacia oleracea) in Pakistan. *Virus Research*, *256*(8), 174–182.

Handelsman, J., Rondon, M. R., Brady, S. F., Clardy, J., & Goodman, R. M. (1998). Molecular biological access to the chemistry of unknown soil microbes: a new frontier for natural products. *Chemistry & Biology*, *5*(10), 245–249.

Hanley-Bowdoin, L., Bejarano, E. R., Robertson, D., & Mansoor, S. (2013). Geminiviruses: Masters at redirecting and reprogramming plant processes. *Nature Reviews Microbiology*, *11*(11), 777–788.

Hanley-Bowdoin, L., Settlage, S. B., Orozco, B. M., Nagar, S., & Robertson, D. (2000). Geminiviruses: Models for plant DNA replication, transcription, and cell cycle regulation. In *Critical Reviews in Biochemistry and Molecular Biology, 35*(2).

Harkins, G. W., Delport, W., Duffy, S., Wood, N., Monjane, A. L., Owor, B. E., Donaldson, L., Saumtally, S., Triton, G., Briddon, R. W., Shepherd, D. N., Rybicki, E. P., Martin, D. P., & Varsani, A. (2009). Experimental evidence indicating that mastreviruses probably did not co-diverge with their hosts. *Virology Journal*, *6*(1), 104.

Harkins, G. W., Martin, D. P., Duffy, S., Monjane, A. L., Shepherd, D. N., Windram, O. P., Owor, B. E., Donaldson, L., van Antwerpen, T., Sayed, R. A., Flett, B., Ramusi, M., Rybicki, E. P., Peterschmitt, M., & Varsani, A. (2009). Dating the origins of the maize-adapted strain of maize streak virus, MSV-A. *Journal of General Virology*, *90*(12), 3066–3074.

Harrison, B. D., & Robinson, D. J. (1999). Natural genomic and antigenic variation in whitefly transmitted geminiviruses (begomoviruses). *Annual Review of Phytopathology*, *37*(1), 369–398.

Hernández-Zepeda, C., Varsani, A., & Brown, J. K. (2013). Intergeneric recombination between a new, spinach-infecting curtovirus and a new geminivirus belonging to the genus Becurtovirus: First New World exemplar. *Archives of Virology*, *158*(11), 2245–2254.

Heyraud, F., Matzeit, V., Schaefer, S., Schell, J., & Gronenborn, B. (1993). The conserved nonanucleotide motif of the geminivirus stem-loop sequence promotes replicational release of virus molecules from redundant copies. *Biochimie*, *75*(7), 605–615.

Hogenhout, S. A., Ammar, E.-D., Whitfield, A. E., & Redinbaugh, M. G. (2008). Insect Vector Interactions with Persistently Transmitted Viruses. *Annual Review of Phytopathology*, *46*(1), 327–359.

Horn, N. M., Reddy, S. V., Roberts, I. M., & Reddy, D. V. R. (1993). Chickpea chlorotic dwarf virus, a new leafhopper-transmitted geminivirus of chickpea in India. *Annals of Applied Biology*, *122*(3), 467–479.

Hou, Y. M., Paplomatas, E. J., & Gilbertson, R. L. (1998). Host adaptation and replication properties of two bipartite geminiviruses and their pseudorecombinants. *Molecular Plant-Microbe Interactions*, *11*(3), 208–217.

Hu, J.-M., Fu, H.-C., Lin, C.-H., Su, H.-J., & Yeh, H.-H. (2007). Reassortment and Concerted Evolution in Banana Bunchy Top Virus Genomes. *Journal of Virology*, *81*(4), 1746–1761.

Huddleston, J., Ranade, S., Malig, M., Antonacci, F., Chaisson, M., Hon, L., Sudmant, P. H., Graves, T. A., Alkan, C., Dennis, M. Y., Wilson, R. K., Turner, S. W., Korlach, J., & Eichler, E. E. (2014). Reconstructing complex regions of genomes using long-read sequencing technology. *Genome Research*, *24*(4), 688–696.

Hugenholtz, P., & Tyson, G. W. (2008). Microbiology: Metagenomics. *Nature*, *455*(7212), 481–483.

Hughes, F. L., Rybicki, E. P., & Von Wechmar, M. B. (1992). Genome typing of southern African subgroup 1 geminiviruses. *Journal of General Virology*, *73*(5), 1031–1040.

Hull, R. (2014) Plant Virology, Fifth Edit. (ed. by Academic Press).

Hunter, J. A., Chamberlain, E. E., & Atkinson, J. D. (1958). Note on transmission of apple mosaic by natural root grafting. *New Zealand Journal of Agricultural Research*, *1*(1), 80–82.

Huse, S. M., Huber, J. A., Morrison, H. G., Sogin, M. L., & Welch, D. M. (2007). Accuracy and quality of massively parallel DNA pyrosequencing. *Genome Biology*, *8*(7), 1–9.

Idris, A., Al-Saleh, M., Piatek, M. J., Al-Shahwan, I., Ali, S., & Brown, J. K. (2014). Viral metagenomics: Analysis of begomoviruses by illumina high-throughput sequencing. *Viruses*, *6*(3), 1219–1236.

Ingwell, L. L., Eigenbrode, S. D., & Bosque-Pérez, N. A. (2012). Plant viruses alter insect behavior to enhance their spread. *Scientific Reports*, *2*.

Inoue-Nagata, A. K., Albuquerque, L. C., Rocha, W. B., & Nagata, T. (2004). A simple method for cloning the complete begomovirus genome using the bacteriophage φ29 DNA polymerase. *Journal of Virological Methods*, *116*(2), 209–211.

Ioannou, N., Kyriakou, A., & Hadjinicolis, A. (1987). Host range and natural reservoirs of tomato yellow leaf curl virus. *Cyprus Agricultural Research Institute Technical Bulletin.*, *1987;*(85), 1-8.

Islam, W., Zhang, J., Adnan, M., Noman, A., Zaynab, M., & Wu, Z. (2017). Plant virus ecology: A glimpse of recent accomplishments. *Applied Ecology and Environmental Research*, *15*(1), 691–705.

Jacquemond, M. (2012). Cucumber Mosaic Virus. In *Advances in Virus Research, 84,* 439-504.

Jeger, M. J., Seal, S. E., & Van den Bosch, F. (2006). Evolutionary Epidemiology of Plant Virus Disease. *Advances in Virus Research*, *67*(06), 163–203.

Jenkins, G. M., Rambaut, A., Pybus, O. G., & Holmes, E. C. (2002). Rates of molecular evolution in RNA viruses: A quantitative phylogenetic analysis. *Journal of Molecular Evolution*, *54*(2), 156–165.

Jeske, H., Lu, M., & Preiû, W. (2001). DNA forms indicate rolling circle and recombination-dependent replication of Abutilon mosaic virus. *The EMBO journal*, *20*(21), 6158-6167.

Jiang, S., Steward, G., Jellison, R., Chu, W., & Choi, S. (2004). Abundance, distribution, and diversity of viruses in alkaline, hypersaline Mono Lake, California. *Microbial Ecology*, *47*(1), 9–17.

Jones, M. S., Kapoor, A., Lukashov, V. V, Simmonds, P., Hecht, F., & Delwart, E. (2005). New DNA viruses identified in patients with acute viral infection symdrome. *Journal of Virology*, *79*(13), 8230–8236.

Jones, R. A. C. (2009). Plant virus emergence and evolution: Origins, new encounter scenarios, factors driving emergence, effects of changing world conditions, and prospects for control. *Virus Research*, *141*(2), 113–130.

Jones, R. A. C. (2018). Plant and Insect Viruses in Managed and Natural Environments: Novel and Neglected Transmission Pathways. In *Advances in Virus Research, 101*, 149-187

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

Kamali, M., Heydarnejad, J., Pouramini, N., Masumi, H., Farkas, K., Kraberger, S., & Varsani, A. (2017). Genome sequences of Beet curly top iran virus, Oat dwarf virus, Turnip curly top virus and Wheat dwarf virus identified in leafhoppers. *Genome Announcements*, *5*(8), 1674–16.

Katoh, K., & Standley, D. M. (2013). MAFFT multiple sequence alignment software version 7: Improvements in performance and usability. *Molecular Biology and Evolution*, *30*(4), 772–780.

Kazlauskas, D., Varsani, A., Koonin, E. V., & Krupovic, M. (2019). Multiple origins of prokaryotic and eukaryotic single-stranded DNA viruses from bacterial and archaeal plasmids. *Nature Communications*, *10*(1), 1–12.

Kearse, M., Moir, R., Wilson, A., Stones-Havas, S., Cheung, M., Sturrock, S., Buxton, S., Cooper, A., Markowitz, S., Duran, C., Thierer, T., Ashton, B., Meintjes, P., & Drummond, A. (2012). Geneious basic: An integrated and extendable desktop software platform for the organization and aalysis of sequence data. *Bioinformatics*, *28*(12), 1647–1649.

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

Keesing, Felicia, Belden, L. K., Daszak, P., Dobson, A., Harvell, C. D., Holt, R. D., Hudson, P., Jolles, A., Jones, K. E., Mitchell, C. E., Myers, S. S., Bogich, T., & Ostfeld, R. S. (2010). Impacts of biodiversity on the emergence and transmission of infectious diseases. *Nature*, *468*(7324), 647–652.

Kettle, D. S. (1984). *Medical and veterinary entomology*. Croom Helm Ltd.

Kilianski, A., Haas, J. L., Corriveau, E. J., Liem, A. T., Willis, K. L., Kadavy, D. R., Rosenzweig, C. N., & Minot, S. S. (2015). Bacterial and viral identification and differentiation by amplicon sequencing on the MinION nanopore sequencer. *GigaScience*, *4*(1).

Kiselev, D., Matsvay, A., Abramov, I., Dedkox, V., Shipulin, G., & Khafizov, K. (2020). Current trends in diagnostics of viral infections of unknown etiology. *Viruses*, *12*(2), 211.

Koonin, E. V., & Ilyina, T. V. (1992). Geminivirus replication proteins are related to prokaryotic plasmid rolling circle DNA replication initiator proteins. *Journal of General Virology*, *73*(10), 2763–2766.

Koonin, Eugene V., & Dolja, V. V. (2018). Metaviromics: a tectonic shift in understanding virus evolution. *Virus Research*, *246*, A1–A3.

Kraberger, S., Argüello-Astorga, G. R., Greenfield, L. G., Galilee, C., Law, D., Martin, D. P., & Varsani, A. (2015). Characterisation of a diverse range of circular replication-associated protein encoding DNA viruses recovered from a sewage treatment oxidation pond. *Infection, Genetics and Evolution*, *31*, 73–86.

Kraberger, S., Geering, A. D. W., Walters, M., Martin, D. P., & Varsani, A. (2017). Novel mastreviruses identified in Australian wild rice. *Virus Research*, *238*(6), 193–197.

Kraberger, S., Harkins, G. W., Kumari, S. G., Thomas, J. E., Schwinghamer, M. W., Sharman, M., Collings, D. A., Briddon, R. W., Martin, D. P., & Varsani, A. (2013). Evidence that dicot-infecting mastreviruses are particularly prone to inter-species recombination and have likely been circulating in Australia for longer than in Africa and the Middle East. *Virology*, *444*(1–2), 282–291.

Kraberger, S., Kumari, S. G., Hamed, A. A., Gronenborn, B., Thomas, J. E., Sharman, M., Harkins, G. W., Muhire, B. M., Martin, D. P., & Varsani, A. (2015). Infection , Genetics and Evolution Molecular diversity of Chickpea chlorotic dwarf virus in Sudan : High rates of intra-species recombination – a driving force in the emergence of new strains. *Infection, Genetics and Evolution*, *29*, 203–215.

Kraberger, S., Mumtaz, H., Claverie, S., Martin, D. P., Briddon, R. W., & Varsani, A. (2015). Identification of an Australian-like dicot-infecting mastrevirus in Pakistan. *Archives of Virology*, *160*(3),

Kraberger, S., Saumtally, S., Pande, D., Khoodoo, M. H. R., Dhayan, S., Dookun-Saumtally, A., Shepherd, D. N., Hartnady, P., Atkinson, R., Lakay, F. M., Hanson, B., Redhi, D., Monjane, A. L., Windram, O. P., Walters, M., Oluwafemi, S., Michel-Lett, J., Lefeuvre, P., Martin, D. P., & Varsani, A. (2017). Molecular diversity, geographic distribution and host range of monocot-infecting mastreviruses in Africa and surrounding islands. *Virus Research*, *238*(J6), 171–178.

Kraberger, S., Thomas, J. E., Geering, A. D. W., Dayaram, A., Stainton, D., Hadfield, J., Walters, M., Parmenter, K. S., van Brunschot, S., Collings, D. A., Martin, D. P., & Varsani, A. (2012). Australian monocot-infecting mastrevirus diversity rivals that in Africa. *Virus Research*, *169*(1), 127–136.

Kraft, F., & Kurth, I. (2019). Long-read sequencing in human genetics. *Medizinische Genetik*, *31*(2), 198–204.

Kreuze, J. F., 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), 1–7.

Kreuzer, K. N., Saunders, M., Weislo, L. J., & Kreuzer, H. W. E. (1995). Recombination-dependent DNA replication stimulated by double-strand breaks in bacteriophage T4. *Journal of Bacteriology*, *177*(23), 6844–6853.

Krishnamurthy, S. R., & Wang, D. (2017). Origins and challenges of viral dark matter. *Virus Research*, *239*, 136–142.

Krupovic, M., Ravantti, J. J., & Bamford, D. H. (2009). Geminiviruses: A tale of a plasmid becoming a virus. *BMC Evolutionary Biology*, *9*(1), 1–11.

Kühnert, D., Wu, C. H., & Drummond, A. J. (2011). Phylogenetic and epidemic modeling of rapidly evolving infectious diseases. *Infection, Genetics and Evolution*, *11*(8), 1825–1841.

Kumar, J., Kumar, J., Singh, S. P., & Tuli, R. (2014). Association of Satellites with a Mastrevirus in Natural Infection: Complexity of Wheat Dwarf India Virus Disease. *Journal of Virology*, *88*(12), 7093–7104.

Kumar, Jitendra, Singh, S. P., Kumar, J., & Tuli, R. (2012). A novel mastrevirus infecting wheat in India. *Archives of Virology*, *157*(10), 2031–2034.

Kumari, S. G., Makkouk, K. M., Attar, N., Ghulam, W., & Lesemann, D. E. (2004). First report of Chickpea chlorotic dwarf virus infecting spring chickpea in Syria. *Plant Disease*, *88*(4), 424-424.

Kunin, V., Copeland, A., Lapidus, A., Mavromatis, K., & Hugenholtz, P. (2008). A Bioinformatician's Guide to Metagenomics. *Microbiology and Molecular Biology Reviews*, *72*(4), 557–578.

Kwok, S., & Higuchi, R. (1989). Avoiding false positives with PCR. *Nature*, *339*(6221), 237–238.

Lasken, R. S., & Stockwell, T. B. (2007). Mechanism of chimera formation during the Multiple Displacement Amplification reaction. *BMC Biotechnology*, *7*, 1–11.

Laufs, J., Schumacher, S., Geisler, N., Jupin, I., & Gronenborn, B. (1995). Identification of the nicking tyrosine of geminivirus Rep protein. *FEBS Letters*, *377*(2), 258–262.

Lawry, R., Martin, D. P., Shepherd, D. N., van Antwerpen, T., & Varsani, A. (2009). A novel sugarcane-infecting mastrevirus from South Africa. *Archives of Virology*, *154*(10), 1699–1703.

Lecuit, M., & Eloit, M. (2013). The human virome: New tools and concepts. *Trends in Microbiology*, *21*(10), 510–515.

Lee, H., Gurtowski, J., & Yoo, S. (2014). Error correction and assembly complexity of single molecule sequencing reads. *BioRxiv*, 1–17.

Lee, H., Gurtowski, J., Yoo, S., Nattestad, M., Marcus, S., Goodwin, S., McCombie, W. R., & Schatz, M. (2016). Third-generation sequencing and the future of genomics. *BioRxiv*, 048603.

Lefeuvre, P., Lett, J. M., Reynaud, B., & Martin, D. P. (2007). Avoidance of protein fold disruption in natural virus recombinants. *PLoS Pathogens*, *3*(11), 1782–1789.

Lefeuvre, P., & Moriones, E. (2015). Recombination as a motor of host switches and virus emergence: Geminiviruses as case studies. *Current Opinion in Virology*, *10*, 14-19

Legg, J. P., & Thresh, J. M. (2000). Cassava mosaic virus disease in East Africa: A dynamic disease in a changing environment. *Virus Research*, *71*(1–2), 135–149.

Legg, J. Q., & Fauquet, C. M. (2004). Cassava mosaic geminiviruses. *Plant Molecular Biology*, *56*(4), 585–599.

Levins, R. (1969). Some demographic and genetic consequences of environmental heterogeneity for biological control. *American Entomologist*, *15*(3), 237–240.

Li, H., & Roossinck, M. J. (2004). Genetic Bottlenecks Reduce Population Variation in an Experimental RNA Virus Population. *Journal of Virology*, *78*(19), 10582–10587.

Li, J. X., Liu, S. S., & Gu, Q. S. (2016). Transmission Efficiency of Cucumber green mottle mosaic virus via Seeds, Soil, Pruning and Irrigation Water. *Journal of Phytopathology*, *164*(5), 300–309.

Liang, P., Navarro, B., Zhang, Z., Wang, H., Lu, M., Xiao, H., Wu, Q., Zhou, X., Di Serio, F., & Li, S. (2015). Identification and characterization of a novel geminivirus with a monopartite genome infecting apple trees. *Journal of General Virology*, *96*(8), 2411–2420.

Liu, H., Boulton, M. I., Oparka, K. J., & Davies, J. W. (2001). Interaction of the movement and coat proteins of Maize streak virus: Implications for the transport of viral DNA. *Journal of General Virology*, *82*(1), 35–44.

Liu, H., Boulton, M. I., Thomas, C. L., Prior, D. A. M., Oparka, K. J., & Davies, J. W. (1999). Maize streak virus coat protein is karyophyllic and facilitates nuclear transport of viral DNA. *Molecular Plant-Microbe Interactions*, *12*(10), 894–900.

Liu, L., Van Tonder, T., Pietersen, G., Davies, J. W., & Stanley, J. (1997). Molecular characterization of a subgroup I geminivirus from a legume in South Africa. *Journal of General Virology*, *78*(8), 2113–2117.

Liu, Li, Saunders, K., Thomas, C. L., Davies, J. W., & Stanley, J. (1999). Bean yellow dwarf virus RepA, but not Rep, binds to maize retinoblastoma protein and the virus tolerates mutations in the consensus binding motif. *Virology*, *256*(2), 270–279.

Loconsole, G., Saldarelli, P., Doddapaneni, H., Savino, V., Martelli, G. P., & Saponari, M. (2012). Identification of a single-stranded DNA virus associated with citrus chlorotic dwarf disease, a new member in the family Geminiviridae. *Virology*, *432*(1), 162–172.

Lucas, W. J. (2006). Plant viral movement proteins: Agents for cell-to-cell trafficking of viral genomes. *Virology*, *344*(1), 169–184.

Lucía-Sanz, A., & Manrubia, S. (2017). Multipartite viruses: adaptive trick or evolutionary treat? *Npj Systems Biology and Applications*, *3*(1).

Lythgoe, K. A., Gardner, A., Pybus, O. G., & Grove, J. (2017). Short-Sighted Virus Evolution and a Germline Hypothesis for Chronic Viral Infections. *Trends in Microbiology*, *25*(5), 336–348.

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. *Journal of General Virology*, *96*(8), 2421–2434.

Mabvakure, B., Martin, D. P., Kraberger, S., Cloete, L., van Brunschot, S., Geering, A. D. W., Thomas, J. E., Bananej, K., Lett, J. M., Lefeuvre, P., Varsani, A., & Harkins, G. W. (2016). Ongoing geographical spread of Tomato yellow leaf curl virus. *Virology*, *498*, 257–264.

Mahé, F., Rognes, T., Quince, C., de Vargas, C., & Dunthorn, M. (2015). Swarm v2: highly-scalable and high-resolution amplicon clustering. *PeerJ*, *3*, 1420.

Malmstrom, C. M., Hughes, C. C., Newton, L. A., & Stoner, C. J. (2005). Virus infection in remnant native bunchgrasses from invaded California grasslands. *New Phytologist*, *168*(1), 217–230.

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

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

Mar, T. B., Mendes, I. R., Lau, D., Fiallo-Olivé, E., Navas-Castillo, J., Alves, M. S., & Zerbini, F. M. (2017). Interaction between the new world begomovirus Euphorbia yellow mosaic virus and its associated alphasatellite: Effects on infection and transmission by the whitefly Bemisia tabaci. *Journal of General Virology*, *98*(6), 1552–1562.

Martin, D. P., Willment, J. A., Billharz, R., Velders, R., Odhiambo, B., Njuguna, J., James, D., & Rybicki, E. P. (2001). Sequence diversity and virulence in Zea mays of Maize streak virus isolates. *Virology*, *288*(2), 247–255.

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

Martin, D.P., Posada, D., Crandall, K. A., & Williamson, C. (2005). A Modified Bootscan Algorithm for Automated Identification of Recombinant Sequences and Recombination Breakpoints. *AIDS Research and Human Retroviruses*, *21*(1), 98–102.

Martin, Darren P., Biagini, P., Lefeuvre, P., Golden, M., Roumagnac, P., & Varsani, A. (2011). Recombination in eukaryotic single stranded DNA viruses. *Viruses*, *3*(9), 1699–1738.

Martin, Darren P., Linderme, D., Lefeuvre, P., Shepherd, D. N., & Varsani, A. (2011). Eragrostis minor streak virus: An Asian streak virus in Africa. *Archives of Virology*, *156*(7), 1299–1303.

Martin, Darren P., Murrell, B., Golden, M., Khoosal, A., & Muhire, B. (2015). RDP4: Detection and analysis of recombination patterns in virus genomes. *Virus Evolution*, *1*(1), 1–5.

Martin, Darren P., Van Walt, E. Der, Posada, D., & Rybicki, E. P. (2005). The evolutionary value of recombination is constrained by genome modularity. *PLoS Genetics*, *1*(4), 0475–0479.

Massart, S., Candresse, T., Gil, J., Lacomme, C., Predajna, L., Ravnikar, M., Reynard, J. S., Rumbou, A., Saldarelli, P., Škoric, D., Vainio, E. J., Valkonen, J. P. T., Vanderschuren, H., Varveri, C., & Wetzel, T. (2017). A framework for the evaluation of biosecurity, commercial, regulatory, and scientific impacts of plant viruses and viroids identified by NGS technologies. *Frontiers in Microbiology*, *8*(1).

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

Matsen, F. A., Kodner, R. B., & Armbrust, V. E. (2010). pplacer: linear time maximum-likelihood and Bayesian phylogenetic placement of sequences onto a fixed reference tree. *BMC Bioinformatics*, *11*, 538.

Mauck, K., Bosque-Pérez, N. A., Eigenbrode, S. D., De Moraes, C. M., & Mescher, M. C. (2012). Transmission mechanisms shape pathogen effects on host-vector interactions: Evidence from plant viruses. *Functional Ecology*, *26*(5), 1162–1175.

Mauck, K. E., De Moraes, C. M., & Mescher, M. C. (2014). Evidence of local adaptation in plant virus effects on host-vector interactions. *Integrative and Comparative Biology*, *54*(2), 193–209.

Mauck, Kerry E., De Moraes, C. M., & Mescher, M. C. (2014). Biochemical and physiological mechanisms underlying effects of Cucumber mosaic virus on host-plant traits that mediate transmission by aphid vectors. *Plant, Cell and Environment*, *37*(6), 1427–1439.

Maynard Smith, J. (1992). Analysing the Mosaic Structure of Genes. *Journal of Molecular Evolution*, *34*, 126–129.

McLeish, M. J., Fraile, A., & García-Arenal, F. (2018). Ecological Complexity in Plant Virus Host Range Evolution. *Advances in Virus Research*, *101*, 293–339.

Mehle, N., Gutiérrez-Aguirre, I., Kutnjak, D., & Ravnikar, M. (2018). Water-Mediated Transmission of Plant, Animal, and Human Viruses. *Advances in Virus Research*, *101*, 85–128.

Metzker, M. L. (2010). Sequencing technologies the next generation. *Nature Reviews Genetics*, *11*(1), 31–46.

Mink, G. I. (1993). Pollern and seed transmitted viruses and viroids. *Annual Review of Phytopathology*, *31*, 375–402.

Mitchell, C. E., & Power, A. O. (2003). Release of invasive plants from fungal and viral pathogens. *Nature*, *421*(6923), 625–627.

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

Monci, F., Sánchez-Campos, S., Navas-Castillo, J., & Moriones, E. (2002). A natural recombinant between the geminiviruses Tomato yellow leaf curl Sardinia virus and Tomato yellow leaf curl virus exhibits a novel pathogenic phenotype and is becoming prevalent in Spanish populations. *Virology*, *303*(2), 317–326.

Monjane, A. L., Harkins, G. W., Martin, D. P., Lemey, P., Lefeuvre, P., Shepherd, D. N., Oluwafemi, S., Simuyandi, M., Zinga, I., Komba, E. K., Lakoutene, D. P., Mandakombo, N., Mboukoulida, J., Semballa, S., Tagne, A., Tiendrebeogo, F., Erdmann, J. B., van Antwerpen, T., Owor, B. E., … Varsani, A. (2011). Reconstructing the History of Maize Streak Virus Strain A Dispersal To Reveal Diversification Hot Spots and Its Origin in Southern Africa. *Journal of Virology*, *85*(18), 9623–9636.

Moonan, F., Molina, J., & Mirkov, T. E. (2000). Sugarcane yellow leaf virus: An emerging virus that has evolved by recombination between luteoviral and poleroviral ancestors. *Virology*, *269*(1), 156–171.

Morales, F. J. (1987). Seed Transmission Characteristics of Selected Bean Common Mosaic Virus Strains in Differential Bean Cultivars. In *Plant Disease, 71*(1), 51.

Morris, C. E., & Moury, B. (2019). Revisiting the Concept of Host Range of Plant Pathogens. *Annual Review of Phytopathology*, *57*(1), 63–90.

Mosig, G., Gewin, J., Luder, A., Colowick, N., & Vo, D. (2001). Two recombination-dependent DNA replication pathways of bacteriophage T4, and their roles in mutagenesis and horizontal gene transfer. *Proceedings of the National Academy of Sciences of the United States of America*, *98*(15), 8306–8311.

Moury, B., Fabre, F., & Senoussi, R. (2007). Estimation of the number of virus particles transmitted by an insect vector. *Proceedings of the National Academy of Sciences of the United States of America*, *104*(45), 17891–17896.

Moury, B., Morel, C., Johansen, E., Guilbaud, L., Souche, S., Ayme, V., Caranta, C., Palloix, A., & Jacquemond, M. (2004). Mutations in Potato virus Y genome-linked protein determine virulence toward recessive resistances in Capsicum annuum and Lycopersicon hirsutum. *Molecular Plant-Microbe Interactions*, *17*(3), 322–329.

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

Moya, A., Holmes, E. C., & González-Candelas, F. (2004). The population genetics and evolutionary epidemiology of RNA viruses. *Nature Reviews Microbiology*, *2*(4), 279–288.

Muhire, B. M., Varsani, A., & Martin, D. P. (2014). SDT: A virus classification tool based on pairwise sequence alignment and identity calculation. *PLoS ONE*, *9*(9).

Muhire, B., Martin, D. P., Brown, J. K., Navas-Castillo, J., Moriones, E., Zerbini, F. M., Rivera-Bustamante, R., Malathi, V. G., Briddon, R. W., & Varsani, A. (2013). A genome-wide pairwise-identity-based proposal for the classification of viruses in the genus Mastrevirus (family Geminiviridae). *Archives of Virology*, *158*(6), 1411–1424.

Mullineaux, P. M., Guerineau, F., & Accotto, G. P. (1990). Processing of complementary sense RNAs of Digitaria streak virus in its host and in transgenic tobacco. *Nucleic Acids Research*, *18*(24), 7259–7265.

Muthukumar, V., Melcher, U., Pierce, M., Wiley, G. B., Roe, B. A., Palmer, M. W., Thapa, V., Ali, A., & Ding, T. (2009). Non-cultivated plants of the Tallgrass Prairie Preserve of northeastern Oklahoma frequently contain virus-like sequences in particulate fractions. *Virus Research*, *141*(2), 169–173.

Nahid, N., Amin, I., Mansoor, S., Rybicki, E. P., Van Der Walt, E., & Briddon, R. W. (2008). Two dicot-infecting mastreviruses (family Geminiviridae) occur in Pakistan. *Archives of Virology*, *153*(8), 1441–1451.

Nault, L. R. (1997). Arthropod transmission of plant viruses: a new synthesis. *Annals of the Entomological Society of America*, *90*(5), 521–541.

Ng, T. F. F., Marine, R., Wang, C., Simmonds, P., Kapusinszky, B., Bodhidatta, L., Oderinde, B. S., Wommack, K. E., & Delwart, E. (2012). High Variety of Known and New RNA and DNA Viruses of Diverse Origins in Untreated Sewage. *Journal of Virology*, *86*(22), 12161–12175.

Ng, Terry Fei Fan, Duffy, S., Polston, J. E., Bixby, E., Vallad, G. E., & Breitbart, M. (2011). Exploring the diversity of plant DNA viruses and their satellites using vector-enabled metagenomics on whiteflies. *PLoS ONE*, *6*(4).

Nikolin, V. M., Osterrieder, K., von Messling, V., Hofer, H., Anderson, D., Dubovi, E., Brunner, E., & East, M. L. (2012). Antagonistic Pleiotropy and Fitness Trade-Offs Reveal Specialist and Generalist Traits in Strains of Canine Distemper Virus. *PLoS ONE*, *7*(12), 1–9.

Novick, R. P. (1998). Contrasting lifestyles of rolling-circle phages and plasmids. *Trends in Biochemical Sciences*, *23*(11), 434–438.

Oksanen, J., Blanchet, F. G., Friendly, M., Kindt, R., Legendre, P., Mcglinn, D., Minchin, P. R., O'hara, R. B., Simpson, G. L., Solymos, P., Henry, M., Stevens, H., Szoecs, E., & Maintainer, H. W. (2019). Package "vegan" Title Community Ecology Package. *Community Ecology Package*, *2*(9), 1–297.

Oluwafemi, S., Alegbejo, M. D., Onasanya, A., & Olufemi, O. (2011). Relatedness of Maize streak virus in maize (Zea mays L.) to some grass isolates collected from different regions in Nigeria. *African Journal of Agricultural Research*, *6*(27), 5878–5883.

Oluwafemi, S., Kraberger, S., Shepherd, D. N., Martin, D. P., & Varsani, A. (2014). A high degree of African streak virus diversity within Nigerian maize fields includes a new mastrevirus from Axonopus compressus. *Archives of Virology*, *159*(10), 2765–2770.

Oluwafemi, S., Varsani, A., Monjane, A. L., Shepherd, D. N., Owor, B. E., Rybicki, E. P., & Martin, D. P. (2008). A new African streak virus species from Nigeria. *Archives of Virology*, *153*(7), 1407–1410.

Owor, B. E., Martin, D. P., Shepherd, D. N., Edema, R., Monjane, A. L., Rybicki, E. P., Thomson, J. A., & Varsani, A. (2007). Genetic analysis of maize streak virus isolates from Uganda reveals widespread distribution of a recombinant variant. *Journal of General Virology*, *88*(11), 3154–3165.

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

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 Pathogens*, *8*(7), 47.

Pande, D., Kraberger, S., Lefeuvre, P., Lett, J. M., Shepherd, D. N., Varsani, A., & Martin, D. P. (2012). A novel maize-infecting mastrevirus from La Réunion Island. *Archives of Virology*, *157*(8), 1617–1621.

Pande, D., Madzokere, E., Hartnady, P., Kraberger, S., Hadfield, J., Rosario, K., Jäschke, A., Monjane, A. L., Owor, B. E., Dida, M. M., Shepherd, D. N., Martin, D. P., Varsani, A., & Harkins, G. W. (2017). The role of Kenya in the trans-African spread of maize streak virus strain A. *Virus Research*, *232*, 69–76.

Paradis, E., Blomberg, S., Bolker, B., Brown, J., Claude, J., Cuong, H. S., Desper, R., Didier, G., Durand, B., Dutheil, J., Ewing, R., Gascuel, O., Guillerme, T., Heibl, C., Ives, A., Jones, B., Krah, F., Lawson, D., Lefort, V., … Vienne, D. de. (2004). *Package 'ape': analysis of phylogenetics and evolution*.

Parizipour, M. H. G., Schubert, J., Behjatnia, S. A. A., Afsharifar, A., Habekuß, A., & Wu, B. (2017). Phylogenetic analysis of Wheat dwarf virus isolates from Iran. *Virus Genes*, *53*(2), 266–274.

Patil, B. L., & Fauquet, C. M. (2010). Differential interaction between cassava mosaic geminiviruses and geminivirus satellites. *Journal of General Virology*, *91*(7), 1871–1882.

Péréfarres, F., Thébaud, G., Lefeuvre, P., Chiroleu, F., Rimbaud, L., Hoareau, M., Reynaud, B., & Lett, J. M. (2014). Frequency-dependent assistance as a way out of competitive exclusion between two strains of an emerging virus. *Proceedings of the Royal Society B: Biological Sciences*, *281*(1781).

Pernet, A., Hoisington, D., Dintinger, J., Jewell, D., Jiang, C., Khairallah, M., Letourmy, P., Marchand, J. L., Glaszmann, J. C., & González De León, D. (1999). Genetic mapping of maize streak virus resistance from the Mascarene source. II. Resistance in line CIRAD390 and stability across germplasm. *Theoretical and Applied Genetics*, *99*(3–4), 540–553.

Pernet, A., Hoisington, D., Franco, J., Isnard, M., Jewell, D., Jiang, C., Marchand, J. L., Reynaud, B., Glaszmann, J. C., & González de León, D. (1999). Genetic mapping of maize streak virus resistance from the Mascarene source. I. Resistance in line D211 and stability against different virus clones. *Theoretical and Applied Genetics*, *99*(3–4), 524–539.

Peterschmitt, M., Granier, M., Frutos, R., & Reynaud, B. (1996). Infectivity and complete nucleotide sequence of the genome of a genetically distinct strain of maize streak virus from Reunion Island. *Archives of Virology*, *141*(9), 1637–1650.

Picard, C., Dallot, S., Brunker, K., Berthier, K., Roumagnac, P., Soubeyrand, S., Jacquot, E., & Thébaud, G. (2017). Exploiting Genetic Information to Trace Plant Virus Dispersal in Landscapes. *Annual Review of Phytopathology*, *55*(1), 139–160.

Pilartz, M., & Jeske, H. (1992). Abutilon mosaic geminivirus double-stranded DNA is packed into minichromosomes. *Virology*, *189*(2), 800–802.

Pita, J. S., Fondong, V. N., Sangaré, A., Otim-Nape, G. W., Ogwal, S., & Fauquet, C. M. (2001). Recombination, pseudorecombination and synergism of geminiviruses are determinant keys to the epidemic of severe cassava mosaic disease in Uganda. *Journal of General Virology*, *82*(3), 655–665.

Pooggin, M. M. (2018). Small RNA-omics for plant virus identification, virome reconstruction, and antiviral defense characterization. *Frontiers in Microbiology*, *9*(11), 1–20.

Power, A. G. (2008). Community Ecology of Plant Viruses. In *Plant Virus Evolution*.

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

Power, A. G., & Flecker, A. S. (2003). Virus Specificity in Disease Systems: Are Species Redundant? *The Importance of Species: Perspectives on Expendability and Triage. Princeton University Press, Princeton*, 330–346.

Power, A. G., & Mitchell, C. E. (2004). Pathogen spillover in disease epidemics. *American Naturalist*, *164*(5).

Preiss, W., & Jeske, H. (2003). Multitasking in Replication Is Common among Geminiviruses. *Journal of Virology*, *77*(5), 2972–2980.

Prendeville, H. R., Ye, X., Jack Morris, T., & Pilson, D. (2012). Virus infections in wild plant populations are both frequent and often unapparent. *American Journal of Botany*, *99*(6), 1033–1042.

Price, M. N., Dehal, P. S., & Arkin, A. P. (2010). FastTree 2 - Approximately maximum-likelihood trees for large alignments. *PLoS O*, *5*(3), e9490.

Prigent, M., Leroy, M., Confalonieri, F., Dutertre, M., & DuBow, M. S. (2005). A diversity of bacteriophage forms and genomes can be isolated from the surface sands of the Sahara Desert. *Extremophiles*, *9*(4), 289–296.

Pyšek, P., Jarošík, V., Pergl, J., Randall, R., Chytrý, M., Kühn, I., Tichý, L., Danihelka, J., Chrtek Jun, J., & Sádlo, J. (2009). The global invasion success of Central European plants is related to distribution characteristics in their native range and species traits. *Diversity and Distributions*, *15*(5), 891–903.

Qiu, W., & Moyer, J. W. (1999). Tomato spotted wilt tospovirus adapts to the TSWV N gene-derived resistance by genome reassortment. *Phytopathology*, *89*(7), 575–582.

Quick, J., Loman, N. J., Duraffour, S., Simpson, J. T., Severi, E., Cowley, L., Bore, J. A., Koundouno, R., Dudas, G., Mikhail, A., Ouédraogo, N., Afrough, B., Bah, A., & Carrol, M. W. (2016). Europe PMC Funders Group Real-time , portable genome sequencing for Ebola surveillance. *Nature*, *530*(7589), 228–232.

Rastrojo, A., & Alcamí, A. (2018). Viruses in Polar Lake and Soil Ecosystems. *Advances in Virus Research*, *101*, 39–54.

Redinbaugh, M. G., & Stewart, L. R. (2018). Maize Lethal Necrosis: An Emerging, Synergistic Viral Disease. *Annual Review of Virology*, *5*(1), 301–322.

Remold, S. (2012). Understanding specialism when the jack of all trades can be the master of all. *Proceedings of the Royal Society B: Biological Sciences*, *279*(1749), 4861–4869.

Remold, S. K. (2002). Unapparent virus infection and host fitness in three weedy grass species. *Journal of Ecology*, *90*(6), 967–977.

Reynaud, B. (1988). *Transmission des virus de la striure, du stripe et de la mosaïque du maïs par leurs vecteurs Cicadulina mbila (Naude, 1924) et Peregrinus maïdis (Ashmead, 1890)(Homoptera): Approches biologique, génétique et épidémiologique de la relation vecteur-virus-plante* (Doctoral dissertation).

Richardson, D. M., Pyšek, P., Rejmánek, M., Barbour, M. G., Dane Panetta, F., & West, C. J. (2000). Naturalization and invasion of alien plants: Concepts and definitions. *Diversity and Distributions*, *6*(2), 93–107.

Roberts, C. A., Dietzgen, R. G., Heelan, L. A., & MaclLean, D. J. (2000). Real-time RT-PCR fluorescent detection of tomato spotted wilt virus. *Journal of Virological Methods*, *88*(1), 1–8.

Rodelo-Urrego, M., Pagán, I., González-Jara, P., Betancourt, M., Moreno-Letelier, A., Ayllõn, M. A., Fraile, A., Piñero, D., & García-Arenal, F. (2013). Landscape heterogeneity shapes host-parasite interactions and results in apparent plant-virus codivergence. *Molecular Ecology*, *22*(8), 2325–2340.

Rohwer, F., & Thurber, R. V. (2009). Viruses manipulate the marine environment. *Nature*, *459*(7244), 207–212.

Rojas, M. R., Hagen, C., Lucas, W. J., & Gilbertson, R. L. (2005). Exploiting Chinks in the Plant's Armor: Evolution and Emergence of Geminiviruses. *Annual Review of Phytopathology*, *43*(1), 361–394.

Roossinck, M. J. (2010). Lifestyles of plant viruses. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *365*(1548), 1899–1905.

Roossinck, M. J. (2011). The big unknown: Plant virus biodiversity. *Current Opinion in Virology*, *1*(1), 63–67.

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

Roossinck, M. J. (2012). Plant Virus Metagenomics: Biodiversity and Ecology. *Annual Review of Genetics*, *46*(1), 359–369.

Roossinck, M. J. (2013). Plant Virus Ecology. *PLoS Pathogens*, *9*(5), e1003304.

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

Roossinck, M. J., Martin, D. P., & Roumagnac, P. (2015). Plant virus metagenomics:Advances in virus discovery. *Phytopathology*, *105*(6), 716–727.

Roossinck, M. J., Saha, P., Wiley, G. B., Quan, J., White, J. D., Lai, H., Chavarría, F., Shen, G., & Roe, B. A. (2010). Ecogenomics: Using massively parallel pyrosequencing to understand virus ecology. *Molecular Ecology*, *19*(1), 81–88.

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

Rosario, K., Marr, C., Varsani, A., Kraberger, S., Stainton, D., Moriones, E., Polston, J. E., & Breitbart, M. (2016). Begomovirus-associated satellite DNA diversity captured through vector-enabled metagenomic (VEM) surveys using whiteflies (Aleyrodidae). *Viruses*, *8*(2), 1–16.

Rosario, K., Padilla-Rodriguez, M., Kraberger, S., Stainton, D., Martin, D. P., Breitbart, M., & Varsani, A. (2013). Discovery of a novel mastrevirus and alphasatellite-like circular DNA in dragonflies (Epiprocta) from Puerto Rico. *Virus Research*, *171*(1), 231–237.

Rosario, K., Seah, Y. M., Marr, C., Varsani, A., Kraberger, S., Stainton, D., Moriones, E., Polston, J. E., Duffy, S., & Breitbart, M. (2015). Vector-enabled metagenomic (VEM) surveys using whiteflies (Aleyrodidae) reveal novel begomovirus species in the new and old worlds. *Viruses*, *7*(10), 5553–5570.

Roshan, P., Kulshreshtha, A., & Hallan, V. (2019). Global Weed-Infecting Geminiviruses. *Geminiviruses*, 103–121.

Rosseel, T., Pardon, B., De Clercq, K., Ozhelvaci, O., & Van Borm, S. (2014). False-positive results in metagenomic virus discovery: A strong case for follow-up diagnosis. *Transboundary and Emerging Diseases*, *61*(4), 293–299.

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

Rúa, M. A., Pollina, E. C., Power, A. G., & Mitchell, C. E. (2011). The role of viruses in biological invasions: Friend or foe? *Current Opinion in Virology*, *1*(1), 68–72.

Sacristan, S., Diaz, M., Fraile, A., & Garcia-Arenal, F. (2011). Contact Transmission of Tobacco Mosaic Virus: a Quantitative Analysis of Parameters Relevant for Virus Evolution. *Journal of Virology*, *85*(10), 4974–4981.

Sacristan, S., Malpica, J. M., Fraile, A., & Garcia-Arenal, F. (2003). Estimation of Population Bottlenecks during Systemic Movement of Tobacco Mosaic Virus in Tobacco Plants. *Journal of Virology*, *77*(18), 9906–9911.

Sánchez-Navarro, J. A., Carmen Herranz, M., & Pallás, V. (2006). Cell-to-cell movement of Alfalfa mosaic virus can be mediated by the movement proteins of Ilar-, bromo-, cucumo-, tobamo- and comoviruses and does not require virion formation. *Virology*, *346*(1), 66–73.

Sanjuán, R., & Domingo-Calap, P. (2016). Mechanisms of viral mutation. *Cellular and Molecular Life Sciences*, *73*(23), 4433–4448.

Sanjuán, R., Moya, A., & Elena, S. F. (2004). The distribution of fitness effects caused by single-nucleotide substitutions in an RNA virus. *Proceedings of the National Academy of Sciences of the United States of America*, *101*(22), 8396–8401.

Sanjuan, R., Nebot, M. R., Chirico, N., Mansky, L. M., & Belshaw, R. (2010). Viral Mutation Rates. *Journal of Virology*, *84*(19), 9733–9748.

Saunders, K., Lucy, A., & Stanley, J. (1991). DNA forms of the geminivirus African cassava mosaic virus consistent with a rolling circle mechanism of replication. *Nucleic Acids Research*, *19*(9), 2325–2330.

Saunders, K., & Stanley, J. (1999). A nanovirus-like DNA component associated with yellow vein disease of Ageratum conyzoides: Evidence for interfamilial recombination between plant DNA viruses. *Virology*, *264*(1), 142–152.

Scholthof, K. B. G., Adkins, S., Czosnek, H., Palukaitis, P., Jacquot, E., Hohn, T., Hohn, B., Saunders, K., Candresse, T., Ahlquist, P., Hemenway, C., & Foster, G. D. (2011). Top 10 plant viruses in molecular plant pathology. *Molecular Plant Pathology*, *12*(9), 938–954.

Scholz, M. B., Lo, C. C., & Chain, P. S. G. (2012). Next generation sequencing and bioinformatic bottlenecks: The current state of metagenomic data analysis. *Current Opinion in Biotechnology*, *23*(1), 9–15.

Schubert, J., Habekuß, A., Kazmaier, K., & Jeske, H. (2007). Surveying cereal-infecting geminiviruses in Germany-Diagnostics and direct sequencing using rolling circle amplification. *Virus Research*, *127*(1), 61–70.

Schubert, J., Habekuß, A., Wu, B., Thieme, T., & Wang, X. (2014). Analysis of complete genomes of isolates of the Wheat dwarf virus from new geographical locations and descriptions of their defective forms. *Virus Genes*, *48*(1), 133–139.

Seal, S. E., VandenBosch, F., & Jeger, M. J. (2006). Factors influencing begomovirus evolution and their increasing global significance: Implications for sustainable control. *Critical Reviews in Plant Sciences*, *25*(1), 23–46.

Shates, T. M., Sun, P., Malmstrom, C. M., Dominguez, C., & Mauck, K. E. (2019). Addressing research needs in the field of plant virus ecology by defining knowledge gaps and developing wild dicot study systems. *Frontiers in Microbiology*, *10*(1), 1–20.

Shepherd, D. N., Martin, D. P., Van Der Walt, E., Dent, K., Varsani, A., & Rybicki, E. P. (2010). Maize streak virus: An old and complex “emerging” pathogen. *Molecular Plant Pathology*, *11*(1), 1–12.

Shepherd, D. N., Varsani, A., Windram, O. P., Lefeuvre, P., Monjane, A. L., Owor, B. E., & Martin, D. P. (2008). Novel sugarcane streak and sugarcane streak Reunion mastreviruses from southern Africa and la Réunion. *Archives of Virology*, *153*(3), 605–609.

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

Shi, M., Zhang, Y. Z., & Holmes, E. C. (2018). Meta-transcriptomics and the evolutionary biology of RNA viruses. *Virus Research*, *243*(10), 83–90.

Sicard, A., Michalakis, Y., Gutiérrez, S., & Blanc, S. (2016). The Strange Lifestyle of Multipartite Viruses. *PLoS Pathogens*, *12*(11), 1–19.

Silva, S. J. C., Castillo-Urquiza, G. P., Hora-Júnior, B. T., Assunção, I. P., Lima, G. S. A., Pio-Ribeiro, G., Mizubuti, E. S. G., & Zerbini, F. M. (2012). Species diversity, phylogeny and genetic variability of begomovirus populations infecting leguminous weeds in northeastern Brazil. *Plant Pathology*, *61*(3), 457–467.

Simmonds, P., Adams, M. J., Benk, M., Breitbart, M., Brister, J. R., Carstens, E. B., Davison, A. J., Delwart, E., Gorbalenya, A. E., Harrach, B., Hull, R., King, A. M. Q., Koonin, E. V., Krupovic, M., Kuhn, J. H., Lefkowitz, E. J., Nibert, M. L., Orton, R., Roossinck, M. J., … Zerbini, F. M. (2017). Consensus statement: Virus taxonomy in the age of metagenomics. *Nature Reviews Microbiology*, *15*(3), 161–168.

Simmonds, P., Aiewsakun, P., & Katzourakis, A. (2019). Prisoners of war — host adaptation and its constraints on virus evolution. *Nature Reviews Microbiology*, *17*(5), 321–328.

Sinha, R., Stanley, G., Gulati, G. S., Ezran, C., Travaglini, K. J., Wei, E., Chan, C. K. F., Nabhan, A. N., Su, T., Morganti, R. M., Conley, S. D., Chaib, H., Red-Horse, K., Longaker, M. T., Snyder, M. P., Krasnow, M. A., & Weissman, I. L. (2017). Index Switching Causes “Spreading-Of-Signal” Among Multiplexed Samples In Illumina HiSeq 4000 DNA Sequencing. *BioRxiv*, 125724.

Smith, O., Clapham, A., Rose, P., Liu, Y., Wang, J., & Allaby, R. G. (2014). A complete ancient RNA genome: Identification, reconstruction and evolutionary history of archaeological Barley Stripe Mosaic Virus. *Scientific Reports*, *4*, 1–6.

Stemmer, C. (1994). DNA shuffling by random fragmentation and. *Genetics*, *91*(10), 10747–10751.

Stobbe, A. H., & Roossinck, M. J. (2014). Plant virus metagenomics: what we know and why we need to know more. *Frontiers in Plant Science*, *5*(4), 1–4.

Storey, H. H. (1928). Transmission Studies of Maize Streak Disease. *Annals of Applied Biology*, *15*(1), 1–25.

Storey, H. H., & McClean, A. P. D. (1930). The transmission of streak disease between maize, sugar cane and wild grasses. *Annals of Applied Biology*, *17*(4).

Sugiura, S. (2010). Species interactions-area relationships: Biological invasions and network structure in relation to island area. *Proceedings of the Royal Society B: Biological Sciences*, *277*(1689), 1807–1815.

Suttle, C. A. (2005). Viruses in the sea. *Nature*, *437*(7057), 356–361.

Suttle, C. A. (2007). Marine viruses - Major players in the global ecosystem. *Nature Reviews Microbiology*, *5*(10), 801–812.

Syller, J. (2012). Facilitative and antagonistic interactions between plant viruses in mixed infections. *Molecular Plant Pathology*, *13*(2), 204–216.

Talbi, C., Lemey, P., Suchard, M. A., Abdelatif, E., Elharrak, M., Jalal, N., Faouzi, A., Echevarría, J. E., Morón, S. V., Rambaut, A., Campiz, N., Tatem, A. J., Holmes, E. C., & Bourhy, H. (2010). Phylodynamics and Human-mediated dispersal of a zoonotic virus. *PLoS Pathogens*, *6*(10).

Tamada, T., & Kondo, H. (2013). Biological and genetic diversity of plasmodiophorid-transmitted viruses and their vectors. *Journal of General Plant Pathology*, *79*(5), 307–320.

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

Tang, M., Tan, M., Meng, G., Yang, S., Su, X., Liu, S., Song, W., Li, Y., Wu, Q., Zhang, A., & Zhou, X. (2014). Multiplex sequencing of pooled mitochondrial genomes - A crucial step toward biodiversity analysis using mito-metagenomics. *Nucleic Acids Research*, *42*(22), 1–13.

Tassin, J., Derroire, G., & Rivière, J. N. (2004). Gradient altitudinal de la richesse spécifique et de l'endémicité de la flore ligneuse indigène à l'île de La Réunion (archipel des Mascareignes). *Acta Botanica Gallica*, *151*(2), 181–196.

Thomas, J. E., Parry, J. N., Schwinghamer, M. W., & Dann, E. K. (2010). Two novel mastreviruses from chickpea (Cicer arietinum) in Australia. *Archives of Virology*, *155*(11), 1777–1788.

Thottapilly, G., Bosque-Perez, N. A., & Rossel, H. W. (1993). Viruses and virus diseases of maize in tropical Africa. *Plant Pathology*, *42*(4), 494–509.

Tiendrébéogo, F., Lefeuvre, P., Hoareau, M., Harimalala, M. A., De Bruyn, A., Villemot, J., Traoré, V. S. E., Konaté, G., Traoré, A. S., Barro, N., Reynaud, B., Traoré, O., & Lett, J. M. (2012). Evolution of African cassava mosaic virus by recombination between bipartite and monopartite begomoviruses. *Virology Journal*, *9*, 1–7.

Traveset, A., Olesen, J. M., Nogales, M., Vargas, P., Jaramillo, P., Antolín, E., Trigo, M. M., & Heleno, R. (2015). Bird-flower visitation networks in the Galápagos unveil a widespread interaction release. *Nature Communications*, *6*(6).

Tṛebicki, P., Harding, R. M., Rodoni, B., Baxter, G., & Powell, K. S. (2010). Vectors and alternative hosts of Tobacco yellow dwarf virus in southeastern Australia. *Annals of Applied Biology*, *157*(1), 13–24.

Uzest, M., Gargani, D., Drucker, M., Hébrard, E., Garzo, E., Candresse, T., Fereres, A., & Blanc, S. (2007). A protein key to plant virus transmission at the tip of the insect vector stylet. *Proceedings of the National Academy of Sciences of the United States of America*, *104*(46), 17959–17964.

Vaghi Medina, C. G., Teppa, E., Bornancini, V. A., Flores, C. R., Marino-Buslje, C., & Lambertini, P. M. L. (2018). Tomato apical leaf curl virus: A novel, monopartite geminivirus detected in tomatoes in Argentina. *Frontiers in Microbiology*, *8*(1), 1–11.

Valverde, S., Vidiella, B., Montañez, R., Fraile, A., Sacristán, S., & García-Arenal, F. (2020). Coexistence of nestedness and modularity in host–pathogen infection networks. *Nature Ecology and Evolution*.

Van Der Walt, E., Palmer, K. E., Martin, D. P., & Rybicki, E. P. (2008). Viable chimaeric viruses confirm the biological importance of sequence specific maize streak virus movement protein and coat protein interactions. *Virology Journal*, *5*, 1–11.

Van Dijk, E. L., Auger, H., Jaszyszyn, Y., & Thermes, C. (2014). Ten years of next-generation sequencing technology. *Trends in Genetics*, *30*(9), 418–426.

Varma, A., & Malathi, V. G. (2003). Emerging geminivirus problems: A serious threat to crop production. *Annals of Applied Biology*, *142*(2), 145–164.

Varsani, A., Oluwafemi, S., Windram, O. P., Shepherd, D. N., Monjane, A. L., Owor, B. E., Rybicki, E. P., Lefeuvre, P., & Martin, D. P. (2008). Panicum streak virus diversity is similar to that observed for maize streak virus. *Archives of Virology*, *153*(3), 601–604.

Varsani, Arvind, Lefeuvre, P., Roumagnac, P., & Martin, D. (2018). Notes on recombination and reassortment in multipartite/segmented viruses. *Current Opinion in Virology*, *33*(9), 156–166.

Varsani, Arvind, Monjane, A. L., Donaldson, L., Oluwafemi, S., Zinga, I., Komba, E. K., Plakoutene, D., Mandakombo, N., Mboukoulida, J., Semballa, S., Briddon, R. W., Markham, P. G., Lett, J. M., Lefeuvre, P., Rybicki, E. P., & Martin, D. P. (2009). Comparative analysis of panicum streak virus and maize streak virus diversity, recombination patterns and phylogeography. *Virology Journal*, *6*, 1–11.

Varsani, Arvind, Navas-Castillo, J., Moriones, E., Hernández-Zepeda, C., Idris, A., Brown, J. K., Murilo Zerbini, F., & Martin, D. P. (2014). Establishment of three new genera in the family Geminiviridae: Becurtovirus, Eragrovirus and Turncurtovirus. *Archives of Virology*, *159*(8), 2193–2203.

Varsani, Arvind, Roumagnac, P., Fuchs, M., Navas-Castillo, J., Moriones, E., Idris, A., Briddon, R. W., Rivera-Bustamante, R., Murilo Zerbini, F., & Martin, D. P. (2017). Capulavirus and Grablovirus: two new genera in the family Geminiviridae. *Archives of Virology*, *162*(6), 1819–1831.

Varsani, Arvind, Shepherd, D. N., Monjane, A. L., Owor, B. E., Erdmann, J. B., Rybicki, E. P., Peterschmitt, M., Briddon, R. W., Markham, P. G., Oluwafemi, S., Windram, O. P., Lefeuvre, P., Lett, J. M., & Martin, D. P. (2008). Recombination, decreased host specificity and increased mobility may have driven the emergence of maize streak virus as an agricultural pathogen. *Journal of General Virology*, *89*(9), 2063–2074.

Vayssier-Taussat, M., Albina, E., Citti, C., Cosson, J.-F., Jacques, M.-A., Lebrun, M.-H., Le Loir, Y., Ogliastro, M., Petit, M.-A., Roumagnac, P., & Candresse, T. (2014). Shifting the paradigm from pathogens to pathobiome: new concepts in the light of meta-omics. *Frontiers in Cellular and Infection Microbiology*, *4*(3), 1–7.

Vigne, E., Marmonier, A., & Fuchs, M. (2008). Multiple interspecies recombination events within RNA2 of Grapevine fanleaf virus and Arabis mosaic virus. *Archives of Virology*, *153*(9), 1771–1776.

Wang, Y., Mao, Q., Liu, W., Mar, T., Wei, T., Liu, Y., & Wang, X. (2014). Localization and distribution of wheat dwarf virus in its vector leafhopper, psammotettix alienus. *Phytopathology*, *104*(8), 897–904.

Webb, B. A., Strand, M. R., Dickey, S. E., Beck, M. H., Hilgarth, R. S., Barney, W. E., Kadash, K., Kroemer, J. A., Lindstrom, K. G., Rattanadechakul, W., Shelby, K. S., Thoetkiattikul, H., Turnbull, M. W., & Witherell, R. A. (2006). Polydnavirus genomes reflect their dual roles as mutualists and pathogens. *Virology*, *347*(1), 160–174.

Webb, M. D. (1987). Species recognition in Cicadulina leafhoppers (Hemiptera: Cicadellidae), vectors of pathogens of Gramineae. *Bulletin of Entomological Research*, *77*(4), 683–712.

Weinbauer, M. G., & Rassoulzadegan, F. (2004). Are viruses driving microbial diversification and diversity? *Environmental Microbiology*, *6*(1), 1–11.

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

White, P. S., Morales, F., & Roossinck, M. J. (1995). Interspecific Reassortment of Genomic Segments in the Evolution of Cucumoviruses. In *Virology, 207*(1), 334–337.

Whitlock, M. C. (1996). The red queen beats the jack-of-all-trades: The limitations on the evolution of phenotypic plasticity and niche breadth. *American Naturalist*, *148*, 65–77.

Whittaker, R. J., & Fernández-Palacios, J. M. (2007). *Island biogeography: ecology, evolution, and conservation*. Oxford University Press.

Wilson, D. S., & Yoshimura, J. (1994). On the coexistence of specialists and generalists. *The American Naturalist*, *144*(4), 692–707.

Woolhouse, M.E.J., Dye, C., Etard, J. F., Smith, T., Charlwood, J. D., Garnett, G. P., Hagan, P., Hii, J. L. K., Ndhlovu, P. D., Quinnell, R. J., Watts, C. H., Chandiwana, S. K., & Anderson, R. M. (1997). Heterogeneities in the transmission of infectious agents. *Proceedings of the National Academy of Sciences of the United States of America*, *94*(1), 338–342.

Woolhouse, Mark E.J., Taylor, L. H., & Haydon, D. T. (2001). Population biology of multihost pathogens. *Science*, *292*(5519), 1109–1112.

Woolhouse, Mark E.J., Webster, J. P., Domingo, E., Charlesworth, B., & Levin, B. R. (2002). Biological and biomedical implications of the co-evolution of pathogens and their hosts. *Nature Genetics*, *32*(4), 569–577.

Wren, J. D., Roossinck, M. J., Nelson, R. S., Scheets, K., Palmer, M. W., & Melcher, U. (2006). Plant virus biodiversity and ecology. *PLoS Biology*, *4*(3), 0314–0315.

Wright, E. S., Heckel, T., Groenendijk, J., Davies, J. W., Boulton, M., & I. (1997). Slincing features in maize streak virus virion and complementary sense gene expression. *The Plant Journal*, *12*(6), 1285–1297.

Wright, P. F., Nilsson, E., Van Rooij, E. M., Lelenta, M., & Jeggo, M. H. (1993). Standardisation and validation of enzyme-linked immunosorbent assay techniques for the detection of antibody in infectious disease diagnosis. *Revue Scientifique et Technique (International Office of Epizootics)*, *12*(2), 435–450.

Wyant, P. S., Strohmeier, S., Schäfer, B., Krenz, B., Assunção, I. P., Lima, G. S. de A., & Jeske, H. (2012). Circular DNA genomics (circomics) exemplified for geminiviruses in bean crops and weeds of northeastern Brazil. *Virology*, *427*(2), 151–157.

Xu, P., Chen, F., Mannas, J. P., Feldman, T., Sumner, L. W., & Roossinck, M. J. (2008). Virus infection improves drought tolerance. *New Phytologist*, *180*(4), 911–921.

Yahaya, A., Dangora, D. B., Alegbejo, M. D., Kumar, P. L., & Alabi, O. J. (2017). Identification and molecular characterization of a novel sugarcane streak mastrevirus and an isolate of the A-strain of maize streak virus from sugarcane in Nigeria. *Archives of Virology*, *162*(2), 597–602.

Zerbini, F. M., Briddon, R. W., Idris, A., Martin, D. P., Moriones, E., Navas-Castillo, J., Rivera-Bustamante, R., & Varsani, A. (2017). ICTV virus taxonomy profile: Geminiviridae. *Journal of General Virology*, *98*(2), 131–133.

Zhang, Y. Z., Shi, M., & Holmes, E. C. (2018). Using Metagenomics to Characterize an Expanding Virosphere. *Cell*, *172*(6), 1168–1172.

Zhou, X., Liu, Y., Calvert, L., Munoz, C., Otim-Nape, G. W., Robinson, D. J., & Harrison, B. D. (1997). Evidence that DNA-A of a geminivirus associated with severe cassava mosaic disease in Uganda has arisen by interspecific recombination. *Journal of General Virology*, *78*(8), 2101–2111.

Zhou, Xueping. (2013). Advances in Understanding Begomovirus Satellites. *Annual Review of Phytopathology*, *51*(1), 357–381.

Zwart, M. P., & Elena, S. F. (2015). Matters of Size: Genetic Bottlenecks in Virus Infection and Their Potential Impact on Evolution. *Annual Review of Virology*, *2*(1), 161–179.

Etude de la diversité et de la structure des communautés virales à l'échelle des agro-écosystèmes – Le modèle épidémiologique des mastrévirus des Poaceae à La Réunion

Ces dernières décennies, le concept de phytovirus en tant qu'entité strictement pathogène a été particulièrement remis en question. L'ubiquité, l'abondance et la diversité des virus ont mis en exergue que ceux-ci font partie intégrante des écosystèmes. Historiquement, la majeure partie des études de phytovirologie ont porté sur une minorité de virus pathogènes des cultures, la diversité et la structure des populations virales à l'échelle des écosystèmes restant largement méconnues. Afin de mieux comprendre comment les phytovirus interagissent entre eux et avec leur environnement et comment les communautés virales s'assemblent et évoluent, il apparait aujourd'hui essentiel de les étudier à l'échelle des écosystèmes. En particulier, les agro-écosystèmes, véritables interfaces entre mondes sauvages et cultivés, apparaissent comme une échelle pertinente. En effet, la promiscuité entre les virus, les insectes vecteurs et de nombreuses espèces de plantes cultivées, sauvages et adventices avec différents mode de vie (annuelles vs pérennes) et aux origines multiples (exotiques vs indigènes) facilitent de nouvelles interactions pouvant aboutir à l'émergence de nouveaux variants viraux. Néanmoins, malgré les avancés liées à la métagénomique, il reste encore difficile de caractériser l'ensemble de la diversité virale d'un environnement. Pour pallier à cette difficulté, nous avons choisi de nous focaliser sur les phytovirus du genre *Mastrevirus*, transmis par cicadelles et responsables de nombreuses maladies sur cultures en Afrique et dans les îles de l'océan Indien.

Dans un premier temps, nos travaux ont porté sur le développement d'une approche de métagénomique dédiée et spécifique aux virus à petits génomes à ADN circulaire tels que ceux des mastrévirus et plus largement ceux de la famille des *Geminiviridae*. Cette approche dénommée RCA-RA-NGS repose sur l'amplification en cercle roulant, le marquage des amplicons par PCR aléatoire et le séquençage haut débit Illumina avant classification des lectures obtenus par recherche de similarité et placement phylogénétique. L'utilisation de 160 étiquettes PCR uniques a permis de multiplexer jusqu'à 1200 échantillons par manipulation. Après validation de la méthodologie par clonage et séquençage Sanger, elle a été appliquée à des échantillons de Poaceae collectés à l'échelle d'un agro-écosystème de La Réunion.

L'analyse de près de 3000 échantillons représentant 30 espèces de Poaceae a permis de démontrer que 18 de ces espèces et globalement 8 % des plantes évaluées étaient infectées par des mastrévirus. De manière importante, la majorité de ces échantillons positifs ne présentaient pas de symptômes visibles de maladie. Alors que cinq espèces de mastrévirus préalablement décrites à La Réunion ont été ré-identifiées durant nos travaux, nous avons aussi pu décrire pour la première fois à La Réunion une espèce déjà connus dans la région mais surtout trois nouvelles espèces de mastrévirus et une espèce d'alphasatellite, inconnus jusqu'à lors et toutes identifiées sur plantes non-cultivées. L'analyse de la structure d'association plante-virus a permis de montrer la présence de virus spécialistes et généralistes. En particulier, la souche B du maize streak virus (MSV-B) a été retrouvée chez 15 des 18 espèces de plantes hôtes identifiées. La structure de la communauté de mastrévirus montre l'imbrication des gammes d'hôte des virus spécialistes dans celle du MSV-B mais l'absence globale de modularité. La présence en co-infection du MSV-B avec la majorité des autres mastrévirus pourrait favoriser les phénomènes de recombinaison. C'est d'ailleurs chez deux variants du MSV-B que des évènements de recombinaison avec le MSV-A ont pu être détectées.

Dans leur ensemble, nos résultats ont permis une première description de l'agencement d'une communauté de mastrévirus à l'échelle d'un agro-écosystème. La stabilité de cette communauté dans le temps et en réponse aux changements tels que ceux associés aux invasions biologiques, sera une prochaine étape dans la compréhension du rôle des virus dans les écosystèmes.

Study of the diversity and structure of viral communities at the scale of agro-ecosystems - The epidemiological model of Poaceae mastreviruses in La Reunion

In the last decades, virologist shift the paradigm from pathogenic virus view to the virosphere concept. The ubiquity, abundance and diversity of viruses have highlighted that viruses are naturally embedded into global ecosystems. Historically, most phytovirology studies have focused on a minority of pathogenic viruses infecting crops, while the diversity of population structure at the ecosystem scales remains largely unknown. In order to understand how phytoviruses interacts with their environment and how viral communities assemble and evolve, it seems cardinal to study viral diversity and distribution at the scale of the ecosystem. In particular, the agrosystems represent genuine interfaces between cultivated settings and natural ecosystems. The promiscuity between viruses, insect vectors and large set of plant species is expected to favour new interactions and emergence of new viral variants. Nevertheless, despite the recent advances of metagenomics, it remains elusive to properly characterise the full viral diversity at the scale of an environment. To bypass this limitation, our work was focused on a unique viral genus, the *Mastrevirus*. These phytoviruses are transmitted to a large set of Poaceae plants by leafhoppers and were historically known for causing disease on a large set of crops in Africa and the South West Indian Ocean islands.

The first part of this work involve the development of a metagenomic method specifically devised for viruses with small circular DNA genome such as those of the mastreviruses and more generally of viruses from the *Geminiviridae* family. This approach, the RCA-RA-NGS procedure, involve a rolling circle amplification step followed with amplicons tagging using random PCR and Illumina high throughput sequencing. Reads are later classified using similarity search and phylogenetic placement. Importantly, the use of 160 distinct PCR tags allow the multiplexing of up to 1,200 samples in a single sequencing run. After the demonstration of the effectiveness of the procedure using classical cloning and Sanger sequencing, it was applied to Poaceae samples collected at the scale of a small agro-ecosytem in Reunion.

The processing of ~ 3000 samples from 30 Poaceae species demonstrates that 18 of these species are hosts of mastreviruses. Globally, 8 % of the samples were found infected with mastreviruses despite most not presenting any discernible disease symptoms. Our work allow the identification of five virus species that have already been described in Reunion and one that have been previously described in the region but not in Reunion. Most notably, three new mastrevirus species and one alphasatellites species were uncovered from uncultivated plants. The analyses of the structure of the plant virus association network demonstrate the presence of both specialist and generalist viruses. In particular, the B strain of the *Maize streak virus* (MSV-B) was found in 15 of the 18 identified plant host species. Whereas no modularity was detected, the community structure was characterised with the nestedness of specialist host range within that of MSV-B. Co-infection of MSV-B with most of the other viruses may be conductive to recombination. In fact, in two MSV-B variants, convergent recombination with MSV-A were detected.

Globally, a first description of the community structure of mastrevirus emerge from our work. The stability of the community structure in time and its robustness facing changes associated through biological invasion would be the next unavoidable step to analyse for our understanding of the role and function of viruses in the environment.