Diversité et écologie des virus associés aux arthropodes : des communautés aux génomes

Sarah François

▶ **To cite this version:**

Sarah François. Diversité et écologie des virus associés aux arthropodes : des communautés aux génomes. Sciences agricoles. Université Montpellier, 2017. Français. NNT : 2017MONTT106 . tel-01697877

HAL Id: tel-01697877
https://theses.hal.science/tel-01697877

Submitted on 31 Jan 2018

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

THÈSE POUR OBTENIR LE GRADE DE DOCTEUR DE L'UNIVERSITÉ DE MONTPELLIER

En Biologie des Interactions - Écologie de la santé

École doctorale GAIA

Unité de recherche Diversité, Génomes & Interactions Microorganismes-Insectes

---

Diversité et écologie des virus associés aux arthropodes : des communautés aux génomes

---

Présentée par Sarah François

Le 28 novembre 2017

Sous la direction de Mylène Ogliastro et de Rémy Froissart

Devant le jury composé de

Thierry CANDRESSE, Directeur de Recherche, INRA, Bordeaux	Président du jury
Christelle DESNUES, Chargée de Recherche, CNRS	Rapporteur
Jean-Christophe SIMON, Directeur de Recherche, INRA, Rennes	Rapporteur
Louis LAMBRECHTS, Chargé de Recherche, CNRS, Paris	Examinateur
Guillaume CASTEL, Chargé de Recherche, INRA, Montferrier	Invité
Philippe ROUMAGNAC, Chargé de Recherche, CIRAD, Montferrier	Invité

UNIVERSITÉ DE MONTPELLIER

INRA SCIENCE & IMPACT

Remerciements

Ma profonde gratitude revient en premier lieu à mes directeurs de thèse, Mylène et Rémy. J'ai eu la chance de profiter de votre complémentarité: la vision globale de Mylène et le sens du détail de Rémy. Je vous remercie pour la grande liberté que vous m'avez accordée, pour votre confiance, pour votre disponibilité et pour votre humanité.

Je remercie sincèrement les membres de mon jury de thèse, Thierry Candresse, Louis Lambrechts, Guillaume Castel, et en particulier mes rapporteurs Jean-Christophe Simon et Christelle Desnues, pour m'avoir fait l'honneur d'accepter d'accorder du temps à l'évaluation de mes travaux de thèse. Je remercie également les membres de mon comité de thèse, Jean-François Guégan, Stéphane Blanc, Jean-Yves Rasplus et Philippe Gayral, pour leur aide dans l'aiguillage de ma thèse.

Je suis reconnaissante à l'Université de Montpellier ainsi qu'à l'INRA pour leur investissement financier.

Je remercie Anne-Nathalie Volkoff pour m'avoir accueillie au sein de l'UMR DGIMI, ainsi que l'ensemble du personnel de cet UMR pour son soutien et sa sympathie, et en particulier les membres de l'équipe DIDI. Merci à Marie pour son indispensable aide concernant les collectes, Doriane pour son aide dans les manips, et Nicolas pour les pauses détentes. Je remercie chaleureusement le personnel de l'UMR BGPI pour son accueil et son soutien logistique. Philippe Roumagnac, Denis Filloux et Emmanuel Fernandez m'ont tout appris sur la préparation et l'analyse des viromes. Je vous remercie profondément pour votre aide, votre enthousiasme, et les discussions enrichissantes pour nous avons partagées. Ce fut un réel plaisir de travailler avec vous.

Je remercie sincèrement les gestionnaires du cluster bioinformatique du CIRAD, en particulier Bertrand Pitollat, sans qui les analyses bioinformatiques auraient été impossibles. Je tiens également à remercier le personnel des secrétariats de DGIMI ainsi que de BGPI, en particulier Marie-Ange, Caroline, Pauline et Florence pour m'avoir permis de réaliser ma thèse dans les meilleures conditions administratives possibles. Je remercie les membres du département d'enseignement « Biologie Écologie » de l'Université de Montpellier pour m'avoir permis de réaliser ma première expérience d'enseignement dans des conditions

exceptionnelles, en particulier Pierrick Labbe, Audrey Caro, Laurent Gavotte, Sylvie et Thierry. Ces trois années de monitorat ont été très enrichissantes.

Mes remerciements vont également à mes collaborateurs. D'abord à Darren P. Martin pour ses discussions instructives, puis Alison Duncan ainsi que les membres du personnel du centre d'écologie, d'évolution et des changements environnementaux du Portugal pour l'envoi d'acariens. Merci à Jean-Christophe Simon pour le partage de collections de pucerons. Merci à Armelle Cœur d'Acier pour son aide concernant la collecte de pucerons. Je tiens également à remercier le personnel du domaine de Restinclières, ainsi que les agriculteurs, de m'avoir autorisée à échantillonner sur leurs parcelles.

Mon affection ainsi que mes encouragements vont également aux doctorants de DGIMI et de BGPI, notamment à Laetitia pour sa sincérité et sa complicité. Je remercie également mon stagiaire, Maximilien, ce fut un plaisir de l'encadrer et de travailler avec lui. Merci à Clément et Laura pour leur aide lors du tri d'arthropodes.

Merci beaucoup aux ANGEs du samedi pour leur accueil chaleureux et leur soutien le long de mes études universitaires, je vous dois une part importante de mon évolution personnelle. Également un grand merci aux camarades rôlistes de tous poils, en particulier Biquette, Nicolas et Rémy, pour les soirées d'évasion et les rires que nous avons partagés. Je remercie mon maître d'armes Gilles pour la finesse de sa lame et de ses paroles.

Un grand merci à Luc et Armand, pour leur amitié ainsi que pour leur aide en informatique et en statistique. Merci à Christopher et Benoit pour les petits voyages qu'ils m'ont permis de réaliser. Merci Antho pour toutes ces balades nocturnes, pour ces discussions enrichissantes, pour toutes tes attentions et pour ton écoute ! Mes remerciements chaleureux et profonds vont à toi, Julie. J'ai beaucoup de chance d'être ton amie.

Enfin, mes remerciements et ma profonde gratitude vont à mon Clan, pour la flamme de leur amour et pour leur noblesse d'âme. À mes ancêtres, que je sois digne de vous. À mes parents, vous qui avez toujours cru en moi et qui m'avez épaulée de manière indéfectible. À mon frère Galaad, dont j'ai toujours admiré la créativité. À Laureline, pour ta confiance et ton soutien. Pour ton ardeur mêlée de douceur. Pour m'avoir apprivoisée tout en me permettant d'être libre. Je vous aime.

Résumé

Les nouvelles technologies de séquençage des génomes ont permis de révéler l'extraordinaire diversité des séquences virales dans des groupes d'hôtes jusque-là largement inexplorés. Ainsi, notre connaissance des virus d'arthropodes, infectant les animaux les plus diversifiés et abondants sur Terre, était jusque-là essentiellement réduite à des espèces d'intérêt économique et médical. Les nouvelles données de diversité virale chez les arthropodes illustrent le besoin d'étendre l'inventaire viral à l'échelle de l'écosystème et d'inclure les virus comme une composante essentielle de leur fonctionnement et de leur évolution.

Dans ces travaux de thèse, j'ai développé et appliqué deux approches d'étude de la diversité virale chez des arthropodes, ainsi que de la circulation des virus dans des écosystèmes, en me focalisant sur des espèces d'intérêt agronomique : i) une approche virus-centrée par fouille de bases de données nucléotidiques, en recherchant la présence d'un groupe de petits virus à ADN inféodés aux arthropodes, les densovirus ii) une approche arthropode-centrée, utilisant une méthode séquençage haut débit de génomes viraux (métagénomique virale) pour analyser des communautés virales associées à des arthropodes de différents niveaux trophiques échantillonnés dans des agroécosystèmes.

Mes résultats ont permis de :

(i) Mettre en évidence que les densovirus sont largement présents dans l'ensemble du règne animal - notamment chez une grande diversité d'arthropodes - et qu'ils sont très diversifiés génétiquement, ce qui a permis de mieux appréhender histoire évolutive de ce groupe de virus ;

(ii) Découvrir de nouveaux virus chez certains ravageurs de cultures : le tétranyque tisserand (*Tetranychus urticae*, Acarien) provenant de populations de laboratoires, ainsi que le puceron vert du pois (*Acyrthosiphon pisum*, Hémiptère), le phytonome de la luzerne (*Hypera postica*, Coléoptère) et l'armigère de la tomate (*Helicoverpa armigera*, Lépidoptère) provenant de populations naturelles échantillonnées dans des cultures de luzerne et des prairies. Ces études ont permis de mettre en évidence la présence de viromes spécifiques de chaque espèce d'arthropode et de caractériser la distribution de certains virus dans des communautés d'arthropodes d'un même écosystème. Une grande diversité de génomes de virus d'arthropodes et de plantes a été mise en évidence. Les liens évolutifs de ces virus avec ceux répertoriés dans les bases de données ont été caractérisés par des analyses phylogénétiques.

(iii) Enfin, les travaux menés en (ii) ont également permis d'optimiser la méthodologie permettant d'obtenir et d'analyser des viromes obtenus à partir d'échantillons multiplexés, optimisant notamment l'étape d'attribution taxonomique des séquences obtenues par séquençage à haut débit, réduisant ainsi leur proportion en « matière noire » inhérente aux analyses des viromes.

Mots-clés : Virus, Arthropodes, Métagénomique, Génomique, Transcriptomique, Diversité, Écologie, Évolution, Ravageurs de cultures, Communautés.

Abstract

High throughput sequencing technologies have revealed the extraordinary diversity of viral sequences in hitherto largely unexplored host groups. Thus, our knowledge about arthropod viruses, infecting the most diverse and abundant animals on Earth, was hitherto essentially reduced to species of economical and medical interest. New data on viral diversity in arthropods illustrate the need to expand viral inventory at the scale of the ecosystem and to include viruses as an essential component of their functioning and their evolution.

In my thesis, I developed and applied two approaches to study the diversity of viruses in arthropods and how virus circulate in ecosystems, focusing on species of agronomic interest: (i) a virus-centered approach by exploring nucleotidic sequence databases, searching for the presence of a group of small DNA viruses infecting arthropods, the densoviruses (ii) an arthropod-centered approach at the scale of the ecosystem, using a viral metagenomic method to analyze viral communities associated with arthropods from different trophic levels from the same agroecosystems.

My results showed that:

(i) Densoviruses are spread throughout the animal kingdom - particularly in a wide diversity of arthropods - and are highly diverse genetically, which led to a better understanding of the evolutionary history of this group of viruses;

(ii) A high diversity of viruses are present in pests: the spider mite (*Tetranychus urticae*, Acari) from laboratory populations, as well as the green pea aphid (*Acyrthosiphon pisum*, Hemiptera), the alfalfa weevil (*Hypera postica*, Coleoptera) and the cotton bollworm (*Helicoverpa armigera*, Lepidoptera) from natural populations sampled from alfalfa crops and grasslands. These studies also highlighted that specific viromes are associated with each pest species, and I characterized the distribution of some of these viruses in arthropod communities. Divergent arthropod and plant virus genomes were discovered. Their evolutionary links with known virus species was characterized by phylogenetic analyzes.

(iii) The work realized in (ii) also contributed to optimize a methodology to prepare and analyze viromes from multiplexed samples, that is particularly suitable to optimize the taxonomic allocation of sequences and thus reduce the "dark matter" that is inherent to viral metagenomics analyses.

Keywords: Viruses, Arthropods, Metagenomics, Genomics, Transcriptomics, Diversity, Ecology, Evolution, Crop Pests, Communities.

Table des matières

Diversité et écologie des virus associés aux arthropodes : des communautés aux génomes

Remerciements ... **1**

Résumé ... **3**

Abstract ... **4**

Table des matières ... **5**

Liste des figures et tableaux ... **7**

Liste des abréviations ... **9**

Introduction ... **10**

I- Les virus : entités biologiques longtemps méconnues aux impacts multiples ... **11**

Historique des découvertes virales ... 11

Définition ... 13

Impacts des virus ... 13

II- Les virus : origines évolutives et diversité ... **17**

Origines des virus ... 17

Mécanismes générateurs de la diversité virale ... 19

Diversité virale et taxonomie ... 23

La métagénomique, outil actuel d'étude de la diversité virale ... 28

III- Les arthropodes : un modèle de choix pour étudier la diversité virale ... **36**

Impacts des arthropodes ... 36

Virus et lutte biologique ... 38

Les virus associés aux arthropodes : une diversité largement méconnue ... 39

IV- Objectifs de la thèse ... **43**

Chapitre I - Les *Parvoviridae* : une famille diversifiée de virus d'animaux à ADN simple brin ... **45**

Des petits virus pour des hôtes très diversifiés ... **46**

Discovery of parvovirus-related sequences in an unexpected broad range of animals ... **49**

Bilan et perspectives ... 78

Chapitre II - Développement d'un protocole dédié à la préparation et à l'analyse de viromes ... **80**

Préparation et analyse de viromes multiplexés d'arthropodes et de plantes ... **81**

Viral metagenomics approaches for high resolution screening of multiplexed arthropod and plant viral communities ... **83**

Bilan et perspectives ... 105

La matière noire, limitation de l'étude des viromes ... **107**

Increase in taxonomic assignment efficiency of viral reads in metagenomic studies ... **109**

Bilan et perspectives 118

Chapitre III - Diversité des communautés virales associées à des arthropodes ravageurs de cultures 119

Impact des arthropodes ravageurs dans les agroécosystèmes 120

Metagenomic analysis of the viral communities associated with the two-spotted mite *Tetranychus urticae*: identification of a novel mini densovirus and nine other new viral species 122

Diversity and composition of arthropod pests' viral communities, and insights about pest viruses distribution in arthropod communities 152

Bilan et perspectives 198

Discussion 200

Résumé des travaux réalisés lors de la thèse 201

Développement de la métagénomique virale : de nombreux défis à relever 202

Limites méthodologiques 202

Limites conceptuelles 204

Conclusion 211

Références bibliographiques 212

Webographie 228

Curriculum vitae scientifique 229

Annexes 231

Les densovirus : une « massive attaque » chez les arthropodes 232

A Novel Itera-Like Densovirus Isolated by Viral Metagenomics from the Sea Barley *Hordeum marinum* 246

Molecular characterization and prevalence of two capulaviruses: Alfalfa leaf curl virus from France and Euphorbia caput-medusae latent virus from South Africa 249

Liste des figures et tableaux

Cette liste ne tient pas compte des figures et des tableaux contenus dans les articles.

Introduction

Figure 1 : Schéma explicatif de l'endogénéisation, en prenant l'exemple d'un rétrovirus.

Figure 2 : Schéma explicatif des trois théories explicatives de l'origine des virus.

Figure 3 : Taux de mutations chez différentes espèces de virus.

Figure 4 : Schémas explicatifs des différents types d'amphimixie.

Figure 5 : Représentation par diagramme de Venn de la forme des capsides de virus infectants archées, bactéries et eucaryotes.

Figure 6 : Représentation de la classification de Baltimore.

Figure 7 : Diversité des familles virales ainsi que des genres viraux selon la classification de Baltimore.

Figure 8 : Proportion d'articles traitant de métagénomique virale (en pourcentage), parmi l'ensemble des articles scientifiques référencés sur PubMed.

Figure 9 : Distribution de l'habitat des virus issus de métagénomes.

Figure 10 : Distribution mondiale de la diversité virale.

Tableau 1 : Tableau récapitulatif de vingt-deux articles de recherche portant sur des viromes d'arthropodes obtenus par métagénomique virale.

Figure 11 : Proportion d'espèces de macroorganismes décrites et proportion d'espèces virales décrites associées à ces macroorganismes.

Discussion

Figure 12 : Nombre de séquences virales et de viroïdes déposées dans la base de données « International Nucleotide Sequence Database Collaboration » (ligne noire), et dans la base de données RefSeq virus (GenBank) (histogramme).

Figure 13 : Actualisation des postulats de Koch pour leur utilisation en métagénomique.

Liste des abréviations

ADN : Acide désoxyribonucléique

ADNc : ADN complémentaire

ADNdb: ADN double brin

ADNsb : ADN simple brin

ARN: Acide ribonucléique

ARNdb : ARN double brin

ARNsb- : ARN simple brin à polarité négative

ARNsb+ : ARN simple brin à polarité positive

ARNsb (RT) : ARN simple brin à retro-transcription

ADNdb (RT) : ADN double brin à retro-transcription

BLAST : Basic Local Alignment Search Tool (Outil de recherche par alignement local)

ICTV: International Committee on Taxonomy of Viruses (Comité international sur la taxonomie des virus)

HTS : High-Throughput Sequencing (Séquençage à haut-débit)

Kb : Kilobase

MDA : Multiple Displacement Amplification (Amplification par déplacements multiples de brin)

NCBI : National Center for Biotechnology Information (Centre américain pour les informations biotechnologiques)

NGS: Next Generation Sequencing (Nouvelle génération de séquençage)

ORF: Open Reading Frame (Cadre de lecture ouvert)

pb : Paire de bases

PCR : Polymerase Chain Reaction (Réaction en chaine par polymérase)

WGA : Whole Genome Amplification (Amplification du génome entier)

Introduction

I- Les virus : entités biologiques longtemps méconnues aux impacts multiples

Historique des découvertes virales

Il y a près de 4500 ans, les chinois avaient identifié la maladie de la variole ainsi que son caractère infectieux. De même, des écrits de l'antiquité grecque relatent la maladie causée par le virus de la rage. Cependant, la découverte, l'isolement et la visualisation des virus n'ont été possibles que grâce à des innovations techniques, dont la première remonte à la fin du XIXème siècle (Berche, 2007; Lwoff, 1957).

Les scientifiques isolaient les agents infectieux à travers des filtres de porcelaine, comme le filtre Chamberland mis au point en 1884, dont les pores de petite taille étaient utilisés pour filtrer les bactéries en suspension dans les liquides, et ainsi les « purifier ». En 1892, Dimitri Ivanovsky démontre que des extraits de tabacs atteints de la mosaïque du tabac, filtrés à travers un filtre de Chamberland, peuvent transmettre la maladie à des plants de tabac sains. En 1898, Martinus Willem Beijerinck dilua la sève de plants de tabac infectés et l'inocula à des plantes saines qui développèrent la maladie. Réitérant la manipulation de multiples fois il put transmettre la maladie, et démontra ainsi que la sève de la dernière plante infectée était aussi virulente que la première, effet qu'une toxine après tant de dilutions n'aurait pas pu produire. De ces travaux émergeait le concept d'agents infectieux de très petite taille, invisibles au microscope et capables de se multiplier dans les cellules vivantes.

Ces agents furent appelés « virus », terme signifiant « poison » en latin. La même année, Friedrich Loeffler et Paul Frosch identifièrent le premier virus animal : l'agent de la fièvre aphteuse des bovidés. En 1901, Walter Reed et James Carrol identifièrent le premier virus humain : le virus de la fièvre jaune. Ils montrèrent également que ce virus était transmis par piqûres de moustiques. De nombreux virus pathogènes de l'homme furent découverts dans les années qui suivirent, comme le virus de la rage en 1903, celui de la poliomyélite en 1908, et celui de la rougeole en 1911. Durant la première guerre mondiale, Frederick William Twort et Félix d'Hérelle mirent en évidence indépendamment et par les mêmes processus de filtration et de dilution, la lyse de colonies bactériennes que d'Hérelle nommera « bactériophages » en 1917 (Berche, 2007).

En 1926, Theodor Svedberg mit au point la cristallisation de virus par ultracentrifugation d'un tube contenant des virus en solution. Les protéines virales concentrées peuvent alors former des cristaux analysables par diffraction des rayons X, permettant de déduire leur structure tridimensionnelle. Le premier virus cristallisé fut celui de la mosaïque du tabac, en 1935. La découverte de la microscopie électronique par Ernst Ruska et Max Knoll en 1932 permit enfin de visualiser les virus, d'abord des poxvirus en 1938, puis le virus de la mosaïque du tabac en 1939 et le virus de la grippe en 1943 (Berche, 2007).

La mise au point des cultures de cellules eucaryotes durant le XXème siècle a permis de caractériser de nombreux virus, d'abord le virus du sarcome de Rous en 1926 par Alexis Carrel. La culture sur œufs de poule embryonnés, mise au point par Alice Miles Woodruff et Ernest William Goodpasture a permis notamment de cultiver des virus de la grippe entre 1931 et 1933. Après la seconde guerre mondiale, la découverte (par Flemming en 1929), la synthèse, la diffusion et l'adjonction des antibiotiques aux cultures cellulaires, empêchant les contaminations bactériennes, furent de grands progrès dans l'isolement des virus. De nombreux nouveaux virus furent alors découverts, comme les herpèsvirus, le virus de la rubéole et les entérovirus. Enfin, dans les années 1980, la découverte des interleukines, molécules stimulant la croissance des lymphocytes, a permis la découverte de virus ne se multipliant que dans ces cellules, comme les rétrovirus, dont le virus responsable du syndrome d'immunodéficience acquise (SIDA) découvert en 1983 (Barré-Sinoussi *et al.*, 1983; Berche, 2007).

L'avènement de la biologie moléculaire a permis l'identification de virus restant incultivables. Le génome de l'hépatite C a notamment été caractérisé par clonage dans des bactéries suivi de séquençage, mis au point indépendamment en 1977 par les équipes de Walter Gilbert et de Frederick Sanger. D'autre part, depuis 1986, l'utilisation de la technique de la polymérisation en chaine (PCR), puis en 1995 celle des puces à ADN contenant des séquences virales synthétiques (microarrays), requérant toutes deux des connaissances préalables sur les séquences d'acides nucléiques recherchées, ont permis de découvrir de nombreux virus, variants de virus connus. Le virus de l'hépatite E en 1990, l'herpèsvirus responsable du syndrome de Kaposi en 1998 et le virus du SARS en 2003 ont ainsi été découverts (Berche, 2007; Mokili *et al.*, 2012). Cependant, ces méthodes ne permettraient pas de détecter des virus divergents de formes déjà connues (Mokili, Rohwer, & Dutilh, 2012).

Aujourd'hui, ces méthodes moléculaires traditionnelles de découverte virale ont été substituées par des méthodes moléculaires sans *a priori* sur les séquences génomiques par des amplifications aléatoires des matrices génétiques couplées aux nouvelles technologies de séquençage à haut débit mises en place à partir de 2005. Cependant, ces deux types d'approches, avec et sans *a priori*, jouent des rôles complémentaires dans l'isolation, l'identification et la caractérisation des virus (Mokili *et al.*, 2012).

Définition

Le terme « virus » regroupe l'ensemble des entités biologiques constituées d'un génome composé d'une ou de plusieurs molécules d'acide nucléique (ADN et/ou ARN) ne contenant pas l'information nécessaire à la synthèse de l'ensemble protéique ribosomal ni des protéines permettant la synthèse d'énergie (ATPase). Le matériel génétique viral est protégé par une coque protéique (capside) éventuellement entourée d'une enveloppe lipidique provenant des membranes des cellules hôtes. Les virus sont des parasites intracellulaires obligatoires qui se répliquent en détournant la machinerie cellulaire des cellules infectées. Le cycle de multiplication des virus implique une alternance entre décapsidation (sortie du génome viral de la capside) et recapsidation (recouvrement du génome viral par autoassemblage des éléments formant la capside) (Holmes, 2009; Lwoff, 1957; Wolkowicz and Schaechter, 2008).

Impacts des virus

Les virus infectent l'ensemble des trois règnes du vivant : bactéries, archées et eucaryotes. Tous les organismes cellulaires connus sont infectés par au moins un ou plusieurs virus. Certains virus sont même capables de détourner les structures cellulaires mises en place par d'autres virus (La Scola *et al.*, 2008). En conséquence de leur ubiquité chez le vivant, les virus sont présents dans l'ensemble des écosystèmes terrestres et aquatiques connus, des fonds marins (Danovaro *et al.*, 2008) jusqu'aux glaciers (Zablocki *et al.*, 2014). Ils y sont extrêmement abondants : par exemple, l'abondance des particules virales de la surface de l'océan est comprise entre 10^5 et 10^8 particules par millilitre (Bergh *et al.*, 1989) (Suttle, 2007). Autre exemple, le sang humain contient 10^5 particules virales par millilitre tandis que l'urine en contient en moyenne 10^7/mL (Rascovan *et al.*, 2016).

Les virus sont également abondamment présents dans le génome de leurs hôtes. Il peut s'agir de virus dont le cycle infectieux passe par une étape d'intégration, ou non. Si les cellules infectées appartiennent à la lignée germinale, le génome viral intégré devient « fixé » au sein du génome de l'hôte et sera transmis verticalement à la descendance de l'hôte. Ce phénomène, appelé endogénéisation, est retrouvé dans une grande diversité de génomes procaryotes (e.g. prophages) et eucaryotes (**Fig. 1**) (Feschotte et Gilbert, 2012). De tels évènements ont eu lieu d'une façon répétée durant des millions d'années d'évolution et de coévolution. Les séquences virales endogénéisées peuvent ainsi représenter une proportion significative du génome des hôtes. Par exemple, 8% du génome humain et 10% du génome de la souris sont composés de dérivés de séquences virales endogénéisées (Horie et Tomonaga, 2011).

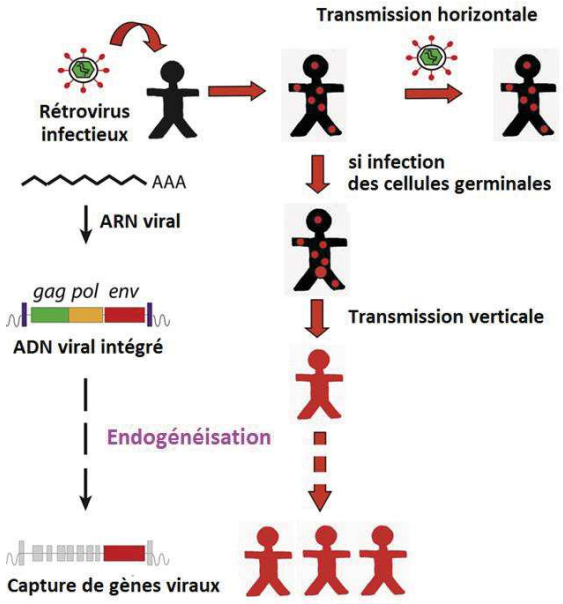


Figure 1 : Schéma explicatif de l'endogénéisation, en prenant l'exemple d'un rétrovirus. Les rétrovirus ont la capacité de transcrire leur ARN en ADN qui est ensuite inséré dans le génome des cellules infectées sous forme d'un provirus, ici porteur de trois gènes (gag, pol et env). La production de nouvelles particules virales permet la transmission horizontale à des nouveaux hôtes. Dans de rares cas, l'intégration du génome viral a lieu dans le génome de cellules germinales, ce qui permet leur transmission à la descendance de l'hôte. La majorité des gènes viraux endogénéisés sont perdus au cours de l'évolution. Cependant, occasionnellement, quelques gènes viraux restent fonctionnels et peuvent impacter le phénotype de leur hôte. Tiré de Dupressoir *et al.*, 2012.

Les virus sont donc abondamment présents au sein des écosystèmes comme de leurs hôtes. Cependant, les virus ne sont pas importants uniquement en raison de leur abondance, mais également par leurs impacts sur les écosystèmes et l'évolution des organismes cellulaires.

En effet, les virus, de par leur cycle de vie incluant un parasitisme obligatoire, comptent parmi les acteurs importants de la dynamique des populations de leurs hôtes (Weinbauer et Rassoulzadegan, 2004). De par les pressions de sélection qu'ils exercent sur leurs hôtes, les virus jouent un rôle majeur dans la régulation de cycles biogéochimiques et biologiques, et contribuent ainsi de façon significative à la stabilité et au fonctionnement de certains écosystèmes. Par exemple, de récentes études montrent que les virus aquatiques ont un impact important sur le fonctionnement des cycles biogéochimiques et écologiques à travers l'infection et la lyse des communautés bactériennes et de micro-algues intervenant dans le recyclage de la matière organique et de la production primaire (Danovaro *et al.*, 2008; Rohwer et Thurber, 2009). En outre, en parasitant l'espèce la plus compétitive dans un écosystème donné, les virus laissent une niche écologique vacante où des espèces moins compétitives pour les ressources nutritives peuvent alors se développer, ce qui aboutit à un maintien de la diversité des communautés d'hôtes (selon le modèle « kill the winner ») (Rodriguez-Brito *et al.*, 2010). La disparition des efflorescences (blooms) de phytoplancton en quelques heures, dues aux virus, illustrent bien ce phénomène (Wilhelm, 1999).

D'autre part, les virus sont responsables d'une forte proportion des maladies virales infectieuses émergentes : ils représentent près de 47% des maladies émergentes chez les plantes (Anderson *et al.*, 2004), et entre 25% et 43 % des maladies infectieuses émergentes chez l'homme (Jones *et al.*, 2008; Woolhouse & Gaunt, 2007). Parmi ces derniers, on peut citer les virus de l'immunodéficience humaine (VIH), les virus de la grippe, le virus responsable du syndrome respiratoire aigu sévère (SRAS), ou le virus Ebola. Ces virus ont des impacts médicaux, économiques et politiques significatifs, le VIH étant à lui seul responsable de plus de 25 millions de morts depuis sa découverte en 1983 jusqu'à l'année 2007 (Jones *et al.*, 2008; Woolhouse & Gowtage-Sequeria, 2005; Woolhouse & Gaunt, 2007). Par ailleurs, certains virus ont été utilisés ou projetés d'être utilisés en tant qu'armes biologiques. En 1763, Jeffrey Amherst, général de l'armée anglaise, fit distribuer des couvertures de varioleux aux tribus indiennes de l'Ohio, ce qui a déclenché des épidémies dévastatrices dans la région. Autre exemple, le virus de la variole a été produit en grandes quantités par l'URSS lors de la guerre froide (Berche, 2007).

Depuis leur découverte, les virus ont été considérés majoritairement en tant qu'agents pathogènes. De nombreuses découvertes récentes ont radicalement changé ce point de vue, en mettant en évidence la présence d'un continuum entre parasitisme et mutualisme obligatoire chez les virus (Roossinck, 2015). Certains virus fournissent des bénéfices à leurs hôtes seulement sous certaines conditions environnementales, leur permettant par exemple de s'adapter rapidement à des changements environnementaux. D'autres virus confèrent à des plantes une résistance accrue à la sécheresse, à la chaleur ou au froid (Roossinck, 2015, 2011a). Un densovirus induirait le développement des ailes chez une espèce de pucerons, ce qui augmenterait leur capacité de dispersion (Ryabov *et al.*, 2009). Enfin, certains virus sont indispensables à la survie de leurs hôtes : certains polydnavirus, dont le génome est intégré dans celui de guêpes parasitoïdes (*Braconidae* et *Ichneumonidae*, Hyménoptères), sont requis pour la survie des œufs dans l'insecte-hôte des guêpes. Ces virus sont devenus partie intégrante de l'hôte : les frontières entre virus et hôte s'effacent (Roossinck, 2011a).

Enfin, l'évolution des organismes cellulaires serait en grande partie sculptée par l'interaction entre les virus et leurs hôtes. La course aux armements et les dynamiques de résistance associées se traduisent par une modification continue des génomes des hôtes et des virus (Stern & Rotem, 2011). En outre, les virus modifient également le génome de leurs hôtes à travers l'endogénéisation ou le transfert horizontal de gènes, ce qui influence l'évolution de leurs génomes (Gilbert *et al.*, 2014). Le transfert horizontal de matériel génétique représente la transmission d'ADN entre organismes non-apparentés. Il représente un facteur important de l'évolution des procaryotes mais également des eucaryotes (Syvanen, 2012). Or, de nombreux virus participent au transfert horizontal de gènes, d'origine cellulaire ou virale, ces gènes peuvant impacter le phénotype de l'hôte. Par exemple, l'intégration d'un bactériophage dans le génome de la bactérie *Vibrio cholerae* permet à cette bactérie de synthétiser la toxine cholérique, induisant les symptômes caractéristiques du choléra (Davis & Waldor, 2003). En outre, l'insertion de génomes viraux au sein de génomes cellulaires est susceptible d'inactiver ou de promouvoir l'expression de gènes cellulaires (Feschotte et Gilbert, 2012). Certains évènements d'endogénéisation ont été à l'origine d'innovations évolutives majeures : la syncytine, protéine d'origine virale, joue un rôle fondamental dans la formation du placenta des Mammifères (Dupressoir *et al.*, 2012).

L'ubiquité des virus chez les organismes cellulaires a amené de nombreuses questions quant à leurs origines ainsi que leur diversité.

II- Les virus : origines évolutives et diversité

Origines des virus

Les virus et les cellules sont les deux principales catégories d'organisation biologique, qui sont respectivement caractérisées comme (1) des informations génétiques parasites et (2) des organismes génétiquement auto-entretenus (Koonin et Wolf, 2012; Raoult et Forterre, 2008). Des modèles mathématiques prédisent que le parasitisme génétique émerge inévitablement dans tout système réplicatif possédant des ressources limitées (Szathmary and Maynard Smith, 1997). Ainsi, l'émergence d'éléments parasites comme les virus ferait partie intégrante de la vie. L'étude de l'origine des virus est compliquée de par le fait qu'il n'existe pas de fossile de virus et que les forts taux d'évolution virale posent des problèmes quant à la reconstitution de leur histoire évolutive. Il existe trois hypothèses explicatives de l'origine des virus.

- L'hypothèse de l'évolution régressive postule que les virus sont issus de cellules parasites ayant subies des réductions génomiques graduelles, afin d'en retirer les fonctions qui leur sont fournies par leurs hôtes, jusqu'à ce qu'ils deviennent les virus actuels (Holmes, 2009) (**Fig. 2**). La réduction évolutive est un processus évolutif fréquemment observé chez d'autres parasites, comme la bactérie *Buchnera sp.* infectant des pucerons qui a subi une réduction d'environ 70% de la taille de son génome (Claverie et Ogata, 2009). Cette hypothèse pourrait expliquer l'apparition des virus géants, observables au microscope optique et dont le génome code un appareil de traduction partiel (Legendre *et al.*, 2013). Cependant, cette hypothèse est remise en cause par des résultats récents sur l'étude des virus géants (Schulz *et al.*, 2017). De plus, elle ne permettrait pas d'expliquer l'origine des virus possédant un génome constitué d'ARN, notamment car leurs gènes diffèrent totalement de ceux des organismes cellulaires (Holmes, 2009).

- L'hypothèse de l'échappement de gènes stipule que les virus seraient les descendants de gènes cellulaires parasites ayant acquis la capacité de se répliquer de façon autonome ainsi que de coder pour une capside protectrice leur permettant de transporter leur matériel génétique et le protéger d'un milieu cellulaire hostile (**Fig. 2**). Des évènements d'échappement multiples, à partir de plasmides ou de transposons, auraient donné lieu à l'apparition de nombreux groupes de virus (Holmes, 2009).

- Enfin, l'hypothèse de la vie pré-cellulaire propose que les virus et les cellules soient apparus en même temps et aient évolué parallèlement, les virus étant les descendants de formes pré-cellulaires ayant adopté leur style de vie parasite plus tard dans leur histoire évolutive (**Fig. 2**). Selon cette hypothèse, les plus anciens systèmes génétiques de réplications auraient été composés d'ARN et seraient devenus plus complexes, s'enveloppant dans un sac lipidique aboutissant aux cellules primitives. Une autre forme réplicative ayant gardé sa simplicité aurait formé les virus (Holmes, 2009).

L'origine des virus n'est actuellement pas résolue. Cependant, les virus auraient vraisemblablement des origines évolutives diverses et ne seraient donc pas monophylétiques (Claverie et Ogata, 2009).

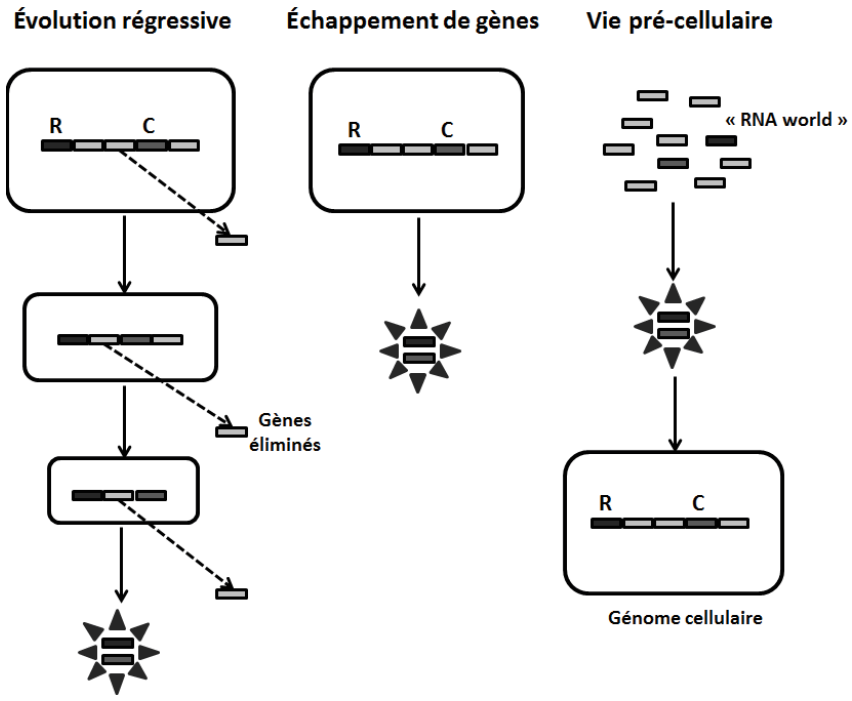


Figure 2 : Schéma explicatif des trois théories explicatives de l'origine des virus. Les cellules sont représentées par des rectangles arrondis. R : protéine de réplication ; C : protéine de capside. Tiré de Holmes, 2009.

Mécanismes générateurs de la diversité virale

Les virus évoluent par des modifications de leur génome susceptibles de créer de nouvelles fonctions. Les moteurs moléculaires de la création de la diversité virale se regroupent en deux catégories : la mutation et l'amphimixie.

Mutation

Les mutations désignent les modifications du génome viral produites lors de sa réplication par les polymérases. Il existe trois formes de mutations :

- La substitution correspond au remplacement d'un nucléotide par un autre ;
- L'insertion correspond à l'insertion d'un ou plusieurs nucléotides entre deux nucléotides préexistants ;
- La délétion correspond à la perte d'un ou plusieurs nucléotides.

Le taux de mutation indique les fréquences de modifications générées par position génomique par cycle de réplication Il est plus élevé chez les virus que chez les organismes cellulaires, et il semblerait y avoir une corrélation négative entre le taux de mutation et la taille du génome (Duffy *et al.*, 2008; Sanjuan *et al.*, 2010). Le taux de mutation généralement observé chez les organismes cellulaires est de 10^{-8} à 10^{-11} mutations /site/cycle de réplication. En comparaison, les taux de mutation viraux vont de 10^{-8} à 10^{-6} mutations /site/cycle de réplication pour les virus à ADN double brin, et sont les plus élevés, de10^{-6} à 10^{-3}, pour les virus à ARN et à ADN simple brin (**Fig. 3**). Parallèlement, ce sont les virus à ARN qui représentent la majorité des virus émergents chez l'homme, ainsi il y aurait un lien entre taux de mutation et capacité d'adaptation chez les virus (Woolhouse & Gaunt, 2007).

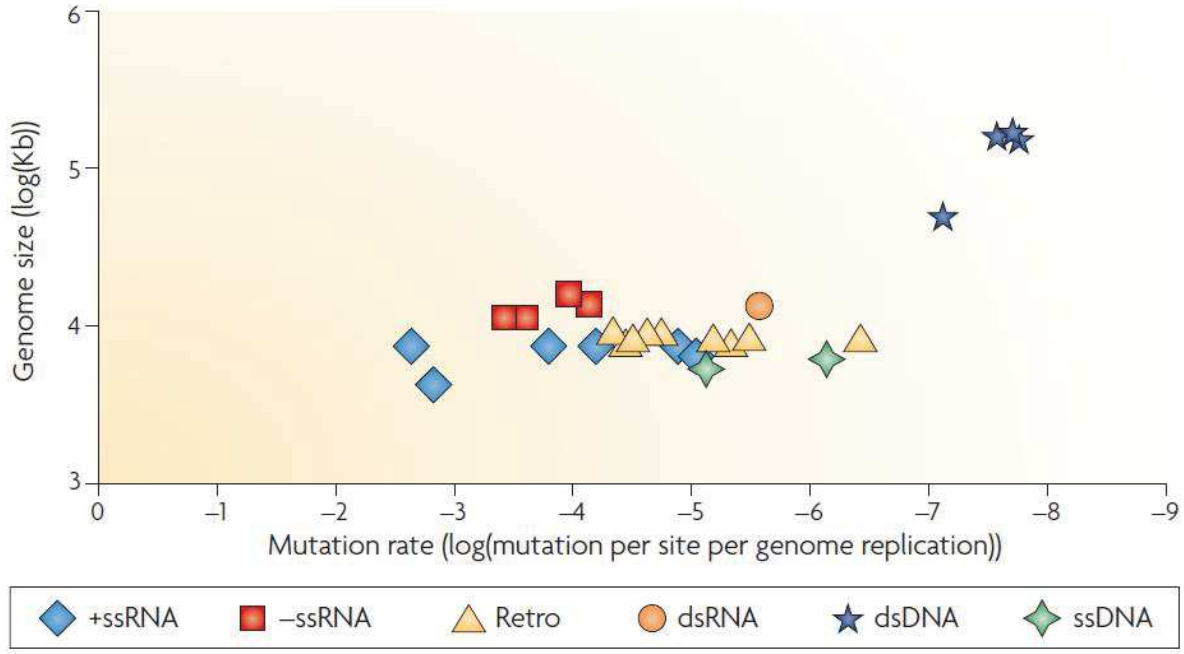


Figure 3 : Taux de mutations chez différentes espèces de virus. Le taux de mutation (en log du nombre de mutations par site par cycle de réplication) est représenté en abscisse, et la taille du génome (en log de kilobases) est représentée en ordonnée. Tiré de Duffy *et al.*, 2008.

Amphimixie

L'amphimixie correspond à la création d'un génome viral à partir de la fusion des génomes de plusieurs virus. Il existe deux catégories d'amphimixie : la recombinaison et le réassortiment (**Fig. 4**).

- La recombinaison est la formation, lors de la réplication virale, de molécules d'acides nucléiques chimériques filles à partir de molécules d'acides nucléiques provenant de génomes parentaux différents. Le génome viral résultant correspond donc à une mise en continuité du matériel génétique provenant de deux chaines différentes d'acide nucléique. La recombinaison se produit lorsque la polymérase effectue des changements de matrices lors de la réplication. Il existe deux types de recombinaison. Lors de la recombinaison homologue, le changement de matrice se produit au niveau de sites homologues. La recombinaison non-homologue désigne une recombinaison qui se produit entre des sites non homologues. Les taux de recombinaison virale peuvent être similaires aux taux de mutations. Par exemple, chez le VIH, $1{,}35.10^{-3}$ évènements de recombinaison se produisent par nucléotide par cycle de réplication virale (Schlub *et al.*, 2010). Des évènements de recombinaison entre des taxa

viraux éloignés, notamment entre des virus à ARN et des virus à ADN, ou entre des virus à ARN et des plasmides, seraient à l'origine de l'apparition et de la diversification de nombreux groupes de virus (Koonin *et al.*, 2015; Krupovic *et al.*, 2015; Lefeuvre *et al.*, 2009; Martin *et al.*, 2011).

- Enfin, le réassortiment (ou pseudo-recombinaison) désigne, lors de la réplication de plusieurs virus possédant un génome segmenté (i.e. constitué de plus d'une molécule d'acide nucléique), l'encapsidation de segments issus de différents génomes dans la même particule virale. Par exemple, des phénomènes de réassortiment sont fréquemment observés entre des souches humaines, porcines et aviaires du virus de la grippe (Li *et al.*, 2010; Zhou *et al.*, 1999).

Afin que des évènements d'amphimixie entre différents types de virus puissent se produire, il est nécessaire qu'ils infectent une même cellule. Cette probabilité d'infection simultanée est fonction du degré de recouvrement entre leur répartition géographique, leurs cycles épidémiologiques, leur spectre d'hôte et leur tropisme cellulaire (Martin *et al.*, 2011).

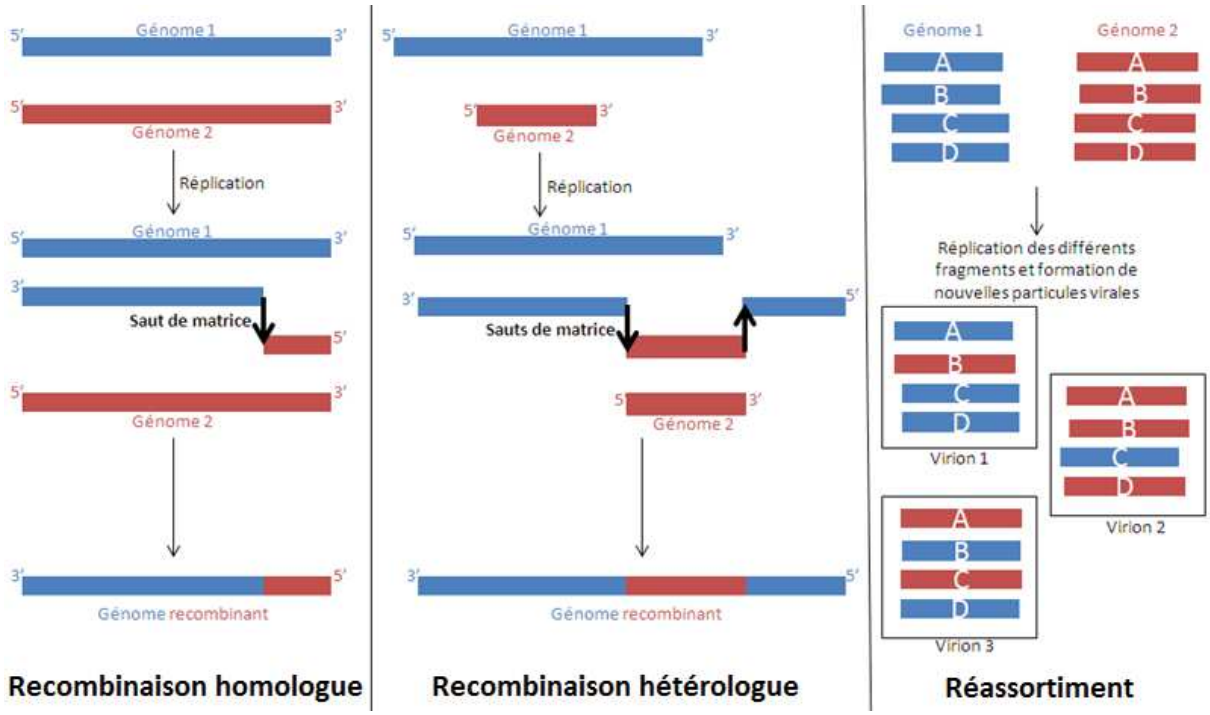


Figure 4 : Schémas explicatifs des différents types d'amphimixie. Image fournie par Pauline Bernardo.

Les modifications du génome viral, obtenues par mutations et par amphimixie, peuvent avoir trois types d'effets en termes évolutifs : délétère, neutre et bénéfique. Les modifications bénéfiques vont apporter un avantage sélectif, contrairement aux modifications délétères qui sont désavantageuses. Les modifications neutres, quant-à-elles, n'apportent ni avantage ni désavantage. Or, la majorité des modifications génétiques ont des impacts négatifs pouvant conduire à un effet létal, les génomes viraux portant ces altérations étant donc inaptes à rivaliser avec les génomes viraux parentaux. Par conséquent, tous les variants produits dans une population virale ne sont pas maintenus (Domingo-Calap *et al.*, 2009; Monjane *et al.*, 2014). Cependant, ces prédictions sont à modérer par deux faits. D'abord, il est à noter qu'il existe des phénomènes de pléiotropie antagoniste, i.e. que certaines variations délétères dans un environnement donné peuvent être bénéfiques dans un autre environnement (Duffy *et al.*, 2006). Enfin, des phénomènes de complémentation peuvent exister au sein de populations virales : l'expression de gènes fonctionnels portés par certains génomes viraux compense les effets des mutations délétères portées par d'autres génomes viraux (Aaskov *et al.*, 2006).

Enfin, les virus se caractérisent généralement par un nombre élevé d'individus par génération et des cycles de réplication très courts. L'accumulation des diversités qui en résulte, observable à l'échelle d'une vie humaine, fait des virus des modèles de choix pour étudier l'évolution du vivant. De plus, comme énoncé précédemment, les virus sont responsables d'une forte proportion des maladies infectieuses émergentes chez l'homme (jusqu'à 43%), bien qu'ils ne soient pas les entités parasites les plus diversifiées en termes d'espèces décrites chez l'homme (aux alentours de 15%). Cette forte proportion de maladies émergentes dues aux virus s'expliquerait notamment par le fait que les virus évoluent plus rapidement que leurs hôtes, grâce à des taux de mutations élevés et un potentiel élevé d'amphimixie (Jones *et al.*, 2008; Woolhouse & Gaunt, 2007). L'évolution virale se traduit notamment par des conséquences sur la pathogénicité, le franchissement de la barrière d'espèce, l'échappement aux antiviraux et aux vaccins (Duffy *et al.*, 2008). Par exemple, des mutations permettent au VIH de résister aux antiviraux (Hirsch *et al.*, 2003). Enfin, le phénomène de recombinaison entre des variants atténués présents dans les vaccins contre la poliomyélite produit des variants neuropathiques, ce qui complique les efforts de lutte contre cette maladie (Kew *et al.*, 2005).

En conséquence de leur évolution rapide, les virus représentent des entités biologiques extrêmement diversifiées.

Diversité virale et taxonomie

Diversité morphologique et génomique des virus

Les virus présentent une grande variabilité en termes de taille, de structure de leur capside, de la diversité de leur organisation génomique et de leurs stratégies de réplication.

Il existe une grande diversité de la taille des génomes viraux et des capsides virales. Les membres de la famille des *Circoviridae* font partie des plus petits virus connus. Leur génome peut mesurer moins de 1800 nucléotides de longueur, et est contenu dans une capside de moins de 20 nm de diamètre (Breitbart *et al.*, 2017). Les plus gros virus appartiennent à la famille des *Pandoraviridae*. Ils possèdent un génome de plus de 2,5 Mb, contenu dans une capside de plus d'1µm (Legendre *et al.*, 2014, 2013). Ces « virus géants » peuvent être observés en microscopie optique. Leur découverte a révélé un continuum dans la taille génomique et la complexité fonctionnelle entre les virus et les organismes cellulaires, les génomes ainsi que les capsides des virus géants étant de taille comparable à celle des bactéries. Entre ces deux extrêmes, une grande diversité de taille existe.

La morphologie des capsides virales, dont les protéines constitutives sont organisées en formes géométriques régulières, est également diversifiée. Leur symétrie peut être icosaédrique, hélicoïdale, ou mixte (tête icosaédrique et queue hélicoïdale). Les capsides peuvent également être en forme de bâtonnets, de filaments, bacilliformes, de balles de fusil, de bouteilles, de citrons ou de gouttes d'eau. Certains virus peuvent également posséder les appendices en forme de queue (**Fig. 5**).

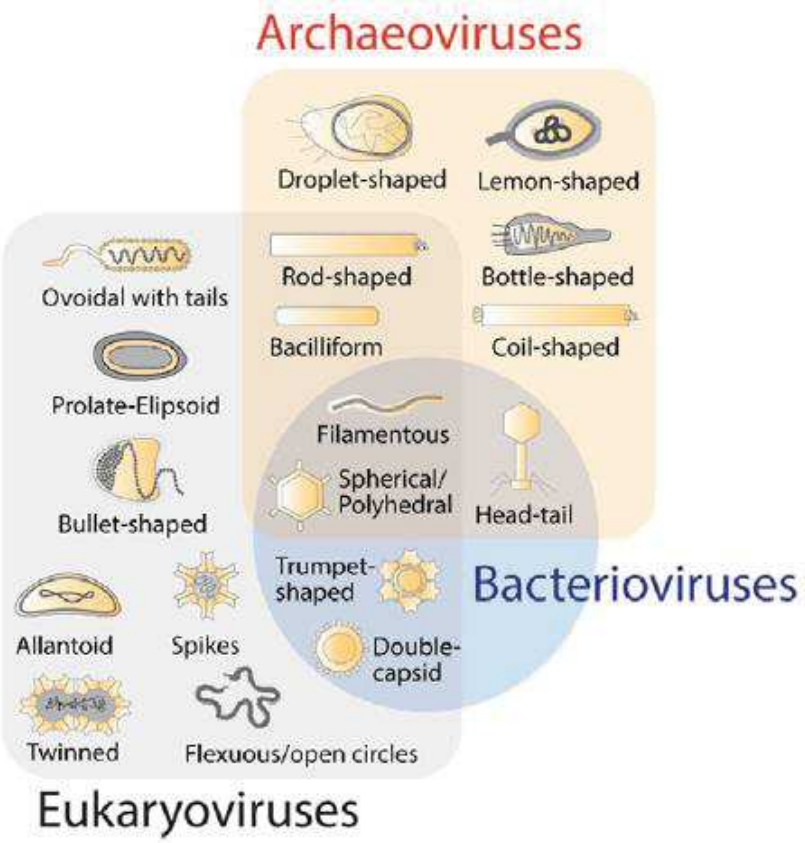


Figure 5 : Représentation par diagramme de Venn de la forme des capsides de virus infectants archées, bactéries et eucaryotes. Tirée de Nasir, Kim, & Caetan-Anolles, 2017.

En outre, les génomes viraux arborent une grande diversité de structure. En effet, contrairement aux génomes cellulaires uniquement constitués d'ADN double brin, il existe chez les virus deux formes de support de l'information génétique, ADN et ARN, déclinées en plusieurs variantes décrites par la classification de Baltimore : ADN double brin (ADNdb), ADN simple brin (ADNsb) ARN simple brin positif (ARNsb+), ARN simple brin négatif (ARNsb-), ARN double brin (ARNdb), et virus à ARN ou à ADN à transcription inverse. De plus, le génome viral peut être linéaire ou circulaire, monocaténaire ou segmenté. Enfin, les virus présentent une grande diversité de stratégies de réplication. Les formes cellulaires utilisent toutes une seule stratégie de réplication et d'expression basée sur la réplication d'ADN double brin, la transcription des gènes en ARN messagers suivie de leur traduction en protéines. Cependant, des processus réplicatifs alternatifs comme la réplication de l'ARN et la transcription inverse sont couramment utilisés par les virus (**Fig. 6**).

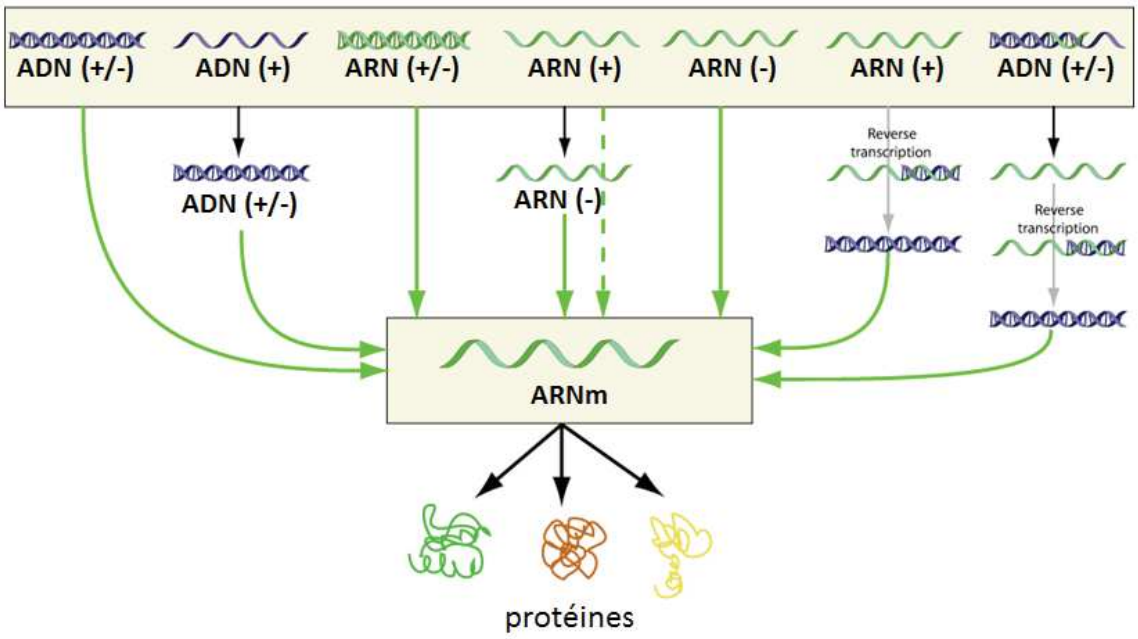


Figure 6 : Représentation de la classification de Baltimore. Tirée du site ViralZone, 2017.

Taxonomie virale

La taxonomie virale a pour but de nommer les entités virales de manière pertinente et universelle et de les classer dans des groupes de la façon la plus rationnelle possible, afin d'illustrer leurs relations évolutives et de simplifier la communication internationale entre scientifiques. Ces entités biologiques n'étant pas immuables et formant un continuum en terme de diversité, leur catégorisation est donc une construction humaine artificielle qui reste néanmoins nécessaire (Kuhn & Jahrling, 2011).

La classification des virus est organisée par le Comité International de Taxonomie des Virus (ICTV), qui édite les critères de démarcation entre les taxa viraux. Les critères dépendent des informations disponibles sur les virus concernés. La taxonomie virale est organisée hiérarchiquement à l'image de la taxonomie animale. Les virus sont tout d'abord séparés en fonction de la nature et de la structure de leur génome ainsi que de leur stratégie de réplication, puis ils sont classés en quatre taxa : ordre, famille, genre et espèce. Les autres propriétés de séparation entre taxa viraux sont le spectre d'hôte, le tropisme tissulaire et cellulaire, la pathogénicité, le mode de transmission, les propriétés chimiques et antigéniques, et les différences de séquences génomiques (Kuhn & Jahrling, 2011).

▪ *Ordre*

Les ordres regroupent des familles partageant une histoire évolutive commune. Actuellement, il existe 8 ordres répertoriés par l'ICTV (*Bunyavirales*, *Nidovirales*, *Mononegavirales*, *Caudovirales, Herpesvirales, Ligamenvirales, Picornavirales* et *Tymovirales*), et 85 familles virales ne sont pas assignées à un ordre (ICTV, mai 2017).

▪ *Famille*

Les genres sont classés en familles selon un groupement basé sur la structure et la stratégie de réplication du génome viral. L'ICTV répertorie actuellement 122 familles virales (**Fig. 7**). Ainsi, 19 genres viraux ne sont actuellement pas assignés à une famille (ICTV, mai 2017) (**Fig. 7**).

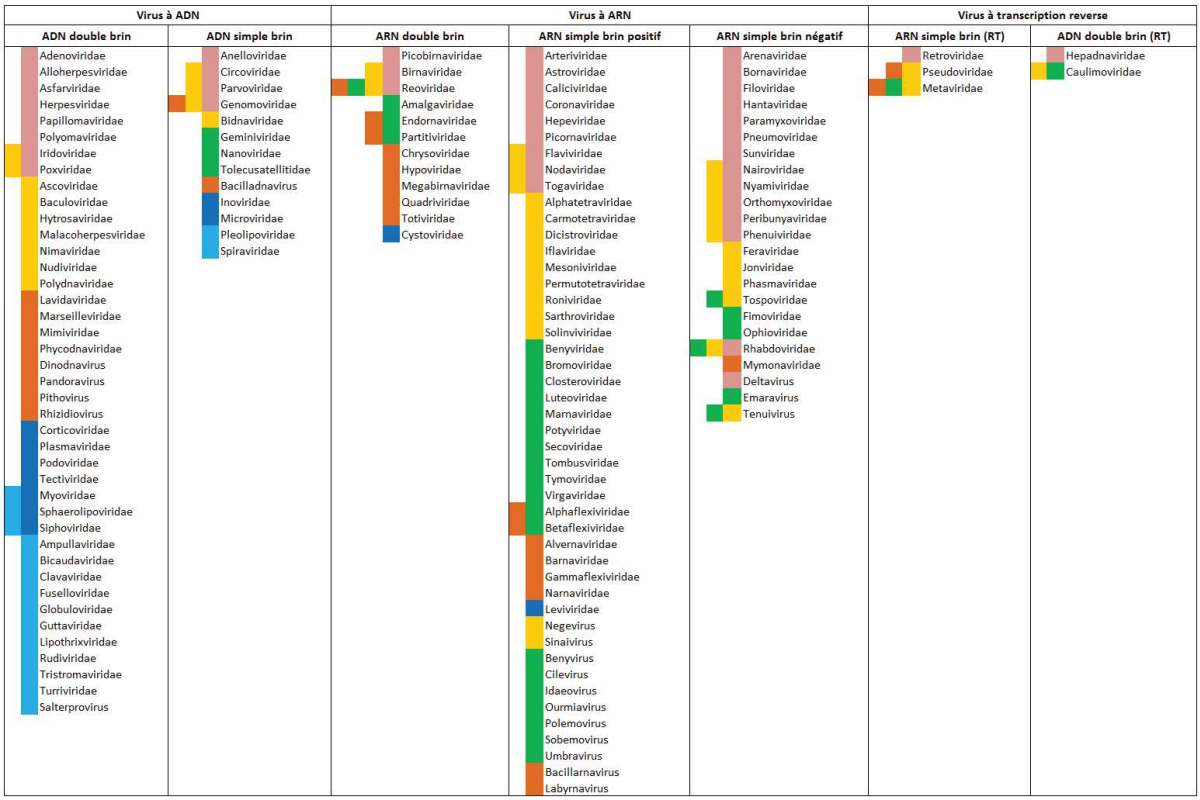


Figure 7 : Diversité des familles virales ainsi que des genres viraux selon la classification de Baltimore. Les couleurs représentent les hôtes : bleu clair : archées, bleu foncé : bactéries, vert : plantes, jaune : invertébrés, rose : vertébrés, orange : autres eucaryotes. Graphique construit d'après les sites de l'ICTV, mai 2017 et de ViralZone, 2017.

- ***Genre***

Chaque genre est défini par un ensemble de caractéristiques communes des espèces qui la constituent : l'organisation du génome, les gènes qu'il contient et le type d'hôte. Un assouplissement des conditions d'intégration à la taxonomie virale a récemment été accepté par l'ICTV. En effet, les genres peuvent à présent être délimités en utilisant seulement la caractérisation des génomes viraux comprenant l'organisation génomique et le degré de différenciation par rapport aux virus répertoriées ; la caractérisation phénotypique devenant optionnelle (Simmonds *et al.*, 2017). Pour chaque genre, l'ICTV a défini une espèce-type qui est la plus caractérisée ou la première décrite. Il existe des espèces virales sans genre attribué mais qui sont tout de même classées au sein d'une famille. Actuellement, l'ICTV reconnait 736 genres viraux (ICTV, mai 2017).

- ***Espèce***

La définition de l'espèce virale adoptée par l'ICTV est la suivante: « L'espèce virale représente un ensemble monophylétique de virus dont les propriétés peuvent être distinguées de celles des autres espèces par de multiples critères », ces critères étant la séquence génomique, des propriétés de réplication, la morphologie du virion, le spectre d'hôte ou la pathogénicité. Une espèce virale est donc considérée comme étant une entité biologique et écologique. Les espèces virales peuvent également être délimitées en utilisant seulement les séquences génomiques (Simmonds *et al.*, 2017). Le dernier rapport de l'ICTV répertorie un total de 4404 espèces virales (ICTV, mai 2017).

Par ailleurs, un isolat désigne généralement un échantillon prélevé sur un organisme (hôte, vecteur, prédateur de l'hôte). Une souche est un ensemble d'isolats ayant en commun plusieurs propriétés qui les caractérisent. Une espèce virale est donc constituée d'une ou de plusieurs souches elles-mêmes constituées d'un ou plusieurs isolats.

Les données actuelles indiquent que la diversité virale reste encore largement inexplorée. Il a été estimé qu'à peine un pour cent de l'ensemble des virus ont été découverts (Mokili *et al.*, 2012).

La métagénomique, outil actuel d'étude de la diversité virale

Historique et définition

Le progrès de la science est ponctué par des avancées technologiques qui révolutionnent les méthodes et/ou les échelles d'analyses et permettent d'aller vers de nouveaux fronts de science et faire avancer nos connaissances. Il y a eu la microscopie électronique, la culture cellulaire et la PCR (Réaction en chaine par polymérase). À présent, des méthodologies liées au développement du séquençage à haut débit (HTS ; également nommé « nouvelle génération de séquençage » (NGS)) bouleversent notre conception de la diversité et de la prévalence virale (Berche, 2007; Mokili *et al.*, 2012).

De nombreuses approches utilisant les données produites par séquençage à haut débit sont utilisées dans le cadre de l'étude de la diversité virale. Elles comprennent notamment le séquençage des ADN ou ARN totaux, le séquençage des petits ARN associés aux mécanismes d'inhibition de l'expression de gènes (silencing) chez les plantes et la fouille de données (Hadidi *et al.*, 2016; Roossinck, 2016). Cependant, la métagénomique virale est la méthode actuelle la plus utilisée pour étudier la diversité virale (Mokili *et al.*, 2012).

La métagénomique, i.e. l'analyse des communautés à travers l'analyse de leurs séquences génomiques, a été définie pour la première fois en 1998 (Handelsman *et al.*, 1998). D'abord utilisée dans le cadre de l'étude des communautés bactériennes, elle fut appliquée à l'étude des communautés virale à partir de 2001 (Allander *et al.*, 2001). La métagénomique virale consiste à concentrer les particules virales d'un échantillon donné, à en extraire et amplifier le contenu génomique sans *a priori*, puis à le séquencer. La puissance de cette méthode est de cibler le contenu génomique viral total, ou virome, sans tenir compte des cibles moléculaires déjà connues, et en s'affranchissant de la culture des virus. Ainsi, l'ensemble des virus contenus dans tout type d'échantillon peut être détecté par métagénomique virale (Delwart, 2007; Edwards & Rohwer, 2005). Les premières études de métagénomique virale se basaient sur le clonage des fragments de séquences virales puis sur leur séquençage par la méthode de Sanger. L'avènement du séquençage à haut débit, permettant une étude exhaustive des viromes, a révolutionné la métagénomique virale (Mokili *et al.*, 2012; Rosario et Breitbart, 2011).

Méthodes de préparation et d'analyse des viromes

Les études de métagénomique virale comportent trois étapes majeures : (1) la préparation des échantillons, (2) le séquençage à haut débit, et (3) l'analyse bioinformatique des viromes.

1- Préparation des échantillons

Théoriquement, tout type d'échantillon peut être analysé par métagénomique virale. Cependant, contrairement à d'autres groupes d'organismes, il n'existe pas de gène commun conservé chez l'ensemble des virus pouvant être utilisé comme cible pour leur amplification (Edwards et Rohwer, 2005). De plus, les génomes viraux, de petite taille, sont souvent noyés dans la masse des génomes cellulaires présents dans les échantillons traités. Il est donc nécessaire d'éliminer au maximum les acides nucléiques non-viraux et d'amplifier les acides nucléiques de manière aléatoire afin d'obtenir des viromes représentatifs des communautés virales présentes dans les échantillons testés.

Purification des particules virales

Afin de concentrer les particules virales présentes dans les prélèvements traités, des étapes d'homogénéisation, de filtration et d'ultracentrifugation sont souvent nécessaires. L'étape de filtration est particulièrement importante. En effet, une des propriétés des virus permettant de les distinguer des organismes cellulaires est leur petite taille. Ainsi, la technique la plus couramment utilisée pour éliminer les cellules est une filtration à travers des filtres de 0,22 ou 0,45 µm de diamètre (Thurber *et al.*, 2009). Cependant, il existe des virus d'une taille comparable à celle de bactéries. Pour ces virus géants, la technique de filtration est inadaptée (Halary *et al.*, 2016). À l'inverse, il est possible que des acides nucléiques d'origine cellulaire traversent les filtres. Il a été cité que les agents de transfert de gènes (GTA), dont la structure ressemble à celle des bactériophages et dont le rôle est le transfert de gènes entre bactéries, pourraient être à l'origine de certaines séquences bactériennes retrouvées dans les viromes (Kristensen *et al.*, 2010; Lang *et al.*, 2012).

Dans un second temps, des traitements via des nucléases (DNases et RNases) permettent de réduire la composition des échantillons en acides nucléiques non-encapsidés présents après l'étape de filtration (Thurber *et al.*, 2009). Les acides nucléiques viraux, protégés par des capsides, restent relativement indemnes (Hall *et al.*, 2014).

Amplification des acides nucléiques viraux

Les acides nucléiques contenus dans les particules virales purifiées sont ensuite extraits. Différentes méthodes permettent l'extraction conjointe d'ADN et d'ARN, l'extraction d'ADN ou d'ARN (Hayes *et al.*, 2017; Thurber *et al.*, 2009).

Après extraction, et rétro-transcription (conduisant à la formation d'ADN, nommé ADNc, à partir d'ARN) suivie ou non de la synthèse du brin complémentaire de l'ADNc par l'utilisation du fragment de Klenow dans le cas des ARN, les acides nucléiques viraux sont amplifiés. Les méthodes d'amplification aléatoires les plus fréquemment utilisées sont (i) l'amplification par déplacement multiple de brins (MDA), (ii) l'amplification du génome entier (WGA) réalisés via la polymérase du bactériophage Phi29, (iii) des dérivés de la PCR aléatoire en utilisant des adaptateurs ou par tagmentation (Brum et Sullivan, 2015; Candresse *et al.*, 2014; Edwards et Rohwer, 2005; Kozarewa *et al.*, 2015; Roossinck *et al.*, 2010).

Après une étape de purification, les produits issus de l'amplification sont séquencés.

2- Séquençage à haut débit

Différentes techniques de séquençage à haut débit (ou HTS) sont utilisables. Elles connaissent une évolution très rapide depuis leur apparition il y a une dizaine d'années. Alors que ces techniques permettaient initialement d'obtenir des centaines de milliers de séquences, elles permettent actuellement d'obtenir plusieurs centaines de millions de séquences. Actuellement, les plateformes de séquençage les plus fréquemment utilisées sont Illumina, Ion Torrent, Pacific Biosystems et SOLID (Genohub-ngs-instrument-guide). Chaque plateforme possède des spécificités de séquençages (et des prix) différents, produisant des masses de données variant en termes de quantité et de qualité. Notamment, chaque technologie de séquençage induit différents taux d'erreurs et des séquences de taille et de nombre variables. Par exemple, la technologie HiSeq 2000 (plateforme Illumina) génère jusqu'à 6 milliards de séquences de 100 bases de longueur avec un taux d'erreur de 0,002% d'insertions-suppressions (indels) ; tandis que la technologie PGM (Plateforme Ion Torrent) génère 2 millions de reads de 120 bases avec 1,5% d'insertions-suppressions (McElroy *et al.*, 2014). La technologie de séquençage est donc choisie en fonction du type de données que l'on souhaite générer.

3- Traitement bioinformatique des données

Les données brutes produites par séquençage à haut débit nécessitent de nombreux traitements afin de permettre *in fine* de déterminer quels virus sont présents dans les échantillons et d'aller jusqu'à prédire les fonctions de leurs gènes. Des outils bioinformatiques ont été mis en place dans ce but. Ils évoluent constamment pour s'adapter à l'évolution des technologies de séquençage, notamment à l'augmentation exponentielle des données qu'elles produisent.

Nettoyage de données

Les technologies de séquençage à haut débit génèrent des lectures de fragments nucléotidiques, nommés reads. Or, comme énoncé ci-dessus, ces reads sont susceptibles de contenir des amorces ainsi que différents taux d'erreurs de séquençage. Les données issues du séquençage sont donc dans un premier temps soumises à des filtres de qualité permettant d'éliminer les adaptateurs de séquençage et de ne garder que les séquences qui se situent au-dessus d'un certain seuil de propreté (McElroy *et al.*, 2014; Oulas *et al.*, 2015).

Assemblage de novo

L'assemblage *de novo* permet de rassembler des reads chevauchants provenant théoriquement du même génome en une seule séquence contigüe (nommée contig). L'assemblage génère donc des contigs dont la longueur est généralement plus longue que celles des reads, et permet parfois d'obtenir des génomes entiers (Fancello, Raoult, & Desnues, 2012). Cependant, l'assemblage peut générer des chimères : en effet, il n'est souvent pas possible de déterminer si deux reads assemblés dans le même contig proviennent du même génome viral ou de deux génomes viraux différents (Charuvaka et Rangwala, 2011).

Attribution taxonomique

Dans la majorité des études de métagénomique virales, les contigs ou les reads sont comparés avec les séquences présentes dans des bases de données, par exemple avec la base de données publiques GenBank, afin de les attribuer taxonomiquement. L'outil d'attribution taxonomique le plus régulièrement utilisé pour ce faire est le « Basic Local Alignment Search Tool » (BLAST), mais d'autres outils existent (Bruder *et al.*, 2016; Fancello *et al.*, 2012).

Apports de la métagénomique virale

Près de quinze ans après sa première utilisation, la métagénomique virale est une sous-discipline scientifique en plein essor (**Fig. 8**). De nombreux types d'échantillons ont ainsi été analysés, incluant notamment de l'eau (Roux *et al.*, 2016), des prélèvements de sol (Han *et al.*, 2017), de glaciers (Zablocki *et al.*, 2014), du corail (Sweet et Bythell, 2017), des plantes (Palanga *et al.*, 2016; Roossinck *et al.*, 2010), des arthropodes (Ng, *et al.*, 2011; Tokarz *et al.*, 2014) et des échantillons de sang, de tissus ou de fèces provenant de vertébrés (Delwart, 2013).

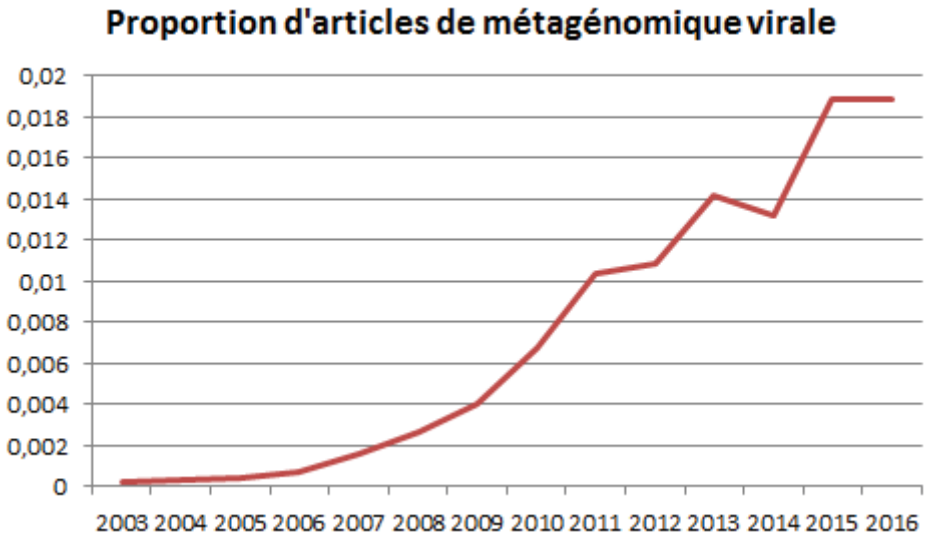


Figure 8 : Proportion d'articles traitant de métagénomique virale (en pourcentage) parmi l'ensemble des articles scientifiques référencés sur PubMed. Ce graphique a été construit en utilisant les mots clé suivants : (virus OR viral OR virome) AND (metagenome OR metagenomics).

Les études comparatives menées à large échelle ont permis l'apport de connaissances inédites sur la distribution des communautés virales. Une métaanalyse de plus de 3 200 métagénomes a révélé une forte spécificité du type d'habitat (caractérisés notamment par des différences de profondeur de la colonne d'eau ou de distance au rivage dans le cas d'écosystèmes marins) pour la grande majorité des virus, mais a également identifié quelques taxa viraux cosmopolites (**Fig. 9**) (Paez-Espino *et al.*, 2016). Ces observations, en accord avec

celles d'autres études se basant sur la métagénomique virale, sont explicables par la distribution des hôtes qui est elle-même explicable par leur potentiel adaptatif aux différents types de niches écologiques (Bettarel *et al.*, 2011; Brum *et al.*, 2015; Roux *et al.*, 2016).

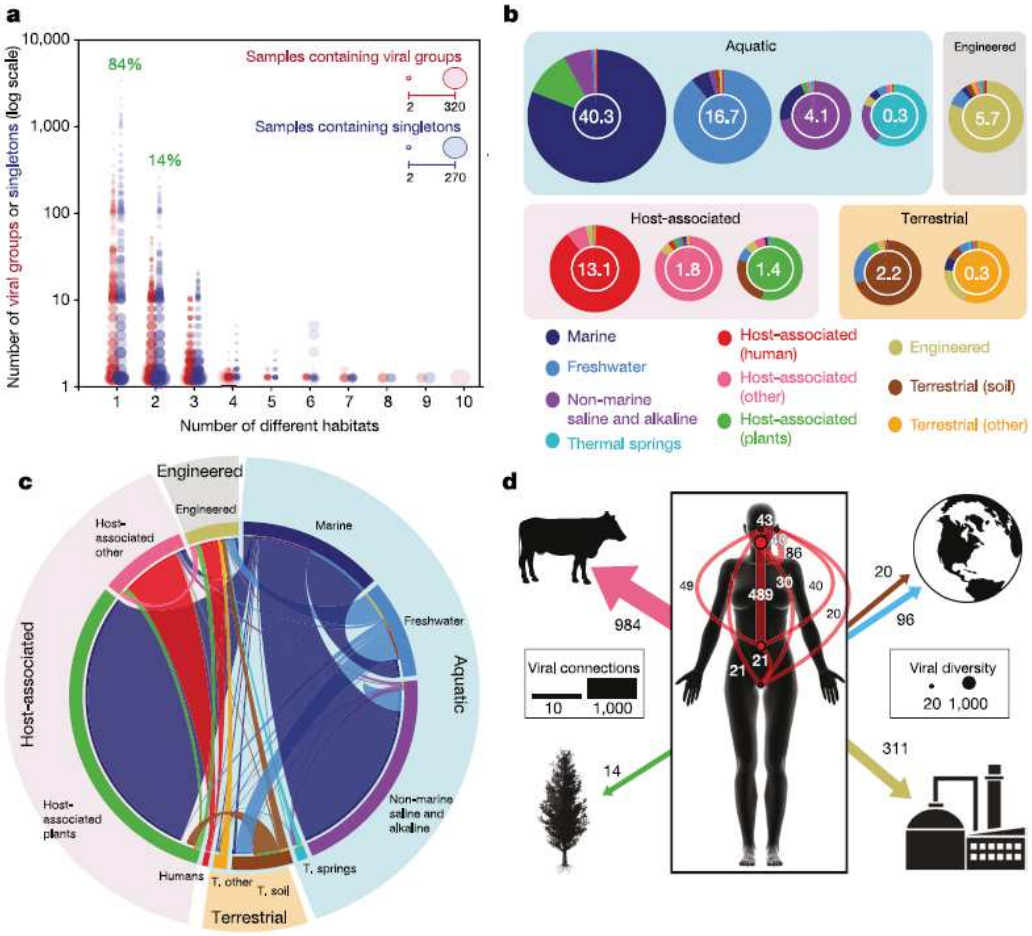


Figure 9 : Distribution de l'habitat des virus issus de métagénomes. **A** : Nombre d'habitats dans lesquels sont distribués différents groupes viraux (cercles rouges) et singletons viraux (cercles bleus). La taille des cercles représente le nombre d'échantillons contenant les séquences virales. **B** : Distribution des séquences virales par types d'habitats et par catégorie environnementale. La taille des camemberts et le nombre à l'intérieur représentent des groupes et singletons viraux par type d'habitat qui est divisé en quatre catégories environnementales ; les valeurs sont données en unités de 10^3. **C** : Connexions entre les types d'habitats basées sur les séquences virales qu'ils partagent. **D** : Répartition des séquences virales selon les différents sites du corps humains et d'autres types d'habitats. Les points rouges indiquent les groupes et les singletons viraux trouvés exclusivement sur un seul site du corps humain. Tirée de Paez-Espino *et al.*, 2016.

En outre, cette analyse montre que de nombreux groupes de virus ont été trouvés dans des niches écologiques similaires séparées par de grandes distances géographiques (jusqu'à plusieurs milliers de kilomètres), ce qui est en accord avec des études antérieures suggérant que certains virus sont passivement transportés par les courants océaniques (**Fig. 10**) (Brum *et al.*, 2015). Ce pattern a également été observé pour des écosystèmes isolés, tels que les lacs et les sols associés aux plantes, où le mode de dispersion des particules virales est moins évident à identifier. Cependant, ce pattern de large distribution n'est pas généralisable, certaines espèces virales n'ayant été retrouvées que dans des zones géographiques restreintes (**Fig. 10**). De manière parallèle, chez l'homme, la répartition virale montre également une spécificité du site du corps avec seulement quelques espèces virales trouvées conjointement dans les échantillons fécaux et oraux. En outre, environ 17% et 9% des espèces virales respectivement intestinales et orales, sont propres à chaque individu, le reste de ces espèces étant partagé par plusieurs individus (Paez-Espino *et al.*, 2016).

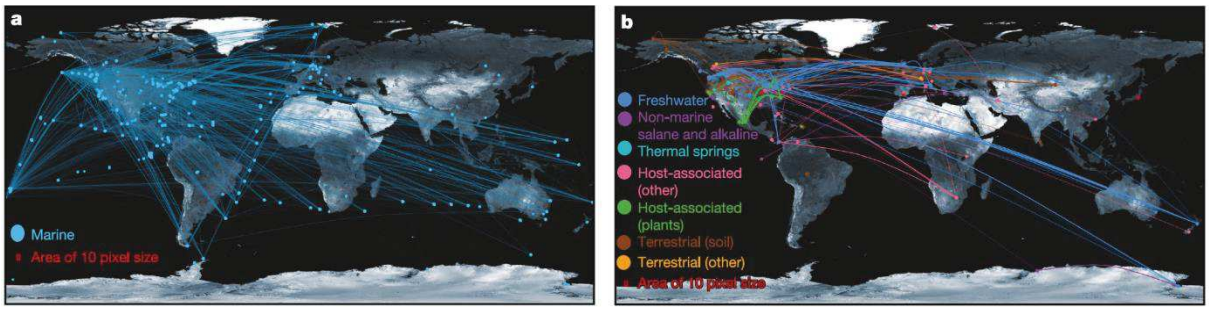


Figure 10 : Distribution mondiale de la diversité virale. La présence des mêmes groupes ou singletons viraux entre les échantillons (cercles) est représentée par des lignes les connectant. Seuls les échantillons partageant deux ou plus de deux groupes ou singletons viraux sont présentés. Les couleurs des cercles indiquent les types d'habitats. **A** : Connexions entre échantillons marins. La transparence des lignes reflète le nombre de groupes viraux partagés. **B** : Connexions entre des échantillons non marins provenant du même type d'habitat. Tirée de Paez-Espino *et al.*, 2016.

De plus, les études menées en métagénomique virale ont mis en lumière une importante diversité de virus et de gènes viraux, de nombreux viromes contenant une proportion importante de virus possédant de faibles similarités avec les virus auparavant connus (Rosario & Breitbart, 2011). La densification des arbres phylogénétiques qui résultent de l'ajout de leurs séquences virales permet d'inférer avec d'avantage de fiabilité les liens évolutifs entre les taxa viraux, et donc d'améliorer substantiellement nos connaissances sur l'évolution virale (Simmonds *et al.*, 2017). La caractérisation de viromes provenant d'échantillons anciens, bien que rarement étudiée, a également permis d'améliorer notre compréhension de l'évolution virale. Par exemple, l'analyse du virome de fèces humaines datant de près de 700 ans a permis de mettre en évidence la présence de bactériophages dont la diversité des fonctions métaboliques est similaire à celle actuelle, contenant déjà des gènes de résistance aux antibiotiques (Appelt *et al.*, 2014). Autre exemple, l'analyse de viromes de fèces de caribous datant de la même période a permis de découvrir deux virus de plantes, liés de manière distante à des virus connus. L'un d'entre eux a été reconstitué dans un clone infectieux et est capable d'infecter des plants de tabac (*Nicotiana benthamiana*) (Fei *et al.*, 2014).

En outre, ces études ont permis d'améliorer de manière substantielle notre conception de l'importance des virus sur la régulation et l'évolution de leurs populations d'hôtes ainsi que dans le fonctionnement de certains écosystèmes (Brum *et al.*, 2015; Roux *et al.*, 2016). Enfin, dans le cadre de la pathologie, les travaux mené en métagénomique virale ont débouché sur la découverte de nouveaux virus pathogènes, et ont permis des avancées sur le plan du diagnostic et de l'épidémiologie (Chiu, 2013; Lipkin, 2013).

La métagénomique virale a ainsi permis des avancées majeures dans notre compréhension de la diversité du monde viral et de son impact sur les organismes cellulaires. Cependant, cette diversité est loin d'être entièrement explorée. En effet, les études de diversité virale se sont principalement basées sur des échantillonnages d'écosystèmes ou d'hôtes réalisés de façon unique. D'autres études ont été réalisées à de plus ou moins grandes échelles spatio-temporelles, se basant sur des échantillons récupérés sur deux lacs (Roux *et al.*, 2012) à ceux récupérés à partir de milliers d'échantillonnages récupérés à des échelles spatiales mondiales, dont ceux obtenus lors des expéditions TARA océans (Brum *et al.*, 2015). Ainsi, les environnements et les hôtes les plus échantillonnés, i.e. marins et humains, sont ceux dont les communautés virales sont le mieux caractérisées.

III- Les arthropodes : un modèle de choix pour étudier la diversité virale

Impacts des arthropodes

Les arthropodes sont apparus il y a plus de 450 millions d'années. Ils ont été parmi les premiers organismes à coloniser les écosystèmes aquatiques et terrestres (Misof *et al.*, 2014). Actuellement, le phylum des arthropodes, comprenant insectes, crustacés, arachnides et myriapodes, regroupe les espèces animales les plus diverses sur Terre, avec plus d'un million d'espèces décrites (Chapman, 2009; Mora *et al.*, 2011). Cinq ordres d'insectes regroupent près de 80% des espèces d'arthropodes recensées : les Coléoptères, les Lépidoptères, les Diptères, les Hyménoptères et les Hémiptères (Chapman, 2009).

Les arthropodes ont un impact important dans une pléthore de processus biologiques et écologiques, comme la pollinisation, la décomposition de la matière organique, et les chaines trophiques. Des centaines d'espèces d'arthropodes sont élevées pour la consommation humaine, comme les crevettes et les vers de farine, ou pour la production de produits comme le miel ou la soie. Enfin, le venin de certains arthropodes est utilisé en pharmacologie**.** Moins de 0,5% du nombre total d'espèces d'arthropodes sont considérées comme étant nuisibles, i.e. dont l'activité est considérée comme négative pour l'homme, ses animaux d'élevage ou ses cultures (FAO, 2002).

Certaines espèces d'arthropodes sont parasites d'animaux, notamment les arthropodes hématophages comme les moustiques (*Aedes sp.*, *Culex sp.*, *Anopheles sp.*, Diptère), les tiques (*Ixodes sp.*, *Rhipicephalus sp.*, Ixodida), les glossines (*Glossina sp.*, Diptère), les réduves (*Triatoma sp.*, Hémiptère) ou les poux (*Pediculus humanus*, Phthiraptera). Ces parasites ont des impacts négatifs directs sur leurs hôtes de par le coût associé au parasitisme. Mais leurs plus forts impacts sont dus à leur rôle de vecteurs d'agents pathogènes. Parmi les maladies vectorielles les plus connues figurent la peste, le paludisme, la maladie du sommeil, les filarioses, la leishmaniose et la maladie de Lyme. Ces arthropodes sont également vecteurs de maladies virales, comme la dengue, la fièvre jaune et la maladie de la langue bleue. Les maladies vectorielles ont été – et restent encore - une des grandes causes de mortalité et de morbidité chez les humains. Près de 17% de l'ensemble des maladies infectieuses humaines sont d'origine vectorielle, la grande majorité d'entre elles étant transmises par des

moustiques. La maladie vectorielle possédant le taux mortalité le plus élevé est le paludisme, qui infecte 216 millions de personnes par an et qui a causé près de 627 000 morts en 2012. Actuellement, les maladies vectorielles humaines sont présentes dans plus de cent pays, et sont prédominantes dans les pays en voie de développement : en Afrique, elles sont responsables de 500 à 2000 morts par million d'habitants (McGraw et O'Neill, 2013).

Enfin, certaines espèces d'arthropodes phytophages induisent des pertes de rendement importantes chez les plantes cultivées, de par l'impact direct de l'herbivorie, ou indirectement dû à leur rôle de vecteurs d'agents phytopathogènes. La création d'habitats manipulés par l'homme correspondant à ses besoins, les agroécosystèmes, où les cultures regroupées en grandes densités et sur de grandes surfaces sont sélectionnées pour un rendement et une valeur nutritive élevés, fournit également des environnements propices aux pullulations d'insectes herbivores de par le déséquilibre créé par la forte densité et la faible diversité des cultures. Dans le processus de sélection artificielle de cultures appropriées à la consommation humaine, des plantes hautement sensibles à l'infestation par des insectes et des agents pathogènes transmis par ceux-ci sont sélectionnées. Par exemple, le virus de la maladie bronzée de la tomate, (Tomato spotted wilt virus, *Bunyaviridae*) transmis par des thrips (Frankliniella occidentalis, Thysanoptère), a été responsable de pertes de rendements estimées à plus d'un milliard de dollars dans les années 1990 (Scholthof *et al.*, 2011). Les insectes phytophages sont responsables de la perte de près d'1/5 de la production agricole annuelle totale (FAO, 2002).

Le changement climatique ainsi que les échanges mondiaux redistribuent les arthropodes nuisibles dans le monde entier avec des conséquences imprévues. Le réchauffement climatique a entrainé l'expansion de l'aire de répartition géographique de certaines espèces de tiques en Europe (*Ixodes sp.*, Ixodida), ce qui a conduit à l'émergence de certaines maladies humaines transmises par ces tiques, comme l'encéphalite à tiques, due à un virus (virus de l'encéphalite à tiques, *Flaviviridae*) (Heinz *et al.*, 2015; Lukan *et al.*, 2010; Ostfeld et Brunner, 2015). Autre exemple, deux introductions dues au commerce international des denrées alimentaires suscitent des inquiétudes concernant les rendements agricoles en Europe et en Afrique: celle du moucheron asiatique (*Drosophila suzukii*, Diptère) et celle de la noctuelle américaine du maïs (*Spodoptera frugiperda*, Lépidoptère), causant respectivement de graves dommages aux fruits et aux cultures (Lee *et al.*, 2011; Wild, 2017).

Virus et lutte biologique

L'utilisation massive de produits chimiques dans le cadre du contrôle des populations d'arthropodes ravageurs de cultures est expliquée par leur rapidité d'action. Cependant, les pesticides posent également de nombreux problèmes dus à leur toxicité, comme la perte de biodiversité et l'impact sur la santé humaine. De plus, leur durée d'utilisation est limitée par la sélection de populations d'arthropodes résistantes (Tilman *et al.*, 2002). Aujourd'hui, l'usage des pesticides est donc limité et l'utilisation de nombreuses molécules a été - et sera - interdite en Europe (Ecophyto 2018 in France, European directives 2009/128/CE). Par conséquent, il existe un besoin urgent de solutions alternatives et durables pour contrôler les populations d'arthropodes ravageurs de cultures.

La lutte biologique, consiste à contrôler les populations d'espèces dites « nuisibles » par l'utilisation de leurs ennemis naturels, qu'il s'agisse de prédateurs ou de parasites, ou de leurs dérivés (composés moléculaires). Elle constitue une alternative à l'utilisation des pesticides pour contrôler les populations d'arthropodes ravageurs des cultures. En effet, les auxiliaires de lutte biologique sont plus sélectifs que les pesticides, ils ont donc des impacts moins forts sur la structuration des communautés des agroécosystèmes ainsi que sur la santé humaine (Lacey, Frutos, Kaya, & Vail, 2001). La lutte biologique a été utilisée la première fois durant l'Égypte ancienne de par l'introduction de chats pour réguler les populations de rongeurs infestant les stocks alimentaires. En 1868, une coccinelle (*Rodolia cardinalis*, Coléoptère) fut introduite dans des vergers d'agrumes de Californie pour contrôler les populations de la cochenille australienne (*Icerya purchasi*, Hémiptère) (Greathead, 1995). De nos jours, les toxines produites par la bactérie *Bacillus thuringiensis* (Bt) représentent les produits les plus utilisés en lutte biologique. Commercialisées depuis 1972, ils représentaient, à la fin du XXème siècle, 2% du marché des insecticides aux États-Unis (Lacey *et al.*, 2015).

Le but de la lutte biologique est ainsi de recréer un équilibre relatif entre les populations d'espèces nuisibles et leurs populations d'antagonistes (OILB-SROP, 1973). Cependant, afin de mettre en place des stratégies de lutte biologique, il est nécessaire d'acquérir une connaissance précise et approfondie non seulement des populations d'arthropodes ravageurs, mais également de celles des auxiliaires et de leurs impacts sur les populations de ravageurs et sur les autres compartiments des écosystèmes (Lacey *et al.*, 2015). Cependant, nos connaissances sur l'écologie des réseaux trophiques ont jusqu'à récemment été centrées sur les macroorganismes, comme le sont les prédateurs, tandis que les

microorganismes, ainsi que les virus, ont été sous-étudiés (Horner-Devine, Carney, & Bohannan, 2004; Horner-Devine & Bohannan, 2006).

Les virus entomopathogènes représentent des ressources de lutte biologique largement inexplorées. En effet, une seule famille virale, celle des *Baculoviridae*, est à ce jour commercialisée en lutte biologique (Lacey *et al.*, 2015). Cependant, ces virus montrent déjà leurs limites dans la mesure où des résistances sont apparues dans des populations d'insectes cibles. Ces faits mettent en évidence le besoin d'explorer de nouvelles ressources virales pathogènes d'arthropodes ravageurs de cultures, ce qui passe par une étape préliminaire d'inventaire de ces virus (Lacey *et al.*, 2015).

Les virus associés aux arthropodes : une diversité largement méconnue

Les arthropodes représentent un bon modèle pour étudier la diversité virale : en plus de regrouper les macroorganismes les plus diversifiés sur Terre et d'être ubiquitaires dans l'environnement, ils seraient susceptibles d'abriter une grande diversité de virus. En effet, les arthropodes interagissent activement avec une grande diversité d'organismes comme les plantes, les champignons et les vertébrés, et peuvent donc agir comme source ou puits de virus dans les écosystèmes (Gall, 2015).

En accord avec cette hypothèse, des études menées en métagénomique virale ont mis en évidence une grande diversité de familles virales chez les arthropodes (**Tab. 1**). En effet, ces études montrent la présence de familles virales présentes conjointement chez les arthropodes et les vertébrés, comme celles des *Parvoviridae*, *Circoviridae*, *Reoviridae*, *Iridoviridae*, *Poxviridae*, ou *Nodaviridae*. Elles ont également mis en évidence une grande diversité de virus infectant uniquement les arthropodes, ces communautés virales étant dominées par les familles des *Baculoviridae*, *Iflaviridae* et *Dicistroviridae*, alors que les *Mesoniviridae*, *Bidnaviridae*, *Ascoviridae*, ainsi que les virus appartenant aux genres *Negevirus* et *Sinaivirus* sont moins présents. Il est également à noter que les viromes de moustiques et de tiques contiennent des représentants de l'ordre des *Bunyavirales* et de la famille des *Flaviviridae*, deux niveaux phylogénétiques au sein desquels on trouve des virus transmis aux vertébrés par ces vecteurs (les arbovirus). Les moustiques et les tiques contiennent également des virus inféodés aux vertébrés, ce qui est dû au régime alimentaire de ces arthropodes. La majorité de ces études font également état de la présence de

bactériophages appartenant aux familles des *Myoviridae*, *Podoviridae*, *Siphoviridae* et *Microviridae* chez les arthropodes. Ces résultats sont explicables par le fait que les arthropodes abritent un microbiote lui aussi susceptible d'être infecté par des virus. Enfin, comme chez les vertébrés, des traces de phytovirus ainsi que de virus de champignons sont retrouvées chez les arthropodes, ceux les plus largement retrouvés appartenant aux *Geminiviridae*, *Partitiviridae*, *Tymoviridae*, *Virgaviridae* (**Tab 1**).

Espèce	Acides nucléiques extraits	Classification Baltimore							Référence
		ADNdb	ADNsb	ARNdb	ARNsb+	ARN sb-	ARNsb-RT	ADNdb-RT	
Copépode	ADN		Circoviridae						(Ng et al., 2013)
Crevette	ADN + ARN		Circoviridae		Nodaviridae				(Dayaram et al., 2013)
Libellule	ADN		Circoviridae						(Dunlap et al., 2013)
Drosophile	ARN	Nudiviridae		Reoviridae Partitiviridae	Iflaviridae, Nodaviridae, Permutotetraviridae, Negevirus Flaviviridae Virgaviridae, Sobemovirus	Peribunyaviridae			(Webster et al., 2015)
Drosophile	ARN			Birnaviridae Reoviridae	Dicistroviridae, Iflaviridae, Nodaviridae, Negevirus Secoviridae, Tymoviridae	Rhabdoviridae			(Webster et al., 2016)
Abeille	ADN + ARN				Dicistroviridae, Sinaivirus Secoviridae, Virgaviridae				(Granberg et al., 2013)
Abeille	ARN				Dicistroviridae, Iflaviridae, Sinaivirus	Rhabdoviridae			(Levin et al., 2016)
Acarien parasite d'abeilles	ARN	Baculoviridae	Circoviridae		Dicistroviridae, Iflaviridae				
Aleurode	ADN		Geminiviridae, Satellites						(Ng et al., 2011)
Aleurode	ADN + ARN		Geminiviridae		Betaflexiviridae, Bromoviridae, Closteroviridae, Virgaviridae	Phenuiviridae			(Rosario et al., 2014)
Aleurode	ADN		Satellites						(Rosario et al., 2015)
Moucheron hématophage	ARN	Faustoviridae, Mimiviridae Siphoviridae		Picobirnaviridae Reoviridae Partitiviridae	Dicistroviridae, Iflaviridae, Mesoniviridae Tymoviridae Flaviviridae	Orthomyxoviridae Rhabdoviridae	Retroviridae		(Temmam et al., 2016)
Moustique	ADN	Herpesviridae, Papillomaviridae Iridoviridae, Poxviridae Myoviridae, Podoviridae, Siphoviridae	Geminiviridae, Nanoviridae Parvoviridae (Densovirinae) Anelloviridae Circoviridae						(Ng et al., 2011b)
Moustique	ADN+ ARN			Reoviridae	Flaviviridae, Togaviridae	Orthomyxoviridae Rhabdoviridae			(Coffey et al., 2014)
Moustique	ADN + ARN			Reoviridae					(Li et al., 2014)
Moustique	ADN + ARN	Myoviridae, Podoviridae, Siphoviridae	Parvoviridae (Densovirinae) Anelloviridae, Parvoviridae (Parvovirinae)	Chrysoviridae	Luteoviridae, Virgaviridae Dicistroviridae, Iflaviridae	Rhabdoviridae			(Shi et al., 2015)
Moustique	ARN				Narnaviridae	Phenuiviridae Rhabdoviridae			(Chandler et al., 2015)
Moustique	ARN	Ascoviridae, Baculoviridae, Nudiviridae, Polydnaviridae Myoviridae, Podoviridae, Siphoviridae	Parvoviridae (Densovirinae) Satellites	Birnaviridae Reoviridae Partitiviridae Hypoviridae	Picornaviridae Dicistroviridae, Iflaviridae Flaviviridae, Nodaviridae Bromoviridae, Luteoviridae, Secoviridae,Tombusviridae, Tymoviridae, Virgaviridae, Ourmiavirus, Sobemovirus	Nyamiviridae, Orthomyxoviridae, Phenuiviridae Rhabdoviridae	Retroviridae		(Frey et al., 2016)
Moustique	ARN					Orthobunyaviridae			(Frey et al., 2016)
Tique	ADN + ARN				Alphatetraviridae Flaviviridae Sobemovirus	Nairoviridae, Nyamiviridae, Phenuiviridae			(Tokarz et al., 2014)
Tique	ADN + ARN	Ascoviridae, Baculoviridae, Polydnaviridae Iridoviridae Mimiviridae, Phycodnaviridae Myoviridae, Podoviridae, Siphoviridae	Anelloviridae Bidnaviridae Circoviridae, Parvoviridae Nanoviridae Microviridae		Picornaviridae Virgaviridae	Rhabdoviridae, Bunyavirales	Retroviridae	Caulimoviridae	(Xia et al., 2015)
Tique	ADN + ARN	Herpesviridae, Papillomaviridae Baculoviridae, Polydnaviridae Iridoviridae Phycodnaviridae Podoviridae	Circoviridae, Parvoviridae	Partitiviridae Hypoviridae	Coronaviridae, Picornaviridae Tetraviridae Flaviviridae Potyviridae	Filoviridae Nairoviridae, Phenuiviridae Rhabdoviridae	Retroviridae		(Sakamoto et al., 2016)
Tique	ARN	Alloherpesviridae ,Herpesviridae, Polyomaviridae Ascoviridae, Baculoviridae, Polydnaviridae Mimiviridae	Parvoviridae (Parvovirinae)	Reoviridae Partitiviridae	Picornaviridae Iflaviridae Flaviviridae, Nodaviridae, Togaviridae Closteroviridae, Luteoviridae, Potyviridae, Tymoviridae,Ourmiavirus, Sobemovirus	Hantaviridae Nairoviridae, Phenuiviridae, Tospoviridae Rhabdoviridae			(Moutailler et al., 2016)

Tableau 1 : Tableau récapitulatif de vingt-deux articles de recherche portant sur des viromes d'arthropodes obtenus par métagénomique virale. Les couleurs désignent le spectre d'hôte des familles virales trouvées dans les viromes : bleu clair : archées ; bleu foncé : bactéries ; vert : plantes ; jaune : arthropodes ; rose : vertébrés ; orange : autres eucaryotes.

Cependant, la diversité des virus infectant les arthropodes est sous-explorée et donc sous-évaluée. En effet, alors qu'ils représentent la majorité des espèces animales décrites, ce qui prédit une très grande diversité des virus pouvant les infecter, seulement 10% des espèces virales décrites infectent les arthropodes (**Fig. 11**). Cette observation est expliquée par le fait que de plus grands efforts de recherche de virus sont investis chez les mammifères et les oiseaux. En outre, la majorité des études de diversité virale associée aux arthropodes a été réalisée chez des espèces hématophages (Junglen et Drosten, 2013) ainsi que chez une espèce d'aleurode vectrice de virus phythopathogènes (*Bemisia tabacci*, Hémiptère) (Ng, *et al.*, 2011; Rosario *et al.*, 2015), à cause de l'impact économique ou médical que peuvent représenter les virus transmis par ces arthropodes. Les arbovirus sont donc les virus d'arthropodes les plus étudiés, le reste de nos connaissances étant encore largement restreintes à des virus ayant causé des épizooties (Junglen et Drosten, 2013).

Figure 11 : Proportion d'espèces de macroorganismes décrites et proportion d'espèces virales décrites associées à ces macroorganismes. En jaune : arthropodes ; en rouge : vertébrés ; en violet : champignons et en vert : plantes. Ce graphique a été construit d'après Chapman, 2009 et d'après Virus-Host Database, 2017.

En 2017, 39 familles de virus infectant les arthropodes sont répertoriées (**Fig. 7**), certains de leurs virus ne sont pas classifiés et beaucoup restent à découvrir si l'on considère la diversité des arthropodes (**Fig. 11**). Dans ce sens, de récentes études de métagénomique virale, portant sur des d'arthropodes et du guano de chauves-souris insectivores, ont permis de

découvrir plus de 1500 nouvelles espèces de virus, dont certains appartiennent à 16 potentielles nouvelles familles virales (Li *et al.*, 2015; Shi *et al.*, 2016; Wu *et al.*, 2012).

Il est donc à prévoir que les études de diversité virale réalisées sur des arthropodes éclaireraient un large pan de la diversité ainsi que de l'évolution virale qui restent encore inexplorées. De manière plus appliquée, ces apports pourraient avoir des répercussions dans la surveillance des virus pathogènes. En effet, surveiller les communautés virales d'arthropodes vecteurs de zoonoses ou de maladies phytopathogènes fournirait une stratégie préventive pour identifier ou tester la présence de ces virus dans les écosystèmes, et donc limiter leurs impacts dans le domaine de la santé humaine, animale et végétale (Rosario *et al.*, 2015; Temmam *et al.*, 2014). Enfin, mener ce genre d'études permettrait d'identifier des virus entomopathogènes éventuellement utilisables en tant qu'agents de lutte biologique contre des arthropodes nuisibles (Levin *et al.*, 2016).

IV- Objectifs de la thèse

Ce travail de doctorat s'est constitué autour de l'étude de la diversité virale associée aux arthropodes dans le but de mieux caractériser leur diversité génétique, depuis l'échelle des communautés jusqu'à l'analyse de génomes spécifiques.

Le premier chapitre traitera de la caractérisation *in silico* de la diversité génétique et du spectre d'hôte potentiel d'une famille de virus connue pour infecter arthropodes, vertébrés et échinodermes : celle des *Parvoviridae*. Cette caractérisation sera réalisée *in silico* par des fouilles de bases de données publiques ainsi que de jeux de données privés, dans le but d'améliorer notre compréhension de l'évolution de cette famille virale.

Le second chapitre sera consacré au développement d'un protocole permettant (i) la préparation de viromes provenant d'arthropodes et (ii) leur analyse bioinformatique. Le but a été de mettre en place un protocole permettant de caractériser l'ensemble des communautés virales associées aux arthropodes tout en permettant le multiplexage d'un grand nombre d'échantillons. Lors de l'analyse bioinformatique, l'efficacité des étapes d'assemblage et de mapping (i.e. l'alignement des reads contre des génomes de référence) sur la réduction de la

proportion de séquences non-attribuées taxonomiquement (appelées « matière noire ») sera évaluée sur différents types de viromes obtenus à partir de prélèvements environnementaux ainsi que réalisés sur des hôtes.

Enfin, le troisième chapitre portera sur l'analyse des communautés virales associées à certaines espèces d'arthropodes ravageurs de cultures. La première étude portera sur la comparaison des communautés virales associées à deux populations d'élevage d'un acarien (*Tetranychus urticae*). La seconde étude analysera les viromes de trois insectes ravageurs (*Helicoverpa armigera*, *Hypera postica* et *Acyrthosiphon pisum*) ayant été échantillonnés dans deux agroécosystèmes adjacents : des champs de luzerne ainsi que des prairies. L'ensemble des viromes analysés auront été obtenus et analysés en utilisant la méthodologie expliquée dans le chapitre précédent. D'autre part, la distribution de certains virus, trouvés dans les ravageurs de cultures, sera étudiée dans les communautés d'arthropodes constitutives des agroécosystèmes échantillonnés.

Chapitre I - Les *Parvoviridae* : une famille diversifiée de virus d'animaux à ADN simple brin

Des petits virus pour des hôtes très diversifiés

La famille des *Parvoviridae* regroupe l'ensemble des virus possédant un génome constitué d'une molécule d'ADN simple brin linéaire de 3,7 à 6,3 kilobases (kb) de longueur, non segmentée, et dont les extrémités sont composées de régions terminales inversées (ITRs) formant des structures secondaires de complexité variable (Bergoin et Tijssen, 2010 ; Cotmore *et al.*, 2014). Leurs virions sont non-enveloppés et possèdent une capside de forme icosaédrique qui mesure entre 18 et 28 nm de diamètre. Le génome des membres de la famille des *Parvoviridae* code pour deux types de protéines: les protéines non structurales (NS) qui permettent la réplication virale, et les protéines structurales (VP) qui forment la capside (Cotmore *et al.*, 2014; Cotmore et Tattersall, 2014)

Cette famille est divisée en deux sous-familles basées sur le spectre d'hôte : la sous famille des *Parvovirinae* regroupe tous les *Parvoviridae* infectant des vertébrés (mammifères et oiseaux), tandis que celle des *Densovirinae* regroupe ceux infectant des invertébrés, essentiellement des arthropodes (crustacés et insectes). Les *Parvovirinae* sont divisés en huit genres ; tandis que cinq genres de *Densovirinae* sont actuellement reconnus par l'ICTV (Cotmore *et al.*, 2014) (**Annexe « les densovirus, une massive attaque chez les arthropodes »**).

Cependant, il existe une différence flagrante quant au nombre d'espèces virales répertoriées entre ces deux sous-familles. En effet, 41 espèces de *Parvovirinae* sont actuellement reconnues par l'ICTV, contre 21 espèces seulement chez les *Densovirinae* (ICTV, 2017). Cette différence serait due au fait que les *Parvovirinae* soient bien plus étudiés que les *Densovirinae* du fait de la nature de leurs hôtes (Vertébrés). En effet, les *Densovirinae* ont été décrits essentiellement dans des arthropodes d'intérêt économique, médical ou vétérinaire, à savoir des noctuelles (El-Far *et al.*, 2012), des pucerons (Ryabov *et al.*, 2009), des blattes (Mukha et Schal, 2003), des moustiques (Ma *et al.*, 2011), et des crevettes (Kaufmann *et al.*, 2010), suggérant d'avantage un biais d'échantillonnage qu'une réalité biologique (Bergoin et Tijssen, 2010). Par exemple, de nouvelles espèces de *Densovirinae* infectant des oursins et des étoiles de mer, chez lesquels elles sont associées à des mortalités

massives, ont été récemment découvertes (Gudenkauf *et al.*, 2013 ; Hewson *et al.*, 2014). Notre manque de connaissance concernant la diversité des hôtes infectés par les *Parvoviridae* pose un frein à notre compréhension de l'impact de ces virus sur les animaux ainsi que de leur histoire évolutive qui remonte à au moins 98 millions d'années (Liu *et al.*, 2011).

Or, les *Parvoviridae* présentent une grande diversité dans leurs séquences et leur organisation génomique qui, mise en parallèle avec la grande diversité des animaux qu'ils infectent, laisse penser que ces virus seraient ubiquitaires dans l'environnement, et que beaucoup resteraient à découvrir (Bergoin et Tijssen, 2010). En accord avec cette hypothèse, des études menées en métagénomique virale ont mis en évidence la présence de nouvelles espèces de *Densovirinae* dans des échantillons d'eaux usées (Cantalupo *et al.*, 2011; Ng *et al.*, 2012), des fèces de chauves-souris (Ge *et al.*, 2012; He *et al.*, 2013; Li *et al.*, 2010; Wu *et al.*, 2012), ou des plantes (**François *et al.*, 2014 ; Annexe « A Novel Itera-Like Densovirus Isolated by Viral Metagenomics from the Sea Barley *Hordeum marinum* »**).

Deux approches permettraient de mieux caractériser le spectre d'hôte de la famille des *Parvoviridae*. La première consiste à échantillonner des taxa animaux chez lesquels la présence de *Parvoviridae* n'a pas été mise en évidence et à y rechercher la présence de ces virus, par exemple en utilisant des amorces dégénérées ou par métagénomique virale. Cependant, ce procédé peut s'avérer coûteux de par la nécessité d'avoir un échantillonnage d'hôtes exhaustif pour statufier de la présence ou de l'absence de ces virus chez les organismes testés.

La seconde, à moindre coût, consiste à rechercher la présence de *Parvoviridae* dans des bases de données « omiques », en partant du postulat que certains organismes utilisés dans le cadre de ce type d'études aient pu être infectés par ces virus, et qu'ils serait donc possible de trouver les traces de ces infections sous forme de séquences. En effet, de par le développement des outils « omiques », des quantités exponentielles de séquences d'acides nucléiques sont présentes dans les bases de données, ce qui augmente la probabilité d'y retrouver la présence de virus, dont des *Parvoviridae* (Radford *et al.*, 2012). De nombreux virus ont ainsi été mis en évidence dans des jeux de données de métagénomique microbienne (Paez-espino, Pavlopoulos, Ivanova, & Kyrpides, 2017; Roux, Enault, Hurwitz, & Sullivan,

2015), de transcriptomique (Clavijo *et al.*, 2016; DeBoever *et al.*, 2013) et de génomique (Bovo *et al.*, 2017; Metegnier *et al.*, 2015).

J'ai effectué une analyse de la diversité génétique ainsi que du spectre d'hôte potentiel de la famille des *Parvoviridae* par une fouille de bases de données publiques de transcriptomique et de génomique. La diversité des séquences virales - appelées ici *Parvoviridae-related* séquences (PRS) car pouvant appartenir aux *Parvoviridae* ou représenter des groupes externes de cette famille virale - ainsi que des hôtes potentiels mis en évidence lors de cette étude pourront apporter un nouvel éclairage sur l'histoire évolutive de la famille des *Parvoviridae*.

Article de recherche 1

Discovery of parvovirus-related sequences in an unexpected broad range of animals

Sarah François[1,2], Denis Filloux[3], Philippe Roumagnac[3], Diane Bigot[4], Philippe Gayral[4,5], Darren P. Martin[6], Rémy Froissart[3,7], Mylène Ogliastro[2]

[1] Université de Montpellier, UMR 1333 DGIMI « Diversité, Génomes et Interactions Microorganismes-Insectes », place Eugène-Bataillon, 34095 Montpellier cedex 5, France

[2] Laboratoire « Diversité, Génomes et Interactions Microorganismes Insectes » (DGIMI) UMR 1333, INRA, Université Montpellier, 34095 Montpellier, France.

[3] Laboratoire « Biologie et Génétique des Interactions Plante-Parasite » UMR BGPI, CIRAD-INRA-SupAgro, Campus International de Montferrier-Baillarguet, Montpellier Cedex-5, France.

[4] Institut de Recherche sur la Biologie de l'Insecte, UMR 7261, CNRS–Université François Rabelais, 37200 Tours, France.

[5] Institut des Sciences de l'Évolution UMR5554, Université Montpellier–CNRS–IRD–EPHE, 34000 Montpellier, France.

[6] Computational Biology Group, Institute of Infectious Disease and Molecular Medicine, Faculty of Health Sciences, University of Cape Town, Observatory, South Africa.

[7] Laboratoire « Maladies Infectieuses et Vecteurs : Écologie, Génétique, Évolution et Contrôle » (MIVEGEC), UMR 5290, CNRS-IRD-UM, 911 avenue Agropolis, 34394, Montpellier, France.

Publié le 7 septembre 2016 dans **Scientific Reports** (6:30880)

SCIENTIFIC REPORTS

OPEN

Discovery of parvovirus-related sequences in an unexpected broad range of animals

Received: 17 May 2016
Accepted: 11 July 2016
Published: 07 September 2016

S. François[1], D. Filloux[2], P. Roumagnac[2], D. Bigot[3], P. Gayral[3,4], D. P. Martin[5], R. Froissart[2,6] & M. Ogliastro[1]

Our knowledge of the genetic diversity and host ranges of viruses is fragmentary. This is particularly true for the *Parvoviridae* family. Genetic diversity studies of single stranded DNA viruses within this family have been largely focused on arthropod- and vertebrate-infecting species that cause diseases of humans and our domesticated animals: a focus that has biased our perception of parvovirus diversity. While metagenomics approaches could help rectify this bias, so too could transcriptomics studies. Large amounts of transcriptomic data are available for a diverse array of animal species and whenever this data has inadvertently been gathered from virus-infected individuals, it could contain detectable viral transcripts. We therefore performed a systematic search for parvovirus-related sequences (PRSs) within publicly available transcript, genome and protein databases and eleven new transcriptome datasets. This revealed 463 PRSs in the transcript databases of 118 animals. At least 41 of these PRSs are likely integrated within animal genomes in that they were also found within genomic sequence databases. Besides illuminating the ubiquity of parvoviruses, the number of parvoviral sequences discovered within public databases revealed numerous previously unknown parvovirus-host combinations; particularly in invertebrates. Our findings suggest that the host-ranges of extant parvoviruses might span the entire animal kingdom.

Recent studies have shown that viruses are the most numerous and diverse genetic entities on Earth: a discovery that has completely changed both our views on their prevalence, and our perception that they are primarily disease-causing agents[1]. Viruses have been discovered infecting organisms throughout the entire tree of life, using a wide array of strategies to move between and infect hosts belonging to either the same or different species. The genomes of many viruses can also ligate to, and become a heritable part of, the genetic material of their hosts[2].

Largely because of their obvious medical and economic importance, the vast majority of viruses that have so far been studied are those that cause recognizable diseases of humans and our domesticated plants (almost exclusively angiosperms) and animals (mainly mammals and birds)[3,4]. One of the greatest achievements of environmental metagenomics has been the discovery that unknown viral species vastly outnumber the known species, and that there probably also remain more unknown genera (and possibly also entire families) than those we have currently discovered[4,5]. For example, it is now estimated that the ~2800 virus species that are currently recognised by the International Committee on Taxonomy of Viruses[6] (ICTV probably account for less than 1% of all viral species on Earth[7]. Our rapidly expanding appreciation of the actual diversity of viruses on Earth is well illustrated in the recent discoveries of viruses with ssDNA genomes that likely belong to multitudes of novel genera/families which are both genetically highly divergent from species in the known ssDNA virus families (such as parvoviruses, circoviruses, microviruses and geminiviruses) and likely infect hosts that span the entire tree of life[8–14].

Parvoviruses illustrate the chasm that likely exists between the known diversity of species within particular virus families, and that which actually exists in all of their potential animal hosts. The linear ssDNA viruses belonging to the family *Parvoviridae* are presently divided into two sub-families: the *Parvovirinae*, which contains

[1]INRA, UMR DGIMI, F-34095, Montpellier, France. [2]CIRAD-INRA-SupAgro, UMR BGPI, Campus International de Montferrier-Baillarguet, Montpellier Cedex-5, France. [3]Institut de Recherche sur la Biologie de l'Insecte, UMR 7261, CNRS–Université François Rabelais, 37200 Tours, France. [4]UMR5554–Institut des Sciences de l'Evolution UMR5554, Université Montpellier–CNRS–IRD–EPHE, 34000 Montpellier, France. [5]Computational Biology Group, Institute of Infectious Disease and Molecular Medicine, Faculty of Health Sciences, University of Cape Town, Observatory, South Africa. [6]CNRS-IRD-UM, UMR 5290, MIVEGEC, 911 avenue Agropolis, 34394, Montpellier, France. Correspondence and requests for materials should be addressed to M.O. (email: ogliastr@supagro.inra.fr)

species infecting vertebrates (which have, to date, mostly been birds and mammals), and the *Densovirinae*, which contains species infecting arthropods[10] (which have, to date, mostly been crustaceans and insects). Parvovirus genomes are characteristically 4 to 6 kb long, with inverted terminal repeats (ITRs) that bracket two sets of genes encoding non-structural (Rep or NS) and structural (VP) proteins[15]. While degrees of sequence identity between viruses from different parvovirus genera are very low (e.g. some pairs of densoviruses share <15% VP sequence identity), most parvoviruses likely express both a NS1 protein with a super family 3 helicase (SF3) domain in the C terminus and a VP containing a unique phospholipase A2 (PLA2) motif in the N terminus[16,17]. These two proteins, even if the PLA2 motif is missing in some parvoviruses[16], are therefore useful for parvovirus phylogenetic inferences.

Our current appreciation of parvovirus diversity is limited to the 41 *Parvovirinae* and 15 *Densovirinae* species which are presently recognized by the ICTV[18]. We estimated that only a few hundred animal species have been reported as hosts of parvoviruses, representing only approximately 0.0001% of the 1.2 million animal species that have presently been described[19]. There is likely a particularly extreme imbalance of sampling between the vertebrate-infecting *Parvovirinae* and the arthropod-infecting *Densovirinae* in that there are likely almost 20 times more arthropod species on Earth than there are species of vertebrates[19]. Given the diversity in sequence and genome organisation displayed by parvoviruses, it is entirely plausible that parvoviruses are ubiquitous in the environment, and that there exist tens of thousands of undiscovered vertebrate parvoviruses, and hundreds of thousands of undiscovered arthropod parvoviruses. Indeed, high throughput sequencing technologies and the advent of routine whole genome sequencing of eukaryotes have revealed the occurrence of "fossil" parvovirus sequences integrated into the genomes of multiple animals including those of humans. While these integrated sequences reflect a history of parvovirus interactions with an unexpectedly broad range of animal hosts[20,21], the recent discovery of a densovirus associated with sea-star wasting disease supports the hypothesis that the host range of extant parvoviruses probably extends far beyond the major animal phyla in which they have already been detected[16]. This discovery also raises questions about whether densoviruses are predisposed to frequently shifting hosts, or whether they may have existed and co-evolved with their hosts during the early evolutionary radiation of multi-cellular animals[22,23].

Parvoviruses belonging to the *Dependovirus* genus may depend on another virus to complete their replication-deficient life cycle. These viruses can be persistent and include a host genome integration step that results in latent infections, as has been shown for *Adeno-Associated-Virus* (AAV) where integration requires the NS1 homologue, Rep[24]. Integration and persistence of densoviruses may in some cases even be beneficial for their hosts in that it could protect them against viral infections[25,26].

We therefore hypothesised that while transcriptome datasets might contain evidence of parvovirus sequences originating either from either *bona fide* transmissible episomal viruses, or from integrated (albeit possibly only transiently) viruses or viral sequence elements (from heritably integrated but still transcriptionally active parvovirus genes), animal genome sequences might contain evidence of heritably integrated, and possibly transcriptionally dormant, parvovirus sequence elements. Crucially, large volumes of transcriptomic and genomic sequence data for a wide array of animal species are currently available in public databases. Several studies have already revealed the presence of a variety of novel RNA and DNA virus sequences within transcriptome datasets[25–28] (including those of animal-model organisms and environmental transcriptomes). We opted to initially focus our search for novel parvoviruses within transcriptome datasets from animals covering vertebrate and invertebrate phyla, spanning the entire animal kingdom and representing an extend of in depth searches that has not been achieved so far for parvoviruses.

The results presented here highlight the extraordinary diversity, abundance and ubiquity of expressed parvoviral sequences in numerous animal phyla, revealing previously unknown parvovirus-host associations–particularly with invertebrates including arthropods, molluscs, annelids, nematodes, and cnidarians–and supporting the hypothesis that the collective host ranges of extant parvoviruses might indeed span the entire animal kingdom.

Results

Identification of parvovirus-related sequences in animal transcriptome datasets. We used 74 representative parvovirus genomes as queries (Supplementary Table S1) to perform BLASTX searches against the National Centre for Biotechnological Information (NCBI) non-redundant (nr) cDNA expressed sequence tag (EST), transcriptome shotgun assembly (TSA) and protein (Uniprot) databases[29,30]. We also used these as queries to screen, eleven new transcriptomes generated from invertebrate datasets provided by N. Galtier (European Research Council advanced grant 232971 (PopPhyl)). All hits were next selected and used as queries to perform BLASTX or BLASTP reciprocal searches of the NCBI non-redundant sequence database (as described in the materials and methods).

Three hundred and fifty-six homologues of parvoviral non-structural protein (NS) sequences and 107 homologues of capsid protein (VP) sequences, from partial to near complete coding sequences (Fig. 1 and Supplementary Table S2), were recovered from the NCBI transcriptome (230 sequences) and genome (206 sequences) databases and from the 11 new invertebrate transcriptome datasets (27 sequences found in 8/11 datasets). These 463 PRSs were found from the transcriptomes and genomes of 118 animal species, including 74 arthropods, 19 platyhelminthes, 12 vertebrates, six molluscs, two echinoderms, two annelids, one tunicate, one nematode, and one cnidarian (Fig. 1).

Overall, 89 potentially new parvovirus host species were identified, including species belonging to animal phyla with no previously known parvovirus host species: *Mollusca*, *Annelida*, *Nematoda* and *Cnidaria* (Fig. 1). Whereas the 88% of PRSs that were recovered from 95 invertebrates (including 69 arthropods) were more similar to viruses belonging to the arthropod-infecting *Densovirinae* subfamily, the remaining 12% of PRSs that were recovered from 12 vertebrate species and 4 molluscs were more similar to vertebrate-infecting viruses in the *Parvovirinae* subfamily (Supplementary Table S2). Among arthropods, PRSs were found in species within classes

Host Taxonomy					Amount of PRS		
Phylum	Class	Order	Family	Species	NS	VP	Total
Annelida	Clitellata	Haplotaxida	Enchytraeidae	***Enchytraeus crypticus***	1		1
	Polychaeta	Canalipalpata	Siboglinidae	***Lamellibrachia sp.***	10	6	16
Arthropoda	Arachnida	Araneae	Theridiidae	***Latrodectus hesperus***	3		3
		Trombidiformes	Tetranychidae	***Tetranychus urticae***	1		1
		Ixodida	Ixodidae	***Amblyomma americanum***	4		4
				Ixodes scapularis [35]	2		2
				Rhipicephalus appendiculatus [21]	2		2
				Rhipicephalus microplus [11]	3	2	5
				Rhipicephalus pulchellus	2		2
		Mesostigmata	Phytoseiidae	***Metaseiulus occidentalis***	1	1	2
	Branchiopoda	Anostraca	Artemiidae	***Artemia franciscana***		1	1
		Diplostraca	Daphniidae	***Daphnia pulex***	1	1	2
	Insecta	Coleoptera	Carabidae	***Pogonus chalceus***		2	2
			Cupedidae	***Priacma serrata***	1		1
			Curculionidae	***Ips typographus***		1	1
				Pissodes strobi	2		2
			Nitidulidae	***Meligethes aeneus***		1	1
			Scarabaeidae	***Onthophagus nigriventris***	1		1
		Collembola	Entomobryidae	***Orchesella cincta***	2	0	2
		Diptera	Culicidae	*Culex pipiens* [36]	1	2	3
			Diopsidae	***Teleopsis dalmanni***	1		1
			Drosophilidae	*Drosophila persimilis* [21*]	2		2
				Drosophila sechellia [21*]	2	2	4
			Syrphidae	***Eristalis tenax***		1	1
			Tephritidae	***Bactrocera dorsalis***	2		2
				Rhagoletis pomonella		1	1
		Hemiptera	Aleyrodidae	***Bemisia tabaci***	7		7
			Aphididae	*Acyrthosiphon pisum* [21]	17	14	31
				Sitobion avenae	1		1
			Cicadellidae	***Graminella nigrifrons***	1		1
			Coreidae	***Anoplocnemis curvipes***		1	1
				Clavigralla tomentosicollis		1	1
			Delphacidae	***Nilaparvata lugens***		3	3
			Kerriidae	***Kerria lacca***	5		5
			Miridae	***Lygus hesperus***	2	9	11
			Pentatomidae	***Chinavia ubica***		1	1
				Halyomorpha halys	1	1	2
			Psyllidae	***Diaphorina citri***	2		2
				Pachypsylla venusta		3	3
			Reduviidae	*Rhodnius prolixus* [21]	4	3	7
				Triatoma infestans	1	1	2
		Hymenoptera	Apidae	***Bombus impatiens***	1		1
			Formicidae	***Acromyrmex echinatior***	5	1	6
				Atta cephalotes	4		4
				Lasius niger		2	2
				Monomorium pharaonis		4	4
				Pogonomyrmex barbatus		3	3
				Messor barbarus	3	1	4
				Messor concolor	1		1
				Tetramorium bicarinatum		1	1
		Lepidoptera	Crambidae	***Chilo suppressalis***	2	1	3
			Micropterigidae	***Micropteix calthella***	2	2	4
			Noctuidae	***Agrotis segetum***		1	1
				Helicoverpa armigera	1		1
			Nymphalidae	*Bicyclus anynana* [21]	1		1
			Thaumetopoeidae	***Thaumetopoea pityocampa***	2	2	4
		Mecoptera	Nannochoristidae	***Nannochorista philpotti***		2	2
		Megaloptera	Corydalidae	***Corydalinae sp.***	1	2	3
		Neuroptera	Chrysopidae	***Chrysopa pallens***		1	1
		Orthoptera	Acrididae	***Schistocerca gregaria***	3		3
			Gryllidae	*Gryllus bimaculatus* [21]	1		1
		Phasmatodea	Heteropterygidae	***Aretaon asperrimus***	11	1	12
			Heteronemiidae	***Sipyloidea sipylus***		1	1
			Phasmatidae	***Extatosoma tiaratum***	16	3	19
				Medauroidea extradentata	3		3
		Raphidioptera	Raphidiidae	***Raphidia ariadne***	1		1
	Malacostraca	Amphipoda	Ampeliscidae	***Ampelisca abdita***	5		5
		Decapoda	Astacidea	***Pontastacus leptodactylus***	1		1
			Palaemonidae	***Macrobrachium nipponense***		1	1
			Penaeidae	*Penaeus monodon* [38*]	2		2
			Portunidae	***Scylla olivacea***	2		2
		Isopoda	Armadillidiidae	*Armadillidium nasatum* [39*]	3		3
				Armadillidium vulgare [39*]	2		2
	Maxillopoda	Arguloida	Argulidae	***Argulus siamensis***	4	4	8
		Siphonostomatoida	Caligidae	***Caligus rogercresseyi***	2		2
				Lepeophtheirus salmonis [21]	8		8
Chordata	Actinopterygii	Cyprinodontiformes	Fundulidae	***Fundulus grandis***	2		2
	Aves	Galliformes	Phasianidae	*Gallus gallus* [37]	1		1
	Mammalia	Artiodactyla	Bovidae	*Bos taurus* [21]		2	2
				Capra hircus		2	2
			Suidae	*Sus scrofa* [21]	1		1
		Diprotodontia	Macropodidae	*Macropus eugenii* [33]		1	1
			Phalangeridae	*Trichosurus vulpecula* [21]	1	1	2
		Lagomorpha	Leporidae	*Oryctolagus cuniculus* [21*]	4	3	7
		Primates	Hominidae	*Homo sapiens* [21]		1	1
		Rodentia	Chinchillidae	*Chinchilla lanigera* [30*]	4		4
			Muridae	*Rattus norvegicus* [30*]	5	4	9
			Octodontidae	*Octodon degus* [30*]	2		2
Cnidaria	Anthozoa	Alcyonacea	Gorgoniidae	***Eunicella cavolinii***	1		1
Echinodermata	Asteroidea	Paxillosida	Luidiidae	***Luidia clathrata***	1		1
		Spinulosida	Echinasteridae	***Echinaster spinulosus***	1		1
Mollusca	Cephalopoda	Octopoda	Octopodidae	***Octopus bimaculoides***	4		4
				Octopus vulgaris		1	1
		Sepiida	Sepiidae	***Sepia officinalis***	1		1
		Sepiolida	Sepiolidae	***Euprymna scolopes***	4		4
	Gastropoda	x	Bithyniidae	***Bithynia siamensis***	1		1
		Neotaenioglossa	Calyptraeidae	***Crepidula fornicata***	1		1
Nematoda	Secernentea	Ascaridida	Ascarididae	***Ascaris suum***	1		1
Platyhelminthes	Cestoda	Cyclophyllidea	Anoplocephalidae	***Moniezia expansa***	1		1
			Hymenolepididae	***Hymenolepis diminua***	14		14
				Hymenolepis microstoma [34]	30		30
				Hymenolepis nana	2		2
			Taeniidae	*Echinococcus granulosus* [34]	3		3
				Echinococcus multilocularis	39		39
				Hydatigera taeniaeformis	2		2
				Taenia asiatica	1		1
				Taenia multiceps	2		2
	Monogenea	Monopisthocotylea	Capsalidae	***Neobenedenia melleni***	1		1
	Trematoda	Echinostomida	Echinostomatidae	***Echinostoma caproni***	5		5
		Opisthorchiida	Opisthorchiidae	*Clonorchis sinensis* [34]	9		9
				Opisthorchis felineus	9		9
				Opisthorchis viverrini	21		21
		Plagiorchiida	Dicrocoeliidae	***Dicrocoelium dendriticum***	4		4
		Strigeatida	Schistosomatidae	*Schistosoma mansoni* [21]	3		3
				Schistosoma margrebowiei	1		1
				Schistosoma mattheei	1		1
	Turbellaria	Tricladida	Dugesiidae	***Schmidtea mediterranea***	1		1
Urochordata	Ascidiacea	Enterogona	Cionidae	*Ciona intestinalis* [21]		1	1
			Total	**118**	**356**	**107**	**463**

Figure 1. Distribution of PRSs among animals. VP and NS refer to viral structural and non-structural proteins respectively. PRSs were identified using the full sequences of 74 representative parvovirus genomes (provided in Table S1) as queries to search the EST (Expressed Sequence Tags), TSA (Transcriptome Shotgun Assembly), nr Nucleotide collection, and Uniprot databases as well as search in 11 un-deposited transcriptomes either produced for this study (*Lamellibrachia spp.*) or already published in another context (*Artemia franciscana*, *Culex pipiens*, *Crepidula fornicata*, *Eunicella cavolinii* and *Messor barbarus and Messor concolor*)[40,41]. Animal species wherein PRSs were first identified in this study are represented in bold, while PRSs that have already been identified are associated with numbers corresponding to the respective references. (*) PRS endogenization has been confirmed by PCRs.

(e.g. *Branchiopoda* and *Arachnida*), orders (*Phasmatodea* and *Coleoptera*) and families (e.g. *Formicidae*) that contained no previously identified parvovirus hosts (PRSs are summarized in Table S2).

It is also noteworthy that several copies of PRSs homologous to both NS and VP encoding genes were found in 72/118 of the animal transcriptomes (Table S2).

Most of the 463 PRSs (77%, corresponding to 79 animal species) potentially encoded proteins sharing between 30–85% aa identity to the NS or VP proteins of a known extant parvovirus, while 17% (derived from 30 animal species) potentially encoded proteins sharing less than 30% identity to any known extant parvoviruses. This degree of similarity is below the parvovirus genus demarcation threshold recommended by the ICTV (i.e. >30% amino acid identity in NS1), suggesting that, if these divergent PRSs are derived from extant viruses, these likely belong to species within as yet uncharacterized parvovirus genera[14]. Finally, 6% of the PRSs found within the transcriptome datasets of nine animal species potentially encoded proteins with more than 85% identity to those expressed by known extant parvoviruses (Supplementary Table S2).

Altogether, we concluded that, while parvoviruses are probably associated with a wider variety of animals than has previously been thought, the PRSs we found were mostly associated with invertebrate species within phyla, classes, orders and families containing no previously known parvovirus host species (Fig. 2).

PRSs likely originate from both fossil viral sequences and extant viruses. The PRSs that we detected could have had a number of different origins including: (1) endogenous "fossil" viral sequences resulting from ancient integration events; (2) endogenous viruses constituting latent infections resulting from recent integration events; or (3) exogenous viruses. Further, it was possible that all three types of PRS elements could have been present either within the cells of the species from which the transcriptome datasets were derived, or from (likely eukaryotic) species either parasitizing, or in some other way associated with, the species from which the transcriptome datasets were derived.

To identify PRSs potentially corresponding to endogenized parvoviral fragments, we used each of the 463 PRSs as queries to screen the publically available eukaryotic genome datasets (both assembled and unassembled) within the NCBI genomic database. This search identified 76 genomic sequences (summarized in Supplementary Table S3) of various sizes (0.05–8 kb) displaying significant matches (cutoff 95% identity, e-value $<10^{-5}$) to PRSs within the genomes of 31 invertebrates from six phyla including 16 arthropods (33 PRSs), 13 platyhelminthes (36 PRSs), one mollusc (2 PRSs) and one nematode (4 PRS), for which endogenization of parvoviruses has never been found before (Table S3). In addition to the possibly endogenized PRSs identified in these 31 invertebrate species, PRSs were also identified in the genomes of 16 animal species for which potential parvovirus endogenization has been previously reported (including nine arthropods, six chordates and one platyhelminthes; Fig. 3)[20,21,31–34].

Ancient integration events are often characterised by degraded integrons with the extent of degradation varying depending on whether the integrated sequences were selectively disadvantageous, beneficial or neutral[2]. We thus searched animal genome sequence databases for degraded PRSs including those containing potential transposable elements (TE) or repeated sequence insertions within the vicinity of the integration site, which may have contributed to their integration. We found 31 PRSs displaying truncations due to the accumulation of internal stop codons and/or adjacent transposable elements in the genomes of fifteen arthropod and seven platyhelminthes species (Table S3): a finding strongly supporting the hypothesis that these PRSs were the product of ancient endogenisation events in these phyla. In arthropods, whereas endogenization has been proposed previously for one of these PRSs–found within the genome of *A. pisum*[21,25]–we detected various other putative PRS endogenization events in a number of other Arthropods, i.e. six in Hymenoptera (*Formicidae*), three in Hemiptera (*Pachypsylla venusta*; family *Psyllidae*, *Halyomorpha halys*; family *Pentatomidae*, *Nilaparvata lugens*; family *Delphacidae*), one in Araneae (*Latrodectus hesperus*; family *Theridiidae*), one in Coleoptera (*Priacma serrata*; family *Cupedidae*) one in Mesostigmata (*Metaseiulus occidentalis*; family *Phytoseiidae*), one in Diplostraca (*Daphnia pulex*; family *Daphniidae*) and one in a Siphonostomatoida (*Caligus rogercresseyi*; family *Caligidae*), (Supplementary Tables S2 and S3); all these PRSs displayed internal stop codons and/or adjacent TE elements.

In platyhelminthes, potential densovirus endogenization has already been reported in one cestode (*Echinococcus granulosis*) and one trematode (*Schistosoma mansoni*), and all platyhelminthes associated PRSs recovered here share >95% nt identity with known platyhelminthes genome sequences[18]. The PRSs were detected in the genomes of 15 species (Fig. 3); mostly cestodes (e.g. *Taenia* sp. and *Echinococcus* sp.) and trematodes (e.g. *Schistosoma* sp.). All the PRSs detected in this phylum displayed ≥30% aa identity corresponding to the same domain of the NS1 protein although the inferred encoded amino acid sequences of this domain contained internal stop codons. For example, 39 PRSs were detected in the genome of the cestode, *Echinococcus multilocularis* that, when translated, displayed approximately 30% aa similarity with the NS1 of the shrimp parvovirus, Decapod penstyldensovirus 1 (PstDV1). These results suggest the endogenization of the NS1-like domain in the genome of several platyhelminthes species. The phylogenetic relationship between these PRSs will be addressed below.

Most genomic PRSs that we detected were, however, located at the extremities of genomic scaffolds that were less than 10 kb in length, and there was therefore limited information regarding their possible genomic context and flanking sequences: a factor which impaired our ability to definitively determine whether these sequences too were ancient degraded integrons associated with transposable elements or repeat-sequence insertions (Supplementary Table S3).

We next searched for PRSs corresponding to non-degraded viral ORFs as these might correspond to extant viruses. Four large PRSs covering both the NS and VP ORFs were found in both the genomes and transcriptomes of four arthropods (mentioned as NS-VP in Supplementary Table S2). One of these large PRS corresponded with the genomic sequence integrated in the pea aphid genome (*A. pisum*) that has been previously characterized by Liu *et al.*[35]. This PRS shares 52% aa identity with both NS and VP of the ambidensovirus infecting the aphid *Dysaphis plantaginea* (DpDV). A similar PRS (sharing >95% identity) was also detected in the genome of the peach-potato aphid (*Myzus persicae*)[23]. The transcription of both the NS and VP encoding genes was also

Host Taxonomy				No. PRS
Phylum	Class	Order	No. Species	
Annelida	**Clitellata**	**Haplotaxida**	1	1
	Polychaeta	**Canalipalpata**	1	16
Arthropoda	Arachnida	**Araneae**	1	3
		Trombidiformes	1	1
		Ixodida	5	15
		Mesostigmata	1	2
	Branchiopoda	**Anostraca**	1	1
		Diplostraca	1	2
	Insecta	**Coleoptera**	6	8
		Collembola	1	2
		Diptera	7	14
		Hemiptera	15	78
		Hymenoptera	9	26
		Lepidoptera	6	14
		Mecoptera	1	2
		Megaloptera	1	3
		Neuroptera	1	1
		Orthoptera	2	4
		Phasmatodea	4	35
		Raphidioptera	1	1
	Malacostraca	**Amphipoda**	1	5
		Decapoda	4	6
		Isopoda	2	5
	Maxillopoda	**Arguloida**	1	8
		Siphonostomatoida	2	10
Chordata	**Actinopterygii**	**Cyprinodontiformes**	1	2
	Aves	Galliformes	1	1
	Mammalia	Artiodactyla	3	5
		Diprotodontia	2	3
		Lagomorpha	1	7
		Primates	1	1
		Rodentia	3	15
Cnidaria	**Anthozoa**	**Alcyonacea**	1	1
Echinodermata	Asteroidea	Paxillosida	1	1
		Spinulosida	1	1
Mollusca	**Cephalopoda**	**Octopoda**	2	5
		Sepiida	1	1
		Sepiolida	1	4
	Gastropoda	**x**	1	1
		Neotaenioglossa	1	1
Nematoda	**Secernentea**	**Ascaridida**	1	1
Platyhelminthes	Cestoda	Cyclophyllidea	9	94
	Monogenea	**Monopisthocotylea**	1	1
	Trematoda	**Echinostomida**	1	5
		Opisthorchiida	3	39
		Plagiorchiida	1	4
		Strigeatida	3	5
	Turbellaria	**Tricladida**	1	1
Urochordata	Ascidiacea	Enterogona	1	1
		Total	**118**	**463**

Figure 2. Summary of the distribution of PRSs among animal transcriptomes and genomes. Animal orders wherein PRSs were first identified in this study are represented in bold.

demonstrated by these authors, suggesting that aphid endogenous parvoviral sequences represent recently integrated persistent viruses that could potentially become exogenous[25]. Remarkably, two other large PRSs (4.2 kb and 5.6 kb covering almost complete viral ORFs) were respectively recovered from the stick insect *Aretaon*

Host Taxonomy					Databases and amount of PRSs	
Phylum	Class	Order	Family	Species	Transcriptomic	Genomic
Annelida	Clitellata	Haplotaxida	Enchytraeidae	***Enchytraeus crypticus***	1	NA
	Polychaeta	Canalipalpata	Siboglinidae	***Lamellibrachia sp.***	16	NA
Arthropoda	Arachnida	Araneae	Theridiidae	***Latrodectus hesperus***	A	5
		Trombidiformes	Tetranychidae	***Tetranychus urticae***	A	1
		Ixodida	Ixodidae	***Amblyomma americanum***	4	NA
				Ixodes scapularis [21]	1	1
				Rhipicephalus appendiculatus [21]	2	NA
				Rhipicephalus microplus [21]	5	2
				Rhipicephalus pulchellus	2	NA
		Mesostigmata	Phytoseiidae	***Metaseiulus occidentalis***	2	1
	Branchiopoda	Anostraca	Artemiidae	***Artemia franciscana***	1	NA
		Diplostraca	Daphniidae	***Daphnia pulex***	A	1
	Insecta	Collembola	Entomobryidae	***Pogonus chalceus***	2	NA
		Coleoptera	Carabidae	***Priacma serrata***	1	A
			Cupedidae	***Ips typographus***	1	NA
			Curculionidae	***Pissodes strobi***	2	NA
				Meligethes aeneus	1	NA
			Nitidulidae	***Onthophagus nigriventris***	1	NA
			Scarabaeidae	***Orchesella cincta***	2	NA
		Diptera	Culicidae	*Culex pipiens* [56]	3	NA
			Diopsidae	***Teleopsis dalmanni***	1	A
			Drosophilidae	*Drosophila persimilis* [21*]	NA	2
				Drosophila sechellia [21*]	4	1
			Syrphidae	***Eristalis tenax***	1	NA
			Tephritidae	***Bactrocera dorsalis***	2	2
				Rhagoletis pomonella	1	NA
		Hemiptera	Aleyrodidae	***Bemisia tabaci***	7	NA
			Aphididae	*Acyrthosiphon pisum* [21]	2	29
				Sitobion avenae	1	NA
			Cicadellidae	***Graminella nigrifrons***	1	NA
			Coreidae	***Anoplocnemis curvipes***	1	NA
			Coreidae	***Clavigralla tomentosicollis***	1	NA
			Delphacidae	***Nilaparvata lugens***	3	1
			Kerriidae	***Kerria lacca***	5	NA
			Miridae	***Lygus hesperus***	11	NA
			Pentatomidae	***Chinavia ubica***	1	NA
				Halyomorpha halys	2	5
			Psyllidae	***Diaphorina citri***	A	6
				Pachypsylla venusta	3	2
			Reduviidae	*Rhodnius prolixus* [71]	A	7
				Triatoma infestans	2	NA
		Hymenoptera	Apidae	***Bombus impatiens***	1	1
			Formicidae	***Acromyrmex echinatior***	NA	9
				Atta cephalotes	NA	6
				Lasius niger	NA	3
				Monomorium pharaonis	4	NA
				Pogonomyrmex barbatus	1	NA
				Messor barbarus	A	6
				Messor concolor	NA	4
				Tetramorium bicarinatum	1	NA
		Lepidoptera	Crambidae	***Chilo suppressalis***	3	A
			Micropterigidae	***Micropteryx calthella***	4	NA
			Noctuidae	***Agrotis segetum***	1	NA
				Helicoverpa armigera	1	NA
			Nymphalidae	*Bicyclus anynana* [21]	1	NA
			Thaumetopoeidae	***Thaumetopoea pityocampa***	4	NA
		Mecoptera	Nannochoristidae	***Nannochorista philpotti***	2	NA
		Megaloptera	Corydalidae	***Corydalinae sp.***	3	NA
		Neuroptera	Chrysopidae	***Chrysopa pallens***	1	NA
		Orthoptera	Acrididae	***Schistocerca gregaria***	3	NA
			Gryllidae	*Gryllus bimaculatus* [21]	1	NA
		Phasmatodea	Heteropterygidae	***Aretaon asperrimus***	12	NA
			Heteronemiidae	***Sipyloidea sipylus***	1	NA
			Phasmatidae	***Extatosoma tiaratum***	19	NA
				Medauroidea extradentata	3	NA
		Raphidioptera	Raphidiidae	***Raphidia ariadne***	1	NA
	Malacostraca	Amphipoda	Ampeliscidae	***Ampelisca abdita***	5	NA
		Decapoda	Astacidea	***Pontastacus leptodactylus***	1	NA
			Palaemonidae	***Macrobrachium nipponense***	1	NA
			Penaeidae	*Penaeus monodon* [58*]	A	2
			Portunidae	***Scylla olivacea***	2	NA
		Isopoda	Armadillidiidae	*Armadillidium nasatum* [30*]	NA	3
				Armadillidium vulgare [30*]	A	2
	Maxillopoda	Arguloida	Argulidae	***Argulus siamensis***	8	NA
		Siphonostomatoida	Caligidae	***Caligus rogercresseyi***	2	4
				Lepeophtheirus salmonis [71]	8	3
Chordata	Actinopterygii	Cyprinodontiformes	Fundulidae	***Fundulus grandis***	2	NA
	Aves	Galliformes	Phasianidae	*Gallus gallus* [57]	1	A
	Mammalia	Artiodactyla	Bovidae	*Bos taurus* [21]	2	A
				Capra hircus	2	A
			Suidae	*Sus scrofa* [21]	1	A
		Diprotodontia	Macropodidae	*Macropus eugenii* [33]	1	12
			Phalangeridae	*Trichosurus vulpecula* [21]	2	NA
		Lagomorpha	Leporidae	*Oryctolagus cuniculus* [21*]	5	2
		Primates	Hominidae	*Homo sapiens* [21]	1	A
		Rodentia	Chinchillidae	*Chinchilla lanigera* [35*]	4	1
			Muridae	*Rattus norvegicus* [33*]	9	42
			Octodontidae	*Octodon degus* [39*]	NA	2
Cnidaria	Anthozoa	Alcyonacea	Gorgoniidae	***Eunicella cavolinii***	1	NA
Echinodermata	Asteroidea	Paxillosida	Luidiidae	***Luidia clathrata***	1	NA
		Spinulosida	Echinasteridae	***Echinaster spinulosus***	1	NA
Mollusca	Cephalopoda	Octopoda	Octopodidae	***Octopus bimaculoides***	NA	6
				Octopus vulgaris	1	NA
		Sepiida	Sepiidae	***Sepia officinalis***	1	NA
		Sepiolida	Sepiolidae	***Euprymna scolopes***	4	NA
	Gasteropoda	x	Bithyniidae	***Bithynia siamensis***	1	NA
		Neotaenioglossa	Calyptraeidae	***Crepidula fornicata***	1	NA
Nematoda	Secernentea	Ascaridida	Ascarididae	***Ascaris suum***	1	1
Platyhelminthes	Cestoda	Cyclophyllidea	Anoplocephalidae	***Moniezia expansa***	1	NA
			Hymenolepididae	***Hymenolepis diminuta***	NA	15
				Hymenolepis microstoma [34]	NA	31
				Hymenolepis nana	NA	4
			Taeniidae	*Echinococcus granulosus* [34]	A	3
				Echinococcus multilocularis	NA	49
				Hydatigera taeniaeformis	NA	7
				Taenia asiatica	NA	2
				Taenia multiceps	2	NA
	Monogenea	Monopisthocotylea	Capsalidae	***Neobenedenia melleni***	1	NA
	Trematoda	Echinostomida	Echinostomatidae	***Echinostoma caproni***	NA	6
		Opisthorchiida	Opisthorchiidae	*Clonorchis sinensis* [34]	A	10
				Opisthorchis felineus	9	NA
				Opisthorchis viverrini	A	28
		Plagiorchiida	Dicrocoeliidae	***Dicrocoelium dendriticum***	NA	6
		Strigeatida	Schistosomatidae	*Schistosoma mansoni* [21]	A	3
				Schistosoma margrebowiei	NA	2
				Schistosoma mattheei	NA	2
	Turbellaria	Tricladida	Dugesiidae	***Schmidtea mediterranea***	1	3
Urochordata	Ascidiacea	Enterogona	Cionidae	*Ciona intestinalis* [21]	1	25
				Number of PRSs	**247**	**376**
				Number of animal species	**88**	**49**

Figure 3. Distribution of PRSs in animal transcriptomes and genomes (in grey). Animal species wherein PRSs were first identified in this study are represented in bold while numbers correspond to references where PRSs were previously described. (*) PRS endogenization proved by PCRs. A: transcriptomic/genomic data (EST and TSA databases) available at NCBI. NA: Not Available, i.e. transcriptomic/genomic data (gss, WGS, chromosome and refseq_genomic databases) are not available at NCBI.

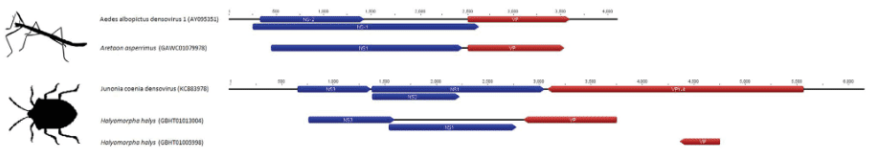


Figure 4. Organization of open reading frames of three PRSs (GAWC01079978 from *Aretaon asperrimus*, GBHT01013004 and GBHT01005998 from *Halyomorpha halys*). Arrowhead boxes indicate viral and predicted viral genes (NS are in blue and VP in red).

asperrimus, belonging to the order *Phasmatodea*, and from the stink-bug, *Halyomorpha halys*, belonging to the order *Hemiptera*: neither species had previously been identified as densovirus hosts. These two large PRSs encompass two ORFs (NS1 and VP; GAWC1079978 recovered from *A. asperrimus*) and three ORFs (NS1, NS3 and VP; GBHT01013004 from *H. halys*; Fig. 4). The arrangement of these ORFs is similar to Brevi- and Ambidensoviruses that respectively infect mosquitoes and caterpillars (Fig. 4). Since only a few reads of the genomes of these potentially new insect hosts are available (for example we only found one contig containing the stink-bug PRS), it is not possible to conclude whether these PRSs are involved in latent infections by undescribed viruses (as is possibly the case for the integrated densovirus in aphids that is described above), or are derived from previously unknown extant exogenous *A. asperrimus* and/or *H. halys* infecting densoviruses[18].

Last, we must emphasize that contamination of genomic datasets may come either from animals that are infected by parvoviruses or from animals that are infected by parasites that are themselves infected by parvoviruses. Intriguingly, we found one ~0.8 kb PRS in the genome sequence dataset of *Gregarina niphandrodes*; a protozoan Apicomplexa that infects a number of invertebrates and which was isolated from the coleopteran, *Tenebrio molitor* according to Genebank data. This PRS shared 100% identity with the VP sequence of a yet undescribed *Blatella germanica* densovirus-like virus found in the metagenome of insectivorous bat faeces[36] (Table S2). We could not find any detectably homologous sequence within either the genome or the transcriptome of its potential host, *T. molitor*. Although we cannot rule out the possibility that the densovirus from which this PRS was derived was able to infect *G. niphandrodes*, we concluded that the PRS probably originated from a contaminant.

In total we discarded 16 PRSs from further analysis due to contamination concerns. Among these were three found in the transcript datasets of plant species (Table S2): all of these plant PRSs had high degrees of identity with extant insect-infecting densoviruses and were therefore likely derived from insects associated with the plants (Table S2). Similarly, PRSs found in the transcriptomes of three vertebrates (all amphibian species), one echinoderm (*Amphiura filiformis*) and one insect (*Drosophila ananassae*) were >85% identical to known mammal-infecting parvoviruses and were thus assumed to be contaminants.

Altogether these results highlight the large diversity of PRSs that can be found by screening publically available transcriptomes and genomes. In total, 623 PRSs were found when adding up all PRSs found in transcriptomic (247) and genomic databases (376) including the newly found and already reported sequences (Fig. 3). Our search identified that parvoviruses are/were associated with a large number of animal species that have never previously been identified as parvovirus host species (Summarized in Fig. 2).

Phylogenetic analyses of PRSs. We next attempted to evaluate the genetic relationships of the PRSs with known exogenous parvoviruses. While the parvovirus protein NS1 contains a SF3 helicase domain that is highly conserved in all known parvoviruses (it can also be found in proteins of viruses in other families)[37], the most conserved domain of parvoviral VPs, PLA2, is missing in certain parvoviruses: a factor which may explain why our search identified more sequences related to NS1 than to VP. The SF3 domain is thus typically used for phylogenetic analyses of divergent parvoviruses[18].

As is shown in the *Parvoviridae* maximum likelihood trees (Figs 2 and 3), all exogenous known parvovirus species (in black text) were clearly placed within the 13 genera recognized by ICTV with bootstrap values >80%. While validating the use of the small SF3 domain of NS1 to study relationships amongst PRSs, it is apparent from the trees that most of the PRSs are situated on long branches that connect to the tree with low degrees of bootstrap support (<70%), basal to clusters of sequences from the established parvovirus genera. This phylogenetic placement is consistent with PRSs belonging to currently undescribed genus-level parvovirus lineages; although we cannot exclude ambiguous alignment of divergent sequences as the cause of low degrees of bootstrap support for the clustering of these PRSs with viruses from the known parvovirus genera.

Nevertheless, the PRSs drawn from the transcriptome datasets of animals belonging to particular genera were frequently monophyletic within clades supported by bootstrap values >50%: consistent with the hypothesis that these PRSs are derived from viruses belonging to undiscovered parvovirus lineages with genus-specific host-ranges. This situation is exemplified with PRSs found within transcriptome datasets of ticks, stick insects and stink-bugs (Fig. 5). Interestingly, the two large transcripts that correspond to almost complete densovirus genomes that were detected in the stink-bug and stick insect transcriptome datasets cluster together with PRSs found in related triatomine insects in either the *Reduviidae* family (for the stink bug PRSs) or the *Phasmatodea* order (for the stick insect PRSs). These large PRSs might correspond to exogenous or persistent viruses belonging to densovirus lineages that infect these insects (Fig. 4). Interestingly, PRSs found in four classes of marine molluscs and more related to vertebrate parvoviruses according to results above, branched out of all known parvovirus genera (represented by the light blue branches in Fig. 5), although we cannot exclude some contamination of these animals. These results suggest that these PRSs belong to new parvovirus lineages yet to be characterized in this phylum.

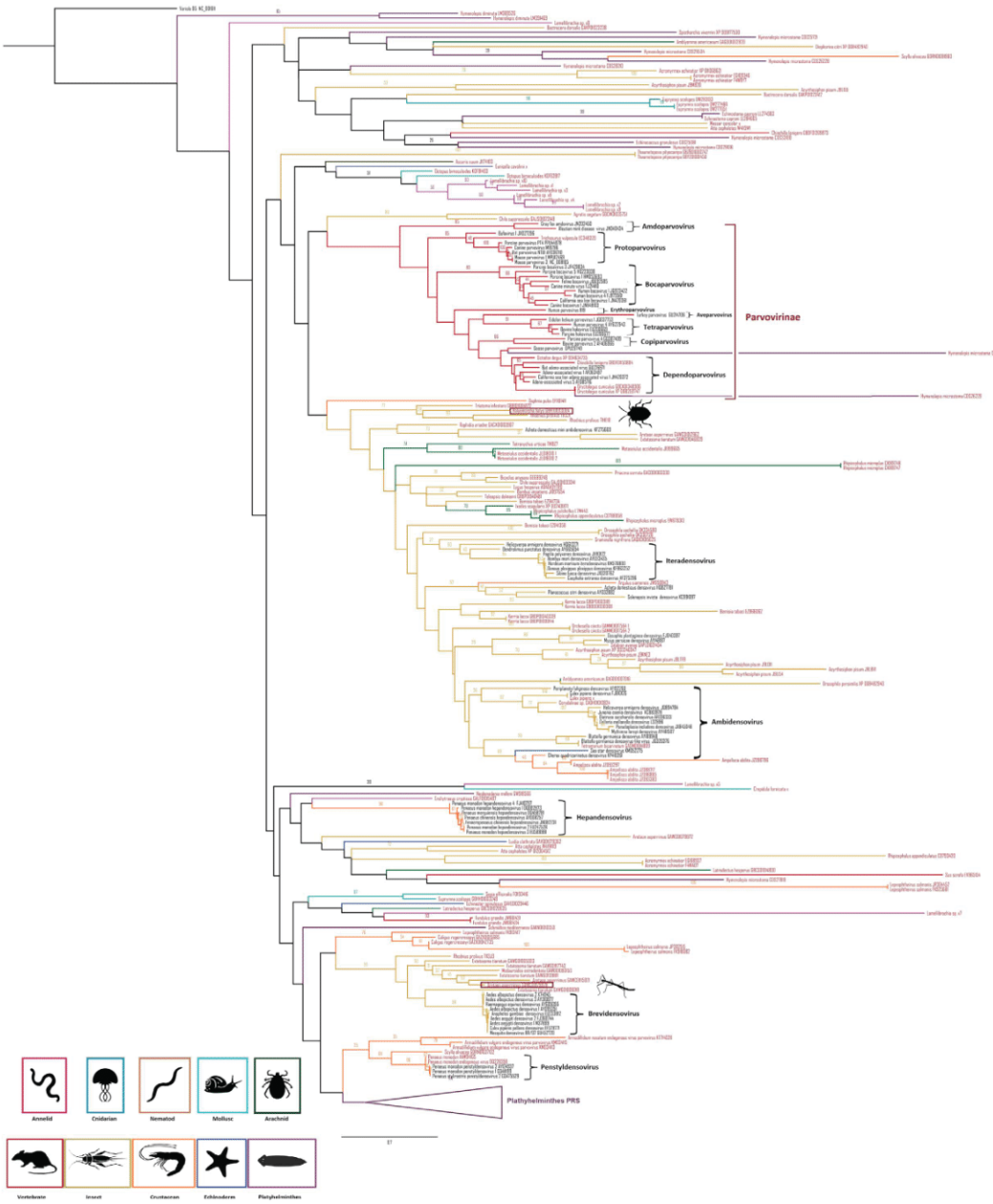


Figure 5. Maximum likelihood phylogenetic tree based on partial SF3 domains of the NS1 protein, including 74 parvovirus species (in black) and 264 PRSs (in red). The alignment was produced using MUSCLE 3.7 with default settings. The tree was rooted with the SF3 domain of the variola virus D5 protein. Bootstrap values >25% are indicated at each node. Scale bars correspond to amino acid substitutions per site. Genera of the *Parvoviridae* family are indicated in brackets. Associated hosts are indicated in the tree by different branch colours and silhouettes at the bottom of the figure.

Such clustering of PRSs according to the evolutionary relationships of the animal species they are associated with is particularly apparent for PRSs found within the platyhelminthes transcriptome datasets (Fig. 6). PRSs found within these datasets form two clades supported by >70% bootstrap values (Fig. 6). The inferred SF3 amino acid sequences encoded by these PRSs share identities of <30% with those of previously known parvoviruses, suggesting that these flatworm-associated PRSs may represent new genera (Fig. 6) of either extant but undiscovered circulating parvoviruses, or integrated (and possibly extinct) parvoviruses.

The data presented here indicate that parvoviruses are likely widespread among multicellular animals. If we consider all clusters with bootstrap values >70%, then we can tentatively estimate that these phylogenetic analyses indicate the existence of around 20 currently undescribed genus-level densovirinae lineages.

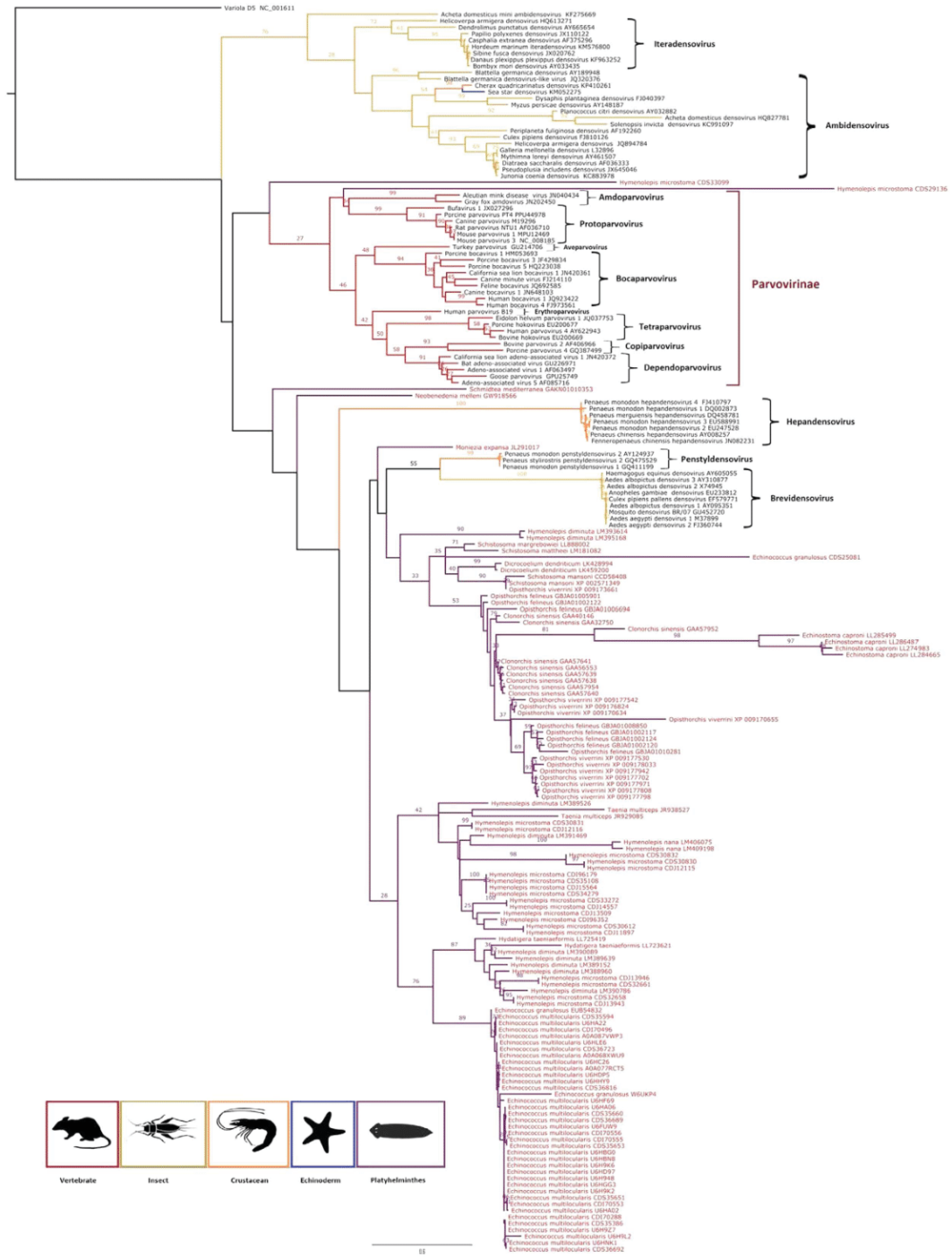


Figure 6. Maximum likelihood phylogenetic tree of platyhelminthes PRSs based on partial SF3 domains of the NS1 protein, including 74 parvovirus species (in black) and 118 platyhelminthes PRSs (in red). The alignment was produced using MUSCLE 3.7 with default settings. The tree was rooted with the variola virus D5 protein SF3 domain. Bootstrap values >25% are indicated at each node. Scale bars correspond to amino acid substitutions per site. The genera of the *Parvoviridae* family are indicated in brackets. Host phyla are represented by different colours and silhouettes at the bottom of the figure.

Discussion

Since the discovery of parvoviruses four decades ago, our perception of the species diversity within this family has been strongly biased by an overwhelming focus on discovering viruses of health and economic interest. While a high diversity of viral genome organisations and sequences has been revealed, few genomes were characterized: a factor which has limited our understanding of the global diversity, prevalence and host-associations of these

viruses. The advent of the genomic era has now provided an alternative way to explore virus diversity in a slightly less biased way via the computational screening of large transcriptome and whole genome sequence datasets[18,21,22].

By scanning public genomic and transcriptomic resources, we have found that parvoviruses likely infect a larger number and wider diversity of animal-hosts, particularly among invertebrate phyla, than has previously been appreciated. These newly discovered potential parvovirus hosts include molluscs, annelids, nematodes, cnidarians and arthropods.

Although these results suggest widespread parvovirus integration into the genomes of diverse invertebrates, the limited number of complete invertebrate genome sequences that are currently available hinders both the definitive identification of these PRSs as integrated sequences, and an accurate estimation of the time-scale of individual PRS integration events. Data summarized in Fig. 3 and Supplementary Table S2 highlight the fact that 60% (376/623) of PRSs were found in animal genomes, among which 10% (37/367) are likely integrated. Integration was thus uncertain for 90% of the PRSs, mostly due to either unavailable or incomplete genomes for most of the invertebrate phyla for which transcriptome data was available. However, it is plausible that some of the endogenous PRSs were integrated millions of years ago since the presence of what appear to be closely related PRSs in different mammalian species have previously suggested that parvoviruses have likely coexisted with mammals for at least 98 million years[21].

It is particularly clear that most of the PRSs identified in this study were more similar to members of the *Densovirinae* sub-family than they were to members of the *Parvovirinae* sub-family. Considering that arthropod species vastly outnumber all other animal species, such over-representation of densoviruses, although unsurprising, contrasts with the similar numbers of described species in both subfamilies in the current ICTV report[18,19].

We cannot exclude the possibility that contamination of transcriptome and genome sequence datasets with unaccounted for eukaryote and/or viral DNA could yield spurious host-virus associations in studies such as that carried out here: as is apparently exemplified by the presence of PRSs in a small number of plant datasets. It is noteworthy, however, that densovirus-plant associations actually do occur in nature. For example, aphid-infecting densoviruses can be injected into, and circulate within, plants[38]. It has, in fact, been speculated that plants might be stationary vectors of some densoviruses that infect plant-feeding insects[38].

It has been proposed that all viruses in a parvovirus genus should be monophyletic and encode NS1 proteins that share >30% amino acid sequence identity to each other[18]. The phylogeny that we have produced for the SF3 helicase domain of NS1 indicated that 15% of the PRSs that contain this domain are highly divergent. The low degrees of similarity shared between these PRSs and both known parvoviruses and the other PRSs meant that they could not be reliably aligned: a factor that could have contributed to these divergent sequences falling on long isolated branches of the phylogenetic tree. While the intermingling of these divergent PRS lineages amongst known members of the *Parvovirinae* and *Densovirinae* genera, suggests that numerous genus-level parvovirus lineages are presently undescribed, we can also not exclude the possibility that some of these PRSs may be derived from virus families, such as *Bidnaviridae*, that are related to the parvoviruses. Like parvoviruses, the bidnaviruses are also ssDNA viruses that express a SF3 domain containing protein[18,39]. Unlike parvoviruses, however, they have two-component genomes and encode DNA polymerases. Although all of the PRSs discovered here were more closely related to known parvovirus sequences than to known bidnaviruses, the possibility remains that some of these PRSs are potentially derived from viruses belonging to currently undescribed families.

Here we have highlighted the extraordinary diversity of PRSs that can be found in public databases. As more animal sequences will be released we can anticipate that our knowledge of the diversity of parvoviruses will also keep improving. Combining such database searches with more directed viral metagenomics approaches and classical etiological survey, will be of great value both for discovering new parvoviruses and, as more endogenous PRSs are discovered within eukaryote genomes, for illuminating the deep evolutionary history of this family in relation to that of the host species that they infect.

Methods

Biological samples and transcriptome datasets. *Lamellibrachia* sp. (marine polychaete annelid) samples were collected in 2007 in the Gulf of Mexico at 1250 m depth for individual, GA27M, and in the Gulf of Guinea at 580–670 m depth for individuals, GA27P, GA27S and GA27U. Vestimentum tissue was dissected and stored immediately in liquid nitrogen. Due to poor yield with standard total RNA isolation methods, we used a modified protocol based on a Trizol-Chloroform method combined with a QIAshredder column (Qiagen) purification step[40] involving the addition of 4 μl of glycogen (Ambion, final concentration = 0.04 mg/μL) to increase RNA yield and a further polyvinylpolypyrrolidone (PVPP) purification step. Five μg of total RNA was reverse-transcribed using the SMART cDNA library construction kit (Clontech, Mountain View, USA). Libraries were sequenced to produce 100 bp paired-end reads on a Genome Analyzer II or Hiseq 2000 (Illumina, Inc.). Low-quality read extremities were trimmed using the SeqClean program (http://compbio.dfci.harvard.edu/tgi/). Reads were deposited in the Sequence Read Archive (SRA) NCBI database under bioproject PRJNA302863, accession numbers SRX1440230, SRX1447229, SRX1447303 and SRX1447300. *Lamellibrachia* transcriptomes produced in this study, as well as individual previously published transcriptomes of *Artemia franciscana* (individual GA17B; SRX565006), *Crepidula fornicata* (individual GA22E; SRX565072), *Eunicella cavolinii* (individual GA31L; SRX565138) and *Messor barbarus* (individual GA40E; SRX565206) were successively assembled using ABYSS V 1.2.0[41] and CAP3[42]. This assembly method was previously found to be suitable for other mollusc and animal transcriptomes[43,44]. Three supplementary transcriptomes of whole individual adults of *Messor barbarus*, *Messor concolor* and *Culex pipiens* were assembled as above and were added to this dataset (N. Galtier, unpublished data). In total eleven transcriptomes have been used in this study, four of which were generated for this study (*Lamellibrachia*) and seven of which were previously used for animal genomics studies that did not involve virus detection (*Artemia franciscana*, *Culex pipiens*, *Crepidula fornicata*, *Eunicella cavolinii*, two *Messor barbarus* and *Messor concolor*).

Homology searches for parvovirus-related sequences (PRSs). We assembled a dataset of NS and VP amino acid sequences derived from each of the 74 parvovirus species, recognized and yet to be approved by the ICTV (all obtained from GenBank; genomes listed in Table S1). These sequences were used as queries to perform BLASTX searches for PRSs within all the non-redundant (nr) nucleotide and protein sequences at the NCBI, including the cDNA EST (http://www.ncbi.nlm.nih.gov/nucest/), TSA (http://www.ncbi.nlm.nih.gov/genbank/tsa), and Uniprot databases[29]. All sequences from these databases that matched parvovirus sequences (E-value $< 10^{-3}$) were selected and used as queries to perform BLASTX or BLASTP reciprocal searches of the cDNA EST, TSA and Uniprot databases. Sequences were considered PRSs when they matched known parvovirus sequences with associated BLASTX or BLASTP E-values $< 10^{-3}$. The eleven new transcriptomes were also screened for the presence of PRSs as described above. Complete and 5′- or 3′-truncated ORFs were detected using Prodigal V2_60[45,46] using the standard genetic code. ORFs displaying internal stretches of undetermined nucleotides (N) were also considered. Putative protein sequences were first annotated by detecting protein homology using the HHblits component of the HHSuite package[47,48] of nr protein sequences of the NCBI database. ORFs matching parvoviral proteins (E-values $< 10^{-3}$) were selected and used as query for reciprocal BLASTP searches of the cDNA EST, TSA and Uniprot databases as described above.

Detection of endogenous parvovirus-related sequences. The 463 PRSs found from BLASTX searches described above were used as queries against the reference genomic sequences (Refseq_genomic, http://www.ncbi.nlm.nih.gov/refseq/), chromosome (http://www.ncbi.nlm.nih.gov/genome/), GSS (Genomic survey sequences) (http://www.ncbi.nlm.nih.gov/nucgss/) and WGS (Whole-Genome Shotgun contigs) (http://www.ncbi.nlm.nih.gov/genbank/wgs) databases using BLASTN and tBLASTN, with a minimum percentage similarity cutoff of 95% and an E-value cutoff of 10^{-5}. Five hundred nucleotide long genomic fragments located up- and down-stream of each PRS were scanned for transposable elements (TE) or repetitive sequences using WSCensor (http://www.girinst.org/censor/).

Phylogenetic analyses. The putative amino acid sequences of the SF3 helicase domains of parvoviral NS1and PRS NS1-like proteins were used for phylogenetic analyses. All PRSs were translated *in silico* using ORF finder (cut off >300 bp) (http://www.ncbi.nlm.nih.gov/projects/gorf/). Putative domains of the resulting proteins were predicted using Interproscan5[49]. Among the 463 PRSs, 264 contained a SF3 helicase domain sequence from which 191 aa-long fragments were aligned with the corresponding SF3 fragments from the 74 known parvoviruses (Table S1) using MUSCLE 3.7 (16 iterations) with default settings[50]. Aligned sequences were manually edited (full alignment of the 191 aa-long partial SF3 helicase domains, is provided in FASTA format, and is presented in Supplementary Figure S1). Maximum likelihood phylogenetic trees were produced from this alignment using PhyML 3.1[51] with a Blossum + G + F + I amino acid substitution model chosen as the best-fit using ProtTest[52]. Five hundred bootstrap replicates were used to test the support of branches. Trees were visualized with FigTree 1.4 (http://tree.bio.ed.ac.uk/software/figtree/). In addition, a second tree focusing on the evolutionary relationships of 118 PRSs derived from platyhelminthes together with the 74 representative parvovirus SF3 domain sequences (Table S1) was constructed using the same approaches described above (full alignment of the 156 aa-long partial SF3 helicase domains is provided in FASTA format and in Supplementary Figure S2). The variola D5 protein (Genbank accession number: P33069) was used in both cases as an outgroup to root the trees[53–58].

References

1. Rosario, K. & Breitbart, M. Exploring the viral world through metagenomics. *Curr Opin Virol* **1,** 289–297 (2011).
2. Feschotte, C. & Gilbert, C. Endogenous viruses: insights into viral evolution and impact on host biology. *Nat Rev Genet* **13,** 283–296 (2012).
3. Wren, J. D. *et al.* Plant virus biodiversity and ecology. *PLoS Biol* **4,** e80 (2006).
4. Temmam, S., Davoust, B., Berenger, J. M., Raoult, D. & Desnues, C. Viral metagenomics on animals as a tool for the detection of zoonoses prior to human infection? *Int J Mol Sci* **15,** 10377–10397 (2014).
5. Roossinck, M. J., Martin, D. P. & Roumagnac, P. 2015. Plant Virus Metagenomics: Advances in Virus Discovery. *Phytopathology* **105(6),** 716–727 (2015).
6. King A. M. Q., Lefkowitz A., Adams M. J., Carstens E. B. (eds). *Virus Taxonomy: Ninth Report of the International Committee on Taxonomy of Viruses*. Elsevier/Academic Press, San Diego, pp 353–369 (2011).
7. Mokili, J. L., Rohwer, F. & Dutilh, B. E. Metagenomics and future perspectives in virus discovery. *Curr Opin Virol* **2,** 63–77 (2012).
8. Bernardo, P. *et al.* 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 (2016).
9. Rosario, K., Duffy, S. & Breitbart, M. A field guide to eukaryotic circular single-stranded DNA viruses: insights gained from metagenomics. *Arch Virol* **157,** 1851–1871 (2012).
10. Roux, S. *et al.* Analysis of metagenomic data reveals common features of halophilic viral communities across continents. *Environ Microbiol* Mar **18(3),** 889–903 (2016).
11. Labonté, J. M. & Suttle, C. A. Previously unknown and highly divergent ssDNA viruses populate the oceans. *ISME J* **7,** 2169–2177 (2013).
12. Ng, T. F. F. *et al.* High variety of known and new RNA and DNA viruses of diverse origins in untreated sewage. *J Virol* **86,** 12161–12175 (2012).
13. Rosario, K. *et al.* Diverse circular ssDNA viruses discovered in dragonflies (Odonata: Epiprocta). *J Gen Virol* **93,** 2668–2681 (2012).
14. Hopkins, M. *et al.* Diversity of environmental single-stranded DNA phages revealed by PCR amplification of the partial major capsid protein. *ISME J* **8,** 2093–2103 (2014).
15. Bergoin, M. & Tijssen, P. *Densoviruses: a Highly Diverse Group of Arthropod Parvoviruses, In Insect Virology*, (eds Asgari, S. & Johnson, K. N.) 59–72 (Caister Academic Press, 2010).
16. Cotmore, S. F. & Tattersall, P. Parvoviruses: Small Does Not Mean Simple. *Annu Rev Virol* **1,** 517–537 (2014).
17. Zádori, Z. *et al.* A viral phospholipase A2 is required for parvovirus infectivity. *Dev Cell* **1,** 291–302 (2001).
18. Cotmore, S. F. *et al.* The family Parvoviridae. *Arch Virol* **159,** 1239–1247 (2014).
19. Chapman, A. D. Numbers of Living Species in Australia and the World. Second edition. Report for the Australian Biological Resources Study (2009).

20. Belyi, V. A., Levine, A. J. & Skalka, A. M. Sequences from ancestral single-stranded DNA viruses in vertebrate genomes: the parvoviridae and circoviridae are more than 40 to 50 million years old. *J Virol* **84,** 12458–12462 (2010).
21. Liu, H. *et al.* Widespread endogenization of densoviruses and parvoviruses in animal and human genomes. *J Virol* **85,** 9863–9876 (2011).
22. Gudenkauf, B. M., Eaglesham, J. B., Aragundi, W. M. & Hewson, I. Discovery of urchin-associated densoviruses (Parvoviridae) in coastal waters of the Big Island, Hawaii. *J Gen Virol.* Mar **95** (Pt 3), 652–658 (2014).
23. Hewson, I. *et al.* Densovirus associated with sea-star wasting disease and mass mortality. *Proc Natl Acad Sci USA* Dec 2 **111(48),** 17278–17283 (2014).
24. Bowles, D., Rabinowitz, J. & Samulski, R. *In Parvoviruses*. (eds Kerr, J. *et al.*) 15–23 (Hodder Arnold, London, 2006).
25. Clavijo, G., van Munster, M., Monsion, B., Bochet, N. & Brault, V. Transcription of densovirus endogenous sequences in Myzus persicae genome. *J Gen Virol. Apr* **97(4),** 1000–1009 (2016).
26. Flegel, T. W. Hypothesis for heritable, anti-viral immunity in crustaceans and insects. *Biol Direct* **4,** 32 (2009)
27. Liu, S., Vijayendran, D. & Bonning, B. C. Next generation sequencing technologies for insect virus discovery. *Viruses* **3,** 1849–1869 (2011).
28. DeBoever, C. *et al.* Whole transcriptome sequencing enables discovery and analysis of viruses in archived primary central nervous system lymphomas. *PLoS One* **8,** e73956 (2013).
29. Consortium, T. U. UniProt: a hub for protein information. *Nucleic Acids Res* **43,** 204–212 (2014).
30. Pruitt, K. D., Tatusova, T. & Maglott, D. R. NCBI Reference Sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res* **33,** D501–D504 (2005).
31. Kapoor, A., Simmonds, P. & Lipkin, W. I. Discovery and characterization of mammalian endogenous parvoviruses. *J Virol* **84,** 12628–12635 (2010).
32. Thézé, J., Leclercq, S., Moumen, B., Cordaux, R. & Gilbert, C. Remarkable diversity of endogenous viruses in a crustacean genome. *Genome Biol Evol* **6,** 2129–2140 (2014).
33. Arriagada, G. & Gifford, R. J. Parvovirus-derived endogenous viral elements in two South American rodent genomes. *J Virol* **88,** 12158–12162 (2014).
34. Metegnier, G. *et al.* Comparative paleovirological analysis of crustaceans identifies multiple widespread viral groups. *Mob DNA 16* **6,** 16 (2015).
35. Liu, H. *et al.* Widespread endogenization of densoviruses and parvoviruses in animal and human genomes. *J Virol* **85,** 9863–9876 (2011).
36. Ge, X. *et al.* Metagenomic analysis of viruses from bat fecal samples reveals many novel viruses in insectivorous bats in China. *J Virol* **86,** 4620–4630 (2012).
37. Iyer, L. M., Koonin, E. V., Leipe, D. D. & Aravind, L. Origin and evolution of the archaeo-eukaryotic primase superfamily and related palm-domain proteins: structural insights and new members. *Nucleic Acids Res* **33,** 3875–3896 (2005).
38. Van Munster, M., Janssen, A., Clérivet, A. & van den Heuvel, J. Can plants use an entomopathogenic virus as a defense against herbivores? *Oecologia* **143,** 396–401 (2005).
39. Hu, Z., Li, G., Li, G., Yao, Q. & Chen, K. Bombyx mori bidensovirus: The type species of the new genus Bidensovirus in the new family Bidnaviridae. *Chinese Sci Bull* **58,** 4528–4532 (2013).
40. Gayral, P. *et al.* Next-generation sequencing of transcriptomes: a guide to RNA isolation in nonmodel animals. *Mol Ecol Resour* **11,** 650–661 (2011).
41. Simpson, J. T., Wong, K., Jackman, S. D., Schein, J. E., Jones, S. J. M. & Birol, I. ABySS: a parallel assembler for short read sequence data. *Genome Res* **19,** 1117–1123 (2009).
42. Huang, X. CAP3: A DNA Sequence Assembly Program. *Genome Res* **9,** 868–877 (1999).
43. Gayral, P. *et al.* Reference-free population genomics from next-generation transcriptome data and the vertebrate-invertebrate gap. *PLoS Genet* **9,** e1003457 (2013).
44. Romiguier, J. *et al.* Comparative population genomics in animals uncovers the determinants of genetic diversity. *Nature* **515,** 261–263 (2014).
45. Hyatt, D. *et al.* Prodigal: prokaryotic gene recognition and translation initiation site identification. BMC *Bioinformatics 8* **11,** 119 (2010).
46. Hyatt, D., Locascio, P. F., Hauser, L. J. & Uberbacher, E. C. Gene and translation initiation site prediction in metagenomic sequences. *Bioinformatics 1* **28(17),** 2223–2230 (2012).
47. Remmert, M., Biegert, A., Hauser, A. & Söding, J. HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment. *Nat Methods* **9,** 173–175 (2011).
48. Söding, J. Protein homology detection by HMM-HMM comparison. *Bioinformatics 1* **21(7),** 951–960 (2005).
49. Jones, P. *et al.* InterProScan 5: genome-scale protein function classification. *Bioinformatics 1* **30,** 1236–1240 (2014).
50. Edgar, R. C. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res 19* **32,** 1792–1797 (2004).
51. Guindon, S. *et al.* New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Syst Biol* **59(3),** 307–321 (2010).
52. Abascal, F., Zardoya, R. & Posada, D. ProtTest: selection of best-fit models of protein evolution. *Bioinformatics 1* **21(9),** 2104–2105 (2005).
53. Chandler, J. A., Liu, R. M. & Bennett, S. N. RNA shotgun metagenomic sequencing of northern California (USA) mosquitoes uncovers viruses, bacteria, and fungi. *Front Microbiol* **24(6),** 185 (2015).
54. Liu, S., Chen, Y. & Bonning, B. C. RNA virus discovery in insects. *Curr Opin Insect Sci* **8,** 54–61 (2015).
55. François, S. *et al.* A Novel Itera-Like Densovirus Isolated by Viral Metagenomics from the Sea Barley Hordeum marinum. *Genome Announce*. Dec 4, **2(6)** (2014).
56. Jousset, F. X., Baquerizo, E. & Bergoin, M. *A new densovirus isolated from the mosquito Culex pipiens (Diptera: Culicidae)* **67(1),** 11–16 (2000).
57. Kisary, J., Avalosse, B., Miller-Faures, A. & Rommelaere, J. The Genome Structure of a New Chicken Virus Identifies It as a Parvovirus. *J Gen Virol* **66** (Pt 10), 2259–2263 (1985).
58. Tang, K. F. & Lightner, D. V. Infectious hypodermal and hematopoietic necrosis virus (IHHNV)-related sequences in the genome of the black tiger prawn Penaeus monodon from Africa and Australia. *Virus Res* **118(1–2)**, 185–191 (2006).

Acknowledgements

We warmly thank N. Galtier, M. Weil, C. Atyame, S. Hourdez, G. Tsagkogeorga, and M. Ballenghien for their help in sample collections, RNA isolation and transcriptome acquisition. The eleven new transcriptomes generated in this study were generously provided by N. Galtier and supported by European Research Council advanced grant 232971 (PopPhyl).

Author Contributions

Data Acquisition S.F., D.F., D.B. and P.G.; Analysis and interpretation of data S.F., P.R., D.P. and M.O.; Manuscript preparation S.F., P.R., D.P. and M.O.; Study supervision S.F., P.R., R.F. and M.O.

Additional Information

Supplementary information accompanies this paper at http://www.nature.com/srep

Competing financial interests: The authors declare no competing financial interests.

How to cite this article: François, S. *et al.* Discovery of parvovirus-related sequences in an unexpected broad range of animals. *Sci. Rep.* **6**, 30880; doi: 10.1038/srep30880 (2016).

This work is licensed under a Creative Commons Attribution 4.0 International License. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in the credit line; if the material is not included under the Creative Commons license, users will need to obtain permission from the license holder to reproduce the material. To view a copy of this license, visit http://creativecommons.org/licenses/by/4.0/

© The Author(s) 2016

Discovery of parvovirus-related sequences in an unexpected broad range of animals

S. François[1], D. Filloux[2], P. Roumagnac[2], D. Bigot[3], P. Gayral[3, 4], D. P. Martin[5], R. Froissart[2,6] & M. Ogliastro[1*]

Supplementary Figures and Tables

SUPPLEMENTARY DATA

Table S1. Reference parvoviral genomes used to search for PRSs and for phylogenetic analyses.

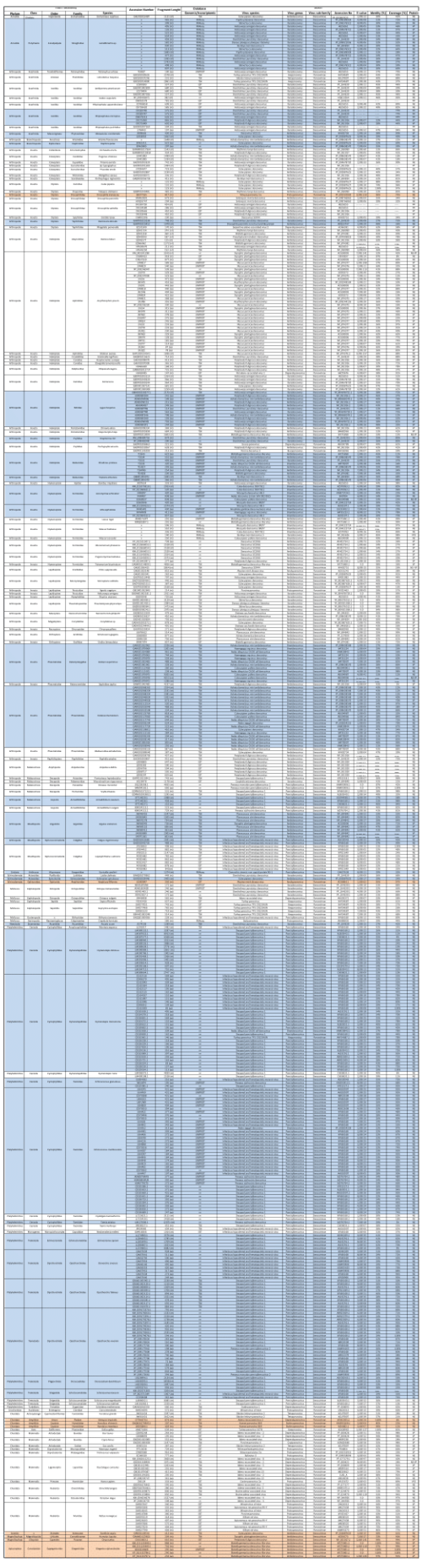
Table S2. Detailed information about PRSs found in both genomic and transcriptomic databases. Host taxonomy and BLAST results are given for each PRS. Animal species containing PRSs displaying < 30% identity with extant parvoviruses are highlighted in blue. Highlighted in yellow are animal species containing PRSs displaying > 85% identity with extant parvoviruses.

Table S3. Detailed information about PRSs that display homologies with genomic databases only (Refseq_genomic, chromosome, gss and WGS).

Figure S1: Amino acid alignment of the SF3 helicase domain of 264 PRSs and 74 parvoviruses. The variola D5 protein SF3 domain was used as outgroup. The alignment was produced using MUSCLE 3.7 with default settings. Aligned sequences were trimmed and columns with gaps were manually removed.

Figure S2: Amino acid alignment of the SF3 helicase domain of 118 platyhelminthes PRSs and 74 parvoviruses. The variola D5 protein SF3 domain was used as an out-group. The alignment was produced using MUSCLE 3.7 with default settings. Aligned sequences were trimmed and columns in the alignment with gaps were manually removed.

Phylum	Class	Order	Family	Organism	Accession No. (PRS)	Genomic Database	Position (L)	Position (R)	Length (bp)	%Coverage	E value	% Identity	Accession No.	Size (bp)	Adjacent transposable element	Degenerated mutations relative to viral genes
Arthropoda	Insecta	Araneae	Theridiidae	**Latrodectus hesperus**	GBCS01037246	WGS	553	346	207	90%	9,00E-102	100%	JIRX01010780.1	2545	Yes	internal stop codons
					GBCS01014800		3499	3610	111	3%	2,00E-41	95%	JIRX01010779.1	3611	Yes	
		Mesostigmata	Phytoseiidae	**Metaseiulus occidentalis**	JI999665	genome (reference only)	2596	3387	791	100%	0.0	99%	NW_003803976.1	536467	Yes	internal stop codons
	Branchiopoda	Diplostraca	Daphniidae	**Daphnia pulex**	EFX61411	WGS	2045	3772	1727	100%	0.0	99%	ACJG01014942.1	6319	No	internal stop codons
		Coleoptera	Cupedidae	**Priacma serrata**	GACO01003330	WGS	564	491	73	9%	6,00E-26	95%	AGRH01000640.1	8293	No	internal stop codons
		Diptera	Tephritidae	**Bactrocera dorsalis**	GAKP01023238	WGS	453219	452802	417	55%	0.0	99%	JFBF01000220.1	459848	No	
							408	378	30	4%	2,00E-06	100%	JFBF01000093.1	1167390	No	
		Hemiptera	Delphacidae	**Nilaparvata lugens**	HS466630	WGS	1335	1502	167	100%	4,00E-81	100%	AOSB01120472.1	2333	No	
					GANM01012737		1548	824	724	100%	0.0	99%	AOSB01120472.1	2333	No	internal stop codons
			Pentatomidae	**Halyomorpha halys**	GBHT01013004	WGS	12819	15765	2946	98%	0.0	98%	JMPT01017517.1	16063	No	
							440	3343	2903	98%	0.0	95%	JMPT01038455.1	18195	Yes	
							41367	43129	1762	60%	0.0	96%	JMPT01084032.1	44282	Yes	
							7323	6481	842	29%	0.0	98%	JMPT01052785.1	7323	No	
							1809	956	853	30%	0.0	97%	JMPT01084370.1	3319	Yes	
			Psyllidae	**Diaphorina citri**	XM_008484718	WGS	3021	3732	711	100%	0.0	100%	AWGM01156254.1	11892	No	
							1	81	80	9%	2,00E-33	100%	AWGM01025324.1	226	No	
							878	817	61	10%	8,00E-19	95%	AWGM01000952.1	1117	No	
							291	243	48	8%	2,00E-14	98%	AWGM01064873.1	293	No	
				Pachypsylla venusta	GAOP01032067	WGS	6232	6859	627	100%	0.0	99%	AZLD01217521.1	7421	Yes	internal stop codons
					GAOP01141099		3944	3729	215	100%	6,00E-100	98%	AZLD01011675.1	4115	No	
		Hymenoptera	Apidae	**Bombus impatiens**	JI097654	WGS	1329	1660	331	99%	2,00E-163	98%	AEQM02016195.1	3848	No	internal stop codons
			Formicidae	**Acromyrmex echinatior**	XM_011070319.1	WGS	27037	25815	1222	100%	0.0	100%	AEVX01004976.1	79951	No	
					EGI68940		1482	1790	308	19%	4,00E-56	96%	AEVX01005448.1	2528	Yes	
							3686	3889	203	8%	9,00E-36	99%	AEVX01005453.1	5028	Yes	
				Atta cephalotes	XM_012209122.1	WGS	20667	21660	993	100%	0.0	96%	ADTU01008063.1	23996	No	internal stop codons
					W4VWX9		112	699	587	47%	6,00E-127	100%	ADTU01004457.1	5143	Yes	
				Lasius niger	KMQ81887.1	WGS	604	2	602	100%	5,00E-124	97%	LBMM01030606.1	776	No	
				Monomorium pharaonis	XM_012666055.1	WGS	20535	17977	2558	100%	0.0	98%	BBSX02014388.1	21843	No	
							1	92	91	3%	3,00E-39	100%	BBSX02013400.1	2878	No	
					XM_012666053.1		1	92	91	3%	3,00E-39	100%	BBSX02013400.1	2878	No	
							20535	17965	2570	100%	0.0	98%	BBSX02014388.1	21843	No	
							1	92	91	3%	3,00E-39	100%	BBSX02013400.1	2878	No	
				Pogonomyrmex barbatus	XM_011645096.1	WGS	4152	6241	2089	100%	0.0	99%	ADIH01015354.1	12755	Yes	
					XM_011645095.1		4152	6238	2086	100%	0.0	100%	ADIH01015354.1	12755	Yes	
	Maxillopoda	Siphonostomatoida	Caligidae	**Caligus rogercresseyi**	GAZX01015685	WGS	71	2189	2118	78%	0.0	99%	LBBV01106530.1	2515	Yes	internal stop codons
							1863	329	1534	57%	0.0	99%	LBBU01242504.1	1863	No	
					GAZX01042735		476	963	487	100%	0.0	100%	LBBV01271803.1	994	No	
							519	32	487	100%	0.0	100%	LBBU01149111.1	2470	Yes	
Mollusca	Cephalopoda	Octopoda	Octopodidae	**Octopus bimaculoides**	KOF81493.1	WGS	5332	5847	515	100%	6,00E-103	93%	LGKD01669633.1	7219	Yes	
					KOF62817.1		935	1672	737	100%	3,00E-174	100%	LGKD01331296.1	2875	Yes	
Nematoda	Secernentea	Ascaridida	Ascarididae	**Ascaris suum**	JI174103	WGS	14090	14131	41	3%	6,00E-12	100%	ANBK01004617.1	39446	Yes	
							4254	4213	41	3%	6,00E-12	100%	AMPH01016759.1	26950	Yes	
							25879	25838	41	3%	6,00E-12	100%	AEUJ02001051.1	53597	Yes	
							36	1	35	2%	5,00E-07	97%	AEUJ02016456.1	219	No	
Platyhelminthe	Cestoda	Cyclophyllidea	Hymenolepididae	**Hymenolepis diminua**	LM390089.1	Genbank nr	1032	2378	1346	100%	0.0	100%	LM390089.1	2383	No	
				Hymenolepis microstoma	CDS32658	WGS	74578	73628	950	93%	0.0	100%	CBLW020003802.1	75037	No	
				Hymenolepis nana	LM406075.1	Genbank nr	1	4839	4838	100%	0.0	100%	LM406075.1	4839	No	
							35	2518	2483	100%	0.0	96%	LM409198.1	2518	No	
			Taeniidae	**Echinococcus multilocularis**	U6HGG3	WGS	5004	6296	1292	100%	0.0	99%	CBLO020000155.1	7682	No	
							3092	4384	1292	100%	0.0	99%	CBLO020000266.1	61081	No	
							890	2182	1292	100%	0.0	97%	CBLO020000868.1	2622	No	
							1842	604	1238	100%	0.0	99%	CBLO020000226.1	4432	No	
							10653	9478	1175	99%	0.0	97%	CBLO020000544.1	33361	No	
							346	1089	743	57%	2,00E-163	99%	CBLO020000490.1	10554	No	
							2422	1733	689	53%	2,00E-147	95%	CBLO020000966.1	5011	No	
							975	412	563	43%	1,00E-119	96%	CBLO020000024.1	3149	No	
							3178	2615	563	43%	4,00E-118	97%	CBLO020000171.1	4825	No	
							3491	3934	446	34%	3,00E-90	97%	CBLO020000092.1	4638	No	
				Hydatigera taeniaeformis	LL723621.1	Genbank nr	1	3111	3110	100%	0.0	100%	LL723621.1	3111	No	internal stop codons
							241	507	266	15%	2,00E-118	97%	LL748207.1	507	No	
							579	724	145	7%	3,00E-64	99%	LL740453.1	724	No	
							1391	1465	74	2%	7,00E-22	95%	LL728283.1	1751	No	
							72	114	42	1%	3,00E-08	98%	LL727136.1	1978	No	
				Taenia asiatica	LM127006.1	Genbank nr	381	620	239	100%	9,00E-99	100%	LM127006.1	12071	No	internal stop codons
	Trematoda	Echinostomida	Echinostomatidae	**Echinostoma caproni**	LL286487.1	Genbank nr	1168	239	929	100%	0.0	100%	LL286487.1	1623	No	internal stop codons
		Opisthorchiida	Opisthorchiidae	**Clonorchis sinensis**	GAAS7638	WGS	1847	1029	818	100%	9,00E-172	95%	BADR02003990.1	27389	No	
				Opisthorchis viverrini	XP_009176824	WGS	97	1590	1493	97%	0.0	97%	JACJ01026580.1	2354	No	
							3	596	593	45%	6,00E-137	95%	JACJ01037170.1	706	No	
							3	527	524	34%	1,00E-117	95%	JACJ01041207.1	527	No	
							1	474	473	30%	1,00E-109	96%	JACJ01043459.1	474	No	
							1	248	247	29%	1,00E-100	97%	JACJ01044501.1	452	No	
							4	246	242	15%	5,00E-50	100%	JACJ01060560.1	248	No	
							3	206	203	13%	2,00E-38	97%	JACJ01069043.1	206	No	
		Plagiorchiida	Dicrocoeliidae	**Dicrocoelium dendriticum**	LK446595.1	Chromosomes	1	1838	1837	100%	0.0	100%	LK446595.1	1838	No	internal stop codons
					LK428994.1		1	2134	2133	100%	0.0	100%	LK428994.1	2134	No	internal stop codons
		Strigeatida	Schistosomatidae	**Schistosoma margrebowiei**	LL888002.1	GenBank nr	1	8085	8084	100%	0.0	100%	LL888002.1	8085	Yes	internal stop codons
				Schistosoma mattheei	LM181082.1	GenBank nr	1	2250	2249	100%	1,00E-96	100%	LM181082.1	2250	No	internal stop codons
	Turbellaria	Tricladida	Dugesiidae	**Schmidtea mediterranea**	GAKN01010353	WGS	8883	9542	659	100%	0.0	100%	AUVC01112254.1	32292	No	
							16546	15886	660	100%	0.0	99%	ASXK01032274.1	18356	No	
							48324	48984	660	100%	0.0	99%	AAWT01002411.1	169447	No	



Sub-family	Genus	Virus species or variant	Accession number
Densovirinae	Ambidensovirus	Acheta domesticus densovirus	HQ827781
		Acheta domesticus mini ambidensovirus	KF275669
		Blattella germanica densovirus	AY189948
		Blattella germanica densovirus-like virus	JQ320376
		Culex pipiens densovirus	FJ810126
		Diatraea saccharalis densovirus	AF036333
		Dysaphis plantaginea densovirus	FJ040397
		Galleria mellonella densovirus	L32896
		Helicoverpa armigera densovirus	JQ894784
		Junonia coenia densovirus	KC883978
		Mythimna loreyi densovirus	AY461507
		Myzus persicae densovirus	AY148187
		Periplaneta fuliginosa densovirus	AF192260
		Planococcus citri densovirus	AY032882
		Pseudoplusia includens densovirus	JX645046
		Solenopsis invicta densovirus	KC991097
	Brevidensovirus	Aedes aegypti densovirus 1	M37899
		Aedes aegypti densovirus 2	FJ360744
		Aedes albopictus densovirus 1	AY095351
		Aedes albopictus densovirus 2	X74945
		Aedes albopictus densovirus 3	AY310877
		Anopheles gambiae densovirus	EU233812
		Culex pipiens pallens densovirus	EF579771
		Haemagogus equinus densovirus	AY605055
		Mosquito densovirus BR/07	GU452720
	Hepandensovirus	Fenneropenaeus chinensis hepandensovirus	JN082231
		Penaeus chinensis hepandensovirus	AY008257
		Penaeus merguiensis hepandensovirus	DQ458781
		Penaeus monodon hepandensovirus 1	DQ002873
		Penaeus monodon hepandensovirus 2	EU247528
		Penaeus monodon hepandensovirus 3	EU588991
		Penaeus monodon hepandensovirus 4	FJ410797
	Iteradensovirus	Bombyx mori densovirus	AY033435
		Casphalia extranea densovirus	AF375296
		Danaus plexippus plexippus densovirus	KF963252
		Dendrolimus punctatus densovirus	AY665654
		Helicoverpa armigera densovirus	HQ613271
		Hordeum marinum itera-like densovirus	KM576800
		Papilio polyxenes densovirus	JX110122
		Sibine fusca densovirus	JX020762
	Penstyldensovirus	Penaeus monodon penstyldensovirus 1	GQ411199
		Penaeus monodon penstyldensovirus 2	AY124937
		Penaeus stylirostris penstyldensovirus 1	AF273215
		Penaeus stylirostris penstyldensovirus 2	GQ475529
Parvovirinae	Amdoparvovirus	Aleutian mink disease virus	JN040434
		Gray fox amdovirus	JN202450
	Aveparvovirus	Turkey parvovirus	GU214706
	Bocaparvovirus	California sea lion bocavirus 1	JN420361
		Canine bocavirus 1	JN648103
		Canine minute virus	FJ214110
		Feline bocavirus	JQ692585
		Human bocavirus 1	JQ923422
		Human bocavirus 4	FJ973561
		Porcine bocavirus 1	HM053693
		Porcine bocavirus 3	JF429834
		Porcine bocavirus 5	HQ223038
	Copiparvovirus	Bovine parvovirus 2	AF406966
		Porcine parvovirus 4	GQ387499
	Dependoparvovirus	Adeno-associated virus 1	AF063497
		Adeno-associated virus 5	AF085716
		Bat adeno-associated virus	GU226971
		California sea lion adeno-associated virus 1	JN420372
		Goose parvovirus	U25749
	Erythroparvovirus	Human parvovirus B19	M13178
	Protoparvovirus	Bufavirus 1	JX027296
		Canine parvovirus	M19296
		Mouse parvovirus 1	U12469
		Mouse parvovirus 3	DQ196318
		Porcine parvovirus PT4	U44978
		Rat parvovirus NTU1	AF036710
	Tetraparvovirus	Bovine hokovirus	EU200669
		Eidolon helvum parvovirus 1	JQ037753
		Human parvovirus 4	AY622943
		Porcine hokovirus	EU200677

Figure S1

>variole_D5_NC_001611
IINDIQPLTKKNRELYEKTLSSCL-CGATKGCLTFFFGETATGKSLLKSAISDLFVETGQTILTVLDKGPNIANMHKRSVFCSELPDFACSGTKKIRCVIGRPCFSNKINNRN-----HATIIIDTNYPVFDRIDNALMR----RIAVVRFRTHFS
>Penaeus_monodon_hepandensovirus_3_EU588991
RPKTIPIVSQNKTVQWIQDFMDIMHGNLPKVNCMMLYGNSNSGKTQLIEALTGLVNT-AIMTNVGDGGTFHFSNITMSTIVVGNETKIRTQTIEQWKGLCGGENVTPMPKYKEHKTHMRKPVFLTNQHPLMDISHYDDRRAIENRSFMYKVELGSE
>Hymenolepis_microstoma_CDS29136
LIIQYFKYR-----GVVFEVGTELHEEWQQAQDMPPIQPHVTCDSELRRARTKSTIA----ERTSRHTKFTLDELKRSDIRQFEEIFSAFSVNEIWI--------NMDKKTTKVNKIAKQQIQSLNAERLQSEKQNSYLQNL--QMMKHDCTVKHP
>Hymenolepis_microstoma_CDS33099
LTKNLFEHDEIVSIEDQEPLINRADDENGRRSEAI---PSTSART-------------GLEQSS--SPSFL-----GH------DPYGAEEGQEIFPQQLG----DMETRTKGALSLR----VAAEDESMENILLGGISITPDQ-----NYHTVII
>Penaeus_monodon_hepandensovirus_2_EU247528
RPKTIPIVSQNKTVQWIQDFMDIMHGNLPKVNCMMLYGNSNSGKTQLIEALTGLVNT-AIMTNVGDGGTFHFSNITMSTIVVGNETKIRTQTIEQWKGLCGGENVTPMPKYKEHKTHMRKPVFLTNQHPLVDISHYDDRRAIENRSFMYKVELGSE
>Penaeus_merguiensis_hepandensovirus_DQ458781
RPKTIPIISKNKTVQWIQDFMDIMHGNLPKINCMMLYGNSNSGKTQLIDALTGLVNT-AVMTNVGDGGTFHFSNITMSTIVVGNETKIRTQTIEQWKGLCGGENVTPMPKYKEHKTHMRKPVFLTNQHPLVDISHYDDRRAIENRSFMYKVELGSE
>Penaeus_chinensis_hepandensovirus_AY008257
RPKTIPIISKNKTVQWIQDFMDIMHGNLPKINCMMLYGNSNSGKTQLIEALTGLVNT-AIMTNVGDGGTFHFSNITMSTIVVGNETKIRTQTIEQWKGLCGGENVTPMPKYKEHKTHMRKPVFLTNQHPLVDISHYDDRRAIENRSFMYKVELGKE
>Fenneropenaeus_chinensis_hepandensovirus_JN082231
RPKTIPIISKNKTVQWIQDFMDIMHGNLPKINCMMLYGNSNSGKTQLIEALTGLVNT-AIMTNVGDGGTFHFSNITMGTIVVGNETKIRTQTIEQWKGLCGGENVTPMPKYKEHKTHMRKPVFLTNQHPLVDISHYDDRRAIENRSFMYKVELGKE
>Penaeus_monodon_hepandensovirus_4__FJ410797
RPKTIPIISKNKTVQWIQDFMDIMHGNLPKINCMMLYGNSNSGKTQLIEALTGLVNT-AVMTNVGDGGTFHFSNITMSTIVVGNETKIRTQTIEQWKGLCGGENVTPMPKYKEHKTHMRKPVFLTNQHPLVEISNYDDRKAIENRCFMYKVELGSE
>Penaeus_monodon_hepandensovirus_1_DQ002873
RPKTIPIVSKNKTVQWIQDFMDIIHGNLPKINCMMLYGNSNSGKTQLIEALTGLINT-AIMTNVGDGGTFHFSNITMSTIVVGNETKIRTQTIEQWKGLCGGENITPMPKYKEHKTHMRKPVFLTNQHPLVEISNYDDRKAIENRCFMYKVELGSE
>Acheta_domesticus_mini_ambidensovirus__KF275669
IIEQLLMYQFESVSDFIHQLYYVLNKQNGKKNCIELIGPPSSYKSTFLHWVAEAFIAVGIINIVNKTAPFGFAPAVNKRVLIIDDYNFSSEYHETLLNILSGTTCNVNVKYKTNGYVTHTPVLMASNYTFTSTRFEH-------RMVTFYWKEYK-
>Acheta_domesticus_densovirus_HQ827781
FLNELILFQEEGVKELLMNIVAWFNLGWPKVNTVCIIGNHNCGKNYFWDAVCCLGLNVGFLGRVNKTNKFALQDCVHKRIVIGNEISLEDGAKEDMKKLCEGCPFNVSVKHQSDGIVARTPVCLISNN---------CIPIACDRMKVFYWRPCSY
>Planococcus_citri_densovirus_AY032882
FLDELLKFQEEIIKPFLSNLRNWFNREGPKMNALCVIGPPNSGKNYFFDAIIALANNVGHIGRVNKTNNFALQDVVNRRIVVGNEISMEEGAKEDFKKLCEGTAFNVRVKFKADCIFTKTPVLLISNNELDICWDPHFKDV---RLKTIRWNTAPL
>Solenopsis_invicta__densovirus_KC991097
FVDDLLRFQEENIVELLTNVRDWFNDGWPKMNALCVIGPPNSGKNYFWDMFCSVAYNVGHIGRVNKTNQFALQECYGRRLVVGNEVSMEDGAKEDFKKLCEGTAFNIRVKYQGDKIFTKAPVLLISNFELDICYDTHFKNV---RLHTIRWVKCPL
>Papilio_polyxenes_densovirus_JX110122
IVEELLDYQFDAKRRFITDLYEILEKKHQKTNTFQIVSPPSAGKNFFIETVLAFYWNTGVIQNFNRYHSFPLMEAVNRRVNYWDEPNFEPDATETLKKLFAGTALKASVKFQKEANVQKTPVIITANRKFTSEVWDD-------RIIKYEWYQCPM
>Casphalia_extranea_densovirus_AF375296
IVEELLDYQEEAKRKFITDLYEILEKKHQKTNTFQIVSPPSAGKNFFIETVLAFYWNTGVIQNFNRYNNFPLMEAVNRRVNYWDEPNFEPDATETLKKLFAGTALKATVKFQKEANVQKTPVIITANHKFTKEVWDD-------RIIKYYWYQCPK
>Sibine_fusca_densovirus_JX020762
IVEELLDFQEEAKRKFITDLYEILEKKHQKTNTFQIVSPPSAGKNFFIETVLAFYWNTGVIQNFNRYNNFPLMEAVNRRVNYWDEPNFEPDAVETLKKLFAGTALKATVKFQKEANVQKTPVIITANYKFTKEVWDD-------RIIKYYWYPCPK
>Hordeum_marinum_iteradensovirus_KM576800
IVEELLDFQEDAKRKFITDLYEILEKKHQKTNTFQIVSPPSAGKNFFIETVLAFYWNTGVIQNFNRYNNFPLMEAVNRRVNYWDEPNFEPDATETLKKLFAGTALKATVKFQKEANVQKTPVIITANFKFTKEVWDD-------RIIKYYWYSCPK
>Bombyx_mori_densovirus_AY033435
IVEELLDFQEDAKKRFITDLYEILEKKHQKTNTFQIVSPPSAGKNFFIETVLAFYWNTGVIQNFNRYNNFPLMEAVNRRVNYWDEPNFEPDATETLKKLFAGTALKATVKFQKEANVQKTPVIITANYKFTKEVWDD-------RIIKYYWYPCPK
>Danaus_plexippus_plexippus_densovirus_KF963252
IVEELLDFQEDAKRKFITDLYEILEKKHQKTNTFQIVSPPSAGKNFFIETVLAFYWNTGVIQNFNRYNNFPLMEAVNRRVNYWDEPNFEPDATETLKKLFAGTALKATVKFQKEANVQKTPVIITANYKFTKEVWDD-------RIIKYYWYPCPK
>Dendrolimus_punctatus_densovirus_AY665654
IIERLLNYQNDAKKVFINDLWDILEKKRPKQNTFQIIGEPSAGKNYFIDAILSFYLNTGIILNFNRFSQFPLMEAVNRRVNVWNEPNFEPAALDTIKMLLAGDPLKVNVKYQKEQYLQKTPIIVMTNKVFPNDLAFHD------RVCTYFWKRAPF
>Helicoverpa_armigera_densovirus_HQ613271
AIEELLDFQTDKKREFVTDLWDVLEKKVPKKNTFQIISPPSAGKNFLLDAIFAFYWNVGMIRNFNKYESFPLMEAVDKRVNCWNEPNFEPSAEDTIKMILGGDPIKAAYKYESERNIARTPVIVMSNKNVFKINDAFED-----RIITYYWDRAEF
>Blattella_germanica_densovirus_AY189948
VVTELLTFQGENLVQFCRNLVDTLECNIPKRNCFVVCSPPSAGKNFFFDGVKDYYLNSGQMNNPNKYNQFAYQDCHNRRIIIWNEPNYEPREMENLKMLFAGDNLSANVKCKPQANVKRTPVIVLTNS-LPNFCQQTAFND---RVITYHWTQATF
>Blattella_germanica_densovirus-like_virus__JQ320376
IINKLLDYQGLAKKQFVTDLINVTDMNIPKCNCFLIVSPPSAGKNFLCDAIKDYYINVGQMQNPNKYNTFAYQDCHNRRLLIWNEPNYEPRETENLKMLFGGDNLSANVKCKPQANVKRTPVICMSNV---------VPRFANHRVITYYWNAAPF
>Cherax_quadricarinatus_densovirus_KP410261
IMDELVNYQEEAIIDFVTTLYNVLERKVPKLNCIVIHSPPSAGKNFFFDAVKDYYLNCGHLCNANKYNNFPFQDAEGRRIVLWNEPNYAPEFLEQIKEILGGDSTSVNVKYQSDTPVYRTPVIVLTNNKVSFMNHSAFID----RIRVFNWMAAPF
>Sea_star_densovirus_KM052275
VMDKLVQYQEDATLDFVNTLYNVLERKIPKLNCICIYSPPSAGKNFFFDAIKDYYINCGHLSNANKYNAFPFQDAEGRRIVLWNEPNYSPEFLEPIKELLGGDSTSVNVKYMHDTPVYRTPMIVLTNNVVSFMTHPAFKD----RVRVFNWMSAPF
>Dysaphis_plantaginea_densovirus_FJ040397
IAEELLCYQGDLISKFLYDVYAIVDKLVPKRNSMCVVSPPSAGKNFFFDAVASYFLNYGMFGTANKTNNFSWADGAGKRLVLWNEPNYETFHVEKIKELLGGDTTRVHVKYKGDQPLQGPPIFLLTNNTLSICNDPAFAD----RLVTYEWKSAPF
>Myzus_persicae_densovirus_AY148187
IAKELLNYQPEVVIEFLTTLYNVIDKRVPKLNSICIKSPPSAGKNFFFDAVASYLLSYGMFGTANKNNNFSWADGAGKRLVLWNEPNYEQYHIEKIKELLGGDTTRIHVKYANDVSVQRVPIIILTNNHLNIISHPAFND----RLRSYEWMSAGF
>Periplaneta_fuliginosa_densovirus_AF192260
ILNKLLAYQVVQIKSFLNTFYNVLERKLPKCNTICVWSPPSAGKNFFFDVYLHYLMNYGQLGIMNKTNNFSLQEATSKRVLLWNEPNYEDAYTDTLKMLTGGDALCVRVKQKKDCHVYKTPLIVLTNNMIGFMHELAFVD----RVKVYRWKQAPF
>Culex_pipiens_densovirus_FJ810126
ILEDFLKFQDEKITQFLVDLIDVLDKRVPKQNAFAVISPPSSGKNFFFDMIMAICLNYGQLGQANRHNLFAFQEAPNKRLLLWNEPNYESAMTDTIKMMMGGDPYTVRVKYQGDTHVTRTPVLILTNNYVSFLQESAFKD----RIKVYKWREAPF
>Helicoverpa_armigera_densovirus__JQ894784
IVTELLTFQESLIVEFLTNLVNVLDRKKPKLNTFVVYSPPTAGKNFFFDMIFGFCLSYGQLGQANKQNLFAFQEAPNKRVLLWNEPNYESSLTDTVKMMLGGDPYTVRVKNRGDQHVKRTPVIVLTNNIVPFMYEIAFNE----RIVQYKWNVAPF
>Galleria_mellonella_densovirus_L32896
IIDELLKYQEGLIVEFLTNLVNVLDRRIPKLNAFLVMSPPSAGKNFFFDMIFGLLLSYGQLGQANRHNLFAFQEAPNKRVLLWNEPNYESSLTDTIKMMFGGDPYTVRVKNRMDAHVKRTPVIILTNNTVPFMYETAFAD----RIIQYKWNAAPF
>Mythimna_loreyi_densovirus_AY461507
IIDELLKYQEGLIVEFLTNLVNVLDRRVPKLNAFLVMSPPSAGKNFFFDMIFGLLLSYGQLGQANRHNLFAFQEAPNKRVLLWNEPNYESSLTDTIKMMFGGDPYTVRVKNRMDAHVKRTPVIILTNNTVPFMYETAFAD----RIIQYKWNAAPF
>Diatraea_saccharalis_densovirus_AF036333
IVNELLKYQEDLIVEFLTNLVNVLDRRIPKLNAFLIISPPSGGKNFFFDMIFGLLLSYGQLGQANRHNLFAFQEAPNKRVLLWNEPNYESSLTDTIKMMFGGDPYTVRVKNRMDAHVKRTPVIILTNNTVPFMYELAFSD----RIIQYKWNAAPF
>Junonia_coenia_densovirus__KC883978
IIIELLKFQEDLIVEFLTNLVNVLDRRIPKLNAFLIISPPSAGKNFFFDMIFGLLLSYGQLGQANRHNLFAFQEAPNKRVLLWNEPNYESSLTDTIKMMFGGDPYTVRVKNRMDAHVKRTPVIILTNNTVPFMYETAFSD----RIIQYKWNAAPF
>Pseudoplusia_includens_densovirus_JX645046
IIVELLKYQEDLIVEFLTNLVNVLDRRIPKLNAFLILSPPSAGKNFFFDMIFGLLLSYGQLGQANRHNLFAFQEAPNKRVLLWNEPNYESSLTDTIKMMFGGDPYTVRVKNRMDAHVKRTPVIILTNNTVPFMYETAFSD----RIIQYKWDAAPF
>Echinococcus_granulosus_CDS25081
-------------QKFLDLAEEIMRRKVRSMNDM-----------------NKKFTYQETFRLMADYGQSY--NMIRKALETVRMMNVAHQRATDYADLSQKEEY--NVRNRSPSHLP-----LSKNQ--------ENHKWAIT------------
>Aleutian_mink_disease__virus_JN040434
LLATIKDM-GLNEQYLKKVLCTILTKQGGKRGCIWFYGPGGTGKTLLASLICKATVNYGMVTTS--NPNFPWTDCGNRNIIWAEECGNFGNWVEDFKAITGGGDVKVDTKNKQPQSI-KGCVIVTSNTNITKVTVGCVETNAHARMIKIRCMKT--
>Gray_fox_amdovirus_JN202450
-LINIINQMGLNEKVLKNIIATVLTKQSGKRGCIWFYGPGGTGKTLLANLICTAVKNFGMVTTS--NQNFPWTDCGNRNMIWLEECGNLGNFIEDFKAITGGGDIKVDTKNKQPQAI-KGVTVITSNKDITKVTIGAVETNVHSRIVKIRCVKT--
>Bufavirus_1_JX027296
-AWKLIKNNNMNPQKVYHAIMCCLNKQMGKRNTILLCGPASTGKSLLAQKIAKLVGNTGCYNPS--NANFPFNDCVNKNLIWIEEAGNLGTQVNSFKAIMSGQAIRLDQKGKGSKSIEPTPVVMTTNEDITRVIIGTELKVEHRRCLRFELKNKLS
>Porcine_parvovirus_PT4_PPU44978
---KIFSMHNWNYIKVCHAITCVLNRQGGKRNTILFHGPASTGKSIIAQHIANLVGNVGCYNAA--NVNFPFNDCTNKNLIWIEEAGNFSNQVNQFKAICSGQTIRIDQKGKGSKQIEPTPVIMTTNEDITKVRIGCEERPEHTRMLNINLTRKLP
>Canine_parvovirus_M19296
---QIFRMHGWNWIKVCHAIACVLNRQGGKRNTVLFHGPASTGKSIIAQAIAQAVGNVGCYNAA--NVNFPFNDCTNKNLIWIEEAGNFGQQVNQFKAICSGQTIRIDQKGKGSKQIEPTPVIMTTNENITIVRIGCEERPEHTRMLNIKLVCKLP
>Mouse_parvovirus_1_MPU12469
---KIFAFHGWNYIKVCHAICCVLNRQGGKRNTVLFHGPASTGKSIIAQAIAQAVGNVGCYNAA--NVNFPFNDCTNKNLIWVEEAGNFGQQVNQFKAICSGQTIRIDQKGKGSKQIEPTPVIMTTNENITVVKIGCEERPEHTRMLNIHLTHTLP
>Mouse_parvovirus_3_NC_008185
---KIFAFHGWNYIKVCHAICCVLNRQGGKRNTVLFHGPASTGKSIIAQAIAQAVGNVGCYNAA--NVNFPFNDCTNKNLIWVEEAGNFGQQVNQFKAICSGQTIRIDQKGKGSKQIEPTPVIMTTNENITVVKIGCEERPEHTRMLNIHLTHTLP
>Rat_parvovirus_NTU1_AF036710
---RIFAGHGWNYIKVCHAIACVLNRQGGKRNTVLFHGPASTGKSIIAQAIAQAVGNVGCYNAA--NVNFPFNDCTNKNLIWVEEAGNFGQQVNQFKAICSGQTIRIDQKGKGSKQIEPTPVIMTTNENITVVRIGCEERPEHTRMLNIHLTNTLP
>Bovine_parvovirus_2_AF406966
-VFDLLQFQGYDPVVAGYIIYAWSIRATGRRGALWFYGPGQTGKSIMARAMATCSVRYGCVNWT--NSNFPFQDLANCQIGWWEEGVITEDIVESAKALLSGGKIRVDRKCRDSVEITPPPFVITSNNDMTLVQGGNQVSFVHKRMIKFNFNKRLP
>Porcine_parvovirus_4_GQ387499
-MHYIFAINNYDPKIASVIMYFWSMKQTGKRNCVWFYGPATTGKTNMAQAICHSSANYGNVNWN--NANFPFQDIAGAQVGWWEEGKMTGDMVEAAKALLGGTALRIDRKCMQSIEVNSPPFLITSNVDMTIVQEGSFVSFEHQRMIKFSFNMTLP
>Eidolon_helvum_parvovirus_1_JQ037753
---DIFRLNGYEPSLVARYMACWAVGHWPKRRALWLWGPASTGKTVIAAAIAAQAPSYGCVNWT--NANFPFNDCHCQPLVWWEEGRMTENIVEVAKAILGGSPVRLDVKNKGSEDFLPTAVIITSNGDLTVTVDGPVISDAHQRMCMIRLERVLP
>Human_parvovirus_4_AY622943
-VAQLFSLNGYNPVDAAWYFAAWARGVWPKRRAIWLWGPASTGKTLLAAAIANLSPSYGCVNWT--NQNFPFNDCHCQSLVWWEEGRMTENIVEVAKAVLGGAPVRLDVKNKGSEDYIPTCVIITSNGDLTVTVDGPVVSTQHQRITMFQFQRMVP
>Bovine_hokovirus_EU200669
-VIELFKLNAYDPEDAAWYFAAWAQGIWPKRRALWLWGPASTGKTLLAAAIANISPSYGCVNWT--NQNFPFNDCHCQSLVWWEEGRMTENIVEVAKAILGGAPVRLDVKNKGSEDFIPTCVIITSNGDLTVTVDGPVVSTVHQRITMFQFNRLVP
>Porcine_hokovirus_EU200677
-VTELFRINGYDPEDAAWYFAAWAAGAWAKRRAMWLWGPASTGKTLLAAAIAAVAPSYGCVNWT--NQNFPFNDCHCQSLVWWEEGRMTENIVEVAKAILGGAPVRLDVKNKGSEDFIPTCVIITSNGDLTVTVDGPVVSTAHQRITMFQFQRLVP
>Human_parvovirus_B19

-IVKLLLCQNYDPLLVGQHVLKWIDKKCGKKNTLWFYGPPSTGKTNLAMAIAKSVPVYGMVNWN--NENFPFNDVAGKSLVVWDEGIIKSTIVEAAKAILGGQPTRVDQKMRGSVAVPGVPVVITSNGDITFVVSGNTTTTVHARMVKLNFTVRCS
>Goose_parvovirus__GPU25749
-VYQILKMNNYNPQYIGSILCGWVKREFNKRNAIWLYGPATTGKTNIAEAIAHAVPFYGCVNWT--NENFPFNDCVDKMLIWWEEGKMTNKVVESAKAILGGSAVRVDQKCKGSVCIEPTPVIITSNTDMCMIVDGNSTTMEHRRMFQIVLSHKLE
>Adeno-associated_virus_1_AF063497
-IYRILELNGYEPAYAGSVFLGWAQKRFGKRNTIWLFGPATTGKTNIAEAIAHAVPFYGCVNWT--NENFPFNDCVDKMVIWWEEGKMTAKVVESAKAILGGSKVRVDQKCKSSAQIDPTPVIVTSNTNMCAVIDGNSTTFEHQRMFKFELTRRLE
>Bat_adeno-associated_virus_GU226971
-IYRLFRMNGYDPAYAGSVLLGWCRTGFGKRNTVWLFGPATTGKTNLAEAISHSVPFYGCVNWT--NENFPFNDCVDKMIIWWEEGKMTSKVVESAKAILGGSKVRVDQKCKNSQQIEPTPVIITSNTNMCEVVDGNSTTFEHRRMFKFELTVRLQ
>California_sea_lion_adeno-associated_virus_1_JN420372
-IYQLFKMNGYDPAYLGSILLGWCQGRFGKRNTVWLYGPATTGKTNLAEAIAHSVPFYGCVNWT--NENFPFNDCVDKMLIWWEEGKMTSKVVESAKAILGGSKVRVDQKCKSSVQIDSTPVIITSNTDMCCVIDGNSTTFEHRRMFRINLEQRLS
>Adeno-associated_virus_5_AF085716
-IWQIFEMNGYDPAYAGSILYGWCQRSFNKRNTVWLYGPATTGKTNIAEAIAHTVPFYGCVNWT--NENFPFNDCVDKMLIWWEEGKMTNKVVESAKAILGGSKVRVDQKCKSSVQIDSTPVIVTSNTNMCVVVDGNSTTFEHQRMFKFELTKRLP
>Turkey_parvovirus__GU214706
-AIRLCSYQGYSPKYVARLILCWLSGQAGKKNALYFHGPANTGKTMMAESICKMVQIYGNVNHN--NKNFPFNDCHNKAVLWWEECSMTEEHVESAKCIMGGSSVRIDKKNQDSVLLCKTPIVITSNNDITQVSSRNAISTVHARCLKFTFNNWLT
>Porcine_bocavirus_3_JF429834
KVVRLLNIQGYNPIQVGHWVATVLSKKAGKQNTICFFGPASTGKTNLAKAIANAVKVYGCVNHL--NKSFVFNDCQNKLICWWEEAVMHNDWVEPAKCLMGGTSFRVDRKHKDSAEQPHTPLILSTNHDIYTVVGGNTVTTVHERVVQFNFMKTLP
>Porcine_bocavirus_5_HQ223038
-VIRLLNFQGYNYWQVGHWLCCVLDKKSGKQNTVSFYGPASTGKTNLAKAIVNAVNLFGNVNHL--NKNFVSNDCSNKLIVWWEEALMHTDWVEPAKCVLGGTTVRVDRKHKDSQLLPQTPCIISTNNNIYECVGGNHVSHVHCRVVQLNFMKTLP
>Human_bocavirus_1_JQ923422
-ALQLLLIQGYNPLAVGHALCCVLNKQFGKQNTVCFYGPASTGKTNMAKAIVQGIRLYGCVNHL--NKGFVFNDCRQRLVVWWEECLMHQDWVEPAKCILGGTECRIDVKHRDSVLLTQTPVIISTNHDIYAVVGGNSVSHVHARVIQLNFMKQLP
>Human_bocavirus_4_FJ973561
-ALKLLLIQGYNPLQVGHAICCVLNKQMGKQNTICFYGPASTGKTNFAKAIVQGVRLYGCVNHL--NKGFVFNDCRQRLIIWWEECLMHQDWVEPAKCILGGTECRIDVKHKDSVLLQQTPVIISTNHDIYSVVGGNTVSHVHARVLQLNFMKQLP
>Feline_bocavirus_JQ692585
-AWRLLLKQGYNPLQVGHWICCVLHKKAGKQNTLNFFGPASTGKTNLAKAIVNAIKLYGCVNHQ--NKNFIFNDCAAKLVWWEECLMHSDWVEQAKCILGGTEFRIDRKHRESHLLPQTPVIISTNNNIYQTLGGNSVSHVHERVVQFNFMTRLE
>Canine_minute_virus_FJ214110
-VWKLLTLQGYNPWQFGHWLLCVLDKKAGKQNTINFYGPASTGKTNLAKAIVNAVQLYGCVNHQ--NKNFVFNDCTAKLICWWEECIMTTDWVEQAKCIMGGTQFRIDRKHKDSHLLPQTPLIISTNHNIYEVTGGNTTTHVHSRVVQFNFMKQLE
>Porcine_bocavirus_1_HM053693
RVFRLLNFQGYNPWQAGHWVCCVLDKKSGKQNTLCFYGPASTGKTNLAKSIVQACKLYGCVNHQ--NKNFVFNDCAAKLIVWWEEALMHNDWVEQAKCILGGTEFRIDRKNRDSMLLPKTPVVISSNNDLYTVVGGNTVSQVHSRIVQWNFMKMLE
>Canine_bocavirus_1_JN648103
-AIQLLIFQGYNPWQVGHWLCCVLHKTAGKQNTVCFFGPASTGKTNFAKAIVNAVKLYGCVNHQ--NKNFVFNDCASKLVNWWEECLMHNDWVEQAKCLLGGTEFRIDRKHKDSQLLPQTPVVISTNHDVYTVVGGNTTTMVHARIVQFNFMKQLS
>California_sea_lion_bocavirus_1_JN420361
-AYQLFAIQGYNAWQAGHWLCCVLNKTAGKQNTVNFYGPASTGKTNMAKAIVQAVKLYGCVNHQ--NKNFLFNDCASKLVCWWEECVMNSEWVEQSKCILGGTEFRIDRKHKDSMLLPQTPVMISTNNDIYTVVGGNTVTGVHSRVVQFNFMKQLP
>Schmidtea_mediterranea_GAKN01010353
----------IYLQGNILYDNLFYSIGPSNTRKTFFVQLLTVDMKV-GCIHNMNNKNTFWTQDVDGKDVVVAEELMLNMENVDYFKLMLEGSNMRIEIKNKPAVNWKRIPVIVTSNQ-WIWQFVSNHQEALQNRMLVIKFNQPID
>Neobenedenia_melleni_GW918566
ILNEIFIKNKINKNFFISSLKNIIFKKYFKINTLLFYGNPNSGKSLIANLISQVYKS-KIVSNVASSSEFYLSNFDDANIIICEELFITPANVDDFKNILGGTSIVVNKKYERRGRINRIPILITSNYDLTHGFCSQLDTSALLRMIVFHFNYDIT
>Aedes_albopictus_densovirus_1_AY095351
WIEYLFKENNIDIIHFLAWNEIIKTKRYKKINGMVLEGITNAGKSLILDNLLAMVKP-EEIPRERDNSGFHLDQVPGAGSILFEEPMITPVNVGTWKLLLEGKTIKTDVKNKDKEPIERTPTWITTATPITNNIDMNETSQILQRIKLYIFKKSIQ
>Aedes_aegypti_densovirus_1_M37899
WIEYLFKENNIIIHFLAWNEIIKTKRYKKINGMVLEGITNAGKSLILDNLLAMVKP-EEIPRERDNSGFHLDQVPGAGSILFEEPMITPVNVGTWKLLLEGKTIKTDVKNKDKEPIERTPTWITTATPITNNIDMNETSQILQRIKLYILKKSIQ
>Anopheles_gambiae__densovirus_EU233812
WIEYLFKENNINIIHFLAWNEIIKTKRYKKINGMVLEGITNAGKSLILDNLLAMVKP-EEIPRERDNSGFHLDQVPGAGSILFEEPMITPVNVGTWKLLLEGKTIKTDVKNKDKEPIERTPTWITTATPITNNIDMNETSQILQRIKLYIFKKSIQ
>Aedes_aegypti_densovirus_2_FJ360744
WIEYLFKENNIIIHFLAWNEIIKTKRYKKINGMVLEGITNAGKSLILDNLLAMVKP-EEIPRERDNSGFHLDQVPGAGSILFEEPMITPVNVGTWKLLLEGKTIKTDVKNKDKEPIERTPTWITTATPITNNIDMNETSQILQRIKLYIFKKSIQ
>Mosquito_densovirus_BR/07_GU452720
WIEYLFKENNINIIHFLAWNEIIKTKRYKKINGMVLEGITNAGKSLILDNLLAMVKP-EEIPRERDNSGFHLDQVPGAGSILFEEPMITPVNVGTWKLLLEGKTIKTDVKNKDKEPIERTPTWITTATPITNNIDMNETSQILQRIKLYIFKKSIQ
>Culex_pipiens_pallens_densovirus_EF579771
WIEYLFKENNINIIHFLAWNEIIKTKRYKKINGMVLEGITNAGKSLILDNLLAMVKP-EEIPRERDNSGFHLDQVPGAGSILFEEPMITPVNVGTWKLLLEGKTIKTDVKNKDKEPIERTPTWITTATPITNNIDMNETSQILQRIKLYIFKKSIQ
>Aedes_albopictus_densovirus_2_X74945
WIEYMFKENNINIIHFLAWNEIIKPKRYKKINGMVLEGITNAGKSLILDNLLAMVKP-EEIPRERDNSGFHLDQLPGAGSVLFEEPMITPVNVGTWKLLLEGKTIKTDVKNKDKEPIERTPTWITTATPITNNVHMNETSQILQRIKLYIFKKSIQ
>Haemagogus_equinus_densovirus_AY605055
WIEYMFKENNINMIHFLAWNEIIKTKRYKKINGMVLEGITNAGKSLILDNLLAMVKP-EEIPRERDNSGFHLDQLPGAGSVLFEEPMITPVNVGTWKLLLEGKTIKTDVKNKDKEPIERTPTWITTATPITNNVDMNETSQILQRIKLYIFKKSIQ
>Aedes_albopictus_densovirus_3_AY310877
WIEYMFKENNINIIHFLAWNEIIKTKRYKKINGMVLEGITNAGKSLILDNLLAMVKP-EEIPRERDNSGFHLDQLPGAGSVLFEEPMITPVNVGTWKLLLEGKTIKTDVKNKDKEPIERTPTWITTATPITNNVDMNETSQILQRIKLYIFKKSIQ
>Taenia_multiceps_JR938527
---MTPFINGFSVLVMPYRHISLS-LIQHFQFDNLLNDAGALMEEPRITMTTKDDFKLLLGGIGLEVDVKYSGHKWLERVLVIQATNEDIIFRLSGQNRKCFTN----------
>Taenia_multiceps_JR929085
---------ITTFLLLYTKVVDVVDEKMDKVNMLCFYRTTNAGKSLLAGLITAPLVV-AAITRCGYQTAFQFDNLLHKTGALME----------------S---
>Penaeus_monodon_penstyldensovirus_2_AY124937
WIKYMLANNDIRVPEILAWILIVADKKLDKINTLVLQGPTGTGKSLTIGALLGKLNT-GLVTRTGDSNTFHLQNLIGKSYALFEEPRISQITVDDFKLLFEGSDLEVNIKHQESEIMGRIPIFISTNKDIDYWVPPADGKALQTRTKTFHLTSQIK
>Penaeus_monodon_penstyldensovirus_1_GQ411199
WIKYMLANNDIRVPEILAWIIIVADKKLDKINTLVLQGPTGTGKSLTIGALLGKLNT-GLVTRTGDSNTFHLQNLIGKSYALFEEPRISQITVDDFKLLFEGSDLEVNIKHQESEIMGRIPIFISTNKDIDYWVPPADGKALQTRTKTFHLTRQIK
>Penaeus_stylirostris_penstyldensovirus_2_GQ475529
WIKYMLANNDIRVPEILAWIIIVADKKLDKINTLVLQGPTGTGKSLTIGALLGKLNT-GLVTRTGDSNTFHLQNLIGKSYALFEEPRISQITVDDFKLLFEGSDLEVNIKHQESEIMGRIPIFISTNKDIDYWVPPADGKALQTRTKTFHLTRQIK
>Hymenolepis_diminuta_LM393614
-LYSIFTNNDIDFGLFLAEVDKIIRMTYPRINALVLRGPTSTGKTLIAKNIVKPYNY-ETVSRDGDATAFYLQNLLDHDVALMEEPHISMTTVQNFKELFAGSPLIVQVKNHAPRELKRIPCIVTTNQSLTDSLIDAESEPIRRRIIEYLLYRPIS
>Hymenolepis_diminuta_LM395168
-LYSNFTNNGIDFGLFIAEVDKIIRMSYPRINALVLRGPTSTGKTLIAKNIVKPYNY-GTVSRDGDATAFYLQNLLDHDVALMEEPHISMTTVQNFKELFAGSPLIVQVKNHAPRELKRIPCIITTNQSLTDSLIDAESEPIKKRIIEYLLYRPIS
>Schistosoma_mansoni_XP_002571349
WFDMLMEKNDIDKKVFCASVSTIMNKKVKRVNTLCLEGPTTTGKSLLLKLICGEYNY-GTVQRSGDHSQFFLQNLLKKTVALMEEPRITPITVQDFKLLLGGEPLDIHVKHQDDCRLPRIPVLISTNHELGFYTSSLDKDALYNRTHYYKFKCQLG
>Opisthorchis_viverrini_XP_009173661
WFDMLMEKNDIDKKVFCASVSTIMNKKVKRVNTLCLEGPTTTGKSLLLKLICGEYNY-GTVQRSGDHSQFFLQNLLKKTVALMEEPRITPITVQDFKLLLGGEPLDIHVKHQDDCRLPRIPVLISTNHELGFYTSSLDKDALYNRTHYYKFKCQLG
>Schistosoma_mansoni_CCD58408
WFDMLMEKNDIDKKVFCASVSTIMNKKVKRVNTLCLEGPTTTGKSLLLKLICGEYNY-GTVQRSGDHSQFFLQNLLKKLLPSWK----------NQE---
>Dicrocoelium_dendriticum_LK428994
FLDNLLAANKIDKDKFCKAIHEIMNKKIDRLNCLCIEGPTTTGKSLVLKMICQNYHC-GTVQRSGDHSQFFLQNLVDKSIALMEEPRMTIATVNDFKELLGGSPFDIHVKHSPDVTLNRLPVLISTNHSLGTYITSIDAAAIYERCHTFHFRVRVG
>Dicrocoelium_dendriticum_LK459200
FLDNLLAANKIDKDKFCKAIHEVMNKKIDRLNTLCIEGPTTTGKSLVLKMICQNYHC-GTVQRSGDHSQFFLQNLVDKSVALMEEPRITLATVNDFKELLGGSPFDIHVKHSPDVTLSRLPVLISTNHSLGAYITSIDAAAIY----------
>Schistosoma_mattheei_LM181082
-------MN----------------RQKFGINTLCIEGPTTTGKTLILKLITQNYTF-GTVRMSGNHSKFFLQNLNNKSVALMEQPRITPVNVNDFKQLLGGSSLDIHIKHQSDVTLQRIPVLISTNSNLGLYINSEDKKAIYKRCITYHFTQSIG
>Schistosoma_margrebowiei_LL888002
WFETFFAANEINPKQFLNDIQTIIDKTKLRVNTFCLEGPTTTGKTLILKLIVQNYTY-GTVQRSGDHSQFFLQNLNNKSVALMEEPRITPITVNDFKELLGGSPFDIHITHQPDVTLPRIPVLISTNSNLGTYINSIDKDAIYKRCITYHLTKQIG
>Clonorchis_sinensis_GAA57952
WLDLLLHVNRINKQNFLTDLTRVMNKQIDRKNAFVLEGPTTTGENIIRKTHC------GQLHIW-----------DRTEIWRSQSVLPNEPTE--
>Opisthorchis_viverrini_XP_009170655
WLDQLMSANKINKQQFLTDLTLVMNKVCTRKNASVIEGPTTTGKTLFVKLIAENYIY-GTVQRPEEQETKPATTTATTTVAPKKEEPETVSIMETKRQRLDGPNLAVAKHAQSLFGIHYKTDVTVTRKHLTLAQYKPADTGITQTCIPYQIFEFWT
>Opisthorchis_felineus_GBJA01006694
WLDLLFHANKINKRDFLVNLTQIMNKDIDGRNAFVIEGPTTTGKTLLVKLIAENYIY-GTVKRSGDRSQCFLMKLLDKALALVEEPRITQLTVNHFKALLGGNAFDIRVKHQKDERLTRLPVLITTNTDLTYQCLGEDAKAIKERCFIYRFFVHVG
>Opisthorchis_felineus_GBJA01010281
WLDQLMQVNNIDKKDILHSLTLIMNKSLKRKNAFVIEGPTTTGKTLFVKLIAENYVY-GTVQRSGDHSQFFLMNLLNKTLALMEEPRITQLSVNDFKELLGGNPFDIHVKHQKDELHY-------GTH----------RRSFWCW-----
>Clonorchis_sinensis_GAA40146
WLDLLLHVNRINKQEFLTNITKIMNREIDRKNAFVIEGPTTTGKTLFVKLIAESYIY-GTVQRSGDHSQFFLMNLLNKALALMEEPRITQLTVNDFKELLGGNAFDIHVKHQKDERLTRFPVLITTNNDLTYYVLGEDDKAIKERCFYYKFFVSMR
>Opisthorchis_felineus_GBJA01005901
WLDLLFYVNKINKQEFLIDITRIMNKEVDRTNAFVIEGPTTTGKTLFVKLIAENYIY-GTVQKSGDHSQFFLMNLLNKALALMEKPRITQLTVNYFKELLGGNPFDIHVKHQKDERLNRLPVLITTNNDLTYYVLGEDGKAIKERCFTYKFFVKVG
>Opisthorchis_felineus_GBJA01002122
WLDLLFQVNKINKQTFLCNITRIMNKEIDRKNAFVIEGPTTTGKTLFVKLIAENYIY-GTVQRSGDHSQFFLMNLLNKALALMEEPRITQLTVNDFKELLGGNPFDIHVKHQKDERLNRLPVLITTNNDLTYYVLGEDGKAIKERCFAYKFYVKVG
>Clonorchis_sinensis_GAA32750
WLDLLLHVNRINKQDFLCNLTKIMNKQIDRKNAFVIEGPTTTGKTLFVKLIAESYIY-GTVQRSGDHSQFFLMNLLNKALALMEEPRITQLTVNDFKELLGGNAFDIHVKHQKDERLTRLPVLITTNNDLTYYVLGEYDIYMTWRLRQLLIRNATK
>Clonorchis_sinensis_GAA56553
WLDLLLFVNRINKQQFLADLTRVMNKEVDRKNAFVLEGPTTTGKTLFVKLIADSYIY-GTVQRSGDHSQFFLMNLLNKALALMEEPRITQLTVNDFK--
>Clonorchis_sinensis_GAA57641
WLDLLLFVNRINKQQFLTDLTRVMNKQIDRKNAFVLEGPTTTGKTLFVKLIADNYIY-GTVQRSGDHSQFFLMNLLNKALALMEEPRITQLTVNDFKELLGGNAFDIHVKHQKDERLTRLPVLITTNNDLTYYVLGEDGKAIKERCFYYKFFVKVG
>Clonorchis_sinensis_GAA57639
WLDRLLFVNRINKQQLLADLTRVMNKQVDRKNAFVLEGPTTTGKTLFVKLIADNYIY-GTVQRSGDHSQFFLMNLLNKALALMEEPRITQLTVNDFKELLGGNAFDIHVKHQKDERLTRLPVLITTNNDLTYYVLGEDGKAIRERCFYYKFFVKVG
>Clonorchis_sinensis_GAA57638
WLDRLLFVNRINKQEFLADLTRVMNKQVDRKNAFVLGGPTTTGKTLFVKLIADNYIY-GTVQRSGDHSQFFLMNLLNKALALMEEPRITQLTVNDFKELLGGNAFDIHVKHQKDERLTRLPVLITTNNDLTYYVLGEDGKAIKERCFYYKFFVKVG

>Clonorchis_sinensis_GAA57954
WLDRLLFVNRINKQKFLADLTRVMNKQVDRKNAFVLEGPTTTGKTLFVKLIADNYIY-GTVQRSGDHSQFFLMNLLNKALALMEEPRITQLTVNDLKELLGGNAFDIHVKHQKDERLTRLPVRITTNNDLTYYVLGEDGKVIKERCFYYKFFVKVG
>Clonorchis_sinensis_GAA57640
WLDRLLFVNRIDKQKFLCDLTRVMNKQVDRKNAFVLEGPTTTGKTLFVKLIADNYIY-GTVQRSGDHSQFFLMNLLNKALALMEEPRITQLTVNDFKELLGGNAFDIHVKHQKDERLTRLPVLITTNNDLTY------------------------
>Opisthorchis_viverrini_XP_009177542
WLDQLMWVNRIDKKQFLIDLTQIMNKAYKRKNAFVIEGPTTTGKTLFVKLIAENYIY-GTVQRSGDHSQFFLMNLLNKSLGLMEEPRITQLTVNDFKELLGGNPFDIHVKHQKDERLQRLPVLITTNNNLVYYVLDPDGKAILERCFYYKFVVKVG
>Opisthorchis_viverrini_XP_009170634
WLDQLMSANKINKQQFLTDLTLVMNKVCTRKNAFVIEGPTTTGKTLFVKLIAENYIY-GTVQRSGDHSQFFLMNLLNKALALMEEPRITQLTVNDFKELLGGNPFDIHVKHQKDERLQRLPVLITTNNNLVYYVLDPDGKAILERCFYYKFLVKVG
>Opisthorchis_viverrini_XP_009176824
WLDQLMAANKINKQQFLTDLTLVMNKVCTRKNAFVIEGPTTTGKTLFVKLIAENYIY-GTVQRSGDHSQFFLMNLLNKALALMEEPRITQLTVNDFKELLGGNPFDIHVKHQKDERLQRLPVLITTNNNLVYYVLDPDGKAILERCFYYKFLIKVG
>Opisthorchis_felineus_GBJA01008850
WLDRLMLVNNINKRELLTSLTLVMNKSMNRKNAFVIEGPTTTGKTLFVKLVAENYVY-GTVQRSGDHSQFFLMNLLNKTLALMEEPRITQLTVNDFKELLGGNPFDIHVKHQKDERLERLPVLITTNNRLTYYVLDSDAKAILERCFYYTFTVKVG
>Opisthorchis_viverrini_XP_009177530
WLDLLINVVNHINKRELLVSLTAVMNKLCTRKNAFVIEGPTTTGKTLFVKLVAENYVY-GTVQRSGDHSQFFLMNLLNKTLALMEEPRITQLTVNNFKELLGGNPFDIHVKHQKDERLERLPVLITTNNPLTYYVLDADGKAILERCFYFKFHVKVG
>Opisthorchis_viverrini_XP_009178033
WLDELIRVVNNINKRELLVSLTAIMNKLCTRKNAFVIEGPTTTGKTLFVKLIAENYVY-GTVQRS-----------------------------------GDH---
>Opisthorchis_viverrini_XP_009177702
WLDQLIQVVNNINKRELLVCLTSVMNKLCTRKNAFVIEGPTTTGKTLFVKLVAENYVY-GTVQRSGDHSQFFLMNLLNKTLALMEEPRITQLTVNDFKELLGGNPFDIHVKHQKDERLERLPVLITTNNPLTYYVMDADGKAIL------------
>Opisthorchis_viverrini_XP_009177971
--MKLLNKTLALMEEPRITQLTVNDFKELLGGNPFDIHVKHQKDERLERLPVLITTNNPLTYYVMDADGKAILE------------
>Opisthorchis_viverrini_XP_009177942
WLDELIRVVNNINKRDLLVALTAVMNKLCTRKNAFVIEGPTTTGKTLFVKLVAENYVY-GTVQRSGDHSQFFLMNLLNKTLALMEEPRITQLTVNDFKELLGGNPFDIHVKHQKDE---
>Opisthorchis_viverrini_XP_009177808
WLDQLIRVVNNINKRELLACLTSVMNKLCTRKNAFVIEGPTTTGKTLFVKLVAENYE----------------------------EPRITQLTVNDFKELLGGNPFDIHVKHQKDERLDRLPVLITTNNPLTYYVMDADGKAILER-----------
>Opisthorchis_viverrini_XP_009177798
---------------------------RKNAFVIEGPTTTGKTLFVKLVAENYVY-GTVQRSGDHSQFFLMNLLNKTLALMEEPRITQLTVNDFKELLGGNAFDIHVKHQKDERLDRLPVLITTNNPLTYYVMDADGKAILERCFYFKFHVKVG
>Opisthorchis_felineus_GBJA01002124
WLDQLMAVNNINKKEMLHSLTLVMNKTMKRKNAFVIEVPTTTGKTLFVKLIAENYVY-GTVQRSADHSPFFLMNLLNKTLALMEEPRITQLTVNDFKELLGGNPFDIHVKHQKDERLDRLPVLITTNNRLTYYVLDSDAKAILERCFYYHFNVKVG
>Opisthorchis_felineus_GBJA01002120
WLDQLMLVNNINKKDMLHALTLVMNKTMKRKNAFVIEGPTTTGKTLFVKLIAENYVY-GTVQRSGDHSQFYLMNLLNKTLALMEEPRITQLTVNDFKELLGGNPFDIHVKHQKDERLERLPVLITTNNRLTYYVLDSDAKAILERCFYYHFTAQLR
>Opisthorchis_felineus_GBJA01002117
WLDQLMLVNNINKKEMLHSLTLVMNKTMKRKNAFVIEGPTTTGKTLFVKLIAENYVY-GTVQRSGDHSQFFLMNLLNKTLGFS---
>Moniezia_expansa_JL291017
---------------------------------VLQGPSNTGKSLLAKLIVSGYNY-ATVARSTESNNFIFQNLLGKTAALMEEPFITKATVNDFKQLLGGERMEIGIKHRDREWLERVPIICTTNDQLADRCNSVDCQAIQNRCVYYRLFKTIK
>Echinostoma_caproni_LL285499
WLDRLLAVNNIDKNKFLTQLTDIMNKKENRVNAFVIQGPTTTGKFQYSSHLSSHRQV-LDVKSP--HGELH----LNSPEEWRPLPVLPHELTKQGSCAYG----RTQDYAINGQRLQRTPGRKTIRYPCETPTRCPAQTPTRPHFYQYRHMSVFT
>Echinostoma_caproni_LL284665
WLDRLLAVNNIDKTRFLTQISNIMNKREQRINALVIQGPTTTGNL---
>Echinostoma_caproni_LL286487
WLDRLLAVNNIDKNRFLTQLTNIMNKKEQRINAFVIQGPTTTGK-FLCSSHLSSHRQILDAKSPHGELHLW-----NSSTKWGPFAILSHEFAEQGC--------RPHGRASNHPAHG-----------------------------------
>Echinostoma_caproni_LL274983
WLDQLLAVNHIDKNRFLTQVTNIMNKKEQRINAFVIQGPTTTGKL---
>Echinococcus_granulosus_W6UKP4
-LEDMFLANEIALVDFAITLRIIMNCEDEKINTLVLYGPTNTGKSLICKLTTTFLEH-GSVMRRQGASAFAYENLLNRKVALMEEPGICAANQQDLKQILGGETFKGPKDMQTTQQAPQ-PVQSTAQHPPAPATPTPH----------------
>Echinococcus_multilocularis_U6HF69
WLEDRFSANEIAVVDFAITLRIIMNCEDEKINALVLYGPTSTGKSLICKLTTTFLEH-GSVMRRQEASAFAHENLLNRKVALMEEPKICAANQQDLKQILGGEPFEVHIKYQNPDLLERLPVIVTTNEPLGVRLSDVDAAATEGRCKIYTLDKQIC
>Echinococcus_granulosus_EUB54832
-LEDMFSANEIAVVDFAITLRIIMNCEDEKINTLVLYGPTNTGKSLICKLTTTFLEH-GCVTRRQEASAFAYENLLNREVALMEEPRICAANQQDLKQILGGEPFEVSVKYQNPDLLERLPVIVTTNEPLGVRLSDVDAAAIEGRCKIYTLDKQIC
>Echinococcus_multilocularis_CDI70496
WLEDMFSANEIAVVDFAITLRIIMNCEDEKINTLVLYGPTNTGKSLICRLTTTFLEH-GSVMRRQEASAFA----------------ICAANQQDLKQILGGEPFEVSVKYQNPDLLERLPVIVTTNEPLDVRLSDVDAAAIEGRCKIYTLDKQIC
>Echinococcus_multilocularis_CDS35594
WLEDMFSANEIAVVDFAITLRIIMNCEDEKINTLVLYGPTNTGKSLICRLTTTFLEH-GSVMRRQEASAFA----------------ICAANQQDLKQILGGEPFEVSVKYQNPDLLERLPVIVTTNEPLDVRLSDVDAAAIEGRCKIYTLDKQIC
>Echinococcus_multilocularis_U6HA22
WLEDMFSANEIAVVDFAITLRIIMNCEDEKINTLVLYGPTNTGKSLICRLTTTFLEH-GSVMRRQEASAFAYENLLNRKVALMEEPRICAANQQDLKQILGGEPFEVSVKYQNPDLLERLPVIVTTNEPLDVRLSDVDAAAIEGRCKIYTLDKQIC
>Echinococcus_multilocularis_A0A087VWP3
WLEDMFSANEIAVVDFAITLRIIMNCEDEKINTLVLYGPTNTGKSLICRLTTTFLEH-GSVMRRQEASAFAYENLLNRKVALMEEPRICAANQQDLKQILGGEPFEVSVKYQNPDLLERLPVIVTTNEPLDVRLSDVDAAAIEGRCKIYTLDKQIC
>Echinococcus_multilocularis_U6HLE6
WLEDMFSANEIAVVDFAITLRIIMNCEDEKINTLVLYGPTNTGKSLICRLTTTFLEH-GSVMRRQEASAFAYENLLNRKVALMEEPRICAANQQDLKQILGGEPFEVHIKYQNPDLLERLPVIVTTNEPLGVRLSDVDAAAIEGRCKIYTLDKQIC
>Echinococcus_multilocularis_A0A068XWU9
WLEDMFSANEIAVVDFAITLRIIMNCEDEKINTLVLYGPTNTGKSLICRLTTTFLEH-GSVMRRQEASAFAYENLLNRKVALMEEPRICAANQQDLKQILGGEPFEVHIKYQNPDLLERLPVIVTTNEPLGVRLSDVDAAAIEGRCKIYTLDKQIC
>Echinococcus_multilocularis_CDS36723
WLEDMFSANEIAVVDFAITLRIIMNCEDEKINTLVLYGPTNTGKSLICRLTTTFLEH-GSVMRRQEASAFAYENLLNRKVALMEEPRICAANQQDLKQILGGEPFEVHIKYQNPDLLERLPVIVTTNEPLGVRLSDVDAAAIEGRCKIYTLDKQIC
>Echinococcus_multilocularis_U6HHY9
WLEDMFSANEIAVVDFAITLRIIMNCEDEKINTLVLYGPTNTGKSLICRLTTSFLEH-GSVMRRQEASAFAYENLLNRKVALMEEPKICAANQQDLKQILGGEPFEVHIKYQNPDLLERLPVIVTTNEPLGVRLSDVDAAAIEGRCKIYTLDKQIC
>Echinococcus_multilocularis_CDS36816
WLEDMFSANEIAVVDFAITLRIIMNCEDEKINTLVLYGPTNTGKSLICRLTTSFLEH-GSVMRRQEASAFAYENLLNRKVALMEEPKICAANQQDLKQILGGEPFEVHIKYQNPDLLERLPVIVTTNEPLGVRLSDVDAAAIEGRCKIYTLDKQIC
>Echinococcus_multilocularis_U6HDP5
WLEDMFSANEIAVVDFAITLRIIMNCEDEKINTLVLYGPTNTGKSLICRLTTSFLEH-GSVMRRQEASAFAYENLLNRKVALMEEPKICAANQQDLKQILGGEPFEVHIKYQNPDLLERLPVIVTTNEPLGVRLSDVDAAAIEGRCKIYTLDKQIC
>Echinococcus_multilocularis_U6HC26
WLEDMFSANEIAVVDFAITLRIIMNCEDEKINTLVLYGPTNTGKSLICRLTTSFLEH-GSVMRRQEASAFAYENLLNRKVALMEEPKICAANQQDLKQILGGEPFEVHIKYQNPDLLERLPVIVTTNEPLGVRLSDVDAAAIEGRCKIYTLDKQIC
>Echinococcus_multilocularis_A0A077RCT5
WLEDMFSANEIAVVDFAITLRIIMNCEDEKINTLVLYGPTNTGKSLICRLTTSFLEH-GSVMRRQEASAFAYENLLNRKVALMEEPKICAANQQDLKQILGGEPFEVHIKYQNPDLLERLPVIVTTNEPLGVRLSDVDAAAIEGRCKIYTLDKQIC
>Echinococcus_multilocularis_U6H9L2
WLEDMFSANAIAVVDFAITLRIIMNCEDEKINTLVLYGPTNTGKSLICKLMTSFLEP-GSVMRR--------------------PGR--------------RQPS---
>Echinococcus_multilocularis_U6HNK1
WLEDMFSANAIAVVDFAITLRIIMNCEDEKINTLVLYGPTNTGKSLICKLMTSFLEH-GSVMRRQEASAFAYEDLLNRKVALMEEPKICAANQQDLKQILGGEPFEVHIKYQNPDLLERLPVVVTTNEPLGVRLSDVDAAATEGRCKSYTLDKQIC
>Echinococcus_multilocularis_CDS36692
WLEDMFSANAIAVVDFAITLRIIMNCEDEKINTLVLYGPTNTGKSLICKLMTSFLEH-GSVMRRQEASAFAYEDLLNRKVALMEEPKICAANQQDLKQILGGEPFEVHIKYQNPDLLERLPVVVTTNEPLGVRLSDVDAAATEGRCKSYTLDKQIC
>Echinococcus_multilocularis_CDI70288
WLEDMFSANAIAVVDFAITLRIIMNCEDEKINTLVLYGPTNTGKSLICKLMTSFLEH-GSVMRRQEASAFAYENLLNRKVALMEEPKICAANQQDLKQILGGEPFEVHIKYQNPDLLERLPVVVTTNEPLGVRLSDVDAAATEGRCKSYTLDKQIC
>Echinococcus_multilocularis_CDS35386
WLEDMFSANAIAVVDFAITLRIIMNCEDEKINTLVLYGPTNTGKSLICKLMTSFLEH-GSVMRRQEASAFAYENLLNRKVALMEEPKICAANQQDLKQILGGEPFEVHIKYQNPDLLERLPVVVTTNEPLGVRLSDVDAAAIEGRCKSYTLDKQIC
>Echinococcus_multilocularis_CDI70556
WLEDMFSANAIAVVDFAITLRIIMNCEDEKINTLVLYGPTNTGKSLICKLMTSFLEH-GSVMRRQEASAFAYENLLNRKVALMEEPKICAANQQDLKQILGGEPFEVHTKYQNPDLLERLPVVVTTNEPLGVRLSDVDAAATEGRCKIYTLDKQIC
>Echinococcus_multilocularis_CDS36689
WLEDMFSANAIAVVDFAITLRIIMNCEDEKINTLVLYGPTNTGKSLICKLMTSFLEH-GSVMRRQEASAFAYENLLNRKVALMEEPKICAANQQDLKQILGGEPFEVHTKYQNPDLLERLPVVVTTNEPLGVRLSDVDAAATEGRCKIYTLDKQIC
>Echinococcus_multilocularis_U6FUW9
WLEDMFSANAIAVVDFAITLRIIMNCEDEKINTLVLYGPTNTGKSLICKLMTSFLEH-GSVMRRQEASAFAYENLLNRKVALMEEPKICAANQQDLKQILGGEPFEVHTKYQNPDLLERLPVVVTTNEPLGVRLSDVDAAATEGRCKIYTLDKQIC
>Echinococcus_multilocularis_CDS35660
WLEDMFSANAIAVVDFAITLRIIMNCEDEKINTLVLYGPTNTGKSLICKLMTSFLEH-GSVMRRQEASAFAYENLLNRKVALMEEPKICAANQQDLKQILGGEPFEVHTKYQNPDLLERLPVVVTTNEPLGVRLSDVDAAATEGRCKIYTLDKQIC
>Echinococcus_multilocularis_U6HA06
WLEDMFSANAIAVVDFAITLRIIMNCEDEKINTLVLYGPTNTGKSLICKLMTSFLEH-GSVMRRQEASAFAYENLLNRKVALMEEPKICAANQQDLKQILGGEPFEVHTKYQNPDLLERLPVVVTTNEPLGVRLSDVDAAATEGRCKIYTLDKQIC
>Echinococcus_multilocularis_U6H9Z7
-----FSANAIAVVDFAITLRIIMNCEDEKINTLVLYGPTNTGKSLICKLMTSFLEH-GSVMRRQEASAFAYENLLNRKVALMEEPKICAANQQDLKQILGGEPFEVHIKYQNPDLLERLPVVVTTNEPLGVRLSDVDAAATEGRCKSYTLDKQIC
>Echinococcus_multilocularis_U6H948
--------------DFAITLRIIMNCEDEKINTLVLYGPTNTGKSLICKLMTSFLEH-GSVMRRQEASAFAYENLLNRKVALMEEPKICAANQQDLKQILGGEPFEVHIKYQNPDLLERLPVVVTTNEPLGVRLSDVDAAATEGRCKIYTLDKQIC
>Echinococcus_multilocularis_U6HBN8
WLEDMFSANAIAVVDFAITLRIIMNCEDEKINTLVLYGPTNTGKSLICKLMTSFLEH-GSVMRRQEASAFAYENLLNRKVALMEEPKICAANQQDLKQILGGEPFEVHIKYQNPDLLERLPVVVTTNEPLGVRLSDVDAAATEGRCKIYTLDKQIC
>Echinococcus_multilocularis_U6H9K2
WLEDMFSANAIAVVDFAITLRIIMNCEDEKINTLVLYGPTNTGKSLICKLMTSFLEH-GSVMRRQEASAFAYENLLNRKVALMEEPKICAANQQDLKQILGGEPFEVHIKYQNPDLLERLPVVVTTNEPLGVRLSDVDAAATEGRCKIYTLDKQIC
>Echinococcus_multilocularis_U6HGG3
WLEDMFSANAIAVVDFAITLRIIMNCEDEKINTLVLYGPTNTGKSLICKLMTSFLEH-GSVMRRQEASAFAYENLLNRKVALMEEPKICAANQQDLKQILGGEPFEVHIKYQNPDLLERLPVVVTTNEPLGVRLSDVDAAATEGRCKIYTLDKQIC
>Echinococcus_multilocularis_U6HBG0
WLEDMFSANAIAVVDFAITLRIIMNCEDEKINTLVLYGPTNTGKSLICKLMTSFLEH-GSVMRRQEASAFAYENLLNRKVALMEEPKICAANQQDLKQILGGEPFEVHIKYQNPDLLERLPVVVTTNEPLGVRLSDVDAAATEGRCKIYTLDKQIC
>Echinococcus_multilocularis_U6HD97

```
-----FSANAIAVVDFAITLRIIMNCEDEKINTLVLYGPTNTGKSLICKLMTSFLEH-GSVMRRQEASAFAYENLLNRKVALMEEPKICAANQQDLKQILGGEPFEVHIKYQNPDLLERLPVVVTTNEPLGVRLSDVDAAATEGRCKIYTLDKQIC
>Echinococcus_multilocularis_U6H9K6
-----FSANAIAVVDFAITLRIIMNCEDEKINTLVLYGPTNTGKSLICKLMTSFLEH-GSVMRRQEASAFAYENLLNRKVALMEEPKICAANQQDLKQILGGEPFEVHIKYQNPDLLERLPVVVTTNEPLGVRLSDVDAAATEGRCKIYTLDKQIC
>Echinococcus_multilocularis_CDI70555
WLEDMFSANAIAVVDFAITLRIIMNCEDEKINTLVLYGPTNTGKSLICKLMTSFLEH-GSVMRRQEASAFAYENLLNRKVALMEEPKICAANQQDLKQILGGEPF-------------------------------------E----------
>Echinococcus_multilocularis_CDS35653
WLEDMFSANAIAVVDFAITLRIIMNCEDEKINTLVLYGPTNTGKSLICKLMTSFLEH-GSVMRRQEASAFAYENLLNRKVALMEEPKICAANQQDLKQILGGEPF-------------------------------------E----------
>Echinococcus_multilocularis_CDI70553
WLEDMFSANAIAVVDFAITLRIIMNCEDEKINTLVLYGPTNTGKSLICKLMTSFLEH-GSVMRRQEASAFAYENLLNGKVALMEEPKICAANQQDLKQILGGEPFEVPVKYQNPDLLERLPVVVTTNEPLGVRLSDVDAAATEGRCKIYTLDKQIC
>Echinococcus_multilocularis_CDS35651
WLEDMFSANAIAVVDFAITLRIIMNCEDEKINTLVLYGPTNTGKSLICKLMTSFLEH-GSVMRRQEASAFAYENLLNGKVALMEEPKICAANQQDLKQILGGEPFEVPVKYQNPDLLERLPVVVTTNEPLGVRLSDVDAAATEGRCKIYTLDKQIC
>Echinococcus_multilocularis_U6HA02
-------MN---------------CEDEKINTLVLYGPTNTGKSLICKLMTSFLEH-GSVMRRQEASAFAYENLLNGKVALMEEPKICAANQQDLKQILGGEPFEVPVKYQNPDLLERLPVVVTTNEPLGVRLSDVDAAATEGRCKIYTLDKQIC
>Hydatigera_taeniaeformis_LL723621
------------------------MHM-------------------------------------------------DRTVAVLEEPRMNAPNVNDMKQLFGGEPFEINVKYKPMQYLNRLPVIVTTNECIGSIKTHIDCAALESRYCQFNFSLADS
>Hymenolepis_microstoma_CDJ13946
-------------MQYLQRLPVIITNEY--LGCRLPDVDAAAGKSLLCKVMTEFLLT-GTISRQSENTNFAFENLLDRSVAILEEPKINASNANDMKQLLGGESFEVAVKYKPMQFLPRLPVIITTNEYLGCRLPDVDAAALESRYYQFTFSTQIA
>Hymenolepis_microstoma_CDS32661
-------------MQYLQRLPVIITNEY--LGCRLPDVDAAAGKSLLCKVMTEFLLT-GTISRQSENTNFAFENLLDRSVAILEEPKINASNANDMKQLLGGESFEVAVKYKPMQFLPRLPVIITTNEYLGCRLPDVDAAALESRYYQFTFSTQIA
>Hydatigera_taeniaeformis_LL725419
WINNLFHSNGISPHAFCRKMECIMDKKDDKVNTLVLYGPTNTGKSLLCKLMTEFLLT-GTISRRSENTAFAFENLLDRSVGILEEPKINSANMNDMKQLLGGEEFEVAVKYKPMQ-------------------------------------
>Hymenolepis_diminuta_LM390786
WLDNLFDANGIIPSEFINKLDRVMNRNDSKINSIVLYGPTNTGKSLLCKIMTEFLLT-GTIIRRSENSNFAYENLLDRSVAILEEPKINAANVNDMKQLLGGESFEVAVKYKPMQFLHRIPVIITTNEYLGCRIPDVDAAALESRYFQYTFSSQIA
>Hymenolepis_diminuta_LM388960
-LDCLFQSNGIVALEFCNQLECVMNKRDNKVNTIVLYGPTNIGKSLVCKIATEFLLT-GTISRRSENTAFAFENLLDRSVAILEEPKINAGNANEMKQLLGGESFEVSVKYKPMQYLPRLPVIITTNEYLGCRLPDVDAAALESRYHQFNFATQIA
>Hymenolepis_microstoma_CDS32658
WLECLFKSNGIEPLEFVSRLERVMNKVDNKVNCIVLYGPTNTGKSLLCKIMTEFLLT-GTINRRSENSNFAFENLLDRSVAILEEPKINAANANDMKQLLGGEAFEVSVKYKPMQYLQRLPVIVTTNEYLGCRLPDVDAAALESRYYQYTFSTQIA
>Hymenolepis_microstoma_CDJ13943
WLECLFKSNGIEPLEFVSRLERVMNKVDNKVNCIVLYGPTNTGKSLLCKIMTEFLLT-GTINRRSENSNFAFENLLDRSVAILEEPKINAANANDMKQLLGGEAFEVSVKYKPMQYLQRLPVIVTTNEYLGCRLPDVDAAALESRYYQYTFSTQIA
>Hymenolepis_diminuta_LM389152
WLNQIFHANGIMPFEFCRQVERVLNRQDNKVNTIVLYGPTNTGKSLLCKIMTEFMLT-GTISRRSENTAFAFENLLDRSVAILEEPRINTGNANDMKQLLGGESFEVAVKYKPMQYLLRLPVIITTNEYLGCRLPDVDAD----------------
>Hymenolepis_diminuta_LM390089
WLDTLFRSNGIVPVEFCKTVEKVLDKKDNKVNTIVLYGPTNTGKSLLCKLMTEFLLT-GTISRRSENTVFAFENLLDRTVAILEEPRINAANANDMKQLLGGEAFEVAVKYKPMQYLNRLPVVITTNEYLGCRLPDVDAAALESRYYQFTFSSQIA
>Hymenolepis_diminuta_LM389639
WLETLFRSNGIVPIEFCKMVEKVLDKKDNKVNTIVLYGPTNTGKSLLCKLMTEFLLT-GTISRRSENTVFAFENLLDRTVAILEEPRINAANANDMKQLLGGEAFEVAVMYKPMQYLNRLPVVITTNE-----YLGCRLPDVDARYYQFTFSSQIA
>Hymenolepis_microstoma_CDS30830
---------KWNL---------------LTSCLLWGEGEEEVGGE----HADQHVHTGGIVCQRGEQKAFHFENLLNRTVALIEEPLITIETKNDYKCLLGGERLEIDIKCGARRFLQRVPVIVTTNEDLGSQLTSLA------------------
>Hymenolepis_microstoma_CDJ12115
---------KWNL---------------LTSCLLWGEGEEEVGGE----HADQHVHTGGIVCQRGEQKAFHFENLLNRTVALIEEPLITIETKNDYKCLLGGERLEIDIKCGARRFLQRVPVIVTTNEDLGSQLTSLA------------------
>Hymenolepis_microstoma_CDS30832
-------------RQLVRNVYRIIQ--------------------QFKHQVL------GIVCQRGEQKAFHFENLLNRTVALIEEPLITIETKNDYKCLLGGERLEVDIKCGARRFLQRVPVIVTTNEDLGSQLTSMDKAALYSRV----------
>Hymenolepis_nana_LM406075
WLIAMLNTNSIDIGALLGDIVQIMDKRKAKINTLCFRGQTNTGKTLLANLLTSHLLVRLHLKYS-----FHFFVFAHKTLILVL-------------ALFN--------RLERCADVVRPPFISTT-----------------------------
>Hymenolepis_nana_LM409198
WLIAMLNANSIDIGALLGDIVQIMDKRKTKINTLCFRGQTNTGKTLLANLVTSHLLVRLHLKYP---SHFF---VFCKTLILVL---------------------ALFYRLERCAVVVRPPFIS-------------------------------
>Hymenolepis_diminuta_LM389526
WLEFMFNMNHIPVYALLTDIIKIMDMKVNKVNCLCFYGQTNTGKTLLANLITSHLTV-------SDHNLFYFCNIF-----------------------------------------------------------------------------
>Hymenolepis_diminuta_LM391469
WLMMLLNQNGINIKELLADIITIMDKKTTKVNTLCFKGQTNTGKTLLANLITSHLTV-----RN-------IDNLI---------------------------SY-------------------------------------------------
>Hymenolepis_microstoma_CDS33272
WLMMMLNQSGININELLTDIINIIYTKTTKMDTFCFKGQTNTGKTLLANLIASHLIL-GPVCRRGDQTAFHFDNLPNRTVALMGKPRITMITKNDY-VSLE----EVDLKS-----------------------------------------
>Hymenolepis_microstoma_CDJ14557
WLMMMLNQSGININELLTDIINIIYTKTTKMDTFCFKGQTNTGKTLLANLIASHLIL-GPVCRRGDQTAFHFDNLPNRTVALMGKPRITMITKNDY-VSLE----EVDLKS-----------------------------------------
>Hymenolepis_microstoma_CDS30612
---MLLNQNGISINELLTDIINIMDKRSTKMNTLCFKGQTNAGKTLCANLITSHLTL-GTVCRRGYQTAFHFDNLLNRTAALMEEPRIAMITLEEVD--------LKAMSIDRAALYS-------------------------GVKQHTLNEPSS
>Hymenolepis_microstoma_CDJ11897
---MLLNQNGISINELLTDIINIMDKRSTKMNTLCFKGQTNAGKTLCANLITSHLTL-GTVCRRGYQTAFHFDNLLNRTAALMEEPRIAMITLEEVD--------LKAMSIDRAALYS-------------------------GVKQHTLNEPSS
>Hymenolepis_microstoma_CDJ13509
----------------------------MNTLCFKGQTNTGKTLLANPITSHLTL-ATACCRGDQTAFHFDNLLNRTVALMEESRNTMITRNDYKCLLGGGRFKIDVNYGAREV-------------------------------------
>Hymenolepis_microstoma_CDI96352
----------------------------MNTLCFKVQTNTGKTLLANLITSHLTL-GTVCRRGYQTAFHFDNLVNRTVALMEELRITMITRNDYKCLLGGGRFESHVKYGARRVCN----------------------VFRGVKQHTLNEPSA
>Hymenolepis_microstoma_CDS35108
WLMMMLSQHGIDIDELLKDIAYIMDKKTTKVNSLCFKGQTNTGKKLLANFITPHLTL-GTVCRRGDRTAFHFDNLLNRTVTLMEEPKITTTTKSAYKCLLGGDRFEIDVNYGARRFLQRIPVVDTTNEDVGALLTSIDRAALYSRVKQYTLNEQDI
>Hymenolepis_microstoma_CDI96179
WLMMMLSQHGIDIDELLKDIAYIMDKKTTKVNSLCFKGQTNTGKKLLANFITPHLTL-GTVCRRGDRTAFHFDNLLNRTVTLMEEPKITTTTKSAYKCLLGGDRFEIDVNYGARRFLQRIPVVDTTNEDVGALLTSIDRAALYSRVKQYTLNEQDI
>Hymenolepis_microstoma_CDS34279
WLMMMLSQHGIDIDELLKDIAYIMDKKTTKVNSLCFKGQTNTGKKLLANFITPHLTL-GTVCRRGDRTAFHFDNLLNRTVTLMEEPKITTTTKSAYKCLLGGDRFEIDVNYGARRFLQRIPVVDTTNEDVGALLTSIDRAALYSRVKQYTLNEQDI
>Hymenolepis_microstoma_CDJ15564
WLMMMLSQHGIDIDELLKDIAYIMDKKTTKVNSLCFKGQTNTGKKLLANFITPHLTL-GTVCRRGDRTAFHFDNLLNRTVTLMEEPKITTTTKSAYKCLLGGDRFEIDVNYGARRFLQRIPVVDTTNEDVGALLTSIDRAALYSRVKQYTLNEQDI
>Hymenolepis_microstoma_CDS30831
WLTLMLNQNGIDIYELLNDIIAIMEKKTTKANTLCFKSQTNTGKTLLANLITSHLTV-GTVCRRGDQTAFHFDNLLNRTAALMEEPRITMTTKNDYKCLFGGDRFEIDVKYGARRFLQRIPVVDTTNEDLGALLALVDRAALYSRVKQYNLNEQIS
>Hymenolepis_microstoma_CDJ12116
WLTLMLNQNGIDIYELLNDIIAIMEKKTTKANTLCFKSQTNTGKTLLANLITSHLTV-GTVCRRGDQTAFHFDNLLNRTAALMEEPRITMTTKNDYKCLFGGDRFEIDVKYGARRFLQRIPVVDTTNEDLGALLALVDRAALYSRVKQYNLNEQIS
```

Figure S2

```
>Fundulus_grandis_JW661431
SSEQNPIVRDILHVMHAYSALSLQEMVTKAPHIMSKHLHRPSLQTIIQNCRLFYMNSSRKANLLVDLPDGRPDPTRNQVLVTQHEQFAHDFTLWCTKRSSKRNTLVLHGPSNTGKTSFIRPLLALYTFSFSGCSRDLIVWENNIDLCKLMFEGSTTELPHKFRDPIKTDRTPILITTNRECAAIQNRCYIY
>Fundulus_grandis_JW661434
SSEQNPIVRDILHVMHAYSALSLQEMVTKAPHIMSKHLHRPSLQTIIQNCRLFYMNSSRKANLLVDLPDGRPDPTRNQVLVTQHEQFAHDFTLWCTKRSSKRNTLVLHGPSNTGKTSFIRALYAEATFSFSGCSRDLIVWENNIDLCKLMFEGSTTELPHKFRDPIKTDRTPILITTNRECAAIQNRCYIY
>Diaphorina_citri_XP_008482940
--------MQYALHELWKITPGYKRYKWLRMQYALHVYHHKRGPARISNEGDLSIIQHTQYRGTVFIIAEHGDHWHCTYTSSQCVKAVTRCFGIRI-KRRTIRSLDFN------TSYATNVIYGEYGPGVGDLSGKSAESGTVENEMRIFLGCTDRGEGAAEGSGTPLSNQPTHKRRKGEAHPCAPVLSTR
>Drosophila_persimilis_XP_008482940
-------------------VQKANQESWKSKR------CKFDNSKSDVPSGRRTFLTAGMCNTSLDTTISIGSECDDNFISSTEDIQCEEMFTTID-GHTLIRTKAVSSPEE--REVFVE------EFSERNLKQRIVHDVNRQRDFRM----DYEQLCRKFTGTI------LIIASHGHDCTYSNRCSFY
>Hymenolepis_microstoma_CDS25721
-----------------------------------------------------MSVHFETAYNSTKTEHEFSCPLRQAQPLPTFGSHHTRHCANVC-----KHPHHEFVCA-GVGN--EAIRVT---GAVAWYHQRCDSTGLLFCAYANCVRLHAAAAVGAIEVSL----QVVVKAMVVFVSDDGDLFVC-----
>Feline_bocavirus_JQ692585
MSKKESLILDCLKRCEDNLWLTYEDLVGGCADLILMLESMPGGSKLIESVLNMLHVRITQTHSALSYLHVRYDMKEADSLHANKLQVGHWICCVLHKKAGKQNTLNFFGPASTGKTNLAKAIVGCVNFIFNDCAKLVWWEDWVEQAKCILGGTEFRIDRKHRESHLLPQTPVIISTNNHEAPLRERVVQF
>Canine_minute_virus_FJ214110
ITKRETLMIDCMNRCEQNYWLTYEDLVNGCADLVVMLECQSGGNKLIEQILSMVHIKMTQKYTALSYILAKYDPIPWKLLTLQGWQFGHWLLCVLDKKAGKQNTINFYGPASTGKTNLAKAIVGCVNFVFNDCTKLICIWWEDWVEQAKCIMGGTQFRIDRKHKDSHLLPQTPLIISTNHHSMPLRERVVQF
>California_sea_lion_bocavirus_1_JN420361
MTKREMLMLDCIDRCEKDNLLTYEDLVDNCSALLVMLESQPGGGRLIENLLQMVHIRICQKHTALSYIQLRYSDEPYQLFAIQGWQAGHWLCCVLNKTAGKQNTVNFYGPASTGKTNMAKAIVGCVNFLFNDCAKLVCWWEEWVEQSKCILGGTEFRIDRKHKDSMLLPQTPVMISTNNHSKPLREQFNFM
>Canine_bocavirus_1_JN648103
ITKKETLMIDCIDRCEKNHLLTYEDLVNECSDLVIMLGSQPGGTKLIETLLQMVHIKICQKYTALSYVLSRYSAIEIQLLIFQGWQVGHWLCCVLHKTAGKQNTVCFFGPASTGKTNFAKAIVGCVNFVFNDCAKLVNWWEDWVEQAKCLLGGTEFRIDRKHKDSQLLPQTPVVISTNHHAKPLRERIVQF
>Scylla_olivacea_GDRN01091983
GNERPPTPPTPRCFKRRRDASTSPDRVRKSKNTAHSPEDDSPEEETTRPTATTSGQNTTRKLGRTTNNKPRGGQRK--------KKNGPKTIDLCERSTRGQDQMDARSEAN--QQVSFY------TFIVRERPPRQGGRADHGDHWHITFASNRTNLNRKRKTICLTSGARVEASASTTLIKMLK-----
>Human_parvovirus_B19
GTKASIKFQTMVNWLCENRVFTEDKWKLVDFNQYTLLSSSHSGSFQIQSALKLAIYKATNLVPTSTFLLHTDFEQVVKLLLCQNQHVLKWIDK----KCGKKNTLWFYGPPSTGKTNLAMAIAGMVNFPFNDVAKSLVVWDTIVEAAKAILGGQPTRVDQKMRGSVAVPGVPVVITSNGERMVKLNRCSPD
>Bufavirus_1_JX027296
VSAKEKSIKETVDSLTDKRIITREKWMLQDPDSYIQFIAQPGGEQTCSAVLDIATLRIATEKSAYDLVKEHPGKLKWKLIKNNNQKVYHAIMCCLNKQMGKRNTILLCGPASTGKSLLAQKIAGCYNFPFNDCVKNLIWIETQVNSFKAIMSGQAIRLDQKGKGSKSIEPTPVVMTTNEHRQPIMDRCLRF
>Bactrocera_dorsalis_GAKP01023238
RNLVGSLLSHESSTTFTEATVTMETSSWILNNYAENFQ-GPSCSSRIMETKSTLSTTAVIPITPVDVHSTISSQTSEDMLEESFRSSTGSICQSIC-RRTQGCTYTSTSPGE--RGFQVV---DVYDFA-KICKPKTEWWK------TFLLQS---PVDPNY----QLVMKLLLVVSNQKEERETK-----
>Bactrocera_dorsalis_GAKP01023412
-----------------------------------------MWTPDFLLHAISTLGKMLPLALVKKEYEA--------------KRVIHTLMEYLLVQKQLLSQKKELGKSS--MRNLVESLESSTTFTTVTMETSSWILNSYAENFQGPSCSRSMATSLS-----MTAVIPTTLVDVHSTMSSQT-----
>Octopus_bimaculoides_KOF81493
-------------------------------------------------------------------------------------------HTFVNNVWYLLNKKHPKCNLLVLRGLPNTGKSMIARSIASTISFPFQNLIVEIGLIEETLQTFKKLAEGTDIKVSVKFKPDQIVHRTPLIITSNNERVAFITRYTE-
>Lamellibrachia_sp._x2
-------------------------------------------------------------------------------------MELASASCGLIE----------EPSFTD----------EALQSFKKLAEGTPTEVSVKNKAAVKIARLPLFVTANYEKQAFSTRMIE-
>Lamellibrachia_sp._x9
-------------------------------------------------------------------------------------MELASASCGLIE----------EPSFTD----------EALQSFKKLAEGTPTEVSVKNKAAVKIARLPLFVTANYEKQAFSTRMIE-
>Lamellibrachia_sp._x4
-------------------------------------------------------------------------------------------------------MYKYPATVQGTTSFPFMELACSLGLIEEALQTFKKLAEGTQTEVSVKNKGAARVPRIPLFVTANYERQAFASRMIE-
>Lamellibrachia_sp._x6
-------------------------------------------------------------------------------------------------------------TIQGTTSFPFMELASSLGLIEETLQTFKKLAEGTPTEVSVKNKGAARVGRIPLFITANYERQAFASRMIE-
>variole_D5_NC_001611
GNKLFNIAQRILDTNSVLLTERGDHIVWINNSWKFNSEITKLILSIRHQLPKEYSSELLCPRKRKTVEANIRDMLVSTGFKFDDINIINDIQPLTDNATKGCLTFFFGGKSTT-KRLLKSDLFGPNVFCLPDFARSDNIKKNNRNHATIIIDTNYKPVRIDN---ALMRRIAVVRFRTHDKVKLLDDGKIQ
>Aretaon_asperrimus_GAWC01079972
STDQ?PICIPYRNLAFW?Q??D?K??N?YDFRGNFFKF?GPSTGIIFHGAKLSISNQATLRQRLLQQ?QTTTTTWDQNLI?AKGPPF?RSDYSFMP-NY?KQDLLTYEGHT---KHMSFHPKMGKYEFQSNNVHKFLIPGRANDSELQDLTVGTYYEP----VELTHYINRPLLLIGPPSENGLMKRLKIQ
>Amblyomma_americanum_GAGD01002820
-------------------------------------------------------------------------------------------MME-NKKSTEEETFYGIKE--SDYIIS---GRSATPSQNNSRSTPHSTKHRDELRRALAT-----QRDYETAQ-------------DEVAKNR-----
>Rhipicephalus_appendiculatus_CD793420
-------------------------------------------------------------------------------------------VLASGCYLFN-EDWGSSGRCIA------ASHLVA---------KPKDSYELSYWN--FEDFCHHYD--------------MVQPNPTFLAMTT------------
>Hymenolepis_microstoma_CDS33100
PTKTVRNVYLLLRYFKSRGQVREAGYELREEWEMAQELTALEPNQSCRSDIRPSRTKSLVAQRSDDRKARN-------------TPLERWTIERWN-RHKHAQELFMQIKDE--KESLTQ---------FRDSLRAIRLHSDKMLSLNDIKGS-----IPRYDTRKPYTRSETATERKSNDASYEPAIR--
>Trichosurus_vulpecula_EC348320
-----QTITSTICNLYNKRIVSIENWMVDEPESYIQHIANPGGEIIIKTIIEMVLLRMAKYMTAYEVIAEKGKITKFKILKMNNIKVTHAVMCVLNKKLGKRNTILLSGPASTGKSLITQNICGCYNFPFNDCCKKHHMDR------------EPINVRTKHQTRPKGKRKQ-INRTNSSNNDIK------
>Porcine_parvovirus_PT4_PPU44978
QTKKEVSIKCTIRDLVNKRCTSIEDWMMTDPDSYIEMMAQTGGENLIKNTLEITTLTLARTKTAYDLILEKAKPSMCKIFSMHNIKVCHAITCVLNRQGGKRNTILFHGPASTGKSIIAQHIAGCYNFPFNDCTKNLIWIENQVNQFKAICSGQTIRIDQKGKGSKQIEPTPVIMTTNEHTQPIRDNINLT
>Canine_parvovirus_M19296
QTKKEVSIKCTLRDLVSKRVTSPEDWMMLQPDSYIEMMAQPGGENLLKNTLEICTLTLARTKTAFELILEKADNTKCQIFRMHGIKVCHAIACVLNRQGGKRNTVLFHGPASTGKSIIAQAIAGCYNFPFNDCTKNLIWIEQQVNQFKAICSGQTIRIDQKGKGSKQIEPTPVIMTTNEHTQPIRDNIKLV
>Rat_parvovirus_NTU1_AF036710
QTKKEVSIKCTLRDLVAKRVTSPEDWMMQQPDSYIEMMAQPGGENLLKNTLEICTLTLARTKTAFDLILEKAEPSKCRIFAGHGIKVCHAIACVLNRQGGKRNTVLFHGPASTGKSIIAQAIAGCYNFPFNDCTKNLIWVEQQVNQFKAICSGQTIRIDQKGKGSKQIEPTPVIMTTNEHTQPIRDNIHLT
>Mouse_parvovirus_1_MPU12469
QTKKEVSIKTTLKELVHKRVTSPEDWMMMQPDSYIEMMAQPGGENLLKNTLEICTLTLARTKTAFDLILEKAETSKCKIFAFHGIKVCHAICCVLNRQGGKRNTVLFHGPASTGKSIIAQAIAGCYNFPFNDCTKNLIWVEQQVNQFKAICSGQTIRIDQKGKGSKQIEPTPVIMTTNEHTQPIRDNIHLT
>Mouse_parvovirus_3_NC_008185
QTKKEVSIKTTLKELVHKRVTSPEDWMMMQPDSYIEMMAQPGGENLLKNTLEICTLTLARTKTAFDLILEKAETSKCKIFAFHGIKVCHAICCVLNRQGGKRNTVLFHGPASTGKSIIAQAIAGCYNFPFNDCTKNLIWVEQQVNQFKAICSGQTIRIDQKGKGSKQIEPTPVIMTTNEHTQPIRDNIHLT
>Gray_fox_amdovirus_JN202450
DEQCSDMADDSQDWDDSQEPPKTEPPKVTTETISNVTYTSTIWASCSTKATKPKTITETAKQPSKKLTTAKSTLDSECIMKFNDEKVLKNIIATVLKQSGKRGCIWFYGPGGTGKTLLANLICGMVNFPWTDCGRNMIWLENFIEDFKAITGGGDIKVDTKNKQPQAIKGVTVITSNKDHSEPLKQRIVKI
>Aleutian_mink_disease_virus_JN040434
MYLDEQSSDIMDANIDWEDGQDAPKVTDQTDSATTKTGTSLIWKSCATKVTSKKEVANPVQQPSKKLYSAQSTLDADCIIKFNEEQYLKKVLCTILKQGGKRGCIWFYGPGGTGKTLLASLICGMVNFPWTDCGRNIIWAENNVEDFKAITGGGDVKVDTKNKQPQSIKGCVIVTSNTNHAEPLKQRMIKI
>Acheta_domesticus_mini_ambidensovirus__KF275669
ISRAGQLIHFLLNNVYSPLDLIKQSNEWYTSPFATCKETGVEGAHTFLKSLSVSEMENFVKGNHIQYKNNIVRYSCEQLLMYQFSDFIHQLYYVLNKQNGKKNCIELIGPPSSYKSTFLHWVAGIIPFGFAPAVKRVLIIDEYHETLLNILSGTTCNVNVKYKTNGYVTHTPVLMASNYTSTRFEHRMVTF
>Aretaon_asperrimus_GAWC01012962
PSQRRYSLEELSMKLTPALIAKKNLCFVCTPLMLLTKLVNLNKTSIQQGLETAKCLINSFTCNELVSYVQGTHPISQLLSYQFLYFKRCFFDVIT-VADKKNVLEIVSLPSAYKTTFITMVAGYCRFGFQNIITRLAIIDPSKDILLCLFSGDKINLNQKNLGDDTAVKTSCVVLSNTIGQRWNDRMYRF
>Extatosoma_tiaratum_GAWG01046839
-------------------------------------------------------------------------------------------MVKLYEVVIG-INNLKNVVEIISPPSAYKSTFATMIAGYCRFGFQNIAVRLAVIDAGKETLLCLFSGDKVNVARKNLPDHVQLNTPFLVLSNR----WDVRMERF
>Penaeus_merguiensis_hepandensovirus_DQ458781
YLENLQVLEKYLVKHKCYTIQDFKMMQRSDDEIWVNYMYDIQNLEKVIEKLNIMEYSLQQADYIEGNTWIGEDLWNKIIENMRVVQWIQDFMDIMHGNLPKINCMMLYGNSNSGKTQLIDALTGLVTFHFSNITMSTIVVGQTIEQWKGLCGGENVTPMKYKEHKHMFRKPVFLTNQHDRRAIENRSFMY
>Penaeus_chinensis_hepandensovirus_AY008257
YLENLQVLEKYLVKHKCYTIQDFKMMQRNEDEIWVNYMYDIQNLEKVIEKLNIMEYSLQQADYIEGNTWIGEDLWNNIIENMKVVQWIQDFMDIMHGNLPKINCMMLYGNSNSGKTQLIEALTGLVTFHFSNITMSTIVVGQTIEQWKGLCGGENVTPMKYKEHKTHMFRKPVFLTNQHDRRAIENRSFMY
>Fenneropenaeus_chinensis_hepandensovirus_JN082231
YLENLQVLEKYLVKHKCYTIQDFKMMQRNEDEIWVNYMYDIQNLEKVIEKLNIMEYSLQQADYIEGNTWIGEDLWNNIIENMKVVQWIQDFMDIMHGNLPKINCMMLYGNSNSGKTQLIEALTGLVTFHFSNITMGTIVVGQTIEQWKGLCGGENVTPMKYKEHKTHMFRKPVFLTNQHDRRAIENRSFMY
>Penaeus_monodon_hepandensovirus_2_EU247528
YLENLQVLEKYLIKHKCYTIQDFKMMQKNEDEIWVNYMYDIQNLEKVIEKLNIMEYSLQQADYIEGNTWIGEDLWNKIIQNMKVVQWIQDFMDIMHGNLPKVNCMMLYGNSNSGKTQLIEALTGLVTFHFSNITMSTIVVGQTIEQWKGLCGGENVTPMKYHKT-HMFRKPVFLTNQHDRRAIENRSFMY
>Penaeus_monodon_hepandensovirus_3_EU588991
YLENLQVLEKYLIKHKCYTIQDFKMMQRNEDEIWVNYMYDIQNLEKVIEKLNIMEYSLQQADYIEGNTWIGEDLWNKIIQNMKVVQWIQDFMDIMHGNLPKVNCMMLYGNSNSGKTQLIEALTGLVTFHFSNITMSTIVVGQTIEQWKGLCGGENVTPMKYHKT-HMFRKPVFLTNQHDRRAIENRSFMY
>Penaeus_monodon_hepandensovirus_1_DQ002873
YLENLQVLEKYLVKHKCYTIQDFKMMQRSDDEIWVNYMYDIQNLEKVIEKLNIMEYSLQQADYIEGNTWIGEDLWNKIIENMKVVQWIQDFMDIIHGNLPKINCMMLYGNSNSGKTQLIEALTGLITFHFSNITMSTIVVGQTIEQWKGLCGGENITPMKYKEHKHMFRKPVFLTNQHDRKAIENRCFMY
>Penaeus_monodon_hepandensovirus_4__FJ410797
YLENLQVLEKYLRKHKCYTIQDFKMMQRNDDEIWVNYMYEIQNLEKVIEKLNIMEYSLQQADYIEGNPWIGEDLWNENMRVESRVQWIQDFMDIMHGNLPKINCMMLYGNSNSGKTQLIEALTGLVTFHFSNITMSTIVVGQTIEQWKGLCGGENVTPMKYKEHKHMFRKPVFLTNQHDRKAIENRCFMY
>Daphnia_pulex_EFX61411
DFRLEEEQKVNSVSSACNFLREISKRYWLSDFSSLLAVSDVMLKEDLQKVLMYNMLKVQQYMNNVFTLEYNIMRNQFSTVRMNDILFRKWFDR----ELGKKNTIILIGEPSCGKTWLAK---GKIQFIFSGCCGRIILHDAYSETLKKIYAGECCPISVKYQSGQLSCGAPVIGTCNSFRQAFDERITFL
>Neobenedenia_melleni_GW918566
EELTNNIILFLIKSVLNLSFHKFDLEIFIKNLDININN--KNILLYIIKNYYDYIKFYYDFYADIQLEPLYLHSFNTKILNEIFNFFISSLKNIIFKKYFKINTLLFYGNPNSGKSLIANQVYKIVEFYLSNFDANIIICEANVDDFKNILGGTSIVVNNKKYERRGINRIPILITSNYDTSALLRRMIV-
>Enchytraeus_crypticus_GALF01015487
LCKRWNAFTVTDMYQAQAEIPKEEDKDWREFWNRMVNK------PSIRTMMDNIRSSVHSDYMYKPFQDLVEEACTRQWLKDQGRDFVQDVVDILDRKLEKINTMCLIGPSNSGNTVMFSATIGRRKFVFQEIPTRLVCVDDNQENWKLLLGGEEMSVPVKFKGNARILRTPVLLTGNKNQDAFMNRMK--
>Planococcus_citri_densovirus_AY032882
EKVVEEVSSLLVKYCPVPLIGIKNLIPPNHADFRTIL-FDPRNKDNYISACDIHALHL-NKYSFKEFHELYSQADNDELLKFQLKPFLSNLRNWFNEGNPKMNALCVIGPPNSGKNYFFDAIIGHINFALQDVVRRIVVGNGAKEDFKKLCEGTAFNVRVKFKADCIFTKTPVLLISNNWDPHFKDRLKTI
>Solenopsis_invicta__densovirus_KC991097
LLNKYFVLPSNHIRDIMLGDSQYNFLFDPSNDKAYMASCEIFQKKFIPLKLKDLRDIFEGSTPVFYANNIDPNLYYDDLLRFQYVELLTNVRDWFNDKNPKMNALCVIGPPNSGKNYFWDMFCGHIQFALQECYRRLVVGNGAKEDFKKLCEGTAFNIRVKYQGDKIFTKAPVLLISNFYDTHFKNRLHTI
>Bemisia_tabaci_EZ941734
-------------------------------------------------------------------------------------------------------------FRNFK----------------TGGDTCNVKVKYCMDHILHRTPVIILSNNKDPAFNSRIIRH
>Lamellibrachia_sp._x7
-------------------------------------------------------------------------------------MYQAVYPWLA-----RDQFHLYVPRVNNNSHWSDPALRMVNITFGDKAMHCCILTVDVATKGATIHSASHKLFVSKRDAHVVYHGP-------NECTDQK-----
>Lamellibrachia_sp._x8
-----------------------------------------------------------LVLLTTPYTQEYREATG--------PTLGSASLKLIS-FINRCDVHVVYGPDTCGKKPWCR------TF-ISKCRGLDTVWANVVHQHKALTKNCQIPSTTKF--PGYLSQSP--------RVTWHM-----
>Argulus_siamensis_JW950843
DKIFKEIRALYAANLVVPIENIVDTQAFMDNSIFWEPNYRLKNEFNQYSLADIRSYLENKVVDPLFTCGWLQTDKYKTLLIHQFKEFLTDVVTIYNNLNPKMNCMIIHGPPNAGKNYFMDTLFGYITFSLSNAYCRVVCANGAIEDFKKLCDG---------SG----------------------------
>Sus_scrofa_EV965154
-------------------------------------------------RHRQLRRDRVRVDHQKERQTQHHRAVGSEQH--------LRPGTETGRQLR-RVRQRTDLLLRGPVR--QGHRHL---GRA---ADQSG-------VRGESQTGLRGG---------------------------SDTGAGQ-----
>Triatoma_infestans_GBBI01004972
ASRKEHAVATVYGWLEKHIIPDHDITSSDKWLSSDFKFITERDGIFQSALKIHRTVLSKLKLNEMIDYINKAEYVHMILSFQMVQFLQFLYKLLNHQTGKMNCLLICGPPSVGKTWFANIIGGYIQFPFQDCPRNLLIFDASLETFKLLFAGTPTPANIKYESQNLINRTPVIVTCNHQNSEYRTRINRF
>Halyomorpha_halys_GBHT01013004
EENIIVPDHDITRHSIWLDHQYFRFITDRSPDFQNALKRTILSNMVLDDLREYIRDKNVYYYPPKETVLTLDESYNQQMTAYQLSEFLHNVYLMLNHRSGKHNCILIKGVPSVGKTWFANLLGGYIQFPFQDCPRNLLIFDSSYENFKLLFAGTPLSANIKYESRNLVNRTPVIVCCNSHNPEFNSRIYRY
>Rhodnius_prolixus_T1I3J1
TDESTDEGTSNVYPITSGIYNIIQRWKGANGSQRRSMATPGSNLVQIRIMPYQQAKTMDPTQRYKCTKKYAQVVLN--------RQLFREMEKVTDTGSGLESNMPLSGPPSVGKTWFANIVGGYIQFPFQDCPRNLQIFDMALENFKLLFAGTPTPANIKYESQNLINRTPVIVTCNRKTPEFQTRIRRF
>Rhodnius_prolixus_T1HE18
----------------MTAHTPTKAVAAD--------LSTQFNNNIKEVTNNTITDESSDEETSNVYPITSGIHNQRWKGANGRSMATPGASLVQRIMPYQQAKTMD--PT--KRYKCT---KRY---AQDCPRNLLIFDMALENFKLLFAGTPTPANIKYESQNLINRTPVIVTCNRKTPEFQKRIRRI
>Graminella_nigrifrons_GAQX01015625
-----------------------------------------MYECRSKSEINSYYYDMKTSTVAI----------LRLLYHQFRNFLVTLYSFLNSDEGKVNTILIVGPPNSGKNFFIDPLCGKIAFPWMDGFRRVNKWNYFEENVLDLMQGNMLSCNIKYEAPQPLLKTPLIVMSNNNQERFNIRHIRY
>Acheta_domesticus_densovirus_HQ827781
SKFQTFAEEAKVLFSKLNCIPVRDARIIIDRKHPMYFEYTAAISLYVRDINEMSLNDFYLQTLENTCVYQAGNIDPNELILFQCKELLMNIVAWFNRLNPKVNTVCIIGNHNCGKNYFWDAVCGFLKFALQDCVKRIVIGNGAKEDMKKLCEGCPFNVSVKHQSDGIVARTPVCLISNNDAHFVDVRMKVF
>Raphidia_ariadne_GACX01003907
QTKGQKLVEFFKQYPCTPIDNVRQTVQYLNSEF-----ADVCVTDKLFMRARELYSCYFLNKTVIELYQVFQKSTPRELLLFQFSLFLQNLYRVLNRTNGKKNAIMLVGPANACKSMFANVVAGQITFSLDDCVKRILIWDEYLETIKLLFSGDPCPANVKYEKQSIIHKTPVIITGNTNNDEFNNRLNIY
>Drosophila_sechellia_DK330728
-------------------------------------------------------------------------------------------------------MSHWNR------SFPLQMCGVRLIFWNSVERTLLRLLGGDALNAAQKNQNDTNISKVPVIVTANHKSDEFKVRIQYH
>Drosophila_sechellia_DK334580
-------------------------------------------------------------------------------------------------------------MLRLLGGDALNAAQKNQNDTNISKVPVIVTANHKSDEFKVRIQYH
>Tetranychus_urticae_T1KBZ7
-------------------------------------------------------------------------------------------------------MVAGHVNFPLNDCDRRILIWNSAYDTVKMLTEGDPFPASVKHRNDATILKTAIIFTANNDSLKQID-----
>Metaseiulus_occidentalis_JI999665
-------------------------------------------------MSTFYYSIEESIVILEEFLRYQYG--------FFFCQKLYNITE-LIPKRNCMYILGPTNAGKTWFMN------CVSYLNVG---------------------------------YVKNPV-------KGARFPY-----
>Metaseiulus_occidentalis_JL018010_1
```

KVTHQTLMCFILTYLPVPFHAACELPAWLDDEDLSYFYKSDRDYCRAIQQVQRRVATMTLHQIHDLTKDCTS----EELLKFQYFFFLQKLYNITERLIPKRNCMYILGPANAGKTWFINCVAGYVHFPYNDCTRRILIWNSALEDVKMLAGGDPFTCNVKYESNVCVNRTPLFFTANSELPEFKSRIYLE
>Metaseiulus_occidentalis_JL018010_2
KVTHQTLMCFILTYLPVPFHAACELPAWLDDEDLSYFYKSDRDYCRAIQQVQRRVATMTLHQIHDLTKDCTS----EELLKFQYFFFLQKLYNITERLIPKRNCMYILGPANAGKTWFINCVAGYVHFPYNDCTRRILIWNSALEDVKMLAGGDPFTCNVKYESNVCVNRTPLFFTANSELPEFKSRIYLE
>Orchesella_cincta_GAMM01017564_1
KSIPDKIIKFLRLYPVVPLYATPDIREYIESPYRLYRAKSPTFQIAVDILENTINRMTICELAELYHNKHCKPIWSESFLLTQFYHFLTILTSILDKSIPKLNTLIVYSPPNAGKNWFFDSVTGQLNFAYQDLYKRVVVWNEELDMLKMVLGGDRTAIKVKHEKGKQSSRVPVIILTNNDDIAFGTRVRMF
>Orchesella_cincta_GAMM01017564_2
KSIPDKIIKFLRLYPVVPLYATPDIREYIESPYRLYRAKSPTFQIAVDILENTINRMTICELAELYHNKHCKPIWSESFLLTQFYHFLTILTSILDKSIPKLNTLIVYSPPNAGKNWFFDSVTGQLNFAYQDLYKRVVVWNEELDMLKMVLGGDRTAIKVKHEKGKQSSRVPVIILTNNDDIAFGTRVRMF
>Papilio_polyxenes_densovirus_JX110122
GNTVEKFVEQFKLRPTWPLEHILDTKDWLESEWCTYIA----GDKQIQRTFQILMKIT----MNMTFEELRMLYQNEELLDYQLRRFITDLYEILEKKHQKTNTFQIVSPPSAGKNFFIETVLGVISFPLMEAVRRVNYWDDATETLKKLFAGTALKASVKFQKEANVQKTPVIITANRTSEVNDDRIIKY
>Bombyx_mori_densovirus_AY033435
GNPVEKFVEQFKLRPTWPLEHILDTKDWLESEWGTYIA----SDKQIQRTFQILMKIT----MNMTYSDLKMLYQNEELLDFQMKRFITDLYEILEKKHQKTNTFQIVSPPSAGKNFFIETVLGVINFPLMEAVRRVNYWDDATETLKKLFAGTALKATVKFQKEANVQKTPVIITANYTKEVNDDRIIKY
>Hordeum_marinum_iteradensovirus_KM576800
GNSVEKFVEQFKLRPTWPLEHILDTKDWLESEWGTYIA----SDKQIQRTFQILMKIT----MNMTYSELKMLYQNEELLDFQMRKFITDLYEILEKKHQKTNTFQIVSPPSAGKNFFIETVLGVINFPLMEAVRRVNYWDDATETLKKLFAGTALKATVKFQKEANVQKTPVIITANFTKEVNDDRIIKY
>Sibine_fusca_densovirus_JX020762
GNPVEKFVEQFKLRPTWPLEHILDTKDWLESEWGTYIA--------SDKQIQRTFQILMKITMNMTYSELKMLYQNEELLDFQLRKFITDLYEILEKKHQKTNTFQIVSPPSAGKNFFIETVLGVINFPLMEAVRRVNYWDDAVETLKKLFAGTALKATVKFQKEANVQKTPVIITANYTKEVNDDRIIKY
>Casphalia_extranea_densovirus_AF375296
GNPVEKFVEQFKLRPTWPLEHILDTKDWLESEWGTYIASDKQIQRTFQILMKITMNMTYSELKMLYQNEHTMPIWGEELLDYQMRKFITDLYEILEKKHQKTNTFQIVSPPSAGKNFFIETVLGVINFPLMEAVRRVNYWDDATETLKKLFAGTALKATVKFQKEANVQKTPVIITANHTKEVNDDRIIKY
>Danaus_plexippus_plexippus_densovirus_KF963252
GNPVEKFVEQFKLRPTWPLEHILDTKDWLESEWGTYIA----SDKQIQRTFQILMKIT----MNMTYDELKMLYQNEELLDFQMRKFITDLYEILEKKHQKTNTFQIVSPPSAGKNFFIETVLGVINFPLMEAVRRVNYWDDATETLKKLFAGTALKATVKFQKEANVQKTPVIITANYTKEVNDDRIIKY
>Priacma_serrata_GACO01003330
--MSNVLYTRQWMQSKYKFLPDTSVLLRSCFKIINYEIANMSVEELWEYIERIEPQNRLFACHSSSSISEVYYSVENEQLLLFQYYVFLYDLYLLLNRSVPKYNCICVIGPPNSGKNFFDAVVHSCINYG----------
>Blattella_germanica_densovirus_AY189948
GGATGIILDLLSKCAVCPLTEIVFTKEYLQDPIV----ACKRLDSKEVKDAIDTRASVINTWEREDFVAFYNNPNTTELLTFQMVQFCRNLVDTLECNIPKRNCFVVCSPPSAGKNFFFDGVKGQMQFAYQDCHRRIIIWNREMENLKMLFAGDNLSANVKCKPQANVKRTPVIVLTNSQQTAFNDRVITY
>Blattella_germanica_densovirus-like_virus__JQ320376
RGDTRVILDLIKRYAVCPLSEIVYTREYLENPIAVKRL-----DDRDVKNAIDCHAAIINTWNREDYVKFYEDPNTNKLLDYQCKQFVTDLINVTDMNIPKCNCFLIVSPPSAGKNFLCDAIKGQMTFAYQDCHRRLLIWNRETENLKMLFGGDNLSANVKCKPQANVKRTPVICMSNVNHEAFADRVITY
>Tetramorium_bicarinatum_GASM01014809
RGDTGVILDLIKRYAVCPLSEIVYTCEYLENPIAVKRL-----DDRDVKNAIDCHAAIINTWNREDYVKFYEDPNTNKLLDYQCKQFVTDLINVTDMNIPKCNCFLIVSPPSAGKNFLCDAIKGQMTFAYQDCHRRLLIWNRETENLKMLFGGDNLSANVKCKPQANVKRTPVICMSNVNHEAFADRVITY
>Dysaphis_plantaginea_densovirus_FJ040397
GLDSERSAGGHQGEEKHEDGPGVRVWPKTLDELLVKYPCCPPEAFYNVKEFYNNPHLNHIDDNCKKIRVALRNWCAEELLCYQFSKFLYDVYAIVDKLVPKRNSMCVVSPPSAGKNFFFDAVAGMFNFSWADGAKRLVLWNFHVEKIKELLGGDTTRVHVKYKGDQPLQGPPIFLLTNNNDPAFADRLVTY
>Acyrthosiphon_pisum_J9LI55
YEDLIVQLGRQLANKFVSGIHRYISTVIVTSGTRGVQECLRELDSLIERYPPAHFYIISEHEDHIHISHVCGYSRQSFLVRGAFRRGLRRISRAVDSATDYRNAQRYMGQAQ--GGLLAREMSGSV---QKHKTKKNSWVDTLIGTYRAYWG-IKLMHFIDFEKVHAYNRLKEQVYYDINELLLYQNSENV
>Acyrthosiphon_pisum_J9M820
DKLVGAPSSFEAHFGAKTEDVDFVGLLEQTTCNQEITSADASISLEGSEKMSDYLIDINIYKRKDIWDKPKGDYWRRGTFPANHVTYASKIVTLLGKMMPMKNAQRYMGQAE--GGLLAW------SYSSWDVPRDVCIENNDADEKRKKDTQPGLQSTLAI----LAPPNEGTANKTNNFASMEARLVLW
>Acyrthosiphon_pisum_J9LBV1
--MEAA-G--------------------KRSVLWNNHVNELKALLGGDSCRI-----SDTDMCNAPNIVHVHIREGVIEE-----
>Acyrthosiphon_pisum_J9LI54
EEHIHIGHICGYTGQTCRCSFLIRGTFWSLYRRRGLRRISRGIDLGPSDYRDILRYLSSGSRRIHCVGGFGENGRL--------KYLSVNLINIIDKRIPKCNSLAILAPPHAGENYFFDAVAGVLNFAFMEAAKRLVLWNNHVNELKSLLGGDSCRISVKYKTDQ-----PI-----------------
>Acyrthosiphon_pisum_J9LDI1
AAKFVSGIHRYVSRGVDGVENCLRELDSLIDRYPPRHFYIISAHGDHVHISHNCGYSGQTCRYPFILKLWRKDIWDDYWRGATAVTYASKRVTLLGKPMKNCNTLAILAPPNAGKNYFFDAVAGCLNFAFMEAAKRLVLWNNHVNELKALLGGDSCRVSVKYKSDQALQGPPIIILTNDGMTAFRPRIKLY
>Acyrthosiphon_pisum_J9LTF9
YTLPTFVDTQAKLTQRHLEQAKGGLLAWSYGEVSRDFPCQPSGSTVRVPNSHINGGENDSVKKSKKNQK-------NTWVDPKNGPVSKNLINIIDKRIPKCNTLAILAPPNAGKNYIFDAVAGVLTINFMEAAKRLVLWNNHVNELKALLGGDSCRISVKYKTDQPIQGPPIIILTNDGMSEFQPRIKLY
>Acyrthosiphon_pisum_XP_003246947
IRPENGPISITELLHQYPCCPPDAYKNIKEYYLNKNLNYLCARDKYVDRDLELWNIKL-MHMSLKDFDNYYKDDNVNELLLYQFFFFLTDVLNVIDKRVPKCNTLAIHAPPNAGKNYFFDAVAGVLNFAFMEAVKRLVLWNNHVNELKALLGGDSCRVAVKYKSDQALQGPPIIILTNDGMSAFQPRIKLY
>Acyrthosiphon_pisum_J9KNE3
----------MPKAQRYLGQAEGGLLAWSDGEVSWDFSCEPSSSSVLVQDSKNAGGENGGPKGRKKDSK-------HAWIRPEN--GPINVLNVIDKRVPKCNTLAIHAPPNAGKNYFFDAVAGVLNFAFMEAVKRLVLWNNHVNELKALLGGDSCRVAVKYKSDQALQGPPIIILTNDGMSAFQPRIKLY
>Helicoverpa_armigera_densovirus_HQ613271
NISTENLLEHITSIVCYPMDHVFQTRDWYTGKYKYLIS----SDKLVKRCYSILQKQL-MNWSYNDFLNHYRKDDTEELLDFQLREFVTDLWDVLEKKVPKKNTFQIISPPSAGKNFLLDAIFGMISFPLMEAVKRVNCWNSAEDTIKMILGGDPIKAAYKYESERNIARTPVIVMSNKINDAFEDRIITY
>Myzus_persicae_densovirus_AY148187
GSQEGPRQKKSRRKQRFRMIREIDVSKPQSIEELIYRYCCPPEAFYNIPEFYANTNINWIDMKDFKVTIPLRNWAAKELLNYQFIEFLTTLYNVIDKRVPKLNSICIKSPPSAGKNFFFDAVAGMFNFSWADGAKRLVLWNYHIEKIKELLGGDTTRIHVKYANDVSVQRVPIIILTNNSHPAFNDRLRSY
>Dendrolimus_punctatus_densovirus_AY665654
GDSVQKIIRFIQQKPTWPLEHVFQTKEWLDGPFGTWVS----SDKTVKRAIEIFSLQT-MHYTYTQFVDYYRSLECERLLNYQCKVFINDLWDILEKKRPKQNTFQIIGEPSAGKNYFIDAILGIIQFPLMEAVRRVNVWNAALDTIKMLLAGDPLKVNVKYQKEQYLQKTPIIVMTNKNDLAFHDRVCTY
>Bemisia_tabaci_EZ966062
--KKTFKKNTLFILSPPNSGKNYFFD-----------CNIQAVINSGHIGNLTKIHLFLWKQLESFY----GMNRIL-------SRMHLRI-----
>Sitobion_avenae_GAPL01021454
GVQSDGSNSGAVEKIEKRRLKRKYNKSVVDVDPNYPKNCCPPEAFYNIPEFYANRNVNWIDLKDFKITIPLRNWAAKNLLMVQFTEFLTTLYNVLDKKIPKLNSICIKSPPSAGKNFFFDAVAGMFNFSWADGAKRLVLWNYHLEKIKELLGGDTTRIHVKYHNDVSIQGVPIIILTNNNHPAFRDRLRTY
>Lygus_hesperus_A0A0A9YTR0
RKLTVNQIGELVNSMAAIPIRNCILSKRWRFNKDVKFI---TEEDKIFRRAVIYVESKCIYRSTRDFVDFYKFINPNDMLLFQFREFLQSCYDIIEKIIPKKNCLEIVSPPSSGKNFFIDFMSGNINFPFQDCARRLLIWNSALDTCKMLFGGDPLPAKIKFCGDATIPRTPVFVLSNKNDTAFNDRMYKF
>Ixodes_scapularis_XP_002401871
----------MDVSASYDAM------VRLLEYQLQEFVHDLYNLLERRVPKKNCMEVVSPPSAGKNFFFDAVVGNINFPLQDAVKRVLIWNAAYETIKKIFGGDVDNVAVKYSPDMIVTRTPVIVLSNNRDEAFNHRMFRY
>Rhipicephalus_microplus_EW679313
----------MSRRRYRKKNCMEIVSPRVQAKTFFSNLFFGTISFPLQDTVRRILVWSSAF----------
>Rhipicephalus_appendiculatus_CD786958
TQPLDFIEMYKNTQPLFDGPHSDLATFYMDVPTSLDAM----------MELLNFQFSAFVNNLYSLVKKAAPKKNCMDIVSPPSAGKKFFLNPIFGTISFPLQDTVRRILLWNSVFDTVKKKFGGDVDSVAVKYSADQTISRTAVIVLSNYDDEAFNHRMWRY
>Rhipicephalus_pulchellus_L7M443
----------MDVPASFEAM----------MELLNFQFSAFVNNLYSLVEKAVPKKNCMEIVSPPSAGKNFFDPILGTISFPLQDTVRRILVWNSAFDTVKKIFGGDVDSVAVKYSADQTISRTPVIVLSNNDDDAFNHRMWRY
>Bemisia_tabaci_EZ941358
----------ASDTPSEYYFGIIKSFDVCMELLSWQMGFENI----------KHFLQDIFDVLEKVRPKVNTLLILGAPNSGKNFFFDSVIGQMNFPLQECVRRVILWN----------ESRSFEN-----
>Bicyclus_anynana_GE699248
----------YVADWNNVGNYYYDIQTSVSVL----------EELVKYQNYIFLRELYWLVDKKIPKKNCMSIVGPPNSGKNYFFDCVSGQIRFPLQDAVRRILLWNGSEEVLKLLFGGDTCPARIKYEGDASIRRTPIIVQSNKNHPAFESRVWKY
>Chilo_suppressalis_GAJS01033341
----------MLFGGDTCPARIKYEGDASIRRTPIIVQTRAQTSAFTSRVWRY
>Cherax_quadricarinatus_densovirus_KP410261
ERIQLKTVELLELYPICPPEAIVKHRIWRCDPELRFKNISSFKDTLTTYSMNEYQKMYNNPDCIPIFSAGYGNYETDELVNYQCIDFVTTLYNVLERKVPKLNCIVIHSPPSAGKNFFDAVKGHLNFPFQDAERRIVLWNEFLEQIKEILGGDSTSVNVKYQSDTPVYRTPVIVLTNNNHSAFIDRIRVF
>Sea_star_densovirus_KM052275
KRKLGAQERIQIRVVEMLQLNPICPPEAIIKHRVWRNDPDLRFKDMTHKEIKTAVNSFSDNLTTFSMSDYQSMYNNDKLVQYQCLDFVNTLYNVLERKIPKLNCICIYSPPSAGKNFFDAIKGHLAFPFQDAERRIVLWNEFLEPIKELLGGDSTSVNVKYMHDTPVYRTPMIVLTNNTHPAFKDRVRVF
>Ampelisca_abdita_JZ096885
----------PIFSAGFLAFENYYYNIENSVKAM----------TEIIMYQCLNFVT?LYNVCERKKPKLNSICIYSPPSSGKNFFFDAITGELVFAFQNMERRIVLWNEYLDLIKKLLGGDTCTVQVKYAHDCVVYRTPVIITTNNNHHAFEDRLRTF
>Ampelisca_abdita_JZ088717
----------MYQCLNFVTTLYNVCERKKPKLNSICIYSPPSSGKNFFFDAITGELVFAFQNMERRIVLWNEYLDLIKKLLGGDTCTVQVKYAHDCVVYRTPVIITTNNNHHAFEDRLRTF
>Ampelisca_abdita_JZ093383
----------TEIIMYQCLNFVTTLYNVCERKKPKLNSICIYSPPSSGKNFFFDAITGELVFAFQNMERRIVLWNEYLDLIKKLLGGDTCTVQVKYAHDCVVYRTPVIITTNNNHHAFEDRLRTF
>Ampelisca_abdita_JZ096796
------AVGLLRKHPICPPEAICKTSIWREDPDLRFKNYCDREIKAAVASFRDELINYTIDDYQKMYNKETC----LPIFSAGFLAFENYYYNIEN-SVKAMTEIIMY----------QCGEEATL-------NFVTTLYNVCER----------KKPKLN-------SICIYSP-----
>Ampelisca_abdita_JZ093297
ERIQLRAVGLLRKHPICPPEAICKTSIWREDPDLRFKNVASFRDELINYTIDDYQKMYNKETCLPIFSAGFLAFENTEIIMYQCLNFVTTLYNVCERKKPKLNSICIYSPPSSGKNFFFDAITGELGTA-----RRNVF----------
>Periplaneta_fuliginosa_densovirus_AF192260
SDIREEMEKIVNENPVCPLAGIISTRQWLTHKDLRYLR----ADNETVKSFLDAKSEEMCYWTLHDFNTFYSQPYCNKLLAYQFKSFLNTFYNVLERKLPKCNTICVWSPPSAGKNFFFDVYLGQLNFSLQEATKRVLLWNAYTDTLKMLTGGDALCVRVKQKKDCHVYKTPLIVLTNNHELAFVDRVKVY
>Bombus_impatiens_JI097654
----------M----------GNFSFPMMDCIKRVIMWNSALETLKMVFGGDTCPVKVKYLGDKLLLCTPIICLSNNQDDAFTCRMFTY
>Teleopsis_dalmanni_GBBP01040481
----------MECVKRIILWNSATETLKMLFGGDTVTAKVKYCNDAVIMRTPIIILSNNNDIAFRSRMFTY
>Kerria_lacca_GBDP01003119
----------MYAKKEAKRNTLFVLSPPNAGKNYFFDAIVGQLNFPMQECVKRIILWNEHFEEIKTLLGGDNANAKIKHQCDSIIPRTPIIMLGNQKDEAFNTRILKY
>Kerria_lacca_GBDO01030308
----------MQAVIRRISIEITEWQIENFMTKYNNCDPYLIFNAPKYYAIDKTYIMYAKKEAKRNTLFVLSPPNAGKNYFFDAIVGQLNFPMQECVKRIILWNEHFEEIKTLLGGDNANAKIKHQCDSIIPRTPIIMLGNQKDEAFNTRILKY
>Corydalinae_sp._GADH01010824
----------LKYYCCPVSAVRDVKEFREDDLL----CDPKNKDYVQAAFDDFGKDL-IDLSLRDFYNMLKDRKPDQLIKFQFCIFLTSLVDVLDKRLPKCNSLCIKSPPSAGKNFFFDMIFGQLVFAFQEAPKRILLWNSLTDTIKMMLGGDPYTVRVKHAMDTHVKRTPVIILTNETNIAFRERIVKF
>Kerria_lacca_GBDP01040339
RSKGQKLYEWFQQFPTGPIKQIINTRHWSQGPYK----FMPKNKPILITCFENMAFEF-LDMNVKDYFKYFNNKDPDELLKFQYTGFLTNVLNIVDKKILKKNALFIHSKPNAGKNYFFDAIIGQLSFPMQECVKRIILRN----------NI----------
>Kerria_lacca_GBDP01000144
RSKGQKLYEWFQQFPTGPIKQIINTRHWSQGPYK----FMPKNKPILITCFENMAFEF-LDMNVKDYFKYFNNKDPDELLKFQYTGFLTNVLNIVDKKILKKNALFIHSKPNAGKNYFFDAIIGQLSFPMQECVKRIILRN----------NI----------
>Culex_pipiens_densovirus_FJ810126
AAIKQETKALLLKYYPCPLSAIRDVSEFRNNNLL----SDPKNREYIDAAFVDFGKDVNQMTIRQIYEILKGHTDPEDFLKFQFTQFLVDLIDVLD-RVPKQNAFAVISPPSSGKNFFFDAICGQLHNLFQEAPKRLLLWNAMTDTIKMMGGDPYTVRVKYQGDTHVTRTPVLI----------
>Culex_pipiens_x
----------AMTDTIKMMGGDPYTVRVKYQGDTHVTRTPVLI----------
>Helicoverpa_armigera_densovirus__JQ894784
AGKFAYIRQKTKALLRKYYCSPIIAIRDVDEFRDDDLLCDPKNRDYVEAACEDFGKDL-NDMSLRQIYNLLTEDYNTELLTFQCVEFLTNLVNVLDRKKPKLNTFVVYSPPTAGKNFFFDMIFGQLLFAFQEAPKRVLLWNSLTDTVKMMLGGDPYTVRVKNRGDQHVKRTPVIVLTNNYEIAFNERIVQY
>Diatraea_saccharalis_densovirus_AF036333
AGKFAYIRQKTKALLRKYYVSPISAICDVPEFRDDDLLCDPKNRDYIQAACEDFGKDL-NAMSLREIYNLLTEDYNNELLKYQCVEFLTNLVNVLDRRIPKLNAFLIISPPSGGKNFFFDMIFGQLLFAFQEAPKRVLLWNSLTDTIKMMFGGDPYTVRVKNRMDAHVKRTPVIILTNNYELAFSDRIIQY
>Junonia_coenia_densovirus__KC883978
AGKFAYIRKKTKALLRKYYVSPVSAICDVPEFRDDDLLCDPKNRDYIQAACDDFGKDL-NAMSLREIYNLLTEDYNIELLKFQCVEFLTNLVNVLDRRIPKLNAFLIISPPSAGKNFFFDMIFGQLLFAFQEAPKRVLLWNSLTDTIKMMFGGDPYTVRVKNRMDAHVKRTPVIILTNNYETAFSDRIIQY
>Pseudoplusia_includens_densovirus_JX645046
AGKFAYIRKKTKALLRKYYVSPVSAICDVPEFRDDDLLCDPKNRDYIQAACDDFGKDLNAMSLREIYNLLTEDYNFVELLKYQCVEFLTNLVNVLDRRIPKLNAFLILSPPSAGKNFFFDMIFGQLLFAFQEAPKRVLLWNSLTDTIKMMFGGDPYTVRVKNRMDAHVKRTPVIILTNNYETAFSDRIIQY
>Galleria_mellonella_densovirus_L32896
AGKFAYIRQKTKALLRKYYVSPISAICDVPEFRDDDLLCDPKNRDYIQAACDDFGKDL-NAMSLREIYDLLTEDYNDELLKYQCVEFLTNLVNVLDRRIPKLNAFLVMSPPSAGKNFFFDMIFGQLLFAFQEAPKRVLLWNSLTDTIKMMFGGDPYTVRVKNRMDAHVKRTPVIILTNNYETAFADRIIQY
>Mythimna_loreyi_densovirus_AY461507
AGKFAYIRQKTKALLRKYYVSPISAICDVPEFRDDDLLCDPKNRDYIQAACDDFGKDLNAMSLREIYDLLTEDCNFDELLKYQCVEFLTNLVNVLDRRVPKLNAFLVMSPPSAGKNFFFDMIFGQLLFAFQEAPKRVLLWNSLTDTIKMMFGGDPYTVRVKNRMDAHVKRTPVIILTNNYETAFADRIIQY
>Agrotis_segetum_GBCW01035751
----------MIMESLVNIHY-NIQRLTGL---TPQS----------SFNYASLLTNACFMDNQFEQWKLLAARQPMPTDVKYKNRQKVTNCILYTASNYTADAEEARTIT-
>Chilo_suppressalis_GAJS01023148
----------MATFSKRNKGIKIMCTKQYNEKIRTYIDIARVLVFQESIKQRFERAKKKNKIRFEDREDVSNEHEDKEIISNNNECFFRNTYEHFMKKTGKRNNLFFLGPPSTGKTMIMNSLVGLVSFNFSGLLRNACLMDNQFEQWKMLASGMPMSADEKG--PM----------
>Acromyrmex_echinatior_EGI68937
RKNVVEAITNLIEENEIWTYVEWELKINYDTKEELLKEFGLQTDTSSVPSDWWLPIQTSQNRRTFDKLCETAWGIDTFDFETSQQKFDATRFDALTRERKHRDSDILTGGRD--RRLSPNEVTKDVTVPWPKRKRRHRKQQSPLEALQRHIQRQALSVPPTSQDPASLGSIDIVIAAPTSDSKPSDRIDGL
>Acromyrmex_echinatior_F4WA07
RKNVVEAITNLIEENEIWTYVEWELKINYDTKEELLKEFGLQTDTSSVPSDWWLPIQTSQNRRTFDKLCETAWGIDTFDFETSQQKFDATRFDALTRERKHRDSDILTGGRD--RRLSPNEVTKDVTVPWPKRKRRHRKQQSPLEALQRHIQRQALSVPPTSQDPASLGSIDIVIAAPTSDSKPSDRIDGL
>Acromyrmex_echinatior_EGI69346
TTEVIEAITQSRQQLESRMEQHVQIGWKKNRVYIILGHLPPESESEEARVSRAAPAKPSGSKAWATIIARRKRKQKTELLEGDEEPTIGHASSIKKRTEKDESTEKEDNPPLE-RTSSSETTTSGTKIGLKDFDKAINELANYEECIQYIEQRREERASGTEKRKNVVEAITSLIEENDKEELLKE-----
>Acromyrmex_echinatior_F4W8Y7
TTEVIEAITQSRQQLESRMEQHVQIGWKKNRVYIILGHLPPESESEEARVSRAAPAKPSGSKAWATIIARRKRKQKTELLEGDEEPTIGHASSIKKRTEKDESTEKEDNPPLE-RTSSSETTTSGTKIGLKDFDKAINELANYEECIQYIEQRREERASGTEKRKNVVEAITSLIEENDKEELLKE-----

```
>Acromyrmex_echinatior_XP_011068621
GSPGTNITTGERKRGSESIARSTSETEWQQSVGDDHRKKKKKTKTNGDGTIGRRRGNSPSSNRTYHWTCILHKKNQKRRIDRKGIVFGATKDNNLVQRGHIAKHMCATGHTE--INITCQHLRGTFKIGLKDFDKAINELANYEECIQYIEQRREERASGTEKRKNVVEAITSLIEENDELKINYDELLKE
>Atta_cephalotes_W4VWX9
RKNVIEVITQLIKKIQNDNLRRLGIKNRFPNERTIIERTQNNTKQKRNNNTSIYKYVSYEEITYRDLTNKYAKILRYEIFEKNRIEFYAWLILIKNMDLVKINGMVLEGPTNAGKSMIIDCTKPVYPITLDNKDQRIKIIT-HEEEYTQIAGA---------ID--RIKRAPTYVTPYDKRDKIEKKLREK
>Atta_cephalotes_W4XSW1
KNRVYITIGRLAEGREERAQPLMETNEHARVARATTTIVAKPGGSKAWRIVQARRKRKREEHQDRVGDTNIRSRRPAHHIRSNAGRIAKHMRAITA-ETAEIAISCQRIKYL--RNFILY---GVETFHYTGAKPELKIIDKKKKERRKNGTTDTHETCVQYIEHRREERATAARLGSDKRKNVVE-----
>Atta_cephalotes_XP_012064512
RKNVVEAITHLIKKYKITTYADXELKIDFQTKEQLLKEFGLQTDTYAHKIIRNKRNNNTSMYKYVSYEEITYRDLTYEIFEKNKVEFYAWLISIKNMDLVKINGMVLEGPMNAGKSMIIDITAGAITYVYDFTTIYLIERDKYQENPERLKTSEKTTIDLEIEMGHATKRAKFTDDGEDLDAANRTTLKRQ
>Hymenolepis_microstoma_CDS26524
-----------------------------------------------------------------------------------------------MEQRHMYLNLL------EELRSR---EVRSFDFQKFT-------EEITAFNAYMGIQWCDVDMIR----ALNCER-------EAKERAN-----
>Hymenolepis_microstoma_CDS33099
GNVVIETELEVTLHSVPDNIAPEDKEVYINDNSIDLLHYMGAEGKQFKWEEPEFKQEEPEIEKVKPEEPEIKKIKKDSIFVDENYGFLGGGLAGTV-VGGIVNALTTGFTTT--SNAVQQTEI---PFPANIADDEIVSIEPLINRADADENGRSEAITSARSSPSFLGHDPEIFPSTQQLGDMET-----
>Thaumetopoea_pityocampa_GBZB01000242
QPNQHQELIFGKSYHFTLTNGLADFRHVLSSPSGYYLAHGIPWERLLMYISEGELLRLYRDYTALKVEEVICEIYSQTAYNGNFTEWSENFPNISASTSRDISNPVIEVEYPI--PYGFSNDVGGTTAYGWEKRFNGIIYAENYMTPIPGLENG-YFMTTITERNDAQIRPIPKALSATKQKCAPQAFMIGF
>Thaumetopoea_pityocampa_GBYZ01001450
QPNQHQELIFGKSYHFTLTNGLADFRHVLSSPSGYYLAHGIPWERLLMYISEGELLRLYRDYTALKVEEVICEIYSQTAYNGNFTEWSENFPNISASTSRDISNPVIEVEYPI--PYGFSNDVGGTTAYGWEKRFNGIIYAENYMTPIPGLENG-YFMTTITERNDAQIRPIPKALSATKQKCAPQAFMIGF
>Porcine_bocavirus_5_HQ223038
MSKRESLVLDCMNRAIESDCLTYEQLVDKHPELIIMMESQAGGNRLIEQTLNMVHIKLTQKYTAMTYLVKRFPTFEIRLLNFQGWQVGHWLCCVLDKKSGKQNTVSFYGPASTGKTNLAKAIVGNVNFVSNDCSKLIVWWEDWVEPAKCVLGGTTVRVDRKHKDSQLLPQTPCIISTNNHCKPLKDRVVQL
>Human_bocavirus_1_JQ923422
PNKKEKLMLNLLDKCKELNLLVYEDLVANCPELLLMLEGQPGGARLIEQVLGMHHINVCSNFTALTYLFHLHPVTSLQLLLIQGLAVGHALCCVLNKQFGKQNTVCFYGPASTGKTNMAKAIVGCVGFVFNDCRRLVVWWEDWVEPAKCILGGTECRIDVKHRDSVLLTQTPVIISTNHHAAPLKERVIQL
>Human_bocavirus_4_FJ973561
ANKKEKLMLNLLDKCDELNLLVYEDLVSACPDLLLMLEGQPGGARLIEQVLGMHHIKVCAKHTALSFLFHLHPDQLLKLLLIQGLQVGHAICCVLNKQMGKQNTICFYGPASTGKTNFAKAIVGCVGFVFNDCRRLIIWWEDWVEPAKCILGGTECRIDVKHKDSVLLQQTPVIISTNHHAAPLKERVLQL
>Porcine_bocavirus_1_HM053693
MTKRGGLMLDCMRRCFTENITSYEEMVRECPDLTIMIESQAGGSRLIEQVMGMTHIKIVQSYTALGYVQKLFGEKAFRLLNFQGWQAGHWVCCVLDKKSGKQNTLCFYGPASTGKTNLAKSIVGCVNFVFNDCAKLIVWWEDWVEQAKCILGGTEFRIDRKNRDSMLLPKTPVVISSNNHSRPLRERIVQW
>Porcine_bocavirus_3_JF429834
LTRKQHVMLDTLQRCQEEFICTKEELTMLHPDLVIMFESTPSGSRTLEEVLEMHRVRITRTETALSYIRRQFPDSKVRLLNIQGIQVGHWVATVLSKKAGKQNTICFFGPASTGKTNLAKAIAGCVSFVFNDCQKLICWWEDWVEPAKCLMGGTSFRVDRKHKDSAEQPHTPLILSTNHHEKPIRDRVVQF
>Lamellibrachia_sp._x5
-----------------------------------------CQIPFTQITKFPGSWANYLSQKPRITWRVPICQMME--------WVFADYIERELV-PERKVDDAVA-GP----SNVSAN------DFQSGKSMYKYLKWLFGHDNRESLIKG-------------ALAAA--------DTSTFEK-----
>Messor_concolor_x
-----------------------------------------------------------------------------------------------RITNFLHTTPAGSAE--ITVTCQLLQGLKKFHYGTKIQELKEIDNLTQELLKECDE---------NDEIDLG----------ECKKYIE-----
>Hymenolepis_microstoma_CDS27819
RADDEPVPEEEVLRRETRDAKTVGGWEN-------------------KTTFMSNPTMPFGKLLTLDYTGH------SVVFRKRYMVTLPGLHKEIP-STENATAFVLN------TTFLESFVC---RYRFKSNKTVSTKWNARIKIIRLIPSTKQATGGVVSVDKVQFDNASYMLVGISNQIQPLQKMSTY
>Eidolon_helvum_parvovirus_1_JQ037753
EGKSAERFLNLIDWLVDQGIATEKKWLEADRDGYRSFLSSSGGVLQAKNALNIARREMVLSQPLVRYLAKGGAWDECDIFRLNGPSLVARYMACWAGHWPKRRALWLWGPASTGKTVIAAAIAGCVNFPFNDCHQPLVWWENIVEVAKAILGGSPVRLDVKNKGSEDFLPTAVIITSNGHQDALRTRMCMI
>Human_parvovirus_4_AY622943
AGKGTQRFMDLIDWLVENGTATEKRWLSVNKLSYRSFLGSSGGVLQATNALQIAKREMVLAHPLLSYLTKNASAFEAQLFSLNGWYFAAWARGVW----PKRRAIWLWGPASTGKTLLAAAIAGCVNFPFNDCHQSLVWWENIVEVAKAVLGGAPVRLDVKNKGSEDYIPTCVIITSNGGLAPLPERSFFK
>Bovine_hokovirus_EU200669
GGKGTDRFLSLIDWLVEQGIATERRWLSADKRSYRSFLGSSGGVLQARNALLVARREMVLAHPLLGYLQRGGSAVAIELFKLNAWYFAAWAQGIW----PKRRALWLWGPASTGKTLLAAAIAGCVNFPFNDCHQSLVWWENIVEVAKAILGGAPVRLDVKNKGSEDFIPTCVIITSNGHQEALQTRITMF
>Porcine_hokovirus_EU200677
AGKGTERFLNLIDWLVEEGIATERRWLATDKKSYRSFLGSSGGVLQARNALQVARREMVLARPLLGYLRRGGSDIATELFRINGWYFAAWAAGAWA----KRRAMWLWGPASTGKTLLAAAIAGCVNFPFNDCHQSLVWWENIVEVAKAILGGAPVRLDVKNKGSEDFIPTCVIITSNGHQEALQTRITMF
>Porcine_parvovirus_4_GQ387499
KCATGEKMLDIVQWLVDNNICSESRWENKNALSLYSFLATQAGGYMAKQCLRIAQQKLLKEKPLGLTLMDFKDMNAHYIFAINNPKIASVIMYYFWSKQTGKRNCVWFYGPATTGKTNMAQAICGNVNFPFQDIAAQVGWWEDMVEAAKALLGGTALRIDRKCMQSIEVNSPPFLITSNVHQQPLEDRMIKF
>Oryctolagus_cuniculus_XP_008259747
RTKISQKYMAVVNWHVENGTTSGKPWMQENSKSYLSFSSTKNARSQIRSALDNFTKVMGLTIMVLDCLVGVLALPDHQILQMSGRACAGSVWAAWCCQYGRRNTTWLFGPVTTRRSSIAEATAGCAIFPFRGCVQMTIWWETVVDAVKATPRGGKVRADQICKSSAQVEPTPAVITGNAHQQRLQDQRSTF
>Oryctolagus_cuniculus_GBCK01348906
RTKISQKYMAVVNWHVENGTTSGKPWMQENSKSYLSFSSTKNARSQIRSALDNFTKVMGLTIMVLDCLVGVLALPDHQILQMSGRACAGSVWAAWCCQYGRRNTTWLFGPVTTRRSSIAEATAGCAIFPFRGCVQMTIWWETVVDAVKATPRGGKVRADQICKSSAQVEPTPAVITGNAHQQRLQDQRSTF
>Goose_parvovirus__GPU25749
SNRAAKNYSNLVDWLIEMGITSEKQWLTENRESYRSFQATSSNNRQVKAALENARAEMLLTKTATDYLIGKDPVLDYQILKMNNPQYIGSILCGWVREFNKRNAIWLYGPATTGKTNIAEAIAGCVNFPFNDCVKMLIWWEKVVESAKAILGGSAVRVDQKCKGSVCIEPTPVIITSNTHRIPLEEQIVLS
>California_sea_lion_adeno-associated_virus_1_JN420372
RNRTSQKYMALVSWLVENGITSEKQWIQEDQESYLSFNAAGSSRSQIKSALDNASRIMSLTKKASDYLVGQSVPEDYQLFKMNGPAYLGSILLGWCGRFGKRNTVWLYGPATTGKTNLAEAIAGCVNFPFNDCVKMLIWWEKVVESAKAILGGSKVRVDQKCKSSVQIDSTPVIITSNTHRQPLEDRINLE
>Adeno-associated_virus_5_AF085716
KSKTSQKYMALVNWLVEHGITSEKQWIQENQESYLSFNSTGNSRSQIKAALDNATKIMSLTKSAVDYLVGSSVPEDWQIFEMNGPAYAGSILYGWCRSFNKRNTVWLYGPATTGKTNIAEAIAGCVNFPFNDCVKMLIWWEKVVESAKAILGGSKVRVDQKCKSSVQIDSTPVIVTSNTHQQPLEDRMFKF
>Adeno-associated_virus_1_AF063497
RSKTSARYMELVGWLVDRGITSEKQWIQEDQASYISFNAASNSRSQIKAALDNAGKIMALTKSAPDYLVGPAPPADYRILELNGPAYAGSVFLGWAKRFGKRNTIWLFGPATTGKTNIAEAIAGCVNFPFNDCVKMVIWWEKVVESAKAILGGSKVRVDQKCKSSAQIDPTPVIVTSNTHQQPLQDKFELT
>Octodon_degus_XP_004634735
RTKRSQRYMDLVKWLVEKGITSEKQWMEEDEESFLSFNATGNTRSQIRAALDNTRQKLIWTKTAKDYLVGSSAPPEYQIFDMNGPLYAGSILVGWCHQFGKRNTIWLFGPATTGKTNMAEAIAGCVNFPFNDCLKMIIWWEKVVESAKAILGGSKVRVDQKGKSSAEVAQTPVIMTSNTPEEHLQPRMFKF
>Bat_adeno-associated_virus_GU226971
STKSASEYMALVRWLVDNGIATEREWMREDSDGYLSYNATGATRAQIKAALDNAARIMVNTKTAADYLVGRNPPLDYRLFRMNGPAYAGSVLLGWCTGFGKRNTVWLFGPATTGKTNLAEAISGCVNFPFNDCVKMIIWWEKVVESAKAILGGSKVRVDQKCKNSQQIEPTPVIITSNTHRQPLEDRMFKF
>Chinchilla_lanigera_GBDF01150844
----------LVQWLVRIGMTSEQAWARERDCDSYVSFHASKDSRGQVKAALRNARTLCLTKTAADYLVAERAPDDIYRLFSLNSPAYAGSLLLGWCREFGKRNTVWLFGPPSTGKTNLAEAIAGCVSFPFRDCVKMLIWWERMLESAKAVLGGSAVRVPQKREVSSQIEPTPVIITSSQHRKALEERMFKL
>Eunicella_cavolinii_x
GRSPRCILKLCLDHGMADTDELLQNLEDLEPNLNIDKFFHPTRARMEKQIDQQLKFKYVNMSWMQVMETVWGKRQGYKWIRANNEQFICDVFALMYKLHDKKNAMLLIGPSNSCKSKILHSITSMIEFPFERCQVAVIQHEEVVDLFKRVAEGAVVDCNKK-GTSFVLKRVPNIWCSNHDSVAMLNRVF--
>Latrodectus_hesperus_GBCS01014800
-----------------------------------------MNLVANSIPSAARTVETALNLVGRPDLAATAHS--------SSAIQGVQLIQE-SITSIPAFLTE------ATNLLS---GST---TTSEPDAAAARTGQGPSIQNLQSSTNPQLPFDS----AVNTDP-------SIPADQE-----
>Ascaris_suum_JI174103
SGRYCRAIWEAIQEILPRSITDISNHLYRSANIEMVID--ERFNAVASKLFDAFRENYLNKSWKDIMNGIPENYFDEKLVNFNGKVIIEEIYNVMDTIHMKKNCILFKGPPNSGKTLLANSIVQQMPFEFQSMMQRVILINITIETIKNICEGQAVSIDIKYMSNQVLPRTPLIITTNATNRAVNELTRSY
>Bovine_parvovirus_2_AF406966
RQASGDRMRDLVEWCIKNHVFTEDQYMAKFEESYYSFACTNQGRHMLQTSLELAAKRTTSMIPLGVYLAGFSNIQEFDLLQFQGPVVAGYIIYAWSRATGRRGALWFYGPGQTGKSIMAR---GCVNFPFQDLACQIGWWEDIVESAKALLSGGKIRVDRKCRDSVEITPPPFVITSNNHKKPLEDRMIKF
>Echinaster_spinulosus_GAVE01029446
SQSSNALMRSVMRHEDDRTKEIVKSMF----------YKQNWNTLFNKVLGLVKIHNCDMNMAQMLDMAERERPSYKKWCAHQGKELYQSVQCILDKVLAKKNTLAFIGESNAGKSWVVETLCGEIGYAWQECPKRLIVFNAVIEKCKVVFEGSDCFVSVKQQGDAFLTRTPCIITSNNEETAIRNRCILY
>Luidia_clathrata_GAVQ01079362
-----------------------------------------MTLDELSDRYRTASHWDEPNRYESLDRSCALVKRWC--------HTFVDDLYSIID-RLPKRNCLALVGQSNAGKTWLIN---------LAEVC-------CHVGEITQGTSS-----------------------------------
>Turkey_parvovirus_-_GU214706
TDWGNGGGETVIKTSIIEKLCMEALRLCRENHIFTMKAFKLKFPDKFMQFSSRNQGLVKLEETINLYCETIINEHNIRLCSYQGPKYVARLILCWLGQAGKKNALYFHGPANTGKTMMAESICGNVNFPFNDCHKAVLWWEEHVESAKCIMGGSSVRIDKKNQDSVLLCKTPIVITSNNHAAAIRARCLKF
>Schmidtea_mediterranea_GAKN01010353
-----------------------------------------------------------------------------------------------RGGKINTIYLQG-NILYDNLFYSIGPSNTRKTFFVQLLTGCITFWTQDVDKDVVVAEENVDYFKLMLEGNMIRIEIKNKPAVNWKRIPVIVTSNQHQEALQNRMLVI
>Rhipicephalus_microplus_CK189748
-----------------------------------------------------------------------------------------------MHSEGRCLHSEGCPG-------------VFARGDTI-----------SMQLFGGPPSRP--------GCTSCL-------QRPILPR-----
>Rhipicephalus_microplus_CK189747
-----------------------------------------------------------------------------------------------MHSEGRCLHSEGCPG-------------VFARGDTI-----------SMQLFGGPPSRP--------GCTSCL-------QRPILPR-----
>Hymenolepis_microstoma_CDS26229
------------MMVHDGYGQLDDTRVKHFPYFLYLASTNGTRFNCRKYHNWGHAEPELLCVAVGNPAICVNGQS--------TDGGAMAWNLAH-QRHGSNELLLC---------------ECAFFPFVCAIQRVSGYE---DVLDNIAGGGKVYAWRSS----GVLGAP--VKARK-----------
>Hymenolepis_diminuta_LM393614
--MFELFQTLNIKSEEEFMQWIKFNPEQLPELWSRYPNWEKKLQQYVSVMIKTKSGLFDEPFHIKPYLVHVQHATCYSIFTNNDGLFLAEVDKIIRMTYPRINALVLRGPTSTGKTLIAKNIVETVAFYLQNLLHDVALMETTVQNFKELFAGSPLIVQVKNHAPRELKRIPCIVTTNQESEPIRRRIIE-
>Hymenolepis_diminuta_LM395168
------------------------MKVIVDMPLLERKFALLQEENIYKDICKFVCILSDPDNIHDNDSLAGAKYLYSNFTNNGGLFIAEVDKIIRMSYPRINALVLRGPTSTGKTLIAKNIVGTVAFYLQNLLHDVALMETTVQNFKELFAGSPLIVQVKNHAPRELKRIPCIITTNQESEPIKKRIIE-
>Taenia_multiceps_JR938527
-----------------------------------------------------------------------------------------------MTPFINGFSVLVMPY--RHISLSLIQ---HFQFDNLLDAGALMETTKDDFKLLLGGIGLEVDVKYSGHKWLERVLVIQATNENRKCFTN-----
>Hymenolepis_nana_LM406075
RMHKYFNLLEELRIRQVRTFDEVFHKFNVDEITFMNSSIGIQWGEVARQQIRTLSNERLAKEKTNPYLQNLHELPHIAMLNTNSGALLGDIVQIMDKRKAKINTLCFRGQTNTGKTLLANLLT---SFHFFVFAKTLILVLALFNRL------ERCAD--------VVIRPPFISTTCS-----------
>Hymenolepis_nana_LM409198
RMHKYFNLLEELRIRQVRTFDEVFHKFNVDEITLMNSSIGIQWGEVARQQIRALNNERLAKEKTNPYLQNLHELPHIAMLNANSGALLGDIVQIMDKRKTKINTLCFRGQTNTGKTLLANLVT---KYP-----SHFFVFATLILVLALFYRLERCAV--------VVIRPPFISTT-----------
>Hymenolepis_microstoma_CDS26228
-----------------------------------------------MPPPSNVRLKLPSWLAAENAK----------------------------PDGELPVKEFH------QRYYLT------LIVYTNVPAISCLYVEHDDHMHVAFQA-SSNVKVER----LLNDCHVPLRTGSSAKVLSN-----
>Hymenolepis_diminuta_LM389526
VNNPVKMIQYFKFRNTPVIIGAELRDIFEIARSMPPLETAEQCYGIQLRTQTKAALNNKVDNRVSKYESLYEELKRYTIFTVHEMHFNANIGLQWR-EIAKQAIQAIN------EDLLLE---QTRTYLLQEK-RHDCYADRGVDWLEFMFNMNHIPVLLTD----IIKIMDMKVNKVNCLCFYGQ-----
>Hymenolepis_microstoma_CDS29136
YTPYRYYYTLVLRNYGKVPTSISCLYVEHEQHMHVLLQAKDNISRKVDRLLNDCLVPADNWWMIKNTRQLVRNVYLIQYFKYRGFEVGTELHEEWQADMPPIQPHVTCRRART-KSTIAE---RTEKFTLDELKSDIRQFEFSVNEIWINMDKKTTKVIAKQ----QIQSLNAERLQSEKQNSYLQ-----
>Hymenolepis_microstoma_CDS30830
------------------MYSVMHEADYLSSRIELIVS-----------KWLPNSALHGSVSSVDYLRLSFAD--RACFRTHC-------------KWNLLTSCLLWGEGE--EEVGGE---GIVAFHFENLLRTVALIEETKNDYKCLLGGERLEIDIKCGARRFLQRVPVIVTTNESQLTSLA-----
>Hymenolepis_microstoma_CDJ12115
------------------MYSVMHEADYLSSRIELIVS-----------KWLPNSALHGSVSSVDYLRLSFAD--RACFRTHC-------------KWNLLTSCLLWGEGE--EEVGGE---GIVAFHFENLLRTVALIEETKNDYKCLLGGERLEIDIKCGARRFLQRVPVIVTTNESQLTSLA-----
>Hymenolepis_microstoma_CDS30832
-------MNELTCYQPVRGDQENRVPNNEVPHKYYYTVVLRKYSKVPTSISCLYVEHDDHMHVLLHAKTNVSRKVNDCLIPPDDRQLVRNVYRIIQFKHQVLGIVCQRGEQK-------------AFHFENLLRTVALIEETKNDYKCLLGGERLEVDIKCGARRFLQRVPVIVTTNEDLG---------
>Hymenolepis_microstoma_CDJ12117
-------MNELTCYQPVRGDQENRVPNNEVPHKYYYTVVLRKYSKVPTSISCLYVEHDDHMHVLLHAKTNVSRKVNDCLIPPDDRQLVRNVYRIIQFKHQVLGIVCQRGEQK-------------AFHFENLLRTVALIEETKNDYKCLLGGERLEVDIKCGARRFLQRVPVIVTTNEDLG---------
>Hymenolepis_microstoma_CDS33272
-----------------------------------------MAKQQIPSLNTERLQVEQASSYPNNLRNLKHDCTMMMLNQSGNELLTDIINIIYTKTTKMDTFCFKGQTNTGKTLLANLIAGPVAFHFDNLPRTVALMGITKNDYVSLEEV-----DLKS----------M-----------------
>Hymenolepis_microstoma_CDJ14557
-----------------------------------------MAKQQIPSLNTERLQVEQASSYPNNLRNLKHDCTMMMLNQSGNELLTDIINIIYTKTTKMDTFCFKGQTNTGKTLLANLIAGPVAFHFDNLPRTVALMGITKNDYVSLEEV-----DLKS----------M-----------------
>Hymenolepis_microstoma_CDS35108
GVVSGSLMEYLSGSGEKLTISICGAAVQEITKFNSS--IGVQLREIAKQQMQSPNTQRLKDEQTKSHLENRRGMKHMMMLSQHGDELLKDIAYIMDKKTTKVNSLCFKGQTNTGKKLLANPHLGTVAFHFDNLLRTVTLMETTKSAYKCLLGGDRFEIDVNYGARRFLQRIPVVDTTNEDRAALYSRVKQY
>Hymenolepis_microstoma_CDI96179
GVVSGSLMEYLSGSGEKLTISICGAAVQEITKFNSS--IGVQLREIAKQQMQSPNTQRLKDEQTKSHLENRRGMKHMMMLSQHGDELLKDIAYIMDKKTTKVNSLCFKGQTNTGKKLLANPHLGTVAFHFDNLLRTVTLMETTKSAYKCLLGGDRFEIDVNYGARRFLQRIPVVDTTNEDRAALYSRVKQY
>Hymenolepis_microstoma_CDS34279
---------------------------------MQSPNTQRLKDEQTKSHLENRRGMKHEYSVGNTR---------MMMLSQHGDELLKDIAYIMDKKTTKVNSLCFKGQTNTGKKLLANFITGTVAFHFDNLLRTVTLMETTKSAYKCLLGGDRFEIDVNYGARRFLQRIPVVDTTNEDRAALYSRVKQY
>Hymenolepis_microstoma_CDJ15564
---------------------------------MQSPNTQRLKDEQTKSHLENRRGMKHEYSVGNTR---------MMMLSQHGDELLKDIAYIMDKKTTKVNSLCFKGQTNTGKKLLANFITGTVAFHFDNLLRTVTLMETTKSAYKCLLGGDRFEIDVNYGARRFLQRIPVVDTTNEDRAALYSRVKQY
>Hymenolepis_diminuta_LM391469
VKNVYLLLQYFKFRGEIHKVGQELSEEWLLAQELPAVEPRQPCFSDLRRTQTKSAATVRLESRHAKFTNLLEELKREDIFSKFSVKEITDLNTSMG-QWREIAKQQIQGMNA--ERLKME---KQCNYLLENLRHDCAVSNDLGANWLMMLLNNGINILLADKKTTKVNTLCFKGQTNTGKTLLAN-----
>Hymenolepis_microstoma_CDS30831
RAEIPGTTSGAGYERPSSPSFIGHDPFGIEEGQQLLPSTQRLGDVGSRTQNTLSSGMAEKDSSVDNILPTERIPQSTLMLNQNGYELLNDIIAIMEKKTTKANTLCFKSQTNTGKTLLANLITGTVAFHFDNLLRTAALMETTKNDYKCLFGGDRFEIDVKYGARRFLQRIPVVDTTNEDRAALYSRVKQ-
>Hymenolepis_microstoma_CDJ12116
RAEIPGTTSGAGYERPSSPSFIGHDPFGIEEGQQLLPSTQRLGDVGSRTQNTLSSGMAEKDSSVDNILPTERIPQSTLMLNQNGYELLNDIIAIMEKKTTKANTLCFKSQTNTGKTLLANLITGTVAFHFDNLLRTAALMETTKNDYKCLFGGDRFEIDVKYGARRFLQRIPVVDTTNEDRAALYSRVKQ-
>Hymenolepis_microstoma_CDS30612
-----------------------------------------------------------------------------------MMLLNQNGNELLTDIINIMDKRSTKMNTLCFKGQTNAGKTLCANLITGTVAFHFDNLLRTAALMEITLEEVDLKAMSDRAALGVKQNEPSSMGCPP-------APCPLAN-----
>Hymenolepis_microstoma_CDJ11897
-----------------------------------------------------------------------------------MMLLNQNGNELLTDIINIMDKRSTKMNTLCFKGQTNAGKTLCANLITGTVAFHFDNLLRTAALMEITLEEVDLKAMSDRAALGVKQNEPSSMGCPP-------APCPLAN-----
>Hymenolepis_microstoma_CDJ13509
```

```
>Echinococcus_multilocularis_U6H9C9
-------------------------------------------------------------------------------MNC-------------EDEKINTLVLYGPTNTGKSLICKLMTGSVAFAYENLLRKVALMEANQQDLKQILGGEPFEVHIKYQNPDLLERLPVVVTTNEDAAATEGRCKI-
>Echinococcus_multilocularis_U6HFK7
-------------------------------------------------------------------------------------------------------------------------------------------MEANQQDLKQILGGEPFEVHTKYQNPDLLERLPVVVTTNEDAAATEGRCKI-
>Echinococcus_multilocularis_CDI70562
-------------------------------------------------------------------------------M-----------------TSFLEH---GSVAFAYENLLRKVALMEANQQDLKQILGGEPFEVHTKYQNPDLLERLPVVVTTNEDAAATEGRCKI-
>Echinococcus_multilocularis_U6HBG0
KRQKYIDLAEEVMRRKVRSMNDMNKKFTYQETVRLMADYGQQSYNMIVRKALETVRMMNVAHQRATDYADLLKEELDEDMFSANAVDFAITLRIIMNCEDEKINTLVLYGPTNTGKSLICKLMTGSVAFAYENLLRKVALMEANQQDLKQILGGEPFEVHIKYQNPDLLERLPVVVTTNEDAAATEGRCKI-
>Echinococcus_multilocularis_U6H9K2
KRQKYIDLAEEVMRRKVRSMNDMNKKFTYQETVRLMADYGQQSYNMIVRKALETVRMMNVAHQRATDYADLLKEELDEDMFSANAVDFAITLRIIMNCEDEKINTLVLYGPTNTGKSLICKLMTGSVAFAYENLLRKVALMEANQQDLKQILGGEPFEVHIKYQNPDLLERLPVVVTTNEDAAATEGRCKI-
>Echinococcus_multilocularis_U6HGG3
KRQKYIDLAEEVMRRKVRSMNDMNKKFTYQETVRLMADYGQQSYNMIVRKALETVRMMNVAHQRATDYADLLKEELDEDMFSANAVDFAITLRIIMNCEDEKINTLVLYGPTNTGKSLICKLMTGSVAFAYENLLRKVALMEANQQDLKQILGGEPFEVHTKYQNPDLLERLPVVVTTNEDAAATEGRCKI-
>Echinococcus_multilocularis_U6FUW9
KRQKYIDLAEEVMRRKVRSMNDMNKKFTYQETVRLMADYGQQSYNMIVRKALETVRMMNVAHQRATDYADLLKEELDEDMFSANAVDFAITLRIIMNCEDEKINTLVLYGPTNTGKSLICKLMTGSVAFAYENLLRKVALMEANQQDLKQILGGEPFEVHTKYQNPDLLERLPVVVTTNEDAAATEGRCKI-
>Echinococcus_multilocularis_CDS35660
KRQKYIDLAEEVMRRKVRSMNDMNKKFTYQETVRLMADYGQQSYNMIVRKALETVRMMNVAHQRATDYADLLKEELDEDMFSANAVDFAITLRIIMNCEDEKINTLVLYGPTNTGKSLICKLMTGSVAFAYENLLRKVALMEANQQDLKQILGGEPFEVHTKYQNPDLLERLPVVVTTNEDAAATEGRCKI-
>Echinococcus_multilocularis_U6HA06
------------------MNDMNKKFTYQETVRLMADYGQQSYNMIVRKALETVRMMNVAHQRATDYADLLKEELDEDMFSANAVDFAITLRIIMNCEDEKINTLVLYGPTNTGKSLICKLMTGSVAFAYENLLRKVALMEANQQDLKQILGGEPFEVHTKYQNPDLLERLPVVVTTNEDAAATEGRCKI-
>Echinococcus_multilocularis_CDI70556
-----------------------------------------------MIVRKALETVRMMNVAHQRATDYADLLKEELDEDMFSANAVDFAITLRIIMNCEDEKINTLVLYGPTNTGKSLICKLMTGSVAFAYENLLRKVALMEANQQDLKQILGGEPFEVHTKYQNPDLLERLPVVVTTNEDAAATEGRCKI-
>Echinococcus_multilocularis_CDS36689
-----------------------------------------------MIVRKALETVRMMNVAHQRATDYADLLKEELDEDMFSANAVDFAITLRIIMNCEDEKINTLVLYGPTNTGKSLICKLMTGSVAFAYENLLRKVALMEANQQDLKQILGGEPFEVHTKYQNPDLLERLPVVVTTNEDAAATEGRCKI-
>Echinococcus_multilocularis_U6HD97
-------------------------------------------------------------------------------MFSANAVDFAITLRIIMNCEDEKINTLVLYGPTNTGKSLICKLMTGSVAFAYENLLRKVALMEANQQDLKQILGGEPFEVHIKYQNPDLLERLPVVVTTNEDAAATEGRCKI-
>Echinococcus_multilocularis_U6H9K6
-------------------------------------------------------------------------------MFSANAVDFAITLRIIMNCEDEKINTLVLYGPTNTGKSLICKLMTGSVAFAYENLLRKVALMEANQQDLKQILGGEPFEVHIKYQNPDLLERLPVVVTTNEDAAATEGRCKI-
>Echinococcus_multilocularis_U6HBN8
-----------------------------------------------MIVRKALETVRMMNVAHQRATDYADLLKEELDEDMFSANAVDFAITLRIIMNCEDEKINTLVLYGPTNTGKSLICKLMTGSVAFAYENLLRKVALMEANQQDLKQILGGEPFEVHTKYQNPDLLERLPVVVTTNEDAAATEGRCKI-
>Echinococcus_multilocularis_U6H9Z7
-------------------------------------------------------------------------------MFSANAVDFAITLRIIMNCEDEKINTLVLYGPTNTGKSLICKLMTGSVAFAYENLLRKVALMEANQQDLKQILGGEPFEVHTKYQNPDLLERLPVVVTTNEDAAATEGRCKS-
>Echinococcus_multilocularis_CDI70288
KRQKYIDLAEEVMRRKVRSMNDMNKKFTYQETVRLMADYGQSYNMIVRKALETVRMMNVAHQRATDYADLLKEELDEDMFSANAVDFAITLRIIMNCEDEKINTLVLYGPTNTGKSLICKLMTGSVAFAYENLLRKVALMEANQQDLKQILGGEPFEVHTKYQNPDLLERLPVVVTTNEDAAATEGRCKS-
>Echinococcus_multilocularis_CDS35386
KRQKYIDLAEEVMRRKVRSMNDMNKKFTYQETVRLMADYGQSYNMIVRKALETVRMMNVAHQRATDYADLLKEELDEDMFSANAVDFAITLRIIMNCEDEKINTLVLYGPTNTGKSLICKLMTGSVAFAYENLLRKVALMEANQQDLKQILGGEPFEVHTKYQNPDLLERLPVVVTTNEDAAATEGRCKS-
>Echinococcus_multilocularis_CDI70555
KRQKYIDLAEEVMRRKVRSMNDMNKKFTYQETVRLMADYGQSYNMIVRKALETVRMMNVAHQRATDYADLLKEELDEDMFSANAVDFAITLRIIMNCEDEKINTLVLYGPTNTGKSLICKLMTGSVAFAYENLLRKVALMEANQQDLKQILGGEPFE--------------------------------
>Echinococcus_multilocularis_CDS35653
KRQKYIDLAEEVMRRKVRSMNDMNKKFTYQETVRLMADYGQSYNMIVRKALETVRMMNVAHQRATDYADLLKEELDEDMFSANAVDFAITLRIIMNCEDEKINTLVLYGPTNTGKSLICKLMTGSVAFAYENLLRKVALMEANQQDLKQILGGEPFE--------------------------------
>Echinococcus_multilocularis_U6H948
-------------------------------------------------------------------------------DFAITLRIIMNCEDEKINTLVLYGPTNTGKSLICKLMTGSVAFAYENLLRKVALMEANQQDLKQILGGEPFEVHIKYQNPDLLERLPVVVTTNEDAAATEGRCKI-
>Hydatigera_taeniaeformis_LL723621
-------------------------------------------------------------------------------------------------------------------M---------------------RTVAVLEPNVNDMKQLFGGEPFEINVKYKPMQYLNRLPVIVTTNEDCAALESRYCQF
>Hydatigera_taeniaeformis_LL725419
-----------------------------------------------MLCREEIMALNGKPRHDCLGRPLNKEGMDWI-NNLFHSNGHAFCRKMECIMDKKDDKVNTLVLYGPTNTGKSLLCKLMTGTIAFAFENLLRSVGILEANMNDMKQLLGGEEFEVAVKY-----------------KPMQFLK-----
>Hymenolepis_diminuta_LM390786
RIEKTYDLARTLLDRRITHPAQLHEKFSLEEMAQLAVDFGNSYKETVSMIIANLRQSRLKEEKEKPYCELLNRELCDNLFDANGSEFINKLDRVMNRNDSKINSIVLYGPTNTGKSLLCKIMTGTINFAYENLLRSVAILEANVNDMKQLLGGESFEVAVKYKPMQFLHRIPVIITTNEDAAALESRYFQ-
>Hymenolepis_microstoma_CDS32658
RIEKTYDLAKTLLERRITHPSQLNEKFSLDEMAQLMLDFGNSYKETVSMVLANLRQSRLKEERTKSYCELLNRELLECLFKSNGLEFVSRLERVMNKVDNKVNCIVLYGPTNTGKSLLCKIMTGTINFAFENLLRSVAILEANANDMKQLLGGEAFEVSVKYKPMQYLQRLPVIVTTNEDAAALESRYYQ-
>Hymenolepis_microstoma_CDJ13943
RIEKTYDLAKTLLERRITHPSQLNEKFSLDEMAQLMLDFGNSYKETVSMVLANLRQSRLKEERTKSYCELLNRELLECLFKSNGLEFVSRLERVMNKVDNKVNCIVLYGPTNTGKSLLCKIMTGTINFAFENLLRSVAILEANANDMKQLLGGEAFEVSVKYKPMQYLQRLPVIVTTNEDAAALESRYYQ-
>Hymenolepis_diminuta_LM389152
RIDKTYDLARTLLDRRITHPSQLNEHFSLEEMTQLMVDFGNSYKETVSMVLANLRQARLKEERELPYKRLLERELLNQIFHANGFEFCRQVERVLNRQDNKVNTIVLYGPTNTGKSLLCKIMTGTIAFAFENLLRSVAILEGNANDMKQLLGGESFEVAVKYKPMQYLLRLPVIITTN------------
>Hymenolepis_diminuta_LM388960
-----------------------------------------------MALAGKSRHDCSGFPFTAEVRSWL-DCLFQSNGLEFCNQLECVMNKRDNKVNTIVLYGPTNIGKSLVCKIATGTIAFAFENLLRSVAILEGNANEMKQLLGGESFEVSVKYKPMQYLPRLPVIITTNEDAAALESRYHQ-
>Hymenolepis_microstoma_CDJ13946
-------------------------------------------------------------------------------MQYLQRLPVIITNEYLGCRLPDVDAAAG--KSLLCKVMTGTINFAFENLLRSVAILESNANDMKQLLGGESFEVAVKYKPMQFLPRLPVIITTNEDAAALES-----
>Hymenolepis_microstoma_CDS32661
-------------------------------------------------------------------------------MQYLQRLPVIITNEYLGCRLPDVDAAAG--KSLLCKVMTGTINFAFENLLRSVAILESNANDMKQLLGGESFEVAVKYKPMQFLPRLPVIITTNEDAAALES-----
>Hymenolepis_diminuta_LM389639
-----------------------------MSHCYFSDDKRLLEREYLMALSGKPRHDCSREPFIPRIEKWL-ETLFRSNGIEFCKMVEKVLDKKDNKVNTIVLYGPTNTGKSLLCKLMTGTIVFAFENLLRTVAILEANANDMKQLLGGEAFEVAVMYKPMQYLNRLPVVITTNEDAAALESRYYQ-
>Hymenolepis_diminuta_LM390089
RIDKTYDLARMLLDRGVTHAAQLNEKFSLEEMTQLMLDFGNSYKETVSMVLGNLRHARLNEERSMSYKRLLDREYLDTLFRSNGVEFCKTVEKVLDKKDNKVNTIVLYGPTNTGKSLLCKLMTGTIVFAFENLLRTVAILEANANDMKQLLGGEAFEVAVKYKPMQYLNRLPVVITTNEDAAALESRYYQ-
>Dicrocoelium_dendriticum_LK428994
RTKRTNLIDERIEVHRCTHLKELKRCLTRKQRLTIYNEFGKQWEESAIMCIEAFNEELFERQTHVSFEEYMLENNHDNLLAANKDKFCKAIHEIMNKKIDRLNCLCIEGPTTTGKSLVLKMICGTVQFFLQNLVKSIALMEATVNDFKELLGGSPFDIHVKHSPDVTLNRLPVLISTNHDAAAIYERCHT-
>Dicrocoelium_dendriticum_LK459200
----------------------MKRCLTRKQRLSIYNEFGMQWEESAKMCIEAFNEELFERQTNISFEEYMQENNHDNLLAANKDKFCKAIHEVMNKKIDRLNTLCIEGPTTTGKSLVLKMICGTVQFFLQNLVKSVALMEATVNDFKELLGGSPFDIHVKHSPDVTLSRLPVLISTNHDAAAIY------
>Schistosoma_mattheei_LM181082
-------------------------------------------------------------------------------------------QKFGINTLCIEGPTTTGKTLILKLITGTVKFFLQNLNKSVALMEVNVNDFKQLLGGSSLDIHIKHQSDVTLQRIPVLISTNSDKKAIYKRCITY
>Schistosoma_margrebowiei_LL888002
RVQCGEHLMDLVTLYNVTHLQQLKMKLSAKDRLNLYYEYGKQWEETAQLCIDVYNETNIKQQMCVPWVTWVSTNHHETFFAANEKQFLNDIQTIIDKTKLRVNTFCLEGPTTTGKTLILKLIVGTVQFFLQNLNKSVALMEITVNDFKELLGGSPFDIHITHQPDVTLPRIPVLISTNSDKDAIYKRCITY
>Taenia_multiceps_JR929085
-----------------------------------------NVAMGMRTIPWKLGASGSIRCYCITTF-----------LLLYTKVVDVVDEKMDKVNMLCFYRTTNAGKSLLAGLITAAIAFQFDNLLKTGALME-----------------------------------------
>Schistosoma_mansoni_XP_002571349
RKQRTDLILDMVRLNDTRQMDAFENSLSTQQRLDLYNDFGNQWQEVAKLCIKIYNEELVRAQIDTPYEQYEGSHDEDMLMEKNDKVFCASVSTIMNKKVKRVNTLCLEGPTTTGKSLLLKLICGTVQFFLQNLLKTVALMEITVQDFKLLLGGEPLDIHVKHQDDCRLPRIPVLISTNHDKDALYNRTHY-
>Opisthorchis_viverrini_XP_009173661
RKQRTDLILDMVRLNDTRQMDAFENSLSTQQRLDLYNDFGNQWQEVAKLCIKIYNEELVRAQIDTPYEQYEGSHDEDMLMEKNDKVFCASVSTIMNKKVKRVNTLCLEGPTTTGKSLLLKLICGTVQFFLQNLLKTVALMEITVQDFKLLLGGEPLDIHVKHQDDCRLPRIPVLISTNHDKDALYNRTHY-
>Schistosoma_mansoni_CCDS8408
RKQRTDLILDMVRLNDTRQMDAFENSLSTQQRLDLYNDFGNQWQEVAKLCIKIYNEELVRAQIDTPYEQYEGSHDEDMLMEKNDKVFCASVSTIMNKKVKRVNTLCLEGPTTTGKSLLLKLICGTVQFFLQNLKLLPSWK----------------------------------------------
>Opisthorchis_viverrini_XP_009170655
RKRRRELMKGLIEAYDARSYNEFYMAMSCDDRDDIYDEYGPQWKETAEHSINNYIAGMLSKQLAMRFEEILMTNYHDQLMSANKQQFLTDLTLVMNKVCTRKNASVIEGPTTTGKTLFVKLIAGTVTKPATTTATTTTVAPQRLDTQEDESYGPNLAVHAQSLFGIHYYKTDVTVTRKQGGDNFRDNLVAV
>Amblyomma_americanum_GAGD01007016
-------------------------------------------------------------------------------------------------------------------------------------------MLLGGDSLSVNKKFEDYVEIYRTPIIMLTNNRDPVWTDRCLFY
>Opisthorchis_felineus_GBJA01006694
RRLVDDVLMNLVEANDARSFNEFYMSLTFEQTNDIYAEYGPQWKHTAKHCIANYTKKILKHQQTSRFEDIIRTNNHDLLFHANKRDFLVNLTQIMNKDIDGRNAFVIEGPTTTGKTLLVKLIAGTVQCFLMKLLKALALVELTVNHFKALLGGNAFDIRVKHQKDERLTRLPVLITTNTDAKAIKERCFI-
>Clonorchis_sinensis_GAA57952
RKRRLDLLRNLVETYDARSFNELNLALTYDERDDIYGEYGSQWKETAEHCIQNYTMRILVEQQTNRFEDHIRTNSHDLLLHVNRQNFLTDLTRVMNKQIDRKNAFVLEGPTTTGENIIRKTHCGQLHIW-----DRTEIWR--------------------------VLPNEP-----------TE-----
>Clonorchis_sinensis_GAA32750
RKRRLDLLRNLVETYDARSFNELNLALTYDERDDIYGEYGPQWKETAEHCISNYTVKILIEQQTSRFEDHIRTNSHDLLLHVNRQDFLCNLTKIMNKQIDRKNAFVIEGPTTTGKTLFVKLIAGTVQFFLMNLLKALALMELTVNDFKELLGGNAFDIHVKHQKDERLTRPVLITTNNSRKFYLDQLSHA
>Clonorchis_sinensis_GAA40146
RKRRLDLLRNLVETYDARSFNELNLALTYDERDDIYGEYGPQWKETAEHCIQNYTMKILLEQQTSRFEDHIRTNSHDLLLHVNRQEFLTNITKIMNREIDRKNAFVIEGPTTTGKTLFVKLIAGTVQFFLMNLLKALALMELTVNDFKELLGGNAFDIHVKHQKDERLTRFPVLITTNNDDKAIKERCFYY
>Clonorchis_sinensis_GAA56553
RKRRLDLLRNLVETYDARSFNELNLALTYDERDDIYGEYGPQWKETAEHCIQNYTMRILVEQQTSRFEDHIRTNSHDLLLFVNRQQFLADLTRVMNKEVDRKNAFVLEGPTTTGKTLFVKLIAGTVQFFLMNLLKALALMELTVNDFK----------------------------------------
>Clonorchis_sinensis_GAA57641
RKRRLDLLRSLVETYDARSFNELNLALTYDERDDIYGEYGPQWKETAEHCIQNYTMRILVEQQTSRFEDHIRTNSHDLLLFVNRQQFLTDLTRVMNKQIDRKNAFVLEGPTTTGKTLFVKLIAGTVQFFLMNLLKALALMELTVNDFKELLGGNAFDIHVKHQKDERLTRLPVLITTNNDGKAIKERCFY-
>Clonorchis_sinensis_GAA57638
RKRRLDLLRNLVETYDARSFNELNLALTYDERDDIYGEYGPQWKETAEHCIQNYTMRILVEQQTSRFEDHIRTNSHDRLLFVNRQEFLADLTRVMNKQVDRKNAFVLGGPTTTGKTLFVKLIAGTVQFFLMNLLKALALMELTVNDFKELLGGNAFDIHVKHQKDERLTRLPVLITTNNDGKAIKERCFY-
>Clonorchis_sinensis_GAA57639
RKRRLDLLRNLVDTYDARSFNELNLALTYDERDDIYGEYGPQWKETAEHCIQNYTMRILVEQQTSRFEDHIRTNSHDRLLFVNRQQLLADLTRVMNKQVDRKNAFVLEGPTTTGKTLFVKLIAGTVQFFLMNLLKALALMELTVNDFKELLGGNAFDIHVKHQKDERLTRLPVLITTNNDGKAIKERCFYY
>Clonorchis_sinensis_GAA57954
RKRRLDLLRNLVETYDARSFNELNLALTYDERDDIYGEYGPQWKETAEHCIQNYTMRILVEQQTSRFEDHIRTNSHDRLLFVNRQKFLADLTRVMNKQVDRKNAFVLEGPTTTGKTLFVKLIAGTVQFFLMNLLKALALMELTVNDLKELLGGNAFDIHVKHQKDERLTRLPVRITTNNDGKVIKERCFY-
>Clonorchis_sinensis_GAA57640
RKRRLDLLRNLVETYDARSFNELNLALTYDERDDIYGEYGPQWKETAEHCIQNYTMRILVEQQTSRFEDHIRTNSHDRLLFVNRQKFLCDLTRVMNKQVDRKNAFVLEGPTTTGKTLFVKLIAGTVQFFLMNLLKALALMELTVNDFKELLGGNAFDIHVKHQKDERLTRLPVLITTNN------------
>Opisthorchis_felineus_GBJA01005901
--MIRPAKRIMIEAYDARSYNELYMSLTYDETDDIYAEYGPQWKETAEHCIANYIKKILISQQTSRFEDVLRSNNHDLLFYVNKQEFLIDITRIMNKEVDRTNAFVIEGPTTTGKTLFVKLIAGTVQFFLMNLLKALALMELTVNYFKELLGGNPFDIHVKHQKDERLNRLPVLITTNNDGKAIKERCFT-
>Opisthorchis_felineus_GBJA01002124
RKRRLDLMRELVTAYDARSYNELYMKLSVEHTDDIYAEYGPTWKETADHAIANYCKKIIIDQQTMTFDEILNSNHHDQLMAVNNKEMLHSLTLVMNKTMKRKNAFVIEVPTTTGKTLFVKLIAGTVPFFLMNLLKTLALMELTVNDFKELLGGNPFDIHVKHQKDERLDRLPVLITTNNDAKAILERCFY-
>Opisthorchis_felineus_GBJA01002120
RKRRLDLMRELVTAYDARSYNELYMKLSVEHTDDIYAEYGPTWKETADHAIANYCKKIIIDQQTMTFDEILRSNHHDQLMLVNNKDMLHALTLVMNKTMKRKNAFVIEGPTTTGKTLFVKLIAGTVQFYLMNLLKTLALMELTVNDFKELLGGNPFDIHVKHQKDERLERLPVLITTNNDAKAILERCFY-
>Opisthorchis_felineus_GBJA01002117
-KRRLDLMRELVTAYDARSYNELYMKLSVEHTDDIYAEYGPTWKETADHAIANYCKKIIIEQQTMTFDEILRSNHHDQLMLVNNKEMLHSLTLVMNKTMKRKNAFVIEGPTTTGKTLFVKLIAGTVQFFLMNLLNKTLGFS----------------------------------------------
>Opisthorchis_viverrini_XP_009177542
GKRRRELMKGLIEAYDARSYNEFYMAMSCEDRDDIYDEYGPQWKETAEHSINNYIAGMLAKQLAMRFEEVLMGNFHDQLMWVNRKQFLIDLTQIMNKAYKRKNAFVIEGPTTTGKTLFVKLIAGTVQFFLMNLLKSLGLMELTVNDFKELLGGNPFDIHVKHQKDERLQRLPVLITTNNDGKAILERCFY-
>Opisthorchis_felineus_GBJA01002122
-------------------------------------------MSRFPDSISLCMDRSDNHNREQCQHPKNTIDGEQDLLFQVNKQTFLCNITRIMNKEIDRKNAFVIEGPTTTGKTLFVKLIAGTVQFFLMNLLKALALMELTVNDFKELLGGNPFDIHVKHQKDERLNRLPVLITTNNDGKAIKERCFA-
>Opisthorchis_viverrini_XP_009176824
RKRRRELMKGLIEAYDARSYNEFYMAMSCDDRDDIYDEYGPQWKETAEHSINNYIAGMLSKQLAMRFEEILMTNYHDQLMAANKQQFLTDLTLVMNKVCTRKNAFVIEGPTTTGKTLFVKLIAGTVQFFLMNLLKALALMELTVNDFKELLGGNPFDIHVKHQKDERLQRLPVLITTNNDGKAILERCFY-
>Opisthorchis_viverrini_XP_009170634
-------MKGLIEAYDARSYNEFYMAMSCDDRDDIYDEYGPQWKETAEHSINNYIAGMLSKQLAMRFEEILMTNYHDQLMSANKQQFLTDLTLVMNKVCTRKNAFVIEGPTTTGKTLFVKLIAGTVQFFLMNLLKALALMELTVNDFKELLGGNPFDIHVKHQKDERLQRLPVLITTNNDGKAILERCFY-
>Opisthorchis_felineus_GBJA01010281
-----------------------------------------------MTFAEILNSNHHSRTCQHPGNTIDGDQLMQVNNKDILHSLTLIMNKSLKRKNAFVIEGPTTTGKTLFVKLIAGTVQFFLMNLLKTLALMELSVNDFKELLGGNPFDIHVKHQKDELHYGTH-------RRSFWCW-----
>Opisthorchis_viverrini_XP_009178033
RKRRLDLMRELVRTYDARSFNELYKRLSVQDTDDIYAEYGPTWKETAEHSISNYCKEIILEQETMTFEQILTSNHHDELIRVVNNRELLVSLTAIMNKLCTRKNAFVIEGPTTTGKTLFVKLIA----------------------------G-----TVQRSGDH-------------------------
>Opisthorchis_felineus_GBJA01008850
RKRRLDLMRELVAAYDARSYNELYMRLSVEHTDDIYAEYGPTWKETADHAITNYCKKIIIEQQTMTFEEILRSNHHDRLMLVNNRELLTSLTLVMNKSMNRKNAFVIEGPTTTGKTLFVKLVAGTVQFFLMNLLKTLALMELTVNDFKELLGGNPFDIHVKHQKDERLERLPVLITTNNDAKAILERCFY-
>Opisthorchis_viverrini_XP_009177530
INKDLTIAGLLSTWDRYVLDPEFREQLSQVLRGIPWDPNDIKYQAHLMWTHATRDVPPYTGNMIDVLQTRRKNYPKDTQEPKTVAKHAQSLFGIHYKTRKQHLTLAQYKPADTGWSINMNFRDGSIPYAAKDSLDKDVFVNKTSDEGLLGLADKTFHQTVYYQMNCLKQAVPAFDNARGERKPVKTMNVKV
>Opisthorchis_viverrini_XP_009177798
-------------------------------------------------------------------------------MNK-------------LCTRKNAFVIEGPTTTGKTLFVKLVAGTVQFFLMNLLKTLALMELTVNDFKELLGGNAFDIHVKHQKDERLDRLPVLITTNNDGKAILERCFYF
>Opisthorchis_viverrini_XP_009177702
```

```
RKRRLDLMRELVRTYDARSFNELYKRLSVQDTDDIYAEYGPTWKETAEHSISNYCKEIILEQETMTFEQILNSNHHDQLIQVNNRELLVCLTSVMNKLCTRKNAFVIEGPTTTGKTLFVKLVAGTVQFFLMNLLKTLALMELTVNDFKELLGGNPFDIHVKHQKDERLERLPVLITTNN----------
>Opisthorchis_viverrini_XP_009177942
-------MRELVRSYDARSFNELYKRLSVQDTDDIYAEYGPTWKETAEHSIANYCKEIILEQETMTFEQILRSNHHDELIRVNNRDLLVALTAVMNKLCTRKNAFVIEGPTTTGKTLFVKLVAGTVQFFLMNLLKTLALMELTVNDFKELLGGNPFDIHVKHQKDE----------
>Bithynia_siamensis_GAQQ01009056
-----------------------------------------------------------------------------------------MCTRKNAFVIEGPTTTGKTLFVKLVAGTVQFFLMNLLKTLALMELTVNDFKELLG----------
>Opisthorchis_viverrini_XP_009177808
RKRRLDLMRELVRSYDARSFNELYKRLSVQDTDDIYAEYGPTWKETAEHSIANYCKEIILEQETMTFEQILNSNHHDQLIRVNNRELLACLTSVMNKLCTRKNAFVIEGPTTTGKTLFVKLVA---NYE----------LTVNDFKELLGGNPFDIHVKHQKDERLDRLPVLITTNNDGKAILER----
>Opisthorchis_viverrini_XP_009177971
-----------------------------------------------------------------------------------------MKLLNKTLALME----------EPRITQ----------LTVNDFKELLGGNPFDIHVKHQKDERLERLPVLITTNN----------
>Moniezia_expansa_JL291017_1
-------------------------------------------------------------------------------------------------------VLQGPSNTGKSLLAKLIVATVNFIFQNLLKTAALMEATVNDFKQLLGGERMEIGIKHRDREWLERVPIICTTNDDCQAIQNRCVY-
>Moniezia_expansa_JL291017_2
-------------------------------------------------------------------------------------------------------VLQGPSNTGKSLLAKLIVATVNFIFQNLLKTAALMEATVNDFKQLLGGERMEIGIKHRDREWLERVPIICTTNDDCQAIQNRCVY-
>Echinostoma_caproni_LL285499
RKARVDLIRELVETYDARTLAELKDALSYDDRLNLYAEHGPSWKETAELACEAYVEKLRKDQETTSLHEYIARNNHDRLLAVNNNKFLTQLTDIMNKKENRVNAFVIQGPTTTGKFQYSSSHRGELNSPWRPLPQGSCAYGINGQRLQRTPGRKTIRYETPTRCPAQTPTRPHFYQYRHGREGHQGAMFL-
>Echinostoma_caproni_LL274983
PDKLWTFIYRGPSPNFTIHTRINVLWADHGDHYHFVFKHPNNKRRTITRIIQAGNLDTNQSMSIFSTCQPVINWEKAYLVRHGRQHLATDARSTTR-AEQDCATLLRNDRAQ-GQKFTNLARVDLIALSYDDRLHGPSWKEAYVERLRKDQETTSLHEYIARNNHNRTCQKPTSTTAGC----------
>Echinostoma_caproni_LL284665
PDKLWTFIYRGRAPTFNSHTRISLLWADHEDHYHFVFKHPNNKKRTIQRIIQAGNLDTAQSVSIFSTCQPVINWDKAYLVRTNRQSLANDARSTTR-AEQDCATLLRNDRAQ-GQKFTNLARVDLIALSYDDRLHGPSWKEAYVERLRKEQETTSLHEYISRNNHSRTCQKP---RSTEEGCRWLDRLLAV
>Echinostoma_caproni_LL286487
RKARVDLIKELVETYDARTLAELKDALSYDDRLNLYAEHGPSWKETAELACEAYVERLRKEQETTTLHEYISRNNHDRLLAVNNNRFLTQLTNIMNKKEQRINAFVIQGPTTT-GKFLCS---SHLSSHILDAKGELHLWNFAILSHEFAEQG---------CRPH--GRAS-------NHPAHGE-----
>Scylla_olivacea_GDRN01037122
-----------------------------------------------------------------------------------------MQKINTLVIKGPTGTGKTLTLATLLGTIQFHLQNLLRNFALFEATVDEYKLLLEGFQFEINVKNSDMEQLHRIPIFISTNRDGQALQSRCKTF
>Penaeus_monodon_AAM94165
NSAEYDYLRHLVKSKAARTVQELVNKLDDEEYKQLWTRTRGQYKDKLRGILTYYNNKKKSSQSQLSLITIYKNISKKYMFANNDPEILAWIIIVADKKLDKINTLVLQGPTGTGKSLTIGALLGLVTFHLQNLIKSYALFEITVDDFKLLFEGSDLEVNIKHQESEIMGRIPIFISTNKDGKALQTRTKT-
>Penaeus_monodon_endogenous_virus_DQ228358_
NSAEYDYLRHLVKSKAARTVQELVNKLDDEEYKQLWTRTRGQYKDKLRGILTYYNNKKKSSQSQLSLITIYKNISKKYMFANNDPEILAWIIIVADKKLDKINTLVLQGPTGTGKSLTIGALLGLVTFHLQNLIKSYALFEITVDDFKLLFEGSDLEVNIKHQESEIMGRIPIFISTNKDGKALQTRTKT-
>Penaeus_monodon_penstyldensovirus_1_GQ411199
NSAEYDYLQHLVKTKSARTVQELVNKLDDEEYKQLWTRTRGQYKDKLRGILTYYNNKKKSNQSQLSLITNLQNISKKYMLANNDPEILAWIIIVADKKLDKINTLVLQGPTGTGKSLTIGALLGLVTFHLQNLIKSYALFEITVDDFKLLFEGSDLEVNIKHQESEIMGRIPIFISTNKDGKALQTRTKTF
>Penaeus_monodon_penstyldensovirus_2_AY124937
NSAEYDYLRHLVKTKSARTVQELVNKLDDEEYKQLWTRTRGQYKDKLRGILTYYNNKKKSNQSQLSLITNLQNISKKYMLANNDPEILAWILIVADKKLDKINTLVLQGPTGTGKSLTIGALLGLVTFHLQNLIKSYALFEITVDDFKLLFEGSDLEVNIKHQESEIMGRIPIFISTNKDGKALQTRTKT-
>Penaeus_stylirostris_penstyldensovirus_2_GQ475529
NSAEYDYLQHLVKTKSARTVQELVNKLDDEEYKQLWTRTRGQYKDKLRGILTYYNNKKKSNQSQLSLITNLQNISKKYMLANNDPEILAWIIIVADKKLDKINTLVLQGPTGTGKSLTIGALLGLVTFHLQNLIKSYALFEITVDDFKLLFEGSDLEVNIKHQESEIMGRIPIFISTNKDGKALQTRTKT-
```

Bilan et perspectives

Les bases de données regorgent d'informations non traitées pouvant permettre de mener des études préliminaires en écologie et en évolution virale. La fouille de transcriptomes et de génomes nous a permis de mettre en lumière une importante diversité d'hôtes potentiels de *Parvoviridae-related* virus chez 29 ordres d'animaux au sein lesquels leur présence n'avait auparavant jamais été décrite, comprenant notamment 14 ordres d'arthropodes (dont des Coléoptères et des Phasmoptères) et 4 ordres de vers plats. Elle a également permis de mettre en évidence pour la première fois leur présence au sein des mollusques, annélides, nématodes et cnidaires. Cette étude montre donc que les *Parvoviridae-related* virus sont présents dans une part importante de la diversité actuelle des animaux. Elle se place dans la continuité d'autres études ayant notamment mis en évidence la présence de *Parvoviridae-related* virus endogénéisés chez un ver plat (ici un schistosome), de tiques, d'une cione et d'un cloporte (Arriagada et Gifford, 2014; Belyi *et al.*, 2010; Kapoor *et al.*, 2010; Liu *et al.*, 2011; Thézé *et al.*, 2014). L'ensemble de ces découvertes suggère que le spectre d'hôte de ces virus est, ou a été, plus important que celui que l'on connait actuellement.

L'ensemble de ces résultats remettent également en question l'existence des sous-familles des *Parvovirinae* et des *Densovirinae* qui est basée sur le fait que les *Parvovirinae* infecteraient les vertébrés tandis que les *Densovirinae* infecteraient les arthropodes. Cette classification est d'autant plus contestable qu'il existe des *Densovirinae* infectant des échinodermes (Gudenkauf *et al.*, 2014; Hewson *et al.*, 2014).

En outre, le fait que certains virus aient été endogénéisés soutient l'hypothèse d'une considérable coévolution entre les *Parvoviridae-related* virus et leurs hôtes. Deux hypothèses se posent sur l'origine de ces virus. D'une part, ils ont pu coévoluer avec les animaux depuis le début de la radiation évolutive de ces derniers, qui remonte au Cambrien, il y a plus de 500 millions d'années. D'autre part, il est possible que leur origine soit postérieure à l'apparition des grands groupes d'animaux actuels, et que ces virus auraient pu être transmis à l'ensemble des animaux via des phénomènes de sauts d'hôtes. Cependant, nos connaissances encore parcellaires de ces virus ne permettent pas de trancher entre ces deux hypothèses. Avec l'augmentation de la masse de données chaque année, il est à prévoir que la fouille de données

permette de continuer à améliorer nos connaissances sur la diversité des hôtes des *Parvoviridae-related* virus.

Cette étude a enfin permis d'avoir une vision plus holistique sur la diversité génétique des *Parvoviridae-related* virus en mettant en lumière près d'une vingtaine de nouveaux genres viraux potentiels. Cependant, les génomes viraux trouvés dans les bases de données sont souvent parcellaires. De plus, notre étude permet seulement de mettre en évidence la présence de ces virus au sein de divers hôtes, ce qui ne signifie pas forcément que ces hôtes soient infectés par des *Parvoviridae-related* virus, mais que la présence de ces virus soit due à des contaminations potentielles. Il serait donc intéressant de reconstituer leur génome entier, puis de tester la virulence de ces virus, en particulier pour ceux retrouvés chez des animaux distants des hôtes actuellement reconnus comme pouvant être infectés par des *Parvoviridae*, comme les vers plats.

De plus, cette vision de la diversité génétique des *Parvoviridae-related* virus reste limitée, notamment par la faible diversité d'espèces animales représentées dans les bases de données génomiques et transcriptomiques. Enfin, étant donné que nos recherches dans les bases de données ont été basées sur la recherche de similarités en comparaison avec des génomes de *Parvoviridae* connus, cette vision est également restreinte par nos connaissances actuelles sur leur diversité génétique. Un changement d'échelle dans notre compréhension de la diversité du monde viral consiste à passer de l'étude de la diversité de certains taxa viraux à l'étude des communautés virales dans leur ensemble. Ce changement d'échelle est rendu possible notamment par l'utilisation de la métagénomique virale.

Chapitre II - Développement d'un protocole dédié à la préparation et à l'analyse de viromes

Préparation et analyse de viromes multiplexés d'arthropodes et de plantes

Durant la dernière décennie, l'essor de la métagénomique virale a entrainé un développement important d'outils moléculaires et bioinformatiques dédiés à la préparation et à l'étude des viromes. La comparaison de l'efficacité de différents protocoles de préparation de viromes (Conceição-neto *et al.*, 2015; Corinaldesi *et al.*, 2017; Hall *et al.*, 2014; Lewandowska *et al.*, 2017; Temmam *et al.*, 2015), de pipelines et de logiciels bioinformatiques a été réalisée (Tangherlini *et al.*, 2016). Cependant, il n'existe pas encore de protocole universel de production et d'analyse de viromes.

De plus, une des limitations principales de l'utilisation de la métagénomique virale sur un grand nombre d'échantillons repose sur le coût relativement élevé du séquençage. Cela a amené au développement de techniques de multiplexage, i.e. permettant le mélange (pool) de plusieurs échantillons préalablement identifiés individuellement, ces échantillons étant ensuite séquencés dans une même réaction de séquençage. Le multiplexage repose sur l'ajout d'un identifiant moléculaire (souvent nommé « tag » ou « barcode ») spécifique de chaque échantillon. Ces tags permettent, lors du traitement bioinformatique des données issues du séquençage, de réassigner chaque séquence générée à son échantillon d'origine. En outre, ils sont utilisés en tant qu'amorces lors des étapes d'amplification. Il est possible d'utiliser des kits de multiplexage provenant de fournisseurs, comme le kit Nextera XT (Illumina). Cependant, leur prix élevé a conduit au développement et la publication de protocoles « maisons » de multiplexage dont l'efficacité varie (Meyer et Kircher, 2010). Par exemple, il a été montré que certains tags ne sont parfois pas fonctionnels, ou amplifient les séquences de manière biaisée (Hamady et Knight, 2009). Par ailleurs, lorsque l'on traite plusieurs échantillons à la fois, il existe un risque inhérent de contamination inter-échantillons (Degnan et Ochman, 2011). Cette contamination amène à la génération de faux positifs difficiles à identifier et qui peuvent amener à une vision biaisée de la diversité des virus présents dans les échantillons traités. Il existe cependant des logiciels permettant d'identifier et de retirer certains contaminants (Schmieder et Edwards, 2011).

Nous avons mis en place un protocole explicatif de la préparation ainsi que de l'analyse bioinformatique de viromes produits à partir d'échantillons d'arthropodes et de plantes. Ce protocole permet de traiter les virus appartenant à l'ensemble de la classification de Baltimore. Il se base sur la filtration des particules virales puis leur concentration via ultracentrifugation. Ces étapes sont suivies d'une digestion par DNase et RNase afin de réduire la proportion d'acides nucléiques non-encapsidés. Ensuite, les acides nucléiques restants (ARN et ADN) sont extraits. Les ADNc sont synthétisés par rétro-transcription suivie de la synthèse de leur brin complémentaire. Ces ADNc double brin sont amplifiés par PCR. Après mélange des produits de PCR et leur purification, ils sont envoyés à séquencer par des services externes (Candresse *et al.*, 2014). Nous avons testé les technologies de séquençage HiSeq et MiSeq (Illumina).

Notre protocole permet de multiplexer jusqu'à 96 échantillons. Ce multiplexage est possible par l'ajout d'adaptateurs, des « linkers » (96 différents) permettant également d'amorcer les réactions de rétro-transcription et de synthèse du brin d'ADNc complémentaire ; et, lors de l'étape de PCR, par l'ajout de tags complémentaires de chaque linker qui sont utilisés comme amorces de la réaction de PCR. Nos linkers sont des séquences de 24 bases de longueur qui contiennent 14 bases uniques à chaque linker et 12 bases aléatoires permettant leur fixation en 5' et 3' sur les acides nucléiques présents dans les échantillons. Nos tags font 22 bases de longueur, ils comportent 12 bases complémentaires d'un linker ainsi que 10 autres bases spécifiques de chaque tag.

Les reads obtenus par séquençage à haut débit sont ensuite démultiplexés, c'est-à-dire réattribués à leur échantillon d'origine par des analyses bioinformatiques se basant sur la reconnaissance des « codes-barres » spécifiques de chaque échantillon. Ensuite, les reads sont « nettoyés » (retrait des séquences des linkers et des tags, et filtrage selon la qualité et la longueur des reads), puis les contig sont créés par assemblage *de novo*. L'abondance relative de chaque contig dans les échantillons est mesurée par mapping (i.e. par alignement des reads contre les contigs). Enfin, une analyse par BLAST contre des bases de données de séquences de référence permet l'assignation taxonomique des contigs ainsi que des reads orphelins (i.e. n'ayant pas pu être assemblés ni mappés).

Chapitre de livre

Viral metagenomics approaches for high resolution screening of multiplexed arthropod and plant viral communities

Sarah François[1,2], Denis Filloux[3], Emmanuel Fernandez[3], Mylène Ogliastro[2], Philippe Roumagnac[3]

[1] Université de Montpellier, UMR 1333 DGIMI « Diversité, Génomes et Interactions Microorganismes-Insectes », place Eugène-Bataillon, 34095 Montpellier cedex 5, France

[2] Laboratoire « Diversité, Génomes et Interactions Microorganismes Insectes » (DGIMI) UMR 1333, INRA, Université Montpellier, 34095 Montpellier, France.

[3] Laboratoire « Biologie et Génétique des Interactions Plante-Parasite » UMR BGPI, CIRAD-INRA-SupAgro, Campus International de Montferrier-Baillarguet, Montpellier Cedex-5, France.

Accepté dans **Methods in Molecular Biology**

Titre du livre: Viral Metagenomics: Methods and Protocols

Éditeur: Springer-Verlag

Viral metagenomics approaches for high resolution screening of multiplexed arthropod and plant viral communities

Sarah François, Denis Filloux, Emmanuel Fernandez, Mylène Ogliastro and Philippe Roumagnac

Abstract

Viral metagenomic approaches have become essential for culture-independent and sequence-independent viral detection and characterization. This chapter describes an accurate and efficient approach to: (i) concentrate viral particles from arthropods and plants, (ii) remove contaminating non-encapsidated nucleic acids, (iii) extract and amplify both viral DNA and RNA and (iv) analyze High Throughput Sequencing (HTS) data by bioinformatics. Using this approach, up to 96 arthropod or plant samples can be multiplexed in a single HTS library.

Key words Metagenomics, Virus discovery, Diagnostic, Arthropod, Plant, Random amplification, High-Throughput Sequencing

1 Introduction

While viruses are the most numerous biological entities on Earth, the number of currently classified virus species is probably dramatically underestimated [1,2]. Several factors can account for this lack of knowledge, including intrinsic characteristics of viruses such as their small size, their rapid rate of evolution, or the lack of universally conserved viral genetic markers [3]. In addition, the genetic material recovered from animal or plant samples is mostly of non-viral origin [4], which renders difficult the study of the host virome (the collection of all viruses that are found in or on the host).

Viral metagenomics, which is the direct genetic analysis of viral genomes contained within a sample, has revolutionized the last decade the field of virus discovery [5–7]. Viral metagenomics approaches have targeted four main classes of nucleic-acids, including total RNA or DNA; virion-associated nucleic acids (VANA) purified from viral particles; double-stranded RNAs (dsRNA) and virus-derived small interfering RNAs (siRNAs) [8–14]. While each of these approaches have advantages and drawbacks, the VANA approach has gained popular because it takes advantage of the hardiness of many viral capsids for concentrating and purifying the viral nucleic acids and allows the detection of both RNA and DNA viruses [15–18].

Here, we provide a detailed protocol for the VANA-based metagenomics determination of arthropods and plants viromes. Starting from arthropod and plant samples, we first concentrate viral particles by filtration and centrifugation prior to partially remove the non-encapsidated material by DNase and RNase digestion [19,20]. Encapsidated DNA and RNA are then extracted and RNA is converted to cDNA using a 26 nt primer (Dodeca Linker) composed by a 14 nt linker linked at 3' end to N_{12} **(Fig. 1)**. Noteworthy, we present here a set of 96 Dodeca Linkers. Double-stranded DNA is synthetized from single-stranded DNA using Large (Klenow) fragment DNA polymerase and the Dodeca Linker used during the reverse transcriptase (RT) step **(Fig. 1)**. Double-stranded DNA are further amplified using one 24 nt PCR multiplex identifier primer composed by the 14 nt linker used during the RT step linked at 5' end to a 10 nt tag **(Fig. 1)**. This PCR yields amplicons that are all tagged at both extremities with the same multiplex identifier primer **(Fig. 1)** [21]. Pools of up to 96 multiplex identifier amplicons can then be mixed, which reduces the cost of library preparation in case of numerous samples. This protocol finally describes the bioinformatic data analysis (data demultiplexing, clean-up, *de novo* assembly, taxonomic assignment and read-mapping).

2 Materials

Prepare and store reagents according to their individual specifications (room temperature if not specified). HBSS 1X solution should be prepared using sterilized ultrapure water. Special care should be taken to keep enzymes at −20 °C until use. Follow all waste disposal regulations when disposing waste materials. To reduce laboratory contaminations during nucleic acid extraction and amplification, working in a clean environment is recommended.

2.1 Purification of viral particles

1. Tissue homogenizer (FastPrep-24 Instrument; MP Biomedicals) and sterile ceramic beads (MP Biomedicals) or sterile mortar, pestle, carborundum (Sigma-Aldrich) and liquid nitrogen.
2. Conical tubes (15 mL) (Falcon).
3. Hanks' Balanced Salt solution (HBSS) 10X (Invitrogen Ref. 14065056).
4. 0.45 μm syringe filters (Sarstedt Filtropur Ref. 83.1826).
5. 5 mL syringes (CarlRoth Omnifix Ref. C537.1).
6. Centrifuge Eppendorf 5810R (rotor A-4-62 and F34-6-38 15mL tubes).
7. Ultracentrifuge polycarbonate bottles (Tube 26.3 mL, Ref. 355654).
8. Ultracentrifuge (Beckman Coulter Optima LE-80 K with rotor 50.2 TI).
9. Deionized water.
10. Microtubes (1.5 mL) (Eppendorf).
11. Single channel pipettes (0.5-10 μL/10-200 μL/1,000-5,000 μL).
12. 0.5 mL PCR 8 tube strips.
13. 10 μL pipette filter tips
14. 200 μL pipette filter tips.
15. 1,000 μL pipette filter tips.
16. 5,000 mL pipette filter tips.
17. DNase I (Biotech CAS#9003-98-9).
18. RNase A (Biotech CAS#9001-99-4).
19. Oven.
20. Ice.

2.2 Viral nucleic acid extraction

1. Nucleospin 96 virus Core kit (Macherey Nagel Ref. 740691.4) (see **Note 1**).
2. Square-well Block (Macherey Nagel Ref. 740481.24).
3. Absolute ethanol.
4. Single channel pipettes (0.5-10 µL/10-200 µL/100-1,000 µL).
5. 10 µL long pipette filter tips (see **Note 2**).
6. 200 µL pipette filter tips.
7. 1,000 µL pipette filter tips.
8. Centrifuge Sigma 4-16K (Qiagen Sigma 09100 Microplate Rotor).
9. Heat block or water bath.

2.3 Reverse transcription

1. Thermocycler.
2. Single channel pipettes (0.5-10 µL/10-200 µL/100-1,000 µL).
3. 10 µL long pipette filter tips (*see* **Note 2**).
4. 200 µL pipette filter tips.
5. 1,000 µL pipette filter tips.
6. PCR plate (96) 0.5 mL (VWR).
7. Microtubes (1.5 mL) (Eppendorf).
8. Nuclease-free water.
9. Dodeca Linkers (10 µM) (**Tab. 1**).
10. SuperScript III reverse transcriptase (200 U/µL) (Invitrogen Ref. 18080093).
11. Reverse transcriptase (RT) buffer (5X; supplied with reverse transcriptase).
12. Dithiothreitol (DTT) (0.1 M).
13. dNTP mix (10 mM).
14. Ice.

2.4 cDNA purification

1. RNase A (Invitrogen).
2. Thermocycler.
3. Column-based PCR purification kit, e.g., QIAquick PCR Purification Kit (Ref. 28181).
4. Absolute ethanol.

5. Microtubes (1.5 mL) (Eppendorf).
6. Single channel pipettes (0.5-10 µL/10-200 µL/100-1,000 µL).
7. 10 µL long pipette filter tips (*see* **Note 2**).
8. 200 µL pipette filter tips.
9. 1,000 µL pipette filter tips.
10. Square-well Block (Macherey Nagel Ref. 740481.24).
11. Centrifuge Sigma 4-16K (Qiagen Sigma 09100 Microplate Rotor).
12. Ice.

2.5 Klenow amplification

1. Single channel pipettes (0.5-10 µL/10-200 µL/100-1,000 µL).
2. 10 µL long pipette filter tips (*see* **Note 2**).
3. 200 µL pipette filter tips.
4. 1,000 µL pipette filter tips.
5. PCR plate (96) 0.5 mL (VWR).
6. Microtubes (1.5 mL) (Eppendorf).
7. Thermocycler.
8. Dodeca Linkers (100 µM) (**Tab. 1**).
9. Exo(−) Klenow DNA polymerase I (5 U/µL; e.g., Promega Ref. M2206).
10. Exo(−) Klenow Buffer (10X; supplied with Exo(−) Klenow DNA polymerase).
11. Nuclease-free water.
12. dNTP mix (10 mM).
13. Ice.

2.6 PCR amplification

1. Single channel pipettes (0.5-10 µL/10-200 µL/100-1,000 µL).
2. 10 µL pipette filter tips.
3. 200 µL pipette filter tips.
4. 1,000 µL pipette filter tips.
5. Microtubes (1.5 mL) (Eppendorf).
6. PCR plate (96-Well) 0.5 mL (VWR).
7. HotStar Taq Plus Master Mix kit (Qiagen Ref. 203645).

8. PCR Primers (10 µM) (see **Tab. 1**).

9. Nuclease-free water.

10. Thermocycler.

2.7-2.8 Verification of the composition and concentration of the PCR products

1. Microtubes (1.5 mL) (Eppendorf).

2. Single channel pipettes (0.5-10 µL/10-200 µL/100-1,000 µL).

3. 10 µL pipette filter tips.

4. 200 µL pipette filter tips.

5. 1,000 µL pipette filter tips.

6. TBE buffer (0.5X): 45 mM Tris–HCl (pH 8.3), 45 mM boric acid, 1 mM EDTA.

7. Agarose (type LE) gel (1 %): Prepare in 0.5X TBE buffer. Add GelRed or ethidium bromide for visualization.

8. DNA ladder.

9. 1 kb Plus DNA Ladder (Invitrogen Ref. 10787026).

10. UV transilluminator.

11. Microtubes (1.5 mL) (Eppendorf).

12. NucleoSpin gel and PCR clean-up (Macherey Nagel Ref. 740609.250).

13. Absolute ethanol.

14. Nuclease-free water.

15. Qubit 2.0 Fluorometer with Qubit Assay HS Kit for dsDNA (ThermoFisher Scientific Ref. Q32851).

2.9 Data handling and Bioinformatics

1. Intel-based server: 4 X 2 Intel Xeon, 256 GB memory per processor or similar capacity.

2. UNIX-based operating system.

3. FastQC software (http://www.bioinformatics.babraham.ac.uk/projects/fastq_screen/).

4. FASTX-Toolkit (http://hannonlab.cshl.edu/fastx_toolkit/).

5. Cutadapt software (https://pypi.python.org/pypi/cutadapt/).

6. SPAdes software (http://cab.spbu.ru/software/spades/).

7. BLAST+ software package (ftp://ftp.ncbi.nlm.nih.gov/blast/executables/blast+/).

8. Bowtie2 software (http://bowtie-bio.sourceforge.net/bowtie2/).

9. Integrative Genomics Viewer (http://www.broadinstitute.org/igv/).

10. Web-based resources: NCBI GenBank (https://www.ncbi.nlm.nih.gov/genbank/).

2.10 Confirmation and retrieval of near-full genome sequences

1. Primer3 software (http://primer3.sourceforge.net/).

2. Single channel pipettes (0.5-10 µL/10-200 µL/100-1,000 µL).

3. 10 µL long pipette filter tips.

4. 200 µL pipette filter tips.

5. 1,000 µL pipette filter tips.

6. Column-based DNA extraction kit with proteinase K: QIAamp DNA Mini Kit (Qiagen Ref. 51304).

7. Column-based RNA extraction kit: RNeasy MiniKit (Qiagen Ref. 74104).

8. First-strand cDNA synthesis kit: SuperScript III reverse transcriptase kit (Invitrogen Ref. 18080093).

9. Nuclease-free water.

10. PCR plate (96) 0.5 mL (VWR).

11. Microtubes (1.5 mL) (Eppendorf).

12. Thermocycler.

13. HotStarTaq Plus Master Mix kit (Qiagen Ref. 203645).

14. Viral-specific PCR primers (10 µM).

15. TBE buffer (0.5X): 45 mM Tris–HCl (pH 8.3), 45 mM boric acid, 1 mM EDTA.

16. Agarose (type LE) gel (1 %): Prepare in 0.5X TBE buffer. Add GelRed or ethidium bromide for visualization.

17. DNA ladder.

18. 1 kb Plus DNA Ladder (Invitrogen).

19. UV transilluminator.

20. Ice.

3 Methods

The following methods outline the extraction of virion-associated nucleic acids (DNA and RNA) from arthropod and plant samples for the construction of NGS sequencing libraries and the basic bioinformatics analysis of the data.

3.1 Purification of viral particles (modified from [15])

1. Grind 200 mg to 800 mg of arthropod or plant material in 15 mL tubes containing 4 sterile ceramic beads using tissue homogenizer. Alternatively, grind material in liquid nitrogen with about 100 mg of carborundum using a pestle and a mortar. Perform all the following steps, until library preparation, on ice.

2. Dilute the Hanks' Balanced Salt solution (HBSS) 10X to 1X with sterile deionized water. Add 8 mL of HBSS 1X and homogenize.

3. Centrifuge at 4,000xg for 5 min at 4 °C.

4. Transfer the supernatant in 15 mL tubes.

5. Centrifuge at 8,000xg for 3 min at 4 °C to pellet debris.

6. Use a 5 mL syringe and a 0.45 µm syringe filter (*see* **Note 3**) to transfer the supernatant in 26.3 mL ultracentrifuge polycarbonate bottles to remove any remaining debris.

7. Fill the ultracentrifuge polycarbonate bottles with HBSS 1X solution.

8. Centrifuge at 148,000xg for 2h30 at 4°C.

9. Discard the supernatant by pipetting. Be careful not to remove the pellet.

10. Add 200 µL of HBSS 1X. Tilt the bottles so that the pellet is immersed in the buffer.

11. Keep the tubes overnight at 4°C to resuspend the pellet.

12. Transfer 150 µL of the viral particles suspension to 0.5 mL PCR 8 tube strips.

13. Dilute the DNase I and the RNase A to the working concentration using nuclease-free water. Gently mix by pipetting or by flicking the tube a few times. Keep on ice until use.

14. Add 1 µL of DNase I (5 mg/mL) and 2 µL of RNase A (10 mg/mL) and 47 µL of HBSS 1X solution. Incubate at 37°C for 1 h 30 min. Store at -80°C or proceed directly to stop DNA and RNA degradation.

3.2 Viral nucleic acid extraction

1. Extract DNA and RNA from the total volume of viral particle digested suspension obtained at the previous step (approximately 200 µL) with the NucleoSpin 96 Virus Core Kit (Macherey-Nagel, Ref: 740452) according to the manufacturer's protocol (*see* **Note 1**).

2. Store at – 80°C or proceed directly.

3.3 Reverse transcription

1. Dilute Dodeca Linkers to 10 µM with nuclease-free water. Add 1 µL of Dodeca Linkers (10 µM) in 10 µL of extracted viral nucleic acid.
2. Denature at 85°C for 2 min in a thermal cycler and chill on ice for 2 min.
3. Add 2 µL of DTT (100 mM), 1.25 µL of dNTP mix (10 mM), 4 µL of 5X SuperScript buffer, and 1 µL of SuperScript III (5 U) and 0.75 µL of nuclease-free water. Mix gently.
4. Perform the reverse transcription by incubating in a thermocycler at 25°C for 10 min, 42°C for 60 min, 70°C for 5 min and at 4°C for 2 min. Store at -80°C or proceed directly.

3.4 cDNA purification

1. Add 1 µL of RNase A (10 mg/mL) and mix gently.

2. Incubate at room temperature for 15 min.

3. Heat at 85°C in a thermocycler during 2 min and keep at room temperature.

4. Purify the cDNA using the QiaQuick PCR cleanup kit (Qiagen) according to the manufacturer's protocol. Store at -80°C or proceed directly.

3.5 Klenow amplification

1. Put 20 µL of cleaned cDNA in PCR plate. Add 0.5 µL of Dodeca Linker (100 µM). Mix.

2. Place the plate in a thermocycler at 95°C for 2 min and immediately in 4°C for 2 min.

3. Add 0.5 µL of Klenow DNA polymerase (5 U/µL), 2.5 µL of Klenow reaction buffer 10X, 1 µL of dNTP mix (10 mM) and 0.5 µL of nuclease-free water.

4. Incubate in a thermocycler at 37°C for 60 min followed by 75°C enzyme heat inactivation for 10 min. Store at -80°C or proceed directly.

3.6 PCR amplification

1. Put 5 µL of the Klenow product in a PCR plate. Add 4 µL of PCR Primer (diluted to 10 µM in nuclease-free water), 10 µL of HotStar Taq Plus Master Mix (Qiagen) and 1 µL of nuclease-free water.

2. Place the plate in a thermocycler and perform the following PCR cycling conditions: 1 cycle of 95°C for 5 min, five cycles of 95°C for 1 min, 50°C for 1 min, 72°C for 1.5 min and 35 cycles of 95°C for 30 sec, 50°C for 30 sec, 72°C for 1.5 min +2 sec at each cycle. Perform an additional final extension for 10 min at 72°C.

3.7 Verification of the composition and concentration of the PCR products

1. Verify the yield of the PCR products by the migration of 6 µL of PCR products loaded with 1 µL of DNA ladder to a 1% agarose gel. Migrate at 100 V for 45 min. Visualize PCR products under UV after staining with ethidium bromide or GelRed (**Fig. 2**).

2. Pool 2 to 6 µL of each PCR product in a 1.5 mL tube according to the smear intensity.

3. Clean the pooled PCR products using the NucleoSpin Gel and PCR clean-up (Macherey-Nagel, Ref: 740609) according to the manufacturer's protocol (*see* **Note 4**).

4. Measure the DNA concentration of cleaned PCR products using the Qubit dsDNA HS Assay kit (Thermo Fisher Scientific) according to the manufacturer’s protocol.

3.8 Library construction and sequencing

1. Send the cleaned PCR products to an external High Throughput Sequencing provider which carries out the library construction and the sequencing. The PCR products can be sequenced on 454 pyrosequencing platform as well as on a number of different Illumina platforms. Most existing HTS platforms have their own protocols to convert PCR products into a sequencing library suitable for subsequent cluster generation and sequencing. These protocols differ in the quantity and quality of the starting material, but they are usually comprised of end repair, modification, and ligation of adapters, which enable DNA amplification by adapter-specific primers and size selection of DNA molecules with a length optimal for the sequencing strategy. The example reported here used the Illumina MiSeq platform 300 pb as paired-end reads. For the example, the sequencing generated about 18 million paired-end reads.

3.9 Data handling and bioinformatics

The bioinformatics data analysis can be divided into four sections: raw data demultiplexing and clean-up, *de novo* assembly, taxonomic assignment and read-mapping (**Fig. 3**). All the presented steps rely on bioinformatical softwares. In most of the cases, standard settings should be sufficient. Consult the software manuals for suggested settings and required preprocessing of data.

3.9.1 Data demultiplexing and clean-up

1. Verify the number of reads and evaluate their average quality using FastQC software.
2. Identify each PCR Primers in each raw reads using the "agrep" command [22] in order to assign them to the particular samples from which they originated (demultiplexing)(*see* **Note 5**).
3. Remove the Illumina adaptors and the PCR Primers, and perform a quality filtering of the reads (remove sequence regions with quality score <q30 and reads smaller than 15 nt) using Cutadapt version 1.9 [23].

3.9.2 De novo assembly

1. Assemble the reads into longer continuous sequences (contigs) using the SPAdes assembler 3.6.2 [24] (or similar software). K-mer length can be modified to improve the assembly. Consult the assembler manual for suggested settings of data.

3.9.3 Taxonomic assignment

1. Perform BLASTn and BLASTx searches [25] against local homologs of NCBI nucleotide and protein databases using NCBI's BLAST program for taxonomic classification (*see* **Note 6**). This can be performed for both reads and contigs.
2. For potential viruses identified during the evaluation of BLAST results, retrieve candidate reference genomes from GenBank in FASTA format.

3.9.4 Read mapping

1. Cleaned unassembled reads can be mapped on viral contigs produced by *de novo* assembly, or on viral reference sequences that can be found in GenBank. Map reads and/or contigs using Bowtie 2.1.0 (options end-to-end very sensitive) [26,27] (or similar software) against the reference genomes or contigs to allow analysis and visualization of similarities and coverage distribution. The results from alignment can be checked using the Integrative Genomics Viewer (IGV) or similar viewers **(Fig. 4)**[28].

3.10 Confirmation and retrieval of near-full genome sequences

1. Based on the results from the alignments, use the Primer3 program [29] (or similar software) to design specific PCR primers to confirm the presence of virus in the original material and to close gaps.

2. Extract DNA and/or RNA from the original material using QIAamp DNA Mini Kit (Qiagen) or RNeasy MiniKit (Qiagen) following manufacturer's protocols. In case or RNA extraction, generate cDNA using a SuperScript III reverse transcriptase kit with random hexamers according to the manufacturer's instructions.

3. Amplify the viral nucleic acid using a PCR kit such as the HotStarTaq Plus Master Mix kit (Qiagen) according to the manufacturer's protocol.

4. Visualize a fraction of the amplified products on an agarose gel and perform Sanger sequencing from the remaining volume of the PCR products. Otherwise, extract the DNA bands of interest by using a UV transilluminator and a scalpel, purify the PCR products using a column-based gel extraction kit and perform Sanger sequencing. PCR with overlapping primers can be used to sequence PCR fragments that are too long to be sequenced in a single round of Sanger sequencing.

4 Notes

1. Nucleospin 96 virus Core kit (Macherey Nagel) is well suited for the simultaneous extraction of encapsidated DNA and RNA nucleic acids.

2. Only 10 µL long pipette filter tips allow pipetting small quantities of solution in MN square well blocks.

3. The use of a 0.45 µm filter may prevent the recovery of giant viruses [30].

4. The NucleoSpin Gel and PCR clean-up (Macherey Nagel) allows size selection of the PCR products. Consult the kit manual for further details.

5. Using the multiplexing method detailed in this chapter, about 50% of Illumina MiSeq 300 pb paired-end raw reads are correctly assigned to their samples of origin.

6. The BLAST software suite released by NCBI can perform a number of different types of homology searches using nucleotide sequences as query against databases of nucleotide and amino acid sequences (e.g. BLASTn, BLASTx). Updated and preformatted nucleotide and protein databases can be obtained from NCBI as compressed archives (ftp://ftp.ncbi.nih.gov/blast/db/). When performing a BLAST search, it is possible to configure parameters of the BLAST search to specify BLAST algorithm, database(s) to be searched, output format, cutoff levels, etc. For more information, see the manual at http://www.ncbi.nlm.nih.gov/books/NBK1763/.

References

1. Suttle, C. (2007) Marine viruses (mdash) major players in the global ecosystem. *Nat.* Rev. Microbiol. 5, 801–812.

2. Brum, J. R. et Sullivan, M. B. (2015) Rising to the challenge: accelerated pace of discovery transforms marine virology. Nat. Rev. Microbiol. 13, 147–159.

3. Koonin, E. V, Dolja, V. V et Krupovic, M. (2015) Origins and evolution of viruses of eukaryotes: The ultimate modularity. Virology 479-480, 2–25.

4. Reyes, A., Semenkovich, N. P., Whiteson, K., Rohwer, F. et Gordon, J. I. (2012) Going viral: next-generation sequencing applied to phage populations in the human gut. Nat Rev Microbiol. 10, 607–617.

5. Chiu, C. Y. (2013) Viral pathogen discovery. Curr. Opin. Microbiol. 16, 468–78.

6. Rosario, K. et Breitbart, M. (2011) Exploring the viral world through metagenomics. Curr. Opin. Virol. 1, 289–97.

7. Mokili, J. L., Rohwer, F. et Dutilh, B. E. (2012) Metagenomics and future perspectives in virus discovery. Curr. Opin. Virol. 2, 63–77.

8. Roossinck, M. J., Martin, D. P. et Roumagnac, P. (2015) Plant Virus Metagenomics : Advances in Virus Discovery. Phytopathology 6, 716-27.

9. Tokarz, R., Williams, S. H., Sameroff, S., Sanchez Leon, M., Jain, K. et Lipkin, W. I. (2014) Virome analysis of Amblyomma americanum, Dermacentor variabilis, and Ixodes scapularis ticks reveals novel highly divergent vertebrate and invertebrate viruses. J. Virol. 88, 11480–92.

10. Alquezar-Planas, D. E. *et al.* (2013) Discovery of a divergent HPIV4 from respiratory secretions using second and third generation metagenomic sequencing. Sci. Rep. 3, 2468.

11. Angly, F. E. *et al.* (2006) The marine viromes of four oceanic regions. PLoS Biol. 4, e368.

12. Zablocki, O., van Zyl, L., Adriaenssens, E. M., Rubagotti, E., Tuffin, M., Cary, S. C. et Cowan, D. (2014) High-level diversity of tailed phages, eukaryote-associated viruses, and virophage-like elements in the metaviromes of antarctic soils. Appl. Environ. Microbiol. 80, 6888–97.

13. Whon, T. W., Kim, M.-S., Roh, S. W., Shin, N.-R., Lee, H.-W. et Bae, J.-W. (2012) Metagenomic characterization of airborne viral DNA diversity in the near-surface atmosphere. J. Virol. 86, 8221–31.

14. Poojari, S., Alabi, O. J., Fofanov, V. Y. et Naidu, R. (2013) A leafhopper-transmissible DNA virus with novel evolutionary lineage in the family geminiviridae implicated in grapevine redleaf disease by next-generation sequencing. PLoS One 8, e64194.

15. Candresse, T. *et al.* (2014) Appearances can be deceptive: revealing a hidden viral infection with deep sequencing in a plant quarantine context. PLoS One 9, e102945.

16. Bernardo, P. *et al.* (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.

17. Palanga, E. *et al.* (2016) Metagenomic-Based Screening and Molecular Characterization of Cowpea- Infecting Viruses in Burkina Faso. PLoS One 11, 1–21.

18. Fancello, L., Raoult, D. et Desnues, C. (2012) Computational tools for viral metagenomics and their application in clinical research. Virology 434, 162–74.

19. Allander, T., Emerson, S. U., Engle, R. E., Purcell, R. H. et Bukh, J. (2001) A virus discovery method incorporating DNase treatment and its application to the identification of two bovine parvovirus species. PNAS 98.

20. Victoria, J. G., Kapoor, A., Li, L., Blinkova, O., Slikas, B., Wang, C., Naeem, A., Zaidi, S. et Delwart, E. (2009) Metagenomic analyses of viruses in stool samples from children with acute flaccid paralysis. J. Virol. 83, 4642–51.

21. Roossinck, M. J., Saha, P., Wiley, G. B., Quan, J., White, J. D., Lai, H., Chavarría, F., Shen, G. et Roe, B. (2010) Ecogenomics: using massively parallel pyrosequencing to understand virus ecology. Mol. Ecol. 19, 81–8.

22. Wu, S. et Manber, U. (1992) A fast approximate pattern-matching tool. Usenix Winter 1992 Tech. Conf.

23. Martin, M. (2011) Cutadapt removes adapter sequences from high-throughput sequencing reads. EMB 17.

24. Bankevich, A. *et al.* (2012) SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. J. Comput. Biol. 19, 455–77.

25. Altschul, S., Gish, W., Miller, W., Myers, E. et Lipman, D. (1990) Basic local alignment search tool. J. Mol. Biol. 5, 403–410.

26. Langmead, B. (2010) Aligning short sequencing reads with Bowtie. Curr Protoc Bioinforma. 11.

27. Toland, A. E., Çatalyürek, Ü. V, Hatem, A. et Bozda, D. (2013) Benchmarking short sequence mapping tools. BMC Bioinformatics 14.

28. Robinson, J.T., Thorvaldsdóttir, Wendy Winckler, W, Guttman, M, Lander, E.S., Getz, G, Mesirov, J.P. (2011) Integrative genomics viewer. Nature Biotechnology 29, 24–26.

29. Untergasser, A., Cutcutache, I., Koressaar, T., Ye, J., Faircloth, B. C., Remm, M. et Rozen, S. G. (2012) Primer3--new capabilities and interfaces. Nucleic Acids Res. 40, e115.

30. Halary, S., Temmam, S., Raoult, D. et Desnues, C. (2016) Viral metagenomics: are we missing the giants? Curr. Opin. Microbiol. 31, 34–43.

Figures

Table 1: 96 Dodeca Linkers and corresponding PCR primers used to tag each arthropod or plant sample.

Dodeca Linker	Sequence	Primer	Sequence
1	**AACAACCTCGGCGG**NNNNNNNNNNNT	1	AACAAGACGT**AACAACCTCGGCGG**
2	AACCGAGTCGCGAGNNNNNNNNNNNT	2	AACACACTCAAACCGAGTCGCGAG
3	ATCCGGCGTGATGGNNNNNNNNNNNT	3	AACATGGAGAATCCGGCGTGATGG
4	AACGCTGTCGCAGGNNNNNNNNNNNT	4	AACCAGCCAGAACGCTGTCGCAGG
5	CGGGTCGTCATAGGNNNNNNNNNNNT	5	AAGGAACGTACGGGTCGTCATAGG
6	AAGTGGTACCCGCGNNNNNNNNNNNT	6	AAGGCATGCGAAGTGGTACCCGCG
7	CGTGGAGACTCTGGNNNNNNNNNNNT	7	AAGGTAGAAGCGTGGAGACTCTGG
8	AACGGCCGCCACTANNNNNNNNNNNT	8	AATACACAGCAACGGCCGCCACTA
9	ACCAATGGTCCGGGNNNNNNNNNNNT	9	AATACAGCCTACCAATGGTCCGGG
10	ACGATCCAGCGTCGNNNNNNNNNNNT	10	AATACCGGTAACGATCCAGCGTCG
11	ACGCCATCACACGGNNNNNNNNNNNT	11	AATACTGTGGACGCCATCACACGG
12	ACGGCGTCGGTAGTNNNNNNNNNNNT	12	AATAGCCACAACGGCGTCGGTAGT
13	ACTACCGCACGCTGNNNNNNNNNNNT	13	AATCCGCTCCACTACCGCACGCTG
14	AGAACACCGCGCAGNNNNNNNNNNNT	14	AATTCCTGCTAGAACACCGCGCAG
15	AGAGGATCTGGCGGNNNNNNNNNNNT	15	ACAATTCGAGAGAGGATCTGGCGG
16	AGATGCACCGAGCGNNNNNNNNNNNT	16	ACAGACGTTAAGATGCACCGAGCG
17	AGCCGGATCGTGAGNNNNNNNNNNNT	17	ACCGAACCGTAGCCGGATCGTGAG
18	AGCGAAGGAAGCGGNNNNNNNNNNNT	18	ACCTGATTCTAGCGAAGGAAGCGG
19	AGCTCTCGATCCGGNNNNNNNNNNNT	19	ACGATGAAGTAGCTCTCGATCCGG
20	AGGACTGGCCGATGNNNNNNNNNNNT	20	ACGCCTCAACAGGACTGGCCGATG
21	AGGCTGGTGCTCAGNNNNNNNNNNNT	21	ACTGGCGCATAGGCTGGTGCTCAG
22	AGTCAACGCGCTCGNNNNNNNNNNNT	22	ACTTACCAAGAGTCAACGCGCTCG
23	AGCTAGGCGCCCTANNNNNNNNNNNT	23	AGAACCACGGAGCTAGGCGCCCTA
24	ATGTCCGCGCTCCTNNNNNNNNNNNT	24	AGAGTGACTTATGTCCGCGCTCCT
25	ATGTGACCGGCTGCNNNNNNNNNNNT	25	AGATAGTGCTATGTGACCGGCTGC
26	CATCCACGCGGTGANNNNNNNNNNNT	26	AGGACATAAGCATCCACGCGGTGA
27	CAAGCGGTAGCCGANNNNNNNNNNNT	27	AGGTGAATAGCAAGCGGTAGCCGA
28	CAAGGCATAGCGCGNNNNNNNNNNNT	28	AGTCCTAATCCAAGGCATAGCGCG
29	CACCTATGCGCCGANNNNNNNNNNNT	29	AGTTGTTGTCCACCTATGCGCCGA
30	CAGAGCGGCACGAANNNNNNNNNNNT	30	ATATCCGCATCAGAGCGGCACGAA
31	CAGCACGTCCGCAANNNNNNNNNNNT	31	ATCTAAGGAGCAGCACGTCCGCAA
32	CAGGAGCGACCTCANNNNNNNNNNNT	32	ATGACGGTAACAGGAGCGACCTCA
33	CCGGCCTACGTCTANNNNNNNNNNNT	33	ATGCGCATATCCGGCCTACGTCTA
34	CCGGTGGTCTCACANNNNNNNNNNNT	34	CAACCGATTGCCGGTGGTCTCACA
35	CACGGGATCGCAGANNNNNNNNNNNT	35	CAACCTCTGACACGGGATCGCAGA
36	CCAAGTACGCGGCANNNNNNNNNNNT	36	CAACGCACAGCCAAGTACGCGGCA
37	CCCGAACGCTGGAANNNNNNNNNNNT	37	CAACGTTAACCCCGAACGCTGGAA
38	CCGACGCTAGGTCANNNNNNNNNNNT	38	CAACTGCTATCCGACGCTAGGTCA
39	CCTGATGCCTACCGNNNNNNNNNNNT	39	CACTACGAATCCTGATGCCTACCG
40	CCCGGTCGCTATCANNNNNNNNNNNT	40	CACTGAGCACCCCGGTCGCTATCA
41	CGAGCTACGCATCGNNNNNNNNNNNT	41	CATTGGCTAACGAGCTACGCATCG
42	CGCCGTTCGCCTTANNNNNNNNNNNT	42	CCACCACACACGCCGTTCGCCTTA
43	CGTCGCCGTATGGANNNNNNNNNNNT	43	CCAGGTCGAACGTCGCCGTATGGA
44	CGTACGAGAGTGCGNNNNNNNNNNNT	44	CCATAACTTGCGTACGAGAGTGCG
45	CGTCCCGTCAACGANNNNNNNNNNNT	45	CCATACTGACCGTCCCGTCAACGA
46	CTAACGGCGGATCGNNNNNNNNNNNT	46	CCGACTTCTCCTAACGGCGGATCG
47	CTCCATAGCGGCCANNNNNNNNNNNT	47	CCGCTATTCGCTCCATAGCGGCCA
48	CTCCCGCTGTACCANNNNNNNNNNNT	48	CCGGAATGCTCTCCCGCTGTACCA
49	CTCGTGCCGGAGTANNNNNNNNNNNT	49	CCGGTCTCTACTCGTGCCGGAGTA
50	GCGCCTGTGGTAGANNNNNNNNNNNT	50	CCGTACGATGGCGCCTGTGGTAGA
51	GAATCCAGCCGCCTNNNNNNNNNNNT	51	CCTCCGTTCTGAATCCAGCCGCCT
52	GACAGATCCGCGCTNNNNNNNNNNNT	52	CCTCTGCGAAGACAGATCCGCGCT
53	GACCACGCACACGTNNNNNNNNNNNT	53	CCTTCCTAGCGACCACGCACACGT
54	GACGCACTGACCGTNNNNNNNNNNNT	54	CGCTTAAGGCGACGCACTGACCGT
55	GCCGTGGACCTAGTNNNNNNNNNNNT	55	CGGACAGAGAGCCGTGGACCTAGT
56	GCAGCGAAGTCGCTNNNNNNNNNNNT	56	CGGTATTAGCGCAGCGAAGTCGCT
57	GCCACACGTGTGCTNNNNNNNNNNNT	57	CGTAGAAGACGCCACACGTGTGCT
58	GCGAACTCACGGCTNNNNNNNNNNNT	58	CGTCCGACTTGCGAACTCACGGCT
59	GCGTGCGACAAGCTNNNNNNNNNNNT	59	CTACTCACTAGCGTGCGACAAGCT
60	GCCGCCTCAGTCATNNNNNNNNNNNT	60	CTCCACTGAAGCCGCCTCAGTCAT
61	GGAACGCCATCGGTNNNNNNNNNNNT	61	CTCCTATTGTGGAACGCCATCGGT
62	GGACGTTCGGGCTTNNNNNNNNNNNT	62	CTGGAAGTAAGGACGTTCGGGCTT
63	GGACTCACCTCCGTNNNNNNNNNNNT	63	CTGGATTGACGGACTCACCTCCGT
64	GGAGCCGTAGCACTNNNNNNNNNNNT	64	CTTGGAGAACGGAGCCGTAGCACT
65	GGATACGCGTACGGNNNNNNNNNNNT	65	GAATCGCCATGGATACGCGTACGG
66	GGCAACACACCGAGNNNNNNNNNNNT	66	GAATGCTGAAGGCAACACACCGAG
67	GGTCGGGATAGACGNNNNNNNNNNNT	67	GAATTACGGCGGTCGGGATAGACG
68	GGTCCCGCTCATCTNNNNNNNNNNNT	68	GACGGCCTATGGTCCCGCTCATCT
69	GGTCTGTCTCGTCGNNNNNNNNNNNT	69	GATATAGCTCGGTCTGTCTCGTCG
70	GTCGGCAGAGGTGTNNNNNNNNNNNT	70	GATCCAAGAAGTCGGCAGAGGTGT
71	GTCGCGGGTAGAGTNNNNNNNNNNNT	71	GATGATGTTGGTCGCGGGTAGAGT
72	GTCTACCGCGACGTNNNNNNNNNNNT	72	GATGTGACAGGTCTACCGCGACGT
73	GTGACCGACACCGTNNNNNNNNNNNT	73	GCAAGATGTAGTGACCGACACCGT
74	GTGCACGACCACCTNNNNNNNNNNNT	74	GCACCTCTTGGTGCACGACCACCT
75	GTGCATCAGCGGGTNNNNNNNNNNNT	75	GCCATGAGAAGTGCATCAGCGGGT
76	GTGTAGCGGGCTCTNNNNNNNNNNNT	76	GCCGTTCCTTGTGTAGCGGGCTCT
77	TCGCGCCATGCCTTNNNNNNNNNNNT	77	GGAATAAGCATCGCGCCATGCCTT
78	TACAGCGCGGTGCTNNNNNNNNNNNT	78	GGAATTCCAATACAGCGCGGTGCT
79	TACGACCGCTGCACNNNNNNNNNNNT	79	GGCATATACCTACGACCGCTGCAC
80	TAGCTAGCGGTGCCNNNNNNNNNNNT	80	GGCGAAGTATTAGCTAGCGGTGCC
81	CGTCCGGACACATCNNNNNNNNNNNT	81	GGCTGTCTTACGTCCGGACACATC
82	TATGCTCGACCGCCNNNNNNNNNNNT	82	GGTCTTACATTATGCTCGACCGCC
83	TCACCACACCTCGCNNNNNNNNNNNT	83	GGTTCCTTAATCACCACACCTCGC
84	TCAGCCGGCATACCNNNNNNNNNNNT	84	GTGATTCTCATCAGCCGGCATACC
85	TCATGGCCGTACGCNNNNNNNNNNNT	85	GTTCATTGCCTCATGGCCGTACGC
86	TCCAGGCGGTAGTCNNNNNNNNNNNT	86	GTTGAGCGTATCCAGGCGGTAGTC
87	TCCTCTGATCGGGCNNNNNNNNNNNT	87	GTTGTATGCTTCCTCTGATCGGGC
88	TCCTGCAACGCCTCNNNNNNNNNNNT	88	TAAGCCTCTTTCCTGCAACGCCTC
89	TCGTCGTACACCGCNNNNNNNNNNNT	89	TAGTCCGCTGTCGTCGTACACCGC
90	TCTCTCCAGGCGACNNNNNNNNNNNT	90	TAGTGCAGTCTCTCTCCAGGCGAC
91	TCACGCGGACGATCNNNNNNNNNNNT	91	TATCGTTACGTCACGCGGACGATC
92	TGAGTCCCGGAGACNNNNNNNNNNNT	92	TCCTCTAGTATGAGTCCCGGAGAC
93	TGCCCGTCTGTCTCNNNNNNNNNNNT	93	TCGAGAGAGCTGCCCGTCTGTCTC
94	TGGCCGGCTACTACNNNNNNNNNNNT	94	TGAGGAGTGGTGGCCGGCTACTAC
95	TGGTGAGGCGCTACNNNNNNNNNNNT	95	TGGAATGGAGTGGTGAGGCGCTAC
96	TGTGGTCCTGGCTCNNNNNNNNNNNT	96	TGTTACCTCATGTGGTCCTGGCTC

Figure 1: Schematic outline of the VANA-based metagenomics method. Top-left: conversion by random priming of viral ssRNA or dsRNA to sequence-ready cDNA, including a reverse transcription step followed by a Klenow reaction step. Top-right: conversion by random priming of viral DNA (e.g. circular ssDNA) to sequence-ready cDNA, including a Klenow reaction step (strand displacement amplification). Bottom: Double-stranded DNA are amplified using one PCR multiplex identifier primer, which yields amplicons that are all tagged at both extremities with the same multiplex identifier primer.

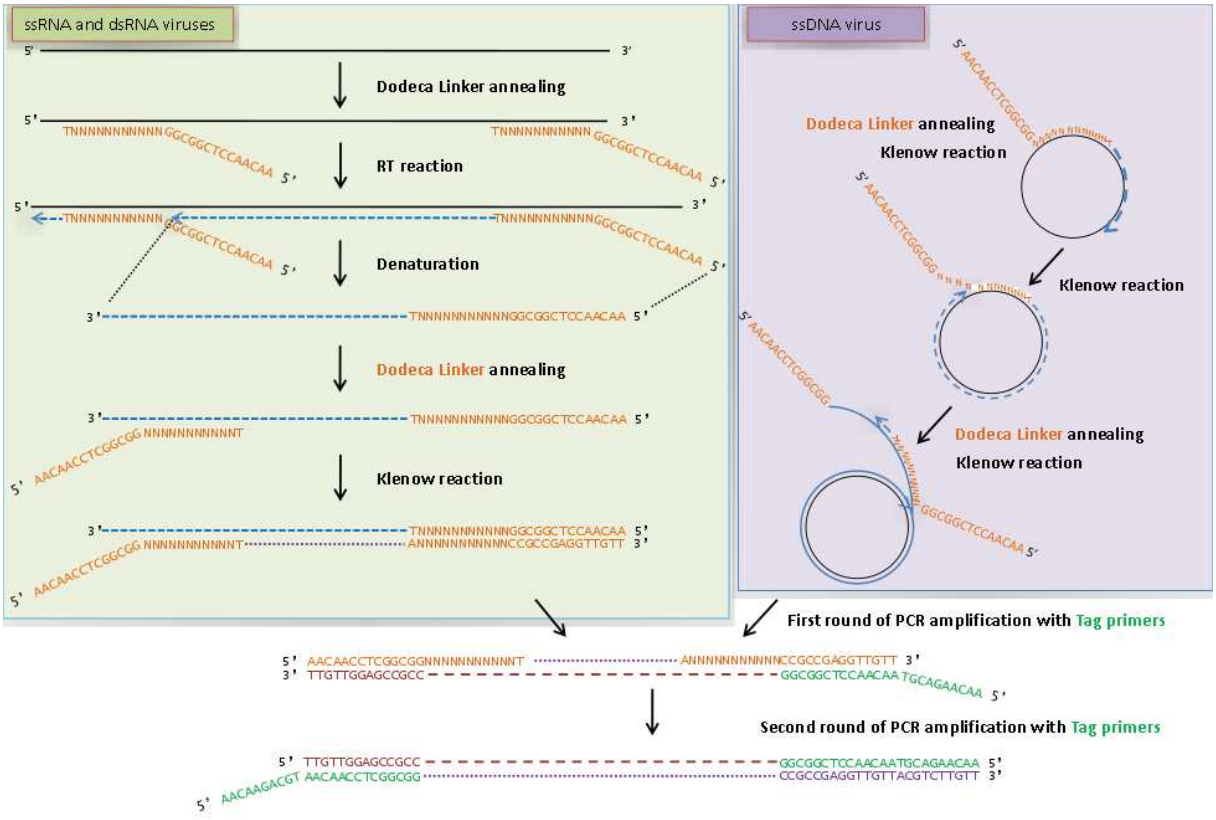


Figure 2: Agarose gel analysis of PCR amplicons obtained from arthropod and plant samples using the VANA-based metagenomics approach. Lane 1 and 25: 1 kb plus ladder size marker (Invitrogen); lane 2 to 24: VANA-based metagenomics amplicons smears.

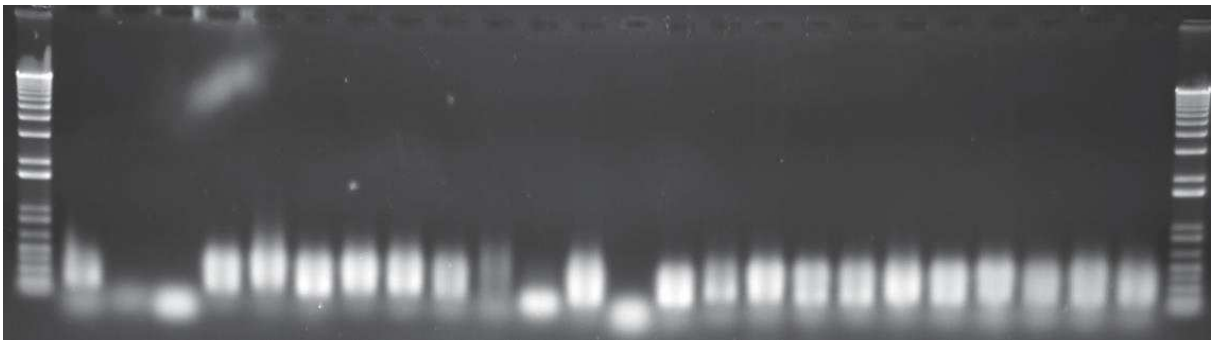


Figure 3: General workflow of the bioinformatics analysis of High Throughput Sequencing data.

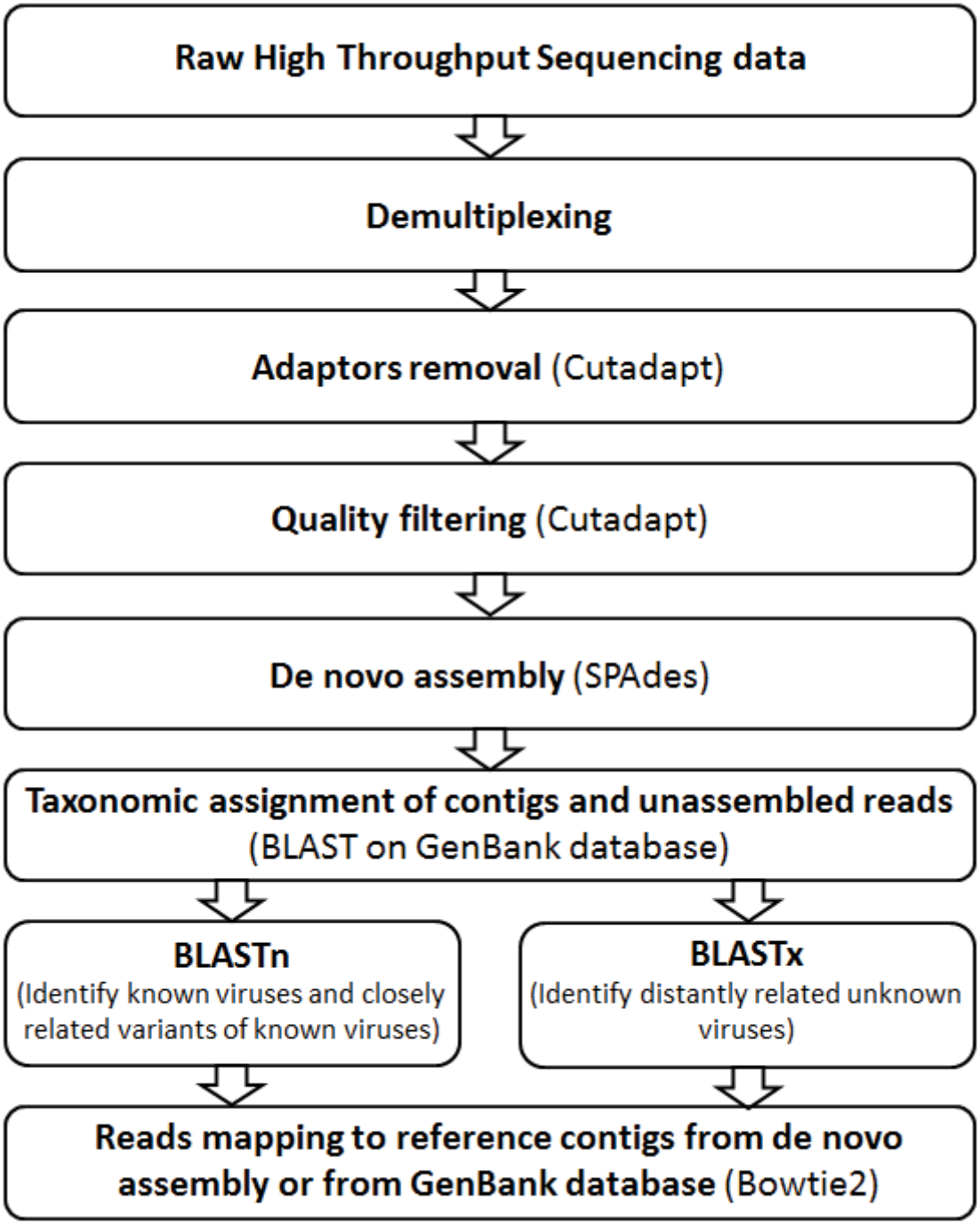


Figure 4: An example of viral genome reconstitution using the protocol presented here. This viral contigs was created using SPAdes 3.6.2 with standard parameters. It represented a nearly complete genome of 9654 nucleotides in length. BLASTn analysis showed that this viral contig shared 98% of nucleotide identity with the Aphid lethal paralysis virus (*Dicitroviridae*, accession number KX884276). 119474 reads from an *Acyrthosiphon pisum* virome were mapped against this viral contig, using Bowtie 2.1.0 options end-to-end very sensitive, which corresponded to an average coverage of 400X.

Acknowledgments

S. F. was supported by a scholarship from the National Institute of Agronomical Research (INRA) and from the University of Montpellier (UM). P.R. has received an EU grant FP7-PEOPLE-2013-IOF (N°PIOF-GA-2013-622571).

Bilan et perspectives

Notre protocole complet de préparation et d'analyse de viromes fonctionne sur les échantillons d'arthropodes et de plantes testés dans le cadre d'un séquençage par la plateforme Illumina. Il permet d'obtenir des contigs pouvant représenter la quasi-totalité de génomes viraux appartenant à l'ensemble de la classification de Baltimore. De plus, il permet le multiplexage de 96 échantillons, ce qui induit une réduction des coûts de séquençage.

Cependant, ce protocole présente plusieurs limitations.

- Étant donné que l'ensemble des virus ne possède pas de marqueur moléculaire commun, il est nécessaire d'amplifier de manière aléatoire leur génome afin d'étudier leur diversité. Or, avant cette amplification, il convient de réduire au maximum la proportion d'acides nucléiques d'origine cellulaire présente dans les échantillons. Cette réduction de la quantité de contaminants passe par la digestion des acides nucléiques non-encapsidés ainsi que par la concentration des particules virales par filtration suivie d'une ultracentrifugation. Si une réduction de la taille du filtre permet de réduire la proportion de contaminants cellulaires, elle réduit conjointement la proportion de virus de grande taille (comme les représentants de la famille des *Mimiviridae*) présents dans les viromes (Halary *et al.*, 2016). Le biais de taille induit par le filtre utilisé ne permet donc pas d'étudier de façon exhaustive la diversité des communautés virale. Enfin, il a été reporté que l'utilisation de certains linkers-tags, peut induire des biais d'amplification, compliquant les analyses quantitatives (Roossinck *et al.*, 2010).

- Il existe également un impact de l'amplification aléatoire sur la représentativité des viromes obtenus dans le cas d'échantillons multiplexés. En effet, avant de réaliser le pool des produits de PCR purifiés des échantillons, le succès de l'amplification a été testé, de manière relative, par visualisation de l'intensité des trainées d'ADN (smears) après migration des produits de PCR. La quantité d'ADN présente après l'amplification aléatoire peut donc être différente entre les échantillons. Cependant, le pool des produits PCR a été réalisé en utilisant un volume identique de produits de PCR par échantillon. Cette méthode privilégie donc, en terme de quantité de reads obtenus après séquençage, les échantillons dont l'amplification a été la plus efficace (Harris *et al.*, 2010). Afin d'éliminer ce biais, il est nécessaire de procéder au pool d'une quantité égale d'ADN entre les échantillons.

- Nous avons enfin remarqué qu'il existe un impact de la technologie de séquençage utilisée sur la qualité du démultiplexage. En effet, le séquençage par MiSeq (générant des séquences de 250pb et de 300pb dans le cas de notre étude) permet de réattribuer de façon sûre 60% des reads à leur échantillon d'origine ; alors que la technologie HiSeq (générant des séquences de 150pb) n'en réattribue que 40%. Cette incertitude dans le démultiplexage d'une grande partie des reads est inhérente au séquençage par haut-débit, notamment du fait que les extrémités des reads sont moins bien séquencées que leur centre. Cela peut causer des erreurs dans les séquences des linkers et des tags situés aux extrémités des reads qui ne seront donc plus reconnus par les scripts permettant leur démultiplexage. Il est cependant important de préciser le fait que des reads ne soient pas demultiplexables n'empêche pas leur utilisation lors de l'étape d'assemblage *de novo* permettant la création de contigs viraux. Enfin, il est nécessaire de relativiser la balance entre efficacité du démultiplexage et profondeur de séquençage en fonction de la technologie de séquençage utilisée et du nombre d'échantillons multiplexés. En effet, dans le cadre du traitement d'un grand nombre d'échantillons, il serait préférable d'utiliser le séquençage par HiSeq, plus profond que le MiSeq, afin de permettre à chaque échantillon de posséder une couverture suffisante permettant l'analyse convenable de leur virome (Genohub-ngs-instrument-guide).

Il est également à noter que notre technique de multiplexage influence le rendement du séquençage par la technologie Illumina. En effet, le fait que les séquences d'ADN possèdent des séquences identiques à leurs deux extrémités auraient une influence négative sur la génération des clusters, ce qui induirait un séquençage moins efficace concernant la quantité de reads produits (ingénieur de la société Genewiz, communication personnelle).

Enfin, les technologies de séquençage évoluent rapidement, si bien que les technologies de séquençage par MiSeq ou HiSeq, donnent aujourd'hui au maximum respectivement jusqu'à 25 000 000 de reads de 600pb de longueur et 375 000 000 reads de 300pb, sont déjà dépassées par la technologie NovaSeq de la plateforme Illumina. Cette dernière permet d'obtenir près de 10 000 000 000 reads de 300pb, ce qui représente un rendement 25 fois supérieur à celle de HiSeq (Genohub-ngs-instrument-guide). Ce gain de profondeur permet aux études de métagénomique virale de traiter un grand nombre d'échantillons en parallèle, ce qui est important dans le cadre de l'étude comparative de viromes.

La matière noire, limitation de l'étude des viromes

Une des limitations principales rencontrées dans l'étude de la majorité des viromes obtenus par métagénomique virale est le pourcentage élevé de séquences ne pouvant être taxonomiquement attribuées (Rosario et Breitbart, 2011). Nommées « matière noire » (dark matter), ces séquences peuvent appartenir à des virus inconnus ou à des contaminants d'origine cellulaire. Elles représentent une part significative, voire la majorité, des séquences présentes dans les viromes (Krishnamurthy et Wang, 2017; Rosario et Breitbart, 2011). Cette matière noire représente un frein important à notre compréhension du monde viral. D'une part, elle représente une source de diversité génétique inexploitable ; et d'autre part, elle peut amener à surestimer la diversité virale présente dans les échantillons testés.

La présence de matière noire dans les viromes est explicable par deux faits : notre connaissance parcellaire de la diversité des virus couplée à un manque d'efficacité des outils d'attribution taxonomique.

- En effet, l'attribution taxonomique des séquences issues des viromes repose majoritairement sur leur comparaison avec les séquences virales présentes dans les bases de données. Or, de par nos connaissances fragmentaires de la diversité des virus, les bases de données virales sont incomplètes. Par exemple, en janvier 2015, 1531 virus entièrement séquencés étaient disponibles dans la base de données NCBI RefSeq, et la majorité d'entre eux (86%) dérivent seulement de 3 phyla hôtes sur 61 (Roux, *et al.*, 2015).

Des études ont été récemment effectuées dans le but de compléter les bases de données virales en y incorporant des séquences de virus infectant des organismes peu étudiés comme les archées et certains groupes de bactéries. Ces études ont été réalisées par métaprotéomique avec annotation de protéines virales à grande échelle à partir d'un virome marin (Brum *et al.*, 2016), par séquençage de cellule unique (single cell genomics) (Rinke *et al.*, 2013), ou par fouille de bases de données (Roux *et al.*, 2015).

- Une approche complémentaire consiste à améliorer l'efficacité de l'attribution taxonomique en développant ainsi qu'en optimisant des outils d'attribution taxonomique. Les outils d'attribution taxonomique regroupent majoritairement ceux se basant sur des alignements (BLAST (Altschul *et al.*, 1990), USEARCH (Edgar, 2010)), sur l'HMM (profile Hidden Markov Models) (Remmert *et al.*, 2012; Skewes-Cox *et al.*, 2014) et l'analyse de la composition des k-mers (motif nucléotidique de longueur k) (Ren *et al.*, 2017; Rosen *et al.*, 2011). Enfin, l'analyse de signatures oligonucléotidiques propres aux virus d'un environnement donné (Willner *et al.*, 2009), et l'analyse de la circularisation des contigs ou de la structure des cadres de lecture ouverts (ORFs) (Labonté et Suttle, 2013) sont également possibles. Il est également possible de combiner plusieurs de ces approches afin d'améliorer l'attribution taxonomique des séquences issues des viromes.

Le BLAST est l'outil le plus fréquemment utilisé en attribution taxonomique. Il a été montré que cette méthode permet d'assigner de manière plus efficaces de grandes séquences que de courtes séquences (Fancello *et al.*, 2012). L'utilisation de l'assemblage *de novo* couplée au mapping permettrait d'augmenter la proportion de reads taxonomiquement attribués par des analyses sans étape de mapping avant l'attribution taxonomique (Krishnamurthy et Wang, 2017). Cependant, l'efficacité de cette approche sur la réduction de la proportion de matière noire n'a pas été clairement quantifiée.

Le but de l'étude qui suit était de quantifier l'efficacité de l'assemblage *de novo* et du mapping avant l'attribution taxonomique par BLAST sur la réduction de la proportion de matière noire. Pour ce faire, deux protocoles, intégrant ou non l'assemblage *de novo* des reads en contigs couplée à une étape de mapping précédant l'attribution taxonomique des contigs par BLAST, ont été testés comparativement sur des viromes provenant d'échantillons diversifiés. Ils ont été testés sur huit de nos viromes d'arthropodes et de plantes, ainsi que sur sept viromes réalisées lors de cinq études de métagénomique virale indépendantes, et provenant d'échantillons de sols, d'eau, de fèces humains et de moustiques. Cette étude montre que l'ajout d'une étape d'assemblage *de novo* et de mapping en amont de l'étape d'attribution taxonomique permet de réduire de près d'un facteur cinq la quantité de matière noire présente dans l'ensemble des jeux de données testés.

Article de recherche 2

Increase in taxonomic assignment efficiency of viral reads in metagenomic studies

Sarah François[1,2], Denis Filloux[3], Marie Frayssinet[2], Philippe Roumagnac[3], Mylène Ogliastro[2*], Rémy Froissart[4*]

[1] Université de Montpellier, UMR 1333 DGIMI « Diversité, Génomes et Interactions Microorganismes-Insectes », place Eugène-Bataillon, 34095 Montpellier cedex 5, France

[2] Laboratoire « Diversité, Génomes et Interactions Microorganismes Insectes » (DGIMI) UMR 1333, INRA, Université Montpellier, 34095 Montpellier, France.

[3] Laboratoire « Biologie et Génétique des Interactions Plante-Parasite » UMR BGPI, CIRAD-INRA-SupAgro, Campus International de Montferrier-Baillarguet, Montpellier Cedex-5, France.

[4] Laboratoire « Maladies infectieuses et vecteurs: écologie, génétique, évolution et contrôle » (MIVEGEC) UMR 5290, CNRS, IRD, Université Montpellier, 911 avenue Agropolis, 34394 Montpellier, France.

Publié le 14 novembre 2017 dans **Virus Research** (244:230-234)

[#] **corresponding authors**: remy.froissart@cnrs.fr, marie-helene.ogliastro@inra.fr

Running Title: Illuminating viral dark matter

Virus Research 244 (2018) 230–234

Contents lists available at ScienceDirect

Virus Research

journal homepage: www.elsevier.com/locate/virusres

Increase in taxonomic assignment efficiency of viral reads in metagenomic studies

S. François[a], D. Filloux[b], M. Frayssinet[a], P. Roumagnac[b], D.P. Martin[c], M. Ogliastro[a,1], R. Froissart[d,*,1]

[a] INRA-Université de Montpellier UMR DGIMI 34095 Montpellier, France
[b] CIRAD-INRA-Supagro, UMR BGPI, Campus International de Baillarguet, 34398 Montpellier, France
[c] Computational Biology Group, Institute of Infectious Disease and Molecular Medicine, Faculty of Health Sciences, University of Cape Town, Observatory, South Africa
[d] CNRS-IRD-Université de Montpellier, UMR MIVEGEC, 911 avenue Agropolis, 34394, Montpellier, France

ARTICLE INFO

Keywords:
Dark matter
Viral metagenomics
BLAST
Mapping

ABSTRACT

Metagenomics studies have revolutionized the field of biology by revealing the presence of many previously unisolated and uncultured micro-organisms. However, one of the main problems encountered in metagenomic studies is the high percentage of sequences that cannot be assigned taxonomically using commonly used similarity-based approaches (e.g. BLAST or HMM). These unassigned sequences are allegorically called « dark matter » in the metagenomic literature and are often referred to as being derived from new or unknown organisms. Here, based on published and original metagenomic datasets coming from virus-like particle enriched samples, we present and quantify the improvement of viral taxonomic assignment that is achievable with a new similarity-based approach. Indeed, prior to any use of similarity based taxonomic assignment methods, we propose assembling contigs from short reads as is currently routinely done in metagenomic studies, but then to further map unassembled reads to the assembled contigs. This additional mapping step increases significantly the proportions of taxonomically assignable sequence reads from a variety –plant, insect and environmental (estuary, lakes, soil, feces) – of virome studies.

1. Introduction

The advent of high throughput sequencing has enabled the cataloguing and enumeration of microbial species without a priori information on their life cycles. When specifically focusing on viruses, this so-called viral metagenomic approach, has so-far revealed the extraordinary diversity and prevalence of viruses in aquatic and terrestrial ecosystems, highlighting the key contributions of these microbes to all ecosystems on Earth (Brum and Sullivan, 2015; Mokili et al., 2012; Suttle, 2007).

One simple but important insight yielded by these astonishing discoveries is that we probably currently know far less than 1% of all viral species that are circulating on Earth (Anthony et al., 2013; Mokili et al., 2012). It is sobering to consider that despite the large numbers of viromes that have been examined over the past 20 years, almost every new viromics project yields large numbers of sequences that have no significant degree of similarity with those referenced in databases. These sequences are often referred to as “dark matter”. Our inability to properly categorize the latter sequences has the potential to strongly bias our view of both the actual diversity of viruses in a given environment and their ecological roles (Krishnamurthy and Wang, 2017; Roossinck et al., 2015; Rosario and Breitbart, 2011).

When attempting to characterize any virome from metagenomic datasets, researchers face two main challenges: i) purifying viral genomes present in heterogeneous materials or biological tissues without introducing biases due to technical processes and ii) accurately assign sequence reads. Whereas solutions to the first of these challenges will vary from environment to environment, the second challenge could be met both with improved computational methods that are capable of accounting for compositionally biased databases, and by vastly increasing the diversity of viral genome sequences within public databases. For instance in most viral metagenomic projects, only approximately 10–20% of sequence reads can be confidently attributed to viruses and, in most cases, the remaining sequence reads are treated as unanalyzable dark matter (Krishnamurthy and Wang, 2017; Rosario and Breitbart, 2011).

* Corresponding author.
E-mail address: remy.froissart@cnrs.fr (R. Froissart).
[1] These authors contributed equally to the work.

https://doi.org/10.1016/j.virusres.2017.11.011
Received 27 September 2017; Received in revised form 10 November 2017; Accepted 10 November 2017
Available online 14 November 2017
0168-1702/ © 2017 Elsevier B.V. All rights reserved.

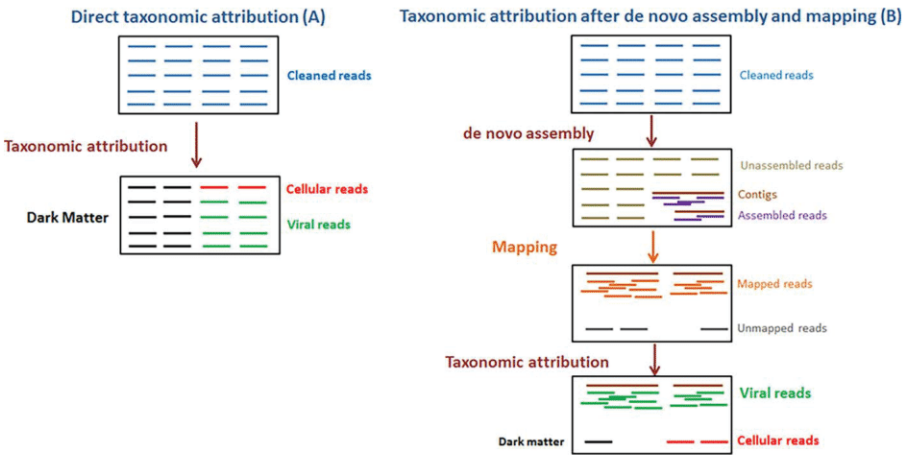


Fig. 1. Comparative analyses of two BLASTx-based methods of taxonomic assignment. A: Direct taxonomic assignment after the classical BLASTx-based approach; B: the *de novo* assembly, mapping and BLASTx approach (AM-BLASTx).

In viral metagenomic studies, the classical bioinformatical workflow consists of *de novo* assembling contigs from short reads generated by high throughput sequencing and then performing homology inferences via alignments of sequences (both reads and contigs) to reference databases using a tool such as BLAST (Allander et al., 2001; Angly et al., 2006; Breitbart et al., 2002). However, this method usually yields low quality taxonomic assignments due, at least in part, to both the length of sequence reads generally being < 500nts, and the low degrees of sequence identity that are commonly shared between query sequences and the virus genomic sequences present in public databases (Tangherlini et al., 2016). Moreover, the classical BLAST workflow most often leads to a high number of reads that cannot be attributed with high confidence to related sequences and are thus considered as unknown sequences.

To decrease the amount of this dark matter, it has been recently proposed to integrate a new step in the computational workflow: a recruitment process consisting of the mapping of unassembled short sequence reads onto assembled contigs prior BLASTx requests (Krishnamurthy and Wang, 2017), a workflow that we will referred to as assembly-mapping-BLAST (AM-BLAST for short) as opposed to the classical BLAST workflow. Although this methodology is used in viral metagenomics (Cotten et al., 2014), no comparative study has ever been made to evaluate how efficiently the use of AM-BLAST reduces the amount of dark matter relative to the classical BLAST workflow.

Alternatives to BLAST have been developed to improve taxonomic assignments of query sequences being compared to a database of reference sequences. One of the most used alternative approaches involves a hidden Markov model (HMM) based classifier where position-specific information on nucleotide variation across a set of related sequences is taken into account when determining whether there are statistically significant matches within a database to query sequences. This approach outperformed BLAST when attempting to find database matches to divergent viral sequences, although it remained less accurate than BLAST with respect to taxonomic assignment (Fancello et al., 2012; Remmert et al., 2012; Skewes-Cox et al., 2014).

The aim of the present study was to quantify improvement in the taxonomic assignment of viral sequences after the use of AM-BLAST relative to classical-BLAST workflows. We thus compared the number of unassigned reads after running these two workflows on fifteen datasets consisting of samples enriched for virus-like particles (VLP). Our results indicate that the AM-BLAST workflow reduced significantly the number of unassigned viral reads compared to the classical-BLAST workflow.

2. Materials and methods

2.1. Sampling, virome preparation and sequencing

Three insect species (*Hypera postica, Acyrthosiphon pisum* and *Coccinella septempunctata*) and one plant species (*Medicago sativa*) were collected in the Montpellier area of Southern France (domaine de Restinclières, Prades le Lez, France, N 43°42′54.362" EO 3°51′31.749"); for each species, several individuals were pooled and constituted one sample. Samples were stored at −80 °C without addition of any preservative solutions. One gram of insect or plant material was processed using a virion-associated nucleic acids (VANA) based metagenomic approach to screen for the presence of viruses (Palanga et al., 2016). Amplified and tagged DNA products of the VANA approach were sequenced using an Illumina platform (MiSeq sequencing: 2 × 300 nt paired-end sequencing with V3 chemistry, Beckman Coulter Genomics, USA).

2.2. Bioinformatic analysis: virome cleaning, read assembly, taxonomic assignment and clustering

Raw reads were first demultiplexed using agrep (Wu and Manber, 1992). Illumina adaptors were removed and we selected reads based on their quality (≥q30 and length elimination of reads < 45 nt) using Cutadapt 1.9 (Martin, 2011). The remaining reads will be hereafter referred to as "cleaned reads". Paired cleaned reads were merged using FLASH 1.2.11 (Magoc and Salzberg, 2011). Then, random subsets of two hundred thousand cleaned reads per virome were used for all the following steps.

First, we performed the classical-BLASTx workflow which involved taxonomically assigning reads using BLASTx searches against the non-redundant GenBank viral protein sequences database for taxonomic attribution (e-value cutoff of $< 10^{-3}$) (Altschul et al., 1990) on 200,000 randomly chosen "cleaned reads" (Fig. 1A).

Second, we performed the AM-BLAST workflow which involved subjecting the cleaned reads to assembly using SPAdes (different kmer sizes: 21, 33, 55, 77, 125) (Bankevich et al., 2012). Contigs and unassembled reads were then assembled using CAP3 with default parameters (Huang, 1999). It is to notice that CAP3 was only used to recruit reads; it should not be used for identification of genomes because this software can result in creation of chimaeras. Mapping of the remaining reads both (i) onto the new contigs obtained after *de novo* assembly and (ii) to the remaining unassembled reads was performed using Bowtie 2.1.0 (using the local and very sensitive option that empirically recruit reads having > 85% of nucleotide identity with corresponding contigs) (Langmead, 2010; Toland et al., 2013). All contigs and unassembled reads were then subjected to BLASTx searches against the non-redundant GenBank viral protein sequences database for taxonomic attribution (e-value cutoff of $< 10^{-3}$) (Altschul et al., 1990) (the whole procedure is summarized in Fig. 1B).

To obtain an overview of genetic diversity across all the metagenomic datasets, 10,000 reads were randomly chosen (3 replicates) and subjected to BLASTx searches against the NCBI non-redundant protein

Table 1
Recapitulative history of the original (our data) and already published metagenomic datasets that were used in this study.

Origin of samples	Technic of virus-like particles enrichment	Technic of sequencing	Cleaned reads median length	Total read number	Reference
Insects 1	0.45 µm filtration, DNA and RNA extraction, random PCR amplification	MiSeq Illumina	219	324 246	Our data
Insects 2			230	399 954	
Insects 3			228	428 089	
Insects 4			217	611 722	
Insects 5			225	203 015	
Insects 6			212	343 955	
Plants 1			195	224 890	
Plants 2			245	440 408	
Estuary	0.2 µm filtration, DNA extraction, RCA amplification	454 pyrosequensing	105	294 068	McDaniel et al. (2008)
Lake Bourget			471	593 084	Roux et al. (2012)
Lake Pavin			445	649 290	
Antarctic open soil		MiSeq Illumina	250	870 687	Zablocki et al. (2014)
Antarctic hypolith			250	1 057 555	
Human feces		454 pyrosequensing	466	504 646	Kim et al. (2011)
Mosquitoes			104	336 760	Ng et al. (2011)

sequences database.

Seven publicly available metagenomic datasets were also analyzed in this study, originating from five independent datasets after enrichment for virus-like particles from mosquitoes (Ng et al., 2011), a human fecal sample (Kim et al., 2011), an estuary sample (McDaniel et al., 2008), two lake samples (Roux et al., 2012), and two Antarctic ecosystem samples (Zablocki et al., 2014) (Table 1). These seven datasets were *de novo* assembled and analyzed as described above.

3. Results

The aim of our study was to compare the efficiency with which classical-BLAST (Fig. 1A) and AM-BLAST (Fig. 1B) workflows taxonomically assign reads from metagenomic sequencing datasets. These datasets were obtained from samples of various origins and enriched for virus-like particles using different procedures (Table 1): (i) eight insects and plants processed for the purpose of the present study (hereafter referred to as viromes 1–8) and (ii) seven datasets from published studies originating from environmental and insect samples (hereafter referred to as viromes 9–15) (Table 1). The viromes represented by these two sets will be hereafter referred to as original and published viromes, respectively.

The classical-BLAST workflow was able to assign, with high level of confidence (according to E-value of BLASTx, see M&M section), between 59% and only 1.3% of reads from the original viromes (Insects 6 and Insects 4, respectively) and between 7.5% and 0.15% of reads for the published viromes (Lake Bourget and Hypolith, respectively), in agreement with published results (Fig. 2). The AM-BLAST workflow on the other hand, allowed the assignment of 89.4% and 42.5% of reads for the original viromes (Insects 4 and Insects 5, respectively) and between 18.6% and 1.4% of reads for the published viromes (Lake Pavin and Estuary, respectively) (Fig. 2). The AM-BLAST workflow yielded a significant improvement in overall taxonomic assignment efficiency ($P = 6.1 \times 10^{-5}$, Wilcoxon comparison test) (Fig. 2, Supplemental Table 1). This improvement was particularly notable for insect virome 4 where the AM-BLAST workflow yielded a 70-fold improvement in the proportion of taxonomically assignable reads.

Proportions of assignable reads varied markedly between the analyzed viromes. For the original datasets an average of 21% of reads were assignable by classical-BLAST and 62% by AM-BLAST. For the published datasets an average of only 3% of reads were assignable by classical-BLAST and 13% by AM-BLAST (Fig. 2 and Supplemental Table 1). On the one hand, the aquatic, marine and fecal environmental viromes were dominated by large dsDNA bacteriophages (> 250 kb) belonging to the *Myoviridae* and *Siphoviridae* families. On the other hand, insect and plant viromes were dominated by small RNA and DNA viruses (< 10 kb) belonging to the *Iflaviridae*, *Dicistroviridae*, *Parvoviridae*, *Amalgaviridae* and *Partitiviridae* families (Supplemental Table 2).

In order to assign the remaining unclassified reads to cellular origin or to dark matter, we taxonomically assigned a subset of ten thousand reads that were randomly sampled from each dataset (i.e. 5% of the total number of reads per dataset) using BLASTx. Despite the datasets all being derived from samples that were processed to enrich for viral-like particles, from 1% to 55% of the reads in both the original and published datasets were most likely of cellular origin (Fig. 3). On the one hand, for the original viromes, up to 22% and 53% of the reads

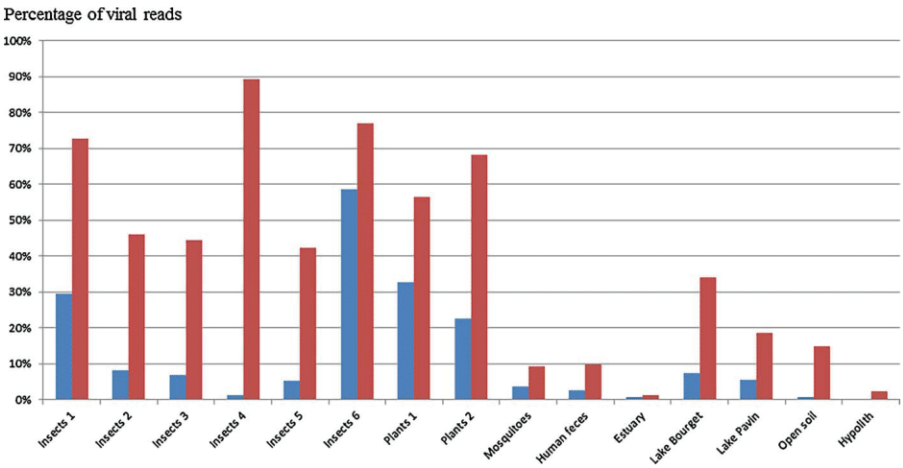


Fig. 2. Comparison on the performance of BLASTx searches. BLASTx searches were performed on raw reads (blue - classical-BLAST workflow) and after mapping on contigs (red - AM-BLAST workflow).

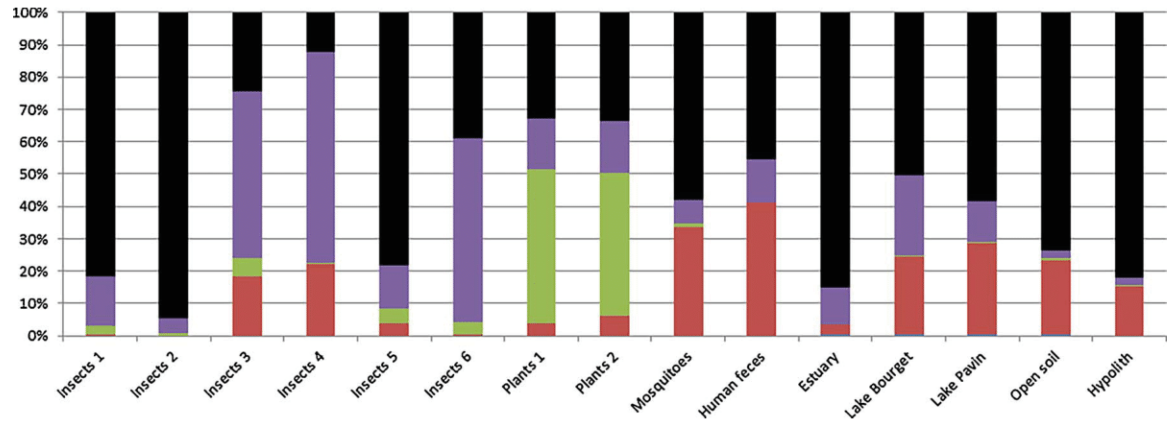


Fig. 3. Average proportions of reads in each virome according to their taxonomic assignment using BLASTx. Query read sequences were assigned to viral (purple), bacterial (red), eukaryotic (green) and unassigned (black) sequences available in online databases. Each analysis has been performed three times on a random sample of 10,000 reads from each dataset.

were respectively assigned to bacteria and eukaryotes. From 47% and 53% of reads from the two plant viromes were assignable to plant genomic sequences, while 18% and 22% of the reads from two of the insect viromes (viromes 3 and 4 from the aphid *A. pisum*) were likely derived from *Candidatus Hamiltonella defensa*, an aphid's endosymbiotic bacteria (Fig. 3). On the other hand, for the published viromes, 3% to 39% and 0–1% of reads were assigned to bacterial and eukaryotic organisms respectively, in agreement with published results (Fig. 3, Supplemental Table 3). Specifically, lake datasets (viromes 11–14), contained similar proportions (about 25%) of bacterial and bacteriophage sequences, indicating the presence of bacteria, bacteriophage particles and prophage nucleic acids as already reported (Enault et al., 2016; Roux et al., 2013). Moreover, the human feces and soil viromes contained a higher proportion of reads assigned to bacteria (from 15% to 39%) than those from other sources (Fig. 3).

4. Discussion

In this study, we propose a modification of the classical BLASTx-based workflow that improves the taxonomic assignment of sequences from metagenomic virome studies. Based on the statement that increasing the length of query sequences could improve the accuracy with which they could be taxonomically assigned using BLASTx, we introduced a recruitment step of remapping unassembled reads onto assembled contigs prior to BLASTx searches (a workflow that we called "assembly-mapping BLAST" or AM-BLAST for short) and tested this on viral metagenomic datasets. These datasets were obtained after different technical procedures, both prior to sequencing (i.e. use of rolling-circle amplification or random PCR amplification) and during the sequencing process (i.e. MiSeq Illumina or 454 Pyrosequencing; Table 1). We found that, when applied to each datasets, the AM-BLAST workflow systematically and substantially increased the numbers of virus-derived sequences that could be taxonomically assigned relative to the numbers that were assignable using the classical-BLAST workflow. Analyses made on fifteen metagenomic datasets lead to an average five-fold increase in the number of assignable reads.

Our analyses thus revealed that one major parameter to improve the performance of BLASTx-based approaches for taxonomically assigning viral reads is likely the lengths of the sequences that will be analyzed by BLASTx. Indeed, viral genomes are more variable than those of cellular organisms because of high mutation rates, large population sizes and short generation times, so longer reads and contigs decrease the impact of point mutations that decrease the degrees of similarity between query and reference sequences within the database that is being searched by BLAST or HMM-based approaches. The lengths of query sequences can be increased both by computational processing of the sequence data prior to performing blast searches (as is done in the AM-BLAST workflow), and by technical procedures during virus-like particle enrichment. Specifically, it is desirable to lengthen the query sequences, either by using sequencing technologies that enable long reads (such as 454 that is no longer used or Pacific BioScience) or by increasing sequencing depth (such as with Illumina) so as to enable the assembly of longer contigs. In fact, with simulated metagenomic datasets, a positive correlation has been found between sequencing depth and the proportions of reads that could be taxonomically assigned (García-López et al., 2015). Interestingly, our analyses did not reveal differences in the degrees of taxonomic improvement between studies using 454 and Illumina sequencing technologies, suggesting that large sequencing depth can compensate shorter read lengths.

Our analyses also revealed that viral metagenomic dataset obtained for the purpose of this study from arthropods and plants (our so called "original viromes") seemed to be dominated by small viruses (< 10 kb), while published environmental viromes contained a high number of reads assigned to prophages and genomic bacterial DNA. The generality of such differential viral communities according to different environments is, however, questionable because only very few studies have reported insect and plant viromes (Junglen and Drosten, 2013) and we can thus not compare our results with those of others. Moreover, technical procedures during the preparation of the original and published viromes differed in that the latter were obtained by rolling-circle amplification, a technique known to induce amplification biases toward circular genomes, while the former viromes were obtained after random PCR, a technique that is not known to have this bias.

Altogether, the AM-BLAST workflow represents a simple and rapid way to improve the taxonomic assignment of viral sequences from metagenomic datasets independently of the origin of the samples. Our results indicate that the proportion of unassigned reads (i.e. the "dark matter") in virome datasets can be significantly reduced by combining the following approaches: (i) the use of purification techniques that rigorously enrich samples for virus-like particles in order to minimize amounts of cellular genomic DNA, (ii) the use of sequencing technologies that maximize the number of reads, and (iii) the use of computational workflows that include steps of mapping of reads to *de novo* assembled contigs prior to BLASTx searches.

Conflict of interest

The authors declare no competing financial interests.

Author contributions

Data acquisition (S.F., D.F., M.F.); Analysis and interpretation of data (S.F., M.O. and R.F.); Manuscript preparation (S.F., M.O, D.M., D.F. and R.F.); Study supervision (S.F., M.O. and R.F.).

Acknowledgments

We are particularly grateful to the Conseil General de l'Hérault for providing us the opportunity to collect insects and plants in the Domaine de Restinclières. We warmly thank Francois Enault and the reviewers for their insightful comments on the manuscript. S. F. is a doctoral fellow from the University of Montpellier and was supported by a scholarship from Institut National de la Recherche Agronomique (INRA).

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at https://doi.org/10.1016/j.virusres.2017.11.011.

References

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. 98, 11609–11614. http://dx.doi.org/10.1073/pnas.211424698.

Altschul, S., Gish, W., Miller, W., Myers, E., Lipman, D., 1990. Basic local alignment search tool. J. Mol. Biol. 215, 403–410. http://dx.doi.org/10.1016/S0022-2836(05)80360-2.

Angly, F.E., Felts, B., Breitbart, M., Salamon, P., Edwards, R., a, Carlson, Chan, C., Haynes, A.M., Kelley, M., Liu, S., Mahaffy, H., Mueller, J.M., Nulton, J.E., Olson, J., Parsons, R., Rayhawk, R., Suttle, S., a, C., Rohwer, F., 2006. The marine viromes of four oceanic regions. PLoS Biol. 4, e368. http://dx.doi.org/10.1371/journal.pbio.0040368.

Anthony, S.J., Epstein, J.H., Murray, K.A., Navarrete-macias, I., Zambrana-torrelio, C.M., Solovyov, A., Ojeda-flores, R., Arrigo, N.C., Islam, A., Khan, A., Hosseini, P., Bogich, T.L., Mazet, J.A.K., Daszak, P., Lipkin, W., 2013. A strategy to estimate unknown viral diversity in mammals. MBio 4, 1–15. http://dx.doi.org/10.1128/mBio.00598-13.

Bankevich, A., Nurk, S., Antipov, D., Gurevich, A.A., Dvorkin, M., Kulikov, A.S., Lesin, V.M., Nikolenko, S.I., Pham, S., Prjibelski, A.D., Pyshkin, A.V., Sirotkin, A.V., Vyahhi, N., Tesler, G., Alekseyev, M.A., Pevzner, P.A., 2012. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. J. Comput. Biol. 19, 455–477. http://dx.doi.org/10.1089/cmb.2012.0021.

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, 14250–134255. http://dx.doi.org/10.1073/pnas.202488399.

Brum, J.R., Sullivan, M.B., 2015. Rising to the challenge: accelerated pace of discovery transforms marine virology. Nat. Rev. Microbiol. 13, 147–159. http://dx.doi.org/10.1038/nrmicro3404.

Cotten, M., Oude Munnink, B., Canuti, M., Deijs, M., Watson, S.J., Kellam, P., Van Der Hoek, L., 2014. Full genome virus detection in fecal samples using sensitive nucleic acid preparation, deep sequencing, and a novel iterative sequence classification algorithm. PLoS One 9, e93269. http://dx.doi.org/10.1371/journal.pone.0093269.

Enault, F., Briet, A., Bouteille, L., Roux, S., Sullivan, M.B., Petit, M.-A., 2016. Phages rarely encode antibiotic resistance genes: a cautionary tale for virome analyses. ISME J. 1–11. http://dx.doi.org/10.1038/ismej.2016.90.

Fancello, L., Raoult, D., Desnues, C., 2012. Computational tools for viral metagenomics and their application in clinical research. Virology 434, 162–174. http://dx.doi.org/10.1016/j.virol.2012.09.025.

García-López, R., Vázquez-Castellanos, J.F., Moya, A., 2015. Fragmentation and coverage variation in viral metagenome assemblies, and their effect in diversity calculations. Front. Bioeng. Biotechnol. 3, 141. http://dx.doi.org/10.3389/fbioe.2015.00141.

Huang, X., 1999. CAP3: a DNA sequence assembly program. Genome Res. 9, 868–877. http://dx.doi.org/10.1101/gr.9.9.868.

Junglen, S., Drosten, C., 2013. Virus discovery and recent insights into virus diversity in arthropods. Curr. Opin. Microbiol. 16, 507–513. http://dx.doi.org/10.1016/j.mib.2013.06.005.

Kim, M.-S., Park, E.-J., Roh, S.W., Bae, J.-W., 2011. Diversity and abundance of single-stranded DNA viruses in human feces. Appl. Environ. Microbiol. 77, 8062–8070. http://dx.doi.org/10.1128/AEM.06331-11.

Krishnamurthy, S.R., Wang, D., 2017. Origins and challenges of viral dark matter. Virus Res. 239, 136–142. http://dx.doi.org/10.1016/j.virusres.2017.02.002.

Langmead, B., 2010. Aligning short sequencing reads with Bowtie. Curr. Protoc. Bioinf. 11.

Magoc, T., Salzberg, S.L., 2011. FLASH: fast length adjustment of short reads to improve genome assemblies. Bioinformatics 27, 2957–2963. http://dx.doi.org/10.1093/bioinformatics/btr507.

Martin, M., 2011. Cutadapt removes adapter sequences from high-throughput sequencing reads. EMB 17, 10–12. http://dx.doi.org/10.14806/ej.17.1.200.

McDaniel, L., Breitbart, M., Mobberley, J., Long, A., Haynes, M., Rohwer, F., Paul, J.H., 2008. Metagenomic analysis of lysogeny in Tampa Bay: implications for prophage gene expression. PLoS One 3, e3263. http://dx.doi.org/10.1371/journal.pone.0003263.

Mokili, J.L., Rohwer, F., Dutilh, B.E., 2012. Metagenomics and future perspectives in virus discovery. Curr. Opin. Virol. 2, 63–77. http://dx.doi.org/10.1016/j.coviro.2011.12.004.

Ng, T.F.F., Willner, D.L., Lim, Y.W., Schmieder, R., Chau, B., Nilsson, C., Anthony, S., Ruan, Y., Rohwer, F., Breitbart, M., 2011. Broad surveys of DNA viral diversity obtained through viral metagenomics of mosquitoes. PLoS One 6, e20579. http://dx.doi.org/10.1371/journal.pone.0020579.

Palanga, E., Filloux, D., Martin, D.P., Fernandez, E., Bouda, Z., Gargani, D., Ferdinand, R., Zabre, J., Neya, B., Sawadogo, M., Traore, O., Peterschmitt, M., Roumagnac, P., 2016. Metagenomic-based screening and molecular characterization of cowpea- infecting viruses in Burkina Faso. PLoS One 11, e0165188. http://dx.doi.org/10.1371/journal.pone.0165188.

Remmert, M., Biegert, A., Hauser, A., Soding, J., 2012. HHblits: lightning-fast iterative protein sequence searching by HMM–HMM alignment. Nat. Methods 9, 173–175. http://dx.doi.org/10.1038/nmeth.1818.

Roossinck, M.J., Martin, D.P., Roumagnac, P., 2015. Plant virus metagenomics: advances in virus discovery. Phytopathology 105, 716–727. http://dx.doi.org/10.1094/PHYTO-12-14-0356-RVW.

Rosario, K., Breitbart, M., 2011. Exploring the viral world through metagenomics. Curr. Opin. Virol. 1, 289–297. http://dx.doi.org/10.1016/j.coviro.2011.06.004.

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, e33641. http://dx.doi.org/10.1371/journal.pone.0033641.

Roux, S., Krupovic, M., Debroas, D., Forterre, P., 2013. Assessment of viral community functional potential from viral metagenomes may be hampered by contamination with cellular sequences. Open Biol. 3, 130160. http://dx.doi.org/10.1038/ismej.2016.90.

Skewes-Cox, P., Sharpton, T.J., Pollard, K.S., DeRisi, J.L., 2014. Profile hidden Markov models for the detection of viruses within metagenomic sequence data. PLoS One 9, e105067. http://dx.doi.org/10.1371/journal.pone.0105067.

Suttle, C.A., 2007. Marine viruses (mdash) major players in the global ecosystem. Nat. Rev. Microbiol. 5, 801–812. http://dx.doi.org/10.1038/nrmicro1750.

Tangherlini, M., Dell'Anno, A., Zeigler Allen, L., Riccioni, G., Corinaldesi, C., 2016. Assessing viral taxonomic composition in benthic marine ecosystems: reliability and efficiency of different bioinformatic tools for viral metagenomic analyses. Sci. Rep. 22, 28428. http://dx.doi.org/10.1038/srep28428.

Toland, A.E., Çatalyürek, Ü.V., Hatem, A., Bozda, D., 2013. Benchmarking short sequence mapping tools. BMC Bioinf. 7, 184. http://dx.doi.org/10.1186/1471-2105-14-184.

Wu, S., Manber, U., 1992. A Fast Approximate Pattern-matching Tool. Usenix Winter 1992 Tech. Conf.

Zablocki, O., van Zyl, L., Adriaenssens, E.M., Rubagotti, E., Tuffin, M., Cary, S.C., Cowan, D., 2014. High-level diversity of tailed phages, eukaryote-associated viruses, and virophage-like elements in the metaviromes of antarctic soils. Appl. Environ. Microbiol. 80, 6888–6897. http://dx.doi.org/10.1128/AEM.01525-14.

Supplementary materials

Supplemental Table 1: Comparison of the performance of BLASTx searches against the viral published sequence database with or without a prior read to contig mapping step. BLASTx searches were performed on raw reads and after mapping of reads to contigs.

	Insects 1	Insects 2	Insects 3	Insects 4	Insects 5	Insects 6	Plants 1	Plants 2	Mosquitoes	Human feces	Estuary	Lake Bourget	Lake Pavin	Open soil	Hypolith
Number of viral reads classical BLASTx analyses	59028	16268	13983	2525	10608	117304	65600	45443	7169	5155	1599	15124	11187	1733	304
Percentage of viral reads classical BLASTx analyses	29,51%	8,13%	6,99%	1,26%	5,30%	58,65%	32,80%	22,72%	3,58%	2,58%	0,80%	7,56%	5,59%	0,87%	0,15%
Number of viral reads AM-BLASTx	145327	91962	89002	178828	84934	154178	112934	136233	18661	19426	2746	68340	37271	29688	4751
Percentage of viral reads AM-BLASTx	72,66%	45,98%	44,50%	89,41%	42,47%	77,09%	56,47%	68,12%	9,33%	9,71%	1,37%	34,17%	18,64%	14,84%	2,38%

Supplemental Table 2: Five most abundant viral families found across the 15 viromes used in this study

Viromes	Viral Taxonomy	Genome length (kb)	Abundance rank	Host range
Insects 1	Iflaviridae	<10	1	Eukaryotic macroorganisms
	Partitiviridae	<10	2	Eukaryotic macroorganisms
	Alphaflexiviridae	<10	3	Eukaryotic macroorganisms
	Luteoviridae	<10	4	Eukaryotic macroorganisms
	Flexviridae	<10	5	Eukaryotic macroorganisms
Insects 2	Iflaviridae	<10	1	Eukaryotic macroorganisms
	Partitiviridae	<10	2	Eukaryotic macroorganisms
	Luteoviridae	<10	3	Eukaryotic macroorganisms
	Alphaflexiviridae	<10	4	Eukaryotic macroorganisms
	Tymoviridae	<10	5	Eukaryotic macroorganisms
Insects 3	Parvoviridae	<10	1	Eukaryotic macroorganisms
	Carmotetraviridae	<10	2	Eukaryotic macroorganisms
	Podoviridae	40	3	Bacteria
	Iflaviridae	<10	4	Eukaryotic macroorganisms
	Mesoniviridae	20	5	Eukaryotic macroorganisms
Insects 4	Podoviridae	40	1	Bacteria
	Parvoviridae	<10	2	Eukaryotic macroorganisms
	Iflaviridae	<10	3	Eukaryotic macroorganisms
	Tymoviridae	<10	4	Eukaryotic macroorganisms
	Luteoviridae	<10	5	Eukaryotic macroorganisms
Insects 5	Parvoviridae	<10	1	Eukaryotic macroorganisms
	Iflaviridae	<10	2	Eukaryotic macroorganisms
	Tymoviridae	<10	3	Eukaryotic macroorganisms
	Dicistroviridae	<10	4	Eukaryotic macroorganisms
	Partitiviridae	<10	5	Eukaryotic macroorganisms
Insects 6	Dicistroviridae	<10	1	Eukaryotic macroorganisms
	Nanoviridae	<10	2	Eukaryotic macroorganisms
Plants 1	Amalgaviridae	<10	1	Eukaryotic macroorganisms
	Partitiviridae	<10	2	Eukaryotic macroorganisms
	Luteoviridae	<10	3	Eukaryotic macroorganisms
	Bromoviridae	<10	4	Eukaryotic macroorganisms
	Iflaviridae	<10	5	Eukaryotic macroorganisms
Plants 2	Amalgaviridae	<10	1	Eukaryotic macroorganisms
	Partitiviridae	<10	2	Eukaryotic macroorganisms
	Tymoviridae	<10	3	Eukaryotic macroorganisms
Mosquitoes	Parvoviridae	<10	1	Eukaryotic macroorganisms
	Anelloviridae	<10	2	Eukaryotic macroorganisms
	Nudiviridae	90-230	3	Eukaryotic macroorganisms
	Microviridae	<10	4	Bacteria
	Circoviridae	<10	5	Eukaryotic macroorganisms
Human feces	Microviridae	<10	1	Bacteria
	Siphoviridae	50	2	Bacteria
	Podoviridae	40	3	Bacteria
	Myoviridae	30-250	4	Bacteria
	Phycodnaviridae	100-550	5	Eukaryotic microorganisms
Estuary	Phycodnaviridae	100-550	1	Eukaryotic microorganisms
	Myoviridae	30-250	2	Bacteria
	Circoviridae	<10	3	Eukaryotic macroorganisms
	Podoviridae	40	4	Bacteria
	Siphoviridae	50	5	Bacteria
Lake Bourget	Microviridae	<10	1	Bacteria
	Phycodnaviridae	100-550	2	Eukaryotic microorganisms
	Myoviridae	30-250	3	Bacteria
	Siphoviridae	50	4	Bacteria
	Podoviridae	40	5	Bacteria
Lake Pavin	Circoviridae	<10	1	Eukaryotic macroorganisms
	Phycodnaviridae	100-550	2	Eukaryotic microorganisms
	Siphoviridae	50	3	Bacteria
	Myoviridae	30-250	4	Bacteria
	Microviridae	<10	5	Bacteria
Open Soil	Phycodnaviridae	100-550	1	Eukaryotic microorganisms
	Podoviridae	40	2	Bacteria
	Myoviridae	30-250	3	Bacteria
	Siphoviridae	50	4	Bacteria
	Mimiviridae	1200	5	Eukaryotic microorganisms
Hypolith	Siphoviridae	50	1	Bacteria
	Myoviridae	30-250	2	Bacteria
	Phycodnaviridae	100-550	3	Eukaryotic microorganisms
	Podoviridae	40	4	Bacteria
	Mimiviridae	1200	5	Eukaryotic microorganisms

Supplemental Table 3: Global diversity in samples (BLASTx searches on 10.000 reads)

		Taxonomy	Insects 1	Insects 2	Insects 3	Insects 4	Insects 5	Insects 6	Plants 1	Plants 2	Mosquitoes	Human feces	Estuary	Lake Bourget	Lake Pavin	Open soil	Hypolith
Number of reads	Classified	Archaea	0	0	0	0	0	0	0	1	0	11	29	50	36	28	6
		Bacteria	44	23	1848	2244	389	51	69	209	3304	3935	294	2507	2920	2356	1528
		Eukaryota	210	41	545	47	427	378	4705	5254	81	21	8	32	31	87	27
		Viruses	2000	714	5360	6515	1389	5695	3495	2143	749	1543	1241	2526	1278	231	232
	Unclassified	Unclassified	7746	9222	2247	1194	7795	3877	1732	2394	5866	4491	8427	4885	5736	7298	8206
Percentage of reads	Classified	Archaea	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	1%	0%	0%	0%
		Bacteria	0%	0%	18%	22%	4%	1%	1%	2%	33%	39%	3%	25%	29%	24%	15%
		Eukaryota	2%	0%	5%	0%	4%	4%	47%	53%	1%	0%	0%	0%	0%	1%	0%
		Viruses	20%	7%	54%	65%	14%	57%	35%	21%	7%	15%	12%	25%	13%	2%	2%
	Unclassified	Unclassified	77%	92%	22%	12%	78%	39%	17%	24%	59%	45%	84%	49%	57%	73%	82%

Bilan et perspectives

Nos analyses bioinformatiques, menées sur des viromes provenant d'échantillons d'origine diverse et ayant été obtenus par différents procédés, montrent que l'association d'une étape de mapping à l'assemblage *de novo* permet d'augmenter en moyenne d'un facteur cinq la proportion de séquences virales attribuées dans les viromes analysés, comparativement à une analyse n'effectuant pas ces étapes. Ce résultat serait explicable par le fait que les reads sont moins longs que les contigs, longueur qui, liée au faible degré de similitudes de certains reads avec les séquences présentes dans les bases de données, réduit en conséquence l'efficacité de l'attribution taxonomique par BLAST (García-López *et al.*, 2015).

Il est à noter que la proportion de matière noire résiduelle dans les viromes est dépendante du type d'échantillon d'où proviennent les viromes, ceux disponibles publiquement contenant plus de matière noire résiduelle que les nôtres. Cette différence pourrait provenir du fait que l'ensemble des viromes publics utilisés lors de cette étude ont été obtenus par amplification par la polymérase phi29 qui est connue pour introduire des biais d'amplification en faveur des génomes constitués d'ADN circulaire (Kim et Bae, 2011). Cette différence pourrait également être expliquée par le fait que les virus environnementaux sont peu présents dans les bases de données, et que donc certains virus contenus dans des viromes provenant d'échantillons environnementaux n'aient pas été identifiés par BLAST.

En conclusion, afin de réduire la quantité de matière noire présente dans les viromes, il est nécessaire :

- De diminuer la contamination des échantillons par des acides nucléiques d'origine cellulaire, notamment en améliorant les étapes de filtration et de digestion des acides nucléiques non-encapsidés ;

- D'utiliser des technologies de séquençage maximisant soit le nombre de reads soit leur longueur en complémentarité, afin de faciliter l'assemblage des reads en contigs ;

- D'améliorer les outils servant à l'attribution taxonomique, et d'utiliser plusieurs outils en combinaison ;

- D'enrichir les bases de données en y déposant des séquences virales divergentes de celles y étant actuellement référencées.

Chapitre III - Diversité des communautés virales associées à des arthropodes ravageurs de cultures

Impact des arthropodes ravageurs dans les agroécosystèmes

Les agroécosystèmes, écosystèmes simplifiés par les activités humaines, représentent un bon modèle d'étude de la diversité ainsi que de la circulation des virus, et en particulier de ceux associés aux arthropodes qui y représentent la composante animale majeure. Améliorer le fonctionnement des agrosystèmes et favoriser le biocontrôle des populations d'insectes ravageurs et/ou vecteurs nécessite de comprendre les équilibres des communautés associées, incluant les communautés virales. Il est donc nécessaire d'établir en préliminaire l'inventaire des virus dans un agrosystème modèle.

La métagénomique virale appliquée aux agroécosystèmes est encore rare, limitée à ce jour aux virus de plantes. Ces études ont permis de mettre en évidence une grande diversité de virus associés aux plantes cultivées et sauvages (Palanga *et al.*, 2016; Roossinck, 2011b; Roossinck *et al.*, 2010; Roumagnac *et al.*, 2015). À notre connaissance, les seuls viromes d'arthropodes ravageurs de cultures publiés sont ceux associés l'aleurode du tabac (*Bemisia tabaci*, Hémiptère), les auteurs de cette étude s'étant principalement intéressés aux phytovirus présents dans cette espèce d'insectes, ne décrivant pas les virus potentiellement entomopathogènes (Ng, Duffy, *et al.*, 2011; Rosario *et al.*, 2014, 2015). Une grande part de la diversité des virus associés aux arthropodes ravageurs de cultures reste donc sous-explorée.

L'objectif des deux études suivantes a été de caractériser, par une approche de métagénomique basée sur la purification de particules virales, l'ensemble de la diversité et de la composition des communautés virales présentes chez certaines espèces d'arthropodes ravageurs de cultures.

- La première étude porte sur le tétranyque tisserand (*Tetranychus urticae*, Acarien). *T. urticae* est un acarien ravageur de cultures possédant une répartition mondiale (Jeppson *et al.*, 1975). Cette espèce est également extrêmement polyphage, et a un impact négatif important sur la production agricole (Jeppson *et al.*, 1975). Enfin, cette espèce présente des records de résistance aux pesticides (Van Leeuwen *et al.*, 2010). Les viromes de deux populations de *T. urticae* provenant de deux élevages différents ont été comparés.

- La seconde étude porte sur la caractérisation des viromes de trois espèces d'insectes ravageurs de cultures possédant également une répartition mondiale : l'armigère de la tomate (*Helicoverpa armigera*, Lepidoptère), le phytonome de la luzerne (*Hypera postica*, Coleoptère) et le puceron vert du pois (*Acyrthosiphon pisum*, Hemiptère), ce dernier étant vecteur de virus phytopathogènes (CABI). Ces insectes ont été échantillonnés dans deux agroécosystèmes adjacents : des champs de luzerne et des prairies. Des plantes, ainsi que deux espèces d'arthropodes prédateurs ont également été échantillonnées. La distribution des espèces virales les plus abondamment présentes dans les viromes obtenus a enfin été examinée dans les communautés d'arthropodes présentes dans les agroécosystèmes testés.

Les résultats de ces deux études ont permis d'améliorer nos connaissances sur la diversité des communautés virales associées aux arthropodes ravageurs de cultures.

Article de recherche 3

Metagenomic analysis of the viral communities associated with the two-spotted mite *Tetranychus urticae*: identification of a novel mini densovirus and nine other new viral species

Sarah François[1,2], Doriane Mutuel[2], Alison Duncan[3], Leonor Rodrigues[4], Denis Filloux[5], Emmanuel Fernandez[5], Philippe Roumagnac[5], Rémy Froissart[6], Mylène Ogliastro[2].

[1] Université de Montpellier, UMR 1333 DGIMI « Diversité, Génomes et Interactions Microorganismes-Insectes », place Eugène-Bataillon, 34095 Montpellier cedex 5, France

[2] Laboratoire « Diversité, Génomes et Interactions Microorganismes Insectes » (DGIMI) UMR 1333, INRA, Université Montpellier, 34095 Montpellier, France.

[3] Institut des Sciences de l'Évolution UMR5554, Université Montpellier–CNRS–IRD–EPHE, 34000 Montpellier, France.

[4] Centre for Ecology, Evolution and Environmental Changes, Faculty of Science, University of Lisbon, P-1749016 Lisbon, Portugal

[5] Laboratoire « Biologie et Génétique des Interactions Plante-Parasite » UMR BGPI, CIRAD-INRA-SupAgro, Campus International de Montferrier-Baillarguet, Montpellier Cedex-5, France.

[6] Laboratoire « Maladies Infectieuses et Vecteurs : Écologie, Génétique, Évolution et Contrôle » (MIVEGEC), UMR 5290, CNRS-IRD-UM, 911 avenue Agropolis, 34394, Montpellier, France.

Soumis dans Scientific Reports

Abstract

The two-spotted spider mite *Tetranychus urticae* is a cosmopolitan agricultural pest that displays an extensive host plant range and extreme records of pesticide resistance. A better understanding of their virome could hence allow developing new biological tools to control this pest using their natural virus-based enemies. Here, we present *T. urticae* viromes obtained by viral metagenomics based on viral particle purification. We recovered contigs that could putatively be attributed to ten new viral species, including a mini-densovirus, representing extant as well as new viral lineages. Two of these viruses were also found in *T. urticae* transcriptomic but not in genomic datasets. These findings offer new insights into arthropod virus evolution and may provide new opportunities for developing biological control agents against this pest.

Keywords: Arthropod, Mite, *Tetranychus urticae*, Viral Metagenomics, virus diversity, phylogeny, agricultural pests.

Introduction

Mites and ticks are small arachnids belonging to *Acari* that are mostly known for their detrimental impact on human, animal and plant health[1]. While ticks represent a relatively small number of taxa (around 900 species) all sharing a parasitic, blood-feeding alimentary regime, mites are extraordinarily diversified (more than 40 000 species) exhibiting a large diversity of lifestyles, including plant feeders, mite-predators or arthropod ectoparasites and are found in all ecosystems, both terrestrial and aquatic. Unlike ticks that can vector a number of pathogenic bacteria and viruses, only a few species of mites present direct threats to plant or animal health. Among them, the two-spotted spider mite, *Tetranychus urticae* is an agricultural pest[2] that produces a silk-like web micro-habitat protecting colonies against predators or abiotic stresses[3]. *T. urticae* disperses actively or passively, it is spread by the wind or through plant movments[2] and displays a worldwide distribution. This mite is highly polyphagous and feeds on more than 1000 plant species corresponding to more than 140 botanical families (http://www1.montpellier.inra.fr/CBGP/spmweb/; https://www.pesticideresistance.org/) making this pest particularly problematic in greenhouses and crops, where they can cause significant damage to a number of food-producing cultures (such as tomato, cucumber, maize, soybean, grape and citrus) and flowers (chrysanthemums and orchids)[2]. As predicted by computational modelling, global warming is expected to worsen this situation by accelerating *T. urticae* development, leading to increase populations and expand their distribution, which raises concerns about control solutions[4]. Indeed, conventional treatments against *T. urticae* mostly involve synthetic acaricides. Although efficient, their heavy use to control *T. urticae* has resulted in one of the highest incidence of pesticides resistance recorded in arthropods[5]. Moreover, acaricides also compromise biocontrol by eliminating *T. urticae* natural predator mite *Phytoseiulus persimilis*. Decreasing the predation pressure has then led to increase even more the use of acaricides, thus increasing

selection of resistance in *T. urticae* populations[6]. In this context, the development of alternative solutions to chemicals is strongly encouraged; one promising way consists in diversifying of the use of *T. urticae* natural enemies and particularly to include their pathogens.

Very little is known about the diversity of mite pathogens, the small size (~1 mm long) of most mites may have been a major obstacle to explore their pathologies and their pathogens; this is particularly true for mite viruses. Only one virus has been described for *T. urticae:* it is an enveloped, rod-shaped virus, discovered in laboratory populations[7], that infects the mite gut cells. Nowadays, the development of whole genome sequencing (WGS) methods allows overcoming the size limitation of organisms, making feasible the exploration without *a priori* knowledge of the microbial communities associated with tiny animals. Indeed, metagenomics has revolutionized microbiology including virology, showing an unexpected diversity and persistence of viruses in all organisms. The next challenge is now to understand these interactions and the roles viruses play in ecosystems functioning at all levels, from the individual organism to populations.

Viral metagenomics applied to *Acari* have been first used to explore the virus communities (so-called the virome) associated with blood-feeding ticks, which revealed the extraordinary diversity of viruses in these small arthropods[8,9]. Recently, a similar approach has been used to analyze the viral population specific to *Varroa destructor*, a bee ectoparasitic mite that also transmits to the insect several viruses that may contribute to colony collapse[10]. This work revealed new viruses infecting *V. destructor*, belonging to *Baculoviridae*, *Circoviridae*, *Dicistroviridae* and *Iflaviridae* families, confirming the interest of metagenomics to analyze viromes in very small animals.

In the study presented here, we investigated the viral community associated with *T. urticae* by in depth sequencing of virion-associated nucleic acids (VANA)[11]. We analyzed the virome composition of two *T. urticae* laboratory populations from different geographic origins and host plants. We found ten new viruses belonging to taxa associated with arthropods, five viruses being shared by both mite populations. In particular, we discovered a new densovirus that was present in abundance in both viromes and in two other *T. urticae* laboratory populations. This highly divergent novel virus, which has an unusual small genome size, is likely to represent a new species in the *Ambidensovirus* genus.

These discoveries illuminate our knowledge of viruses associated with an arthropod taxon which has been long neglected by virologists. In addition to, this work outlines the importance of understanding virus association and dynamics in *T. urticae* for developing biocontrol strategies.

Results

Overview of the Spider mite virome. To explore the viral diversity of *T. urticae*, we semi-purified viral particles from two laboratory populations (respectively called the Portuguese (P) and French (F) population) (see Methods). A total of 680 600 cleaned reads were obtained, including 219 982 reads from the French population and 460 618 reads from the Portuguese population.

Fourteen viral contigs (1.6 to 8.6 kb in length) were obtained by *de novo* assembly of the VANA reads in both populations (**Tab.1**). Thirteen out of the fourteen contigs were mostly related to non-enveloped DNA and RNA viruses belonging to clades infecting arthropods, including *Dicistroviridae*, *Parvoviridae*, *Birnaviridae, Nodaviridae* and

unclassified picornavirales. In addition, one viral contig (2481nt) had similarity with yeast and fungi-infecting viruses of the *Narnaviridae* family, which might come from environment contamination (e.g. food), even though we cannot exclude its replication in spider mites. While nine viral contigs were isolated in the P population only, five viral contigs were found in both populations (**Tab.2** ; **Fig. 1A**).

Among the five viral contigs that are common to both viromes, one assigned to the *Parvoviridae* family largely dominates in terms of reads abundance (>28% of viral reads in both populations), the others being found at much lower frequencies (0.1% to 10.6%) (**Fig.1 B**). Sequence analysis and open reading frame (ORF) prediction showed that this parvovirus has an ambisense genomic organization, and its NS1 shares 39% aa identity with its closest relative *Lupine feces-associated densovirus 2* (accession number: ASM93489. According to the new species demarcation threshold in the *Densovirinae* sub-family proposed to the ICTV (i.e. <85% related by NS1 amino acid sequence identity[12], this virus might represent a new divergent species in the *Ambidensovirus* genus and the first densovirus isolated from Arachnids (**Fig.2**, **Tab.1**). This virus is hereafter referred to as *Tetranychus urticae associated ambidensovirus* (TuaDV).

Although the *Birnaviridae* family has been poorly investigated so far, the phylogenetic analyses showed that two contigs found in *T. urticae* viromes (representing of 3.8% of reads) clustered within the insect-infecting entomobirnaviruses (**Fig.3**). The polymerase and capsid proteins of this putative novel entomobirnavirus share 30% aa identity with the *Infectious bursal disease virus* virus; accession number AAS10174.1 and 33% aa identity with the *Blotched snakehead virus* virus; accession number YP_052864.1, respectively. Therefore, this putative novel entomobirnavirus species-level lineage could represent the first entomobirnavirus isolated from Arachnids (**Tab.1**).

Concerning the three viral contigs clustering within the *Nodaviridae* family (**Fig.4**, **Tab.1**), coding for capsid protein, and found in the P population only (representing 1.35% of reads), they share up to 58% aa identity with *Hubei noda-like virus 9* capsid protein (accession number: YP_009337880.1), an unclassified RNA virus. According to the ICTV species demarcation threshold (<87% of capsid protein aa identity) and to their position in the phylogenetic tree, these contigs may correspond to three novel species-level lineages in a novel genus-level lineage in the *Nodaviridae* family (**Fig.4**, **Tab.1**).

Five contigs were assigned to the *Picornavirales* order. While two contigs cluster in the *Dicistroviridae* family (hereafter referred to as *Tetranychus urticae* associated dicistrovirus 1 and 2 (Tuad1 and Tuad2), the other three remain unclassified. The capsids of Tuad1 and Tuad2 both share 20% aa identity with *Beihai picorna-like virus* 70 virus (accession number APG78062.1; **Tab.1**). Based on the current species demarcation criteria used by the ICTV *Dicistroviridae* study group (less than 90% aa identity of capsid protein identity with closest relatives) and the phylogenetic analyses (**Fig.5**, **Tab.1**), it is likely that both contigs could represent two novel species-level lineages of the *Dicistroviridae* family.

Interestingly, the CP of one of the three unclassified picorna-like viruses shared >99% aa identity with the unclassified picorna-like virus *Aphis glycines virus 1* (**Tab.1**). In addition, the proteins of the two remaining unclassified picorna-like viruses (hereafter referred to as *Tetranychus urticae* associated picorna-like virus 1 and 2) share 53% to 75% aa identity with *Aphis glycines virus 1* and *Hubei picorna-like virus 80* (**Tab.1**). Their phylogenetic trees showed that they might belong to a highly divergent lineage within the *Picornavirales* order (**Fig.5**). Moreover, *Tetranychus urticae associated picorna-like virus 1* and *2* contigs could represent new species-level lineages according to the species demarcation criteria defined by the ICTV (<90% of capsid protein identity with closest relatives) (**Tab.1**).

Finally, the *Tetranychus urticae associated narnavirus*, present in 0.15% and 0.2% of reads in P and F populations respectively, was the only one in this study that clustered with fungi viruses. Although these viruses are poorly known, the phylogenetic position of this contig within the *Narnavirus* genus suggests that it might represent a new species-level lineage according to the ICTV species demarcation threshold (<50% of protein sequence identity compared to the closest relative) (**Fig.6**, **Tab.1**).

Discovery of a new densovirus species in spider mites. As pointed above, the size of the densovirus contig (i.e. 2.8 kb) found in the *T. urticae* viromes was smaller than the size of the viruses characterized so far in the *Ambidensovirus* genus and typically ranging from 5.3-6 kb[12]. Ambidensoviruses usually display a single ORF encoding for four structural proteins (VP1-4) that are produced by leaky scanning, and three ORFs encoding for non-structural (NS) proteins. Viruses in the *Parvoviridae* family are characterized by two typical domains, i) a phospholipase A2 (PLA2) motif located in VP1 of most parvoviruses, including in all species described so far in the *Ambidensovirus* genus. ii) A Super Family 3 (SF3) helicase domain located in the NS1 protein and common to all parvoviruses.

Analysis of the new densovirus contig predicted three open reading frames, one encoding a VP protein and one a NS1 protein with respective sizes of 506 and 354 aa, which would be the smallest proteins described so far among densoviruses. As expected, the NS1 sequence included the SF3 domain and displayed 39% identity with *Lupine feces-associated densovirus 2* [13]. One additional incomplete ORF was also predicted although with lower confidence, corresponding to 164 aa. Interestingly, the VP sequence lacked the typical PLA2 domain and displayed 30% identity with the VP of *Lupine feces-associated densovirus 2*.

Based on the current species demarcation criteria used by the ICTV, such features suggest that the novel densovirus represents a new species among the *Ambidensovirus* genus.

To confirm the presence of this new ambidensovirus in the P and F populations and its sequence, we realized overlapping PCRs with specific primers and sequenced the PCR products by the Sanger method. Our results showed that this densovirus was confirmed in both populations and displayed identical consensus sequences.

Genomic and transcriptomic database screening. To gain insights into the diversity and the distribution of viruses in *T. urticae*, we further screened genomic and transcriptomic datasets of this mite using as queries all ten viruses identified in this study. Our search highlighted that *T. urticae* transcriptomes contained sequences displaying >95% of nucleotide identity to *Aphis glycines* virus 1 and *Tetranychus urticae* associated picorna-like virus 1 **(Tab.3)**. Interestingly, one sequence related to an ambidensovirus was found in the genome of *T. urticae* but this sequence was different from TuaDV (70% of nt identity); no sequence corresponding to this virus was found in any of the transcriptomes analyzed suggesting that TuaDV corresponds to an extant virus, which origin remains to be clarified.

Last, it has to be mentioned that one transcriptomic sequence (accession number GW017620.1) matched with the *Tetranychus urticae associated nodavirus* segment B1, but could not be assigned with high confidence because of its small size **(Tab.3)**.

Discussion

In this work, we explored the diversity of viruses associated with the spider mite *Tetranychus urticae*, an Acari that has been long neglected by virologists. We compared the virus communities of two laboratory populations with different spatial (geographical) origins and rearing history using a viral metagenomics approach and we combined this approach with virus screening in *T. urticae* genomic and transcriptomic databases.

We discovered a panel of 10 potentially new virus species belonging to seven families of small, non-enveloped viruses, which considerably increased our knowledge of viruses associated with this mite. Interestingly, most of the virus genotypes found in this study were classified within arthropod-infecting taxa supporting their direct association with mites; although, the host range of viruses cannot be assigned with high confidence with this approach, as viral sequences can result from trophic (e.g. feeding behavior) or laboratory contaminations. However, the presence of the same viruses (for which the closest phylogenetic taxa are associated to arthropods) within different independent host populations suggest that these viruses could infect spider mites.

The phylogenetic trees we obtained also highlighted the poor knowledge we have on arthropod viruses[14–16]. Indeed, the viruses revealed in the present study were often located at the base of the phylogenetic trees of the largest taxa, such as for *Picornavirales* and *Parvoviridae*, or were grouped into poorly documented viral families, suggesting that many more viruses remain to be discovered in arthropods, and particularly in spider mites.

The comparative analyses of the viromes highlighted the unexpected feature that both mite populations display a (rather small) set of identical viruses (5). The probability is clearly low that five phylogenetically distant viruses randomly infect two independent host populations. Considering on one hand the different geographical origin and rearing history of

the populations and on the other hand the propensity of viruses to diversify, these observations suggest at simplest that cross-contamination might have occurred between the two laboratory populations, due to material exchanges that have not been traced. Alternatively, this community may highlight some properties of this mite-virus system: i) these five virus species might have a uniform and wide distribution in mite populations; ii) some behavior and/or rearing conditions (e.g. feeding, social organization, host density) may select virus communities[17]. To verify the reality of the spider mite common virome, further studies are needed to explore virus prevalence and diversity more thoroughly: i) by improving mite sampling (spatial and temporal) in ecosystems and ii) by increasing sequencing coverage to better determine the genetic structure of spider mite viromes.

Among the five viruses that are common to both mite populations, we found a new densovirus, related to members of the *Ambidensovirus* genus and tentatively called TuaDV. Interestingly, this new *Ambidensovirus* species (noticeably infecting the two F and P host populations) has never been described in databases. Moreover, TuaDV displays original features, in particular an unusual small size of its coding sequence and the absence of a PLA2 motif so far found in all the related species in this genus[12]. Database screening showed no sequence corresponding to TuaDV, although a number of other viral sequences could be found, including one corresponding to another densovirus in the genome of *T. urticae*. However, we cannot exclude that TuaDV genome could have been eliminated during *T. urticae* genome assembly or was not detected in transcriptomes due to low depth sequencing.

Metagenomics allows to make inventories of within-host viral communities, which is the first step towards the understanding of viral associations[18]. It is becoming clear that multi-infections and persistent viruses are important to be considered to better understand infection processes and the pathogenicity of specific strains[18]. In experimental colonies, mite density can vary, due to variation in mortality levels which can occur without knowledge of the causal

agent(s)/conditions (A. Duncan, personal communication). Whether or not the viruses found in this study are pathogenic for mites remains to be assessed by experimental infections. Alternatively, we cannot exclude that these viruses could represent a non-virulent "core virome", which role if any, remains to be determined. Experimental mite colonies could provide a powerful system to combine descriptive and manipulative experiments to test virus pathogenicity in individual hosts and their dynamics (prevalence and persistence) and evolution in host populations.

Methods

Mite populations. Two laboratory *T. urticae* populations from Portugal and France were processed using the VANA metagenomics-based approach. The Portuguese (P) population was established at the University of Lisbon from around 200 females collected in January 2014 in Spain (Almeria) on roses and kept on bean plants (*Phaseolus vulgaris*, Fabaceae, var. *Enana;* Germisem Sementes Lda, Oliveira do Hospital, Portugal) ever since. This population also contained a mix of two other experimental lines, London S and EtoxR. The London strain, originally collected in the Vineland region, Ontario, Canada, originates from the culture used in the *T. urticae* genome project[19], maintained at the University of Logroño and later transferred to the University of Ghent. The EtoxR strain was originally collected in Japan and maintained for 5 years in the laboratory at Bayer CropScience before being transferred to the University of Ghent's, where it was further maintained on potted bean plants and sprayed until runoff with 1,000 mg active ingredient per liter of etoxazole. Both strains were established at the University of Lisbon, from approximately 2000 individuals sampled from the Ghent stock. They were maintained on bean (*Phaseolus vulgaris*, Fabaceae, var. *Enana*; Germisem Sementes Lda, Oliveira do Hospital, Portugal) at the University of Lisbon, since

July 2013. The Montpellier population (so-called the French (F) population here) originated from Netherlands (Pijnacker) and was collected in May 1994 on cucumber plants. This population was maintained at the University of Amsterdam until September 2007, and it was then transferred to Montpellier. In 2011, it was transferred to tomato plants. London S was also used for virus screening in this study. The London strain, originally collected in the Vineland region, Ontario, Canada, originates from the culture used in the *T. urticae* genome project[19], maintained at the University of Logroño and later transferred to the University of Ghent. The London strain was maintained on bean (*Phaseolus vulgaris*, Fabaceae, var. *Enana*; Germisem Sementes Lda, Oliveira do Hospital, Portugal) at the University of Lisbon, since July 2013.

Viromes preparation and sequencing. Samples were made of pools of around 200 *T. urticae* individuals from where viral particles were purified using the method described by Palanga et al. [11]. Briefly, *T. urticae* pools were ground in HBSS buffer with beads using a tissue homogenizer. The homogenized extracts were filtered through a 0.45 µm filter, and centrifuged at 148.000xg for 2.5hrs at 4°C to concentrate viral particles. Then, non-encapsidated nucleic acids were partially eliminated by DNase and RNase digestion for 1.5h at 37°C. Encapsidated DNA and RNA were then extracted and RNA was converted to cDNA using a 26nt primer (Dodeca Linker) composed of a 14nt linker linked at the 3' end to N_{12}. Double-stranded DNA was synthetized from single-stranded DNA using Large (Klenow) fragment DNA polymerase and the Dodeca Linker. Double-stranded DNA was further amplified using one 24nt PCR multiplex identifier primer composed of the 14nt linker used during the RT step linked at the 5' end to a 10nt tag that allowed sample identification. PCR products were cleaned and libraries were sequenced on Illumina MiSeq 2x300nt (Genewiz, USA).

Bioinformatics analyses. After a demultiplexing step using agrep command-line tool in order to assign reads to the samples from which they originated[20], adaptors were removed and reads were filtered for quality (q30 quality and read length >45nt) using Cutadapt 1.9[21]. The cleaned reads were *de novo* assembled into contigs using SPAdes 3.6.2 (K-mer lengths 21,33,55,77,125)[22]. Mapping was performed on contigs by cleaned reads alignment using Bowtie 2.1.0 (options local very sensitive)[23]. Taxonomic assignment was obtained through searches against the NCBI RefSeq viral database and against the non-redundant (nr) GenBank database using BLASTx with an e-value cutoff of $<10^{-3}$ [24].

Viral diversity analyses. Viral contigs were classified as viral operational taxonomic units (vOTU). In order to remove inter-sample and laboratory contamination, we focused on the most abundant vOTU by using an arbitrary abundance cutoff of <0.1% that was applied for each vOTU in all samples. To assess whether vOTUs represent novel or already described viruses, their full-length proteins were aligned and compared with their closest relative viral proteins (found in GenBank database) using MUSCLE 3.7 (16 iterations) according to the species demarcation thresholds recommended within the online (9th or 10th) Reports of the ICTV (https://talk.ictvonline.org/ictv-reports/ictv_online_report/).

Phylogenetic analyses. The putative amino acid sequences of *T. urticae* associated vOTU were used for phylogenetic analyses. All ORFs were translated *in silico* using ORF finder (cut off >300nt, ATG start codon) on Geneious 1.7[25] and aligned with the corresponding proteins fragments of relative viruses deposited on GenBank nr database using MUSCLE 3.7 (16 iterations) with default settings[26]. Aligned sequences were manually edited to remove gaps. Maximum likelihood phylogenetic trees were produced from these alignments using PhyML

3.1[27,28] with substitution models chosen as the best-fit using Prottest 2.4[29]. One thousand bootstrap replicates were used to test the support of branches. Trees were visualized with FigTree 1.4 (http://tree.bio.ed.ac.uk/software%20/figtree/). Outgroups were used when possible, otherwise trees were mid-point rooted.

Database screening. Sequences of all viral contigs were used as queries to perform BLASTn searches within the *T. urticae* genome: RefSeq genomic database (GCF_000239435.1, 641 sequences), WGS (CAEY00000000.1, 2035 sequences) and transcriptomes: EST (txid32264, 80855 sequences) and TSA (BioProject 78685, 9614 sequences; BioProject 78689, 17739 sequences; BioProject 6829 sequences) with and an percentage of identity cutoff of >95% and a e-value cutoff of $<10^{-3}$.

Validation of a full-length viral sequence. To confirm the presence of *T. urticae associated Parvoviridae* in our samples and to validate its complete coding sequence, we screened our pools of mite populations with PCR specific overlapping primers (Table S2). PCR amplicons were sequenced using the Sanger method. In order to obtain the ITRs, RACE PCR, using the 5'/3' RACE kit, 2nd Generation (Roche), was used unsuccessfully due to the difficulty to amplify and sequence *Parvoviridae* ITRs.

Acknowledgments

S. F. was supported by a scholarship from the National Institute of Agronomical Research (INRA) and from the University of Montpellier (UM).

References

1. Walter, D. E. & Proctor, H. C. *Mites: Ecology, Evolution & Behaviour*. (2013).

2. Jeppson, L. R., Keifer, H. H. & Baker, E. W. *Mites Injurious to Economic Plants*. (University of California Press, 1975).

3. Gerson, U. in *Spider Mites: their Biology, Natural Enemies and Control Vol. 1A* (eds. Helle, W. & Sabelis, M. W.) 223–232 (1985).

4. Migeon, A. *et al.* Modelling the potential distribution of the invasive tomato red spider mite, Tetranychus evansi (Acari: Tetranychidae). *Exp Appl Acarol* **48,** 199–212 (2009).

5. Van Leeuwen, T., Vontas, J., Tsagkarakou, A., Dermauw, W. & Tirry, L. Acaricide resistance mechanisms in the two-spotted spider mite Tetranychus urticae and other important Acari: a review. *Insect Biochem. Mol. Biol.* **40,** 563–572 (2010).

6. Rhodes, E. M. & Liburd, O. E. Evaluation of predatory mites and Acramite for control of twospotted spider mites in strawberries in north central Florida. *J. Econ. Entomol.* **99,** 1291–1298 (2006).

7. Poinar, G. & Poinar, R. Parasites and pathogens of mites. *Annu. Rev. Entomol.* **43,** 449–469 (1998).

8. Tokarz, R. *et al.* Virome analysis of Amblyomma americanum, Dermacentor variabilis, and Ixodes scapularis ticks reveals novel highly divergent vertebrate and invertebrate viruses. *J. Virol.* **88,** 11480–11492 (2014).

9. Moutailler, S., Popovici, I., Devillers, E. & Eloit, M. Diversity of viruses in Ixodes ricinus, and characterization of a neurotropic strain of Eyach virus. *New Microbes New Infect.* **11,** 71–81 (2016).

10. Levin, S., Sela, N. & Chejanovsky, N. Two novel viruses associated with the Apis mellifera pathogenic mite Varroa destructor. *Sci. Rep.* **24,** 6:37710 (2016).

11. Palanga, E. *et al.* Metagenomic-Based Screening and Molecular Characterization of

Cowpea- Infecting Viruses in Burkina Faso. *PLoS One* **11,** e0165188 (2016).

12. Cotmore, S. F. *et al.* The family Parvoviridae. *Arch. Virol.* **159,** 1239–1247 (2014).

13. Conceição-neto, N. *et al.* Viral gut metagenomics of sympatric wild and domestic canids , and monitoring of viruses : Insights from an endangered wolf population. *Ecol Evol* **7,** 4135–4146 (2017).

14. Marklewitz, M., Zirkel, F., Kurth, A., Drosten, C. & Junglen, S. Evolutionary and phenotypic analysis of live virus isolates suggests arthropod origin of a pathogenic RNA virus family. *Proc. Natl. Acad. Sci. U. S. A.* **112,** 7536–41 (2015).

15. Li, C.-X. *et al.* Unprecedented genomic diversity of RNA viruses in arthropods reveals the ancestry of negative-sense RNA viruses. *Elife* **4,** 1–26 (2015).

16. Shi, M. *et al.* Divergent Viruses Discovered in Arthropods and Vertebrates Revise the Evolutionary History of the Flaviviridae and Related Viruses. *J. Virol.* **90,** 659–669 (2016).

17. Clotuche, G. *et al.* The formation of collective silk balls in the spider mite Tetranychus urticae Koch. *PLoS One* **6,** e18854 (2011).

18. Alizon, S., de Roode, J. C. & Michalakis, Y. Multiple infections and the evolution of virulence. *Ecol. Lett.* **16,** 556–567 (2013).

19. Grbić, M. *et al.* The genome of Tetranychus urticae reveals herbivorous pest adaptations. *Nature* **479,** 487–92 (2011).

20. Wu, S. & Manber, U. A fast approximate pattern-matching tool. *Usenix Winter 1992 Tech. Conf.* (1992).

21. Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMB* **17,** 10–12 (2011).

22. Zerbino, D. R. & Birney, E. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. *Genome Res.* **18,** 821–9 (2008).

23. Langmead, B. Aligning short sequencing reads with Bowtie. *Curr Protoc Bioinforma.* **11,** (2010).

24. Altschul, S., Gish, W., Miller, W., Myers, E. & Lipman, D. Basic local alignment search tool. *J. Mol. Biol.* **215,** 403–410 (1990).

25. Kearse, M. *et al.* Geneious Basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data. *Bioinformatics* **28,** 1647–9 (2012).

26. Edgar, R. C. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* **32,** 1792–7 (2004).

27. Dereeper, a *et al.* Phylogeny.fr: robust phylogenetic analysis for the non-specialist. *Nucleic Acids Res.* **36,** W465-9 (2008).

28. Guindon, S. *et al.* New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Syst. Biol.* **59,** 307–21 (2010).

29. Abascal, F., Zardoya, R. & Posada, D. ProtTest: selection of best-fit models of protein evolution. *Bioinformatics* **21,** 2104–5 (2005).

Figures

Table 1: Description of the fourteen contigs found in *T. urticae* viromes.

Baltimore classification	Viral Family	Viral contigs	Contig length (nt)	Putative protein	Virus best hit (BLASTx)	Accession Number	Protein identity (%)
ssDNA	*Parvoviridae*	*Tetranychus urticae associated ambidensovirus*	2859	polymerase (NS1)	*Lupine feces-associated densovirus 2*	ASM93489.1	39%
				capsid		ASM93488.1	30%
dsRNA	*Birnaviridae*	*Tetranychus urticae associated entomobirnavirus* Segment A	2899	polymerase	*Infectious bursal disease virus*	AAS10174.1	30%
		Tetranychus urticae associated entomobirnavirus Segment B	2501	polyprotein	*Blotched snakehead virus*	YP_052864.1	33%
ssRNA	*Nodaviridae*	*Tetranychus urticae associated nodavirus* Segment A1	3230	polymerase	*Hubei noda-like virus 9*	APG76321.1	58%
		Tetranychus urticae associated nodavirus Segment A2	2538	polymerase	*Hubei noda-like virus 8*	YP_009337881.1	62%
		Tetranychus urticae associated nodavirus Segment B1	1661	capsid	*Hubei noda-like virus 9*	YP_009337880.1	21%
		Tetranychus urticae associated nodavirus Segment B2	1658	capsid			53%
		Tetranychus urticae associated nodavirus Segment B3	1570	capsid			58%
	Unclassified Picornavirales	*Tetranychus urticae associated dicistrovirus 1*	8290	polymerase	*Beihai picorna-like virus 70*	APG78061.1	24%
				capsid		APG78062.1	20%
		Tetranychus urticae associated dicistrovirus 2	8449	polymerase		APG78061.1	23%
				capsid		APG78062.1	20%
		Aphis glycines virus 1	8590	polymerase	*Aphis glycines virus 1*	AHC72013.1	96%
				capsid		AHC72012.1	99%
		Tetranychus urticae associated picorna-like virus 1	8151	polymerase		AHC72013.1	71%
				capsid		AHC72012.1	75%
		Tetranychus urticae associated picorna-like virus 2	6432	polyprotein	*Hubei picorna-like virus 80*	YP_009337381.1	53%
	Narnaviridae	*Tetranychus urticae associated narnavirus*	2481	polymerase	*Hubei narna-like virus 3*	YP_009337787.1	45%

Table 2: Abundance of the viral contigs found in *T. urticae* viromes. In bold: viral contigs representing > 0.1 of read abundance.

Baltimore classification	Viral Family	Viral contig	Portuguese population	French population
ssDNA	*Parvoviridae*	*Tetranychus urticae associated ambidensovirus*	**28,84%**	**28,29%**
dsRNA	*Birnaviridae*	*Tetranychus urticae associated entomobirnavirus* Segment A	**0,34%**	0,00%
		Tetranychus urticae associated entomobirnavirus Segment B	**3,46%**	0,00%
ss+RNA	*Nodaviridae*	*Tetranychus urticae associated nodavirus* Segment A1	**0,15%**	0,00%
		Tetranychus urticae associated nodavirus Segment A2	**0,36%**	0,00%
		Tetranychus urticae associated nodavirus Segment B1	**0,31%**	0,00%
		Tetranychus urticae associated nodavirus Segment B2	**0,10%**	0,00%
		Tetranychus urticae associated nodavirus Segment B3	**0,43%**	0,00%
	Dicistroviridae	*Tetranychus urticae associated dicistrovirus 1*	**4,47%**	**0,38%**
		Tetranychus urticae associated dicistrovirus 2	**3,01%**	0,00%
	Unclassified Picornavirales	*Aphis glycines virus 1*	**9,36%**	**0,13%**
		Tetranychus urticae associated picorna-like virus 1	**10,60%**	**2,70%**
		Tetranychus urticae associated picorna-like virus 2	**0,30%**	0,00%
	Narnaviridae	*Tetranychus urticae associated narnavirus*	**0,15%**	**0,25%**
		Total	**31,75%**	**61,89%**

Table 3: Summary of the number of sequences found in found in *Tetranychus urticae* genomic and/or transcriptomic databases (Refseq_genomic, EST and TSA) that display homologies with viral species discovered in the viromes generated in this study (for further details, see **Table S1**).

		***Tetranychus uticae* genomic/transcriptomic datasets**		
Viral species	**Contig length (nt)**	**ref_seq genomics**	**EST**	**TSA**
Tetranychus urticae associated Parvoviridae	2859	0	0	0
Tetranychus urticae associated Birnaviridae Segment A	2899	0	0	0
Tetranychus urticae associated Birnaviridae Segment B	2501	0	0	0
Tetranychus urticae associated Nodaviridae Segment A 1	3230	0	0	0
Tetranychus urticae associated Nodaviridae Segment A 2	2538	0	0	0
Tetranychus urticae associated Nodaviridae Segment B 1	1661	0	**1**	0
Tetranychus urticae associated Nodaviridae Segment B 2	1658	0	0	0
Tetranychus urticae associated Nodaviridae Segment B 3	1570	0	0	0
Tetranychus urticae associated Dicistroviridae 1	8290	0	0	0
Tetranychus urticae associated Dicistroviridae 2	8449	0	0	0
Tetranychus urticae associated Picorna-like virus 1	8590	0	**105**	0
Tetranychus urticae associated Picorna-like virus 2	8151	0	0	**14**
Tetranychus urticae associated Picorna-like virus 3	6432	0	0	0
Tetranychus urticae associated Narnaviridae	2481	0	0	0

Figure 1. (**A**) Viruses common and uncommon to *T. urticae* laboratory and wild populations (overlapping and not overlapping circles, respectively). Numbers of virus species are indicated in the circles. (**B**) Relative abundance of viral species in *T. urticae* laboratory and wild populations.

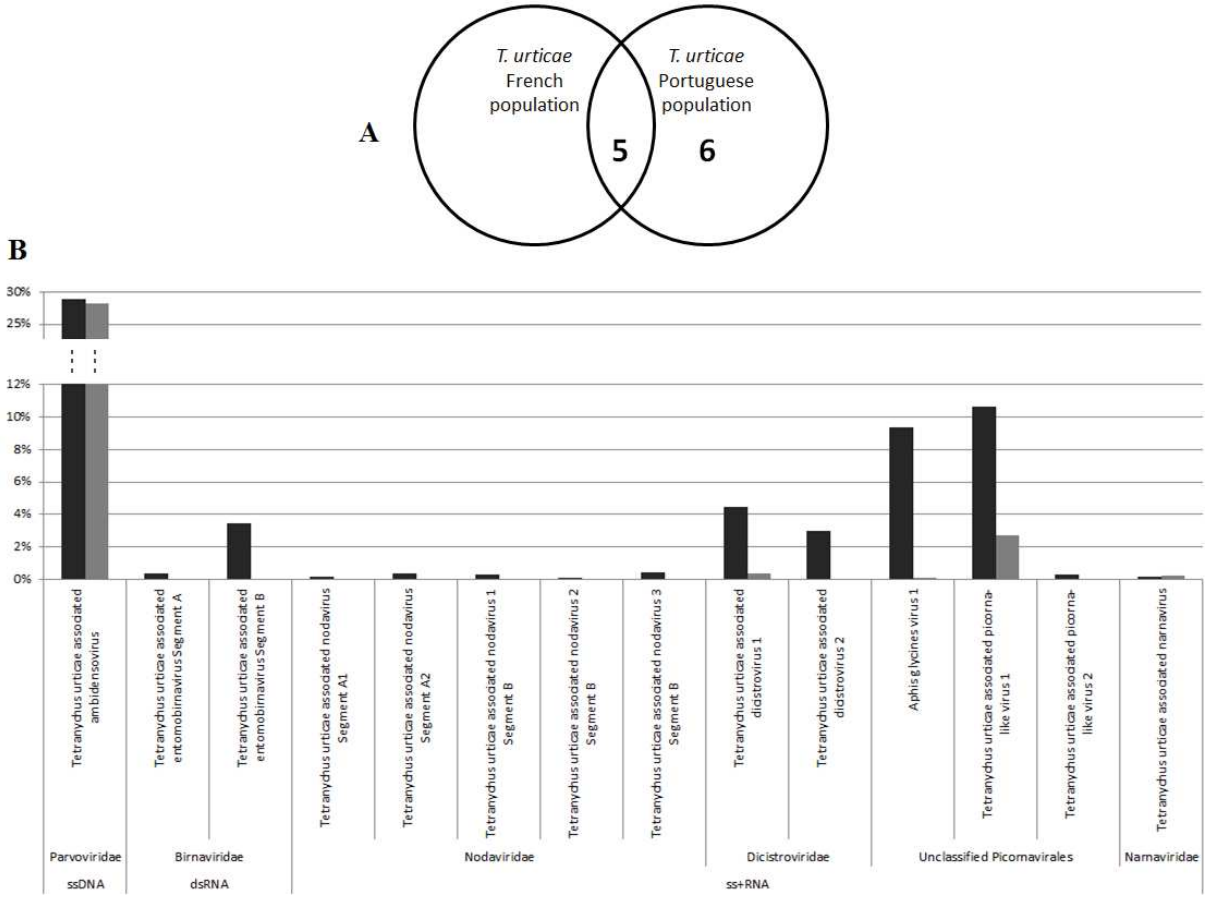


Figure 2. Maximum *Parvoviridae* likelihood phylogenetic tree based on partial SF3 domain of the NS1 protein, including 77 parvovirus species and *Tetranychus urticae associated ambidensovirus* (in bold). The alignment of 126 amino acids in length was produced using MUSCLE 3.7 (16 iterations) and was hand ungapped. The tree was rooted with the SF3 domain of the Variola virus D5 protein. Bootstrap values > 50% are indicated at each node. Scale bars correspond to amino acid substitutions per site. Genera of the *Parvoviridae* family are indicated in brackets. Genomic organization of *Tetranychus urticae associated Ambidensovirus* is also indicated. Grey arrows and rectangles: predicted open reading frames (ORF), Light grey: putative non-structural protein; dark grey: putative structural protein. Arrow: complete ORF; Rectangle: truncated ORF.

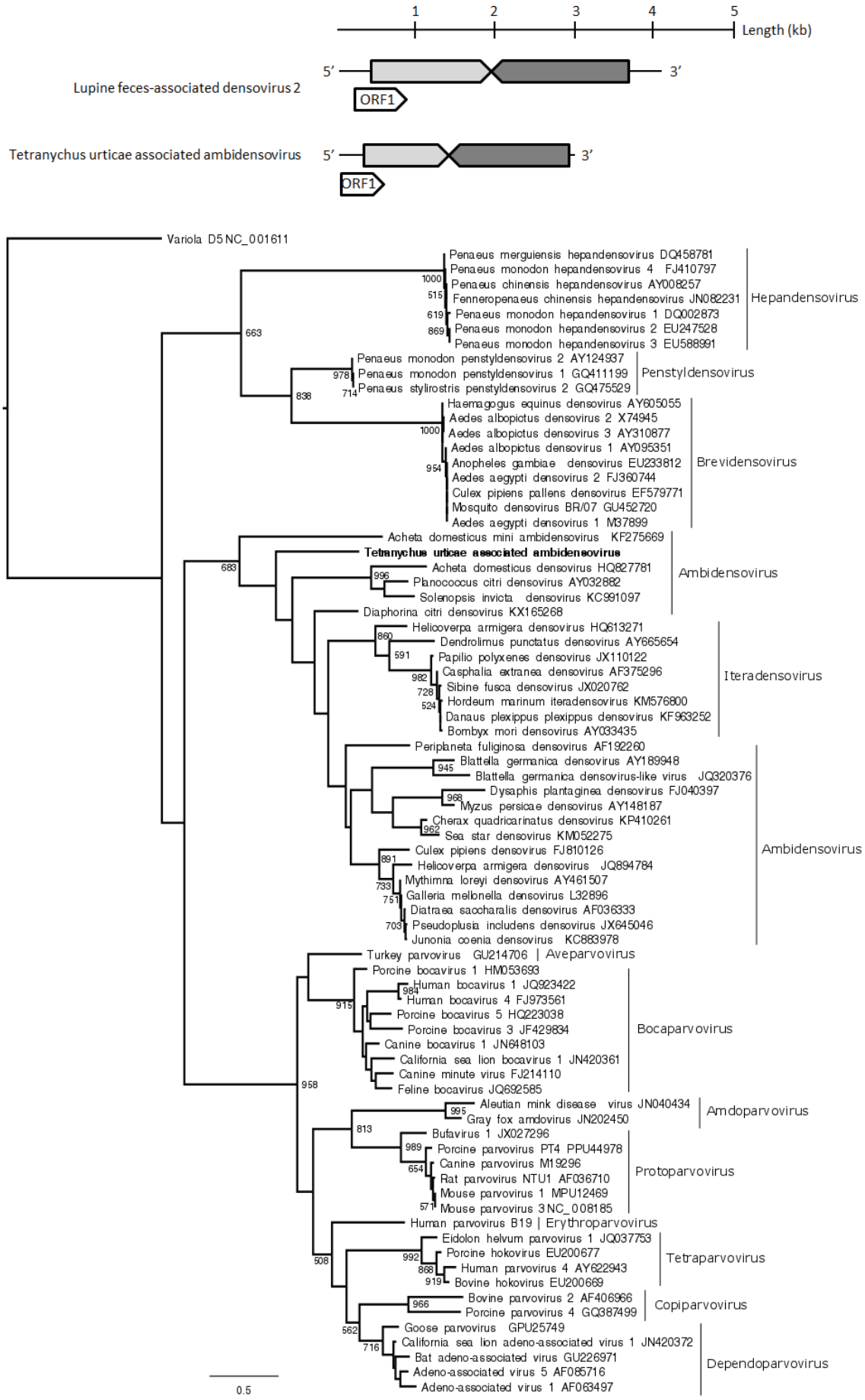

1
2
3
4
5
Length (kb)
5′
3′
Lupine feces-associated densovirus 2
ORF1
Tetranychus urticae associated ambidensovirus
5′
3′
ORF1
Variola D5 NC_001611
Penaeus merguiensis hepandensovirus DQ458781
Penaeus monodon hepandensovirus 4 FJ410797
Penaeus chinensis hepandensovirus AY008257
Fenneropenaeus chinensis hepandensovirus JN082231
Penaeus monodon hepandensovirus 1 DQ002873
Penaeus monodon hepandensovirus 2 EU247528
Penaeus monodon hepandensovirus 3 EU588991
Hepandensovirus
Penaeus monodon penstyldensovirus 2 AY124937
Penaeus monodon penstyldensovirus 1 GQ411199
Penaeus stylirostris penstyldensovirus 2 GQ475529
Penstyldensovirus
Haemagogus equinus densovirus AY605055
Aedes albopictus densovirus 2 X74945
Aedes albopictus densovirus 3 AY310877
Aedes albopictus densovirus 1 AY095351
Anopheles gambiae densovirus EU233812
Aedes aegypti densovirus 2 FJ360744
Culex pipiens pallens densovirus EF579771
Mosquito densovirus BR/07 GU452720
Aedes aegypti densovirus 1 M37899
Brevidensovirus
Acheta domesticus mini ambidensovirus KF275669
Tetranychus urticae associated ambidensovirus
Acheta domesticus densovirus HQ827781
Planococcus citri densovirus AY032882
Solenopsis invicta densovirus KC991097
Ambidensovirus
Diaphorina citri densovirus KX165268
Helicoverpa armigera densovirus HQ613271
Dendrolimus punctatus densovirus AY665654
Papilio polyxenes densovirus JX110122
Casphalia extranea densovirus AF375296
Sibine fusca densovirus JX020762
Hordeum marinum iteradensovirus KM576800
Danaus plexippus plexippus densovirus KF963252
Bombyx mori densovirus AY033435
Iteradensovirus
Periplaneta fuliginosa densovirus AF192260
Blattella germanica densovirus AY189948
Blattella germanica densovirus-like virus JQ320376
Dysaphis plantaginea densovirus FJ040397
Myzus persicae densovirus AY148187
Cherax quadricarinatus densovirus KP410261
Sea star densovirus KM052275
Culex pipiens densovirus FJ810126
Helicoverpa armigera densovirus JQ894784
Mythimna loreyi densovirus AY461507
Galleria mellonella densovirus L32896
Diatraea saccharalis densovirus AF036333
Pseudoplusia includens densovirus JX645046
Junonia coenia densovirus KC883978
Ambidensovirus
Turkey parvovirus GU214706 | Aveparvovirus
Porcine bocavirus 1 HM053693
Human bocavirus 1 JQ923422
Human bocavirus 4 FJ973561
Porcine bocavirus 5 HQ223038
Porcine bocavirus 3 JF429834
Canine bocavirus 1 JN648103
California sea lion bocavirus 1 JN420361
Canine minute virus FJ214110
Feline bocavirus JQ692585
Bocaparvovirus
Aleutian mink disease virus JN040434
Gray fox amdovirus JN202450
Amdoparvovirus
Bufavirus 1 JX027296
Porcine parvovirus PT4 PPU44978
Canine parvovirus M19296
Rat parvovirus NTU1 AF036710
Mouse parvovirus 1 MPU12469
Mouse parvovirus 3 NC_008185
Protoparvovirus
Human parvovirus B19 | Erythroparvovirus
Eidolon helvum parvovirus 1 JQ037753
Porcine hokovirus EU200677
Human parvovirus 4 AY622943
Bovine hokovirus EU200669
Tetraparvovirus
Bovine parvovirus 2 AF406966
Porcine parvovirus 4 GQ387499
Copiparvovirus
Goose parvovirus GPU25749
California sea lion adeno-associated virus 1 JN420372
Bat adeno-associated virus GU226971
Adeno-associated virus 5 AF085716
Adeno-associated virus 1 AF063497
Dependoparvovirus
0.5

Figure 3. Maximum *Birnaviridae* likelihood phylogenetic tree based on the partial polyprotein, including 8 birnavirus species and *Tetranychus urticae associated entomobirnavirus* (in bold). The alignment of 328 amino acids in length was produced using MUSCLE 3.7 (16 iterations) and hand ungapped. The tree was mid-point rooted. Bootstrap values > 50% are indicated at each node. Scale bars correspond to amino acid substitutions per site. Genera of the *Birnaviridae* family are indicated in brackets. Genomic organization of *Tetranychus urticae associated entomobirnavirus* is also indicated. Grey arrows and rectangles: predicted open reading frames (ORF), Light grey: putative non-structural protein; dark grey: putative structural protein. Arrow: complete ORF; Rectangle: truncated ORF.

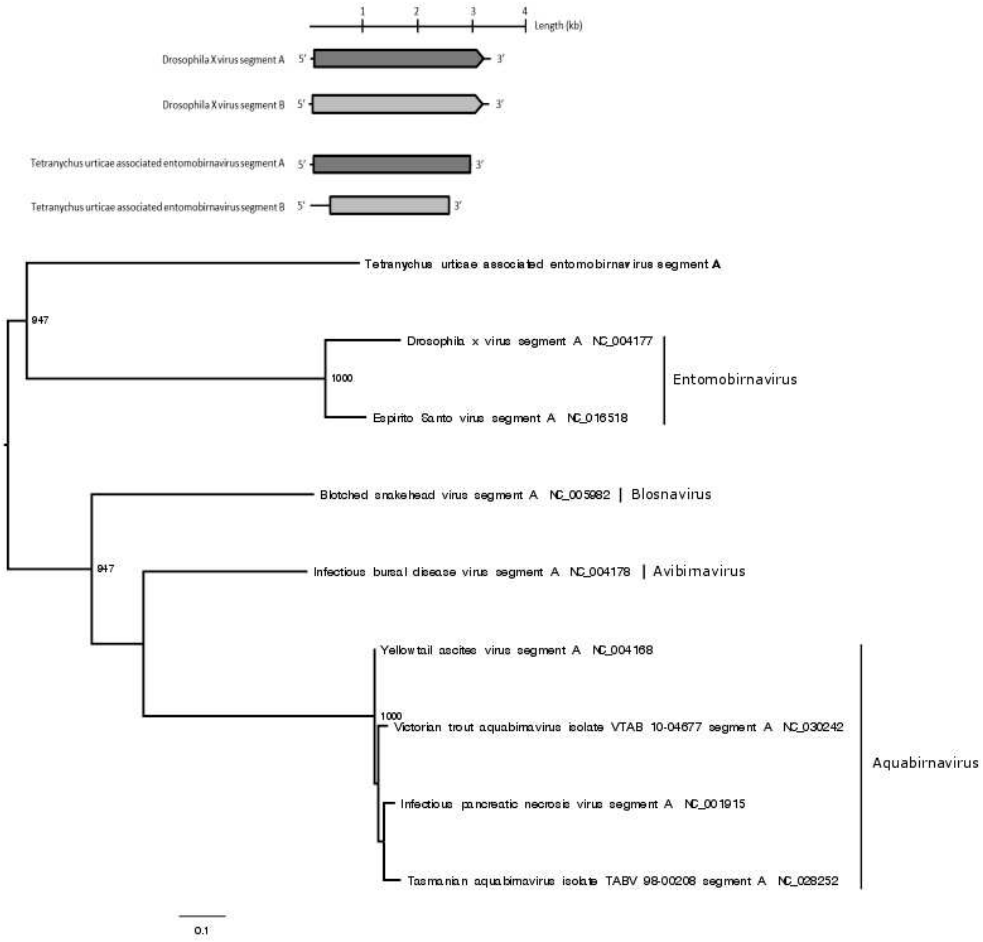
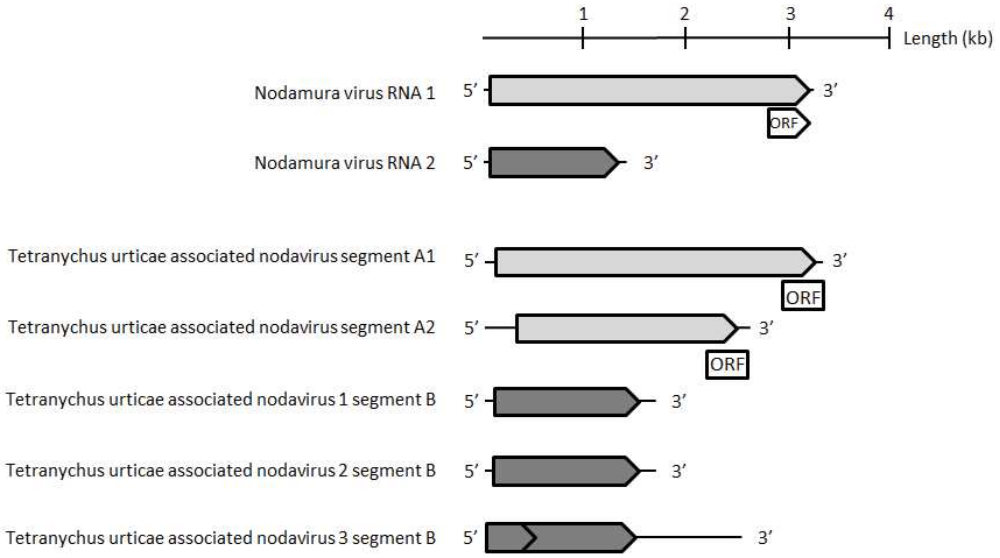


Figure 4. Maximum *Nodaviridae* likelihood phylogenetic tree based on the partial capsid protein, including 14 nodavirus species and *Tetranychus urticae* associated nodaviruses (in bold). The alignment of 122 amino acids in length was produced using MUSCLE 3.7 (16 iterations) and hand ungapped. The tree was mid-point rooted. Bootstrap values > 50% are indicated at each node. Scale bars correspond to amino acid substitutions per site. Genera of the *Nodaviridae* family are indicated in brackets. Genomic organization of *Tetranychus urticae* associated nodaviruses is also indicated. Grey arrows and rectangles: predicted open reading frames (ORF), Light grey: putative non-structural protein; dark grey: putative structural protein. Arrow: complete ORF; Rectangle: truncated ORF.


1
2
3
4
Length (kb)
Nodamura virus RNA 1
5′
3′
ORF
Nodamura virus RNA 2
5′
3′
Tetranychus urticae associated nodavirus segment A1
5′
3′
ORF
Tetranychus urticae associated nodavirus segment A2
5′
3′
ORF
Tetranychus urticae associated nodavirus 1 segment B
5′
3′
Tetranychus urticae associated nodavirus 2 segment B
5′
3′
Tetranychus urticae associated nodavirus 3 segment B
5′
3′

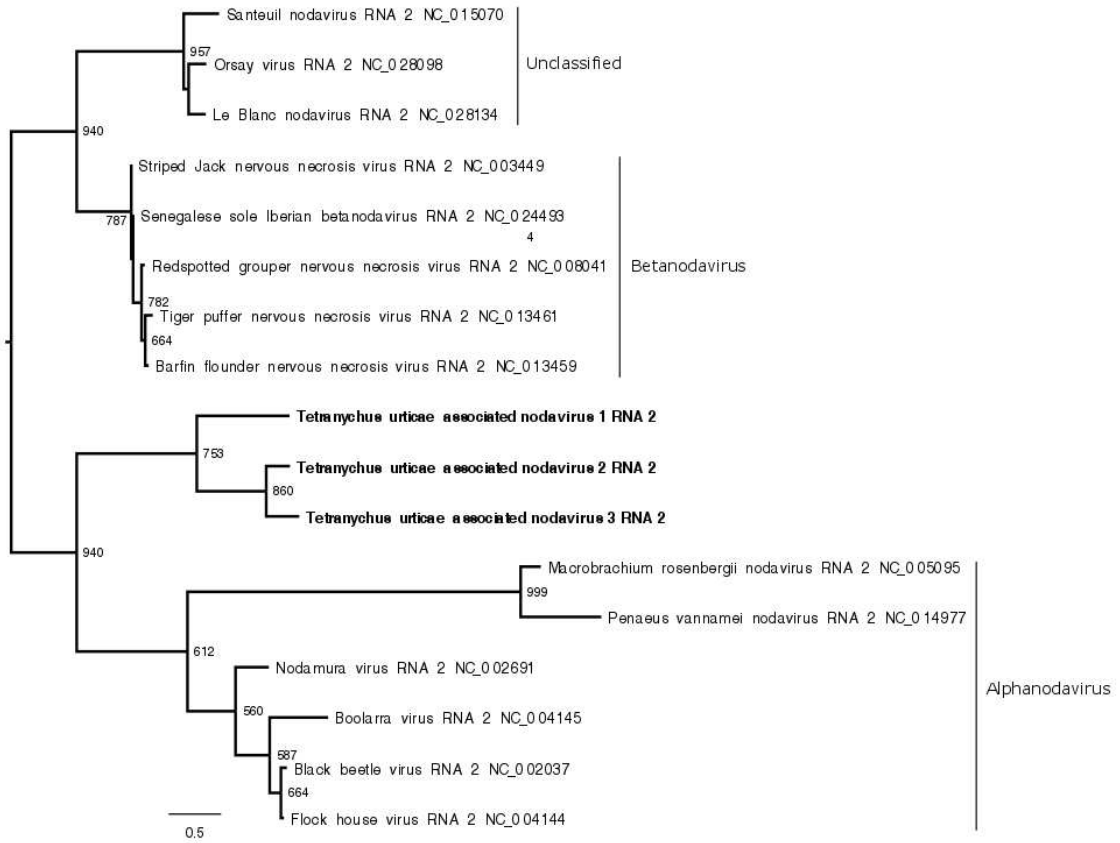

Santeuil nodavirus RNA 2 NC_015070
957
Orsay virus RNA 2 NC_028098
Le Blanc nodavirus RNA 2 NC_028134
Unclassified
940
Striped Jack nervous necrosis virus RNA 2 NC_003449
787
Senegalese sole Iberian betanodavirus RNA 2 NC_024493
4
Redspotted grouper nervous necrosis virus RNA 2 NC_008041
Betanodavirus
782
Tiger puffer nervous necrosis virus RNA 2 NC_013461
664
Barfin flounder nervous necrosis virus RNA 2 NC_013459
Tetranychus urticae associated nodavirus 1 RNA 2
753
Tetranychus urticae associated nodavirus 2 RNA 2
860
Tetranychus urticae associated nodavirus 3 RNA 2
940
Macrobrachium rosenbergii nodavirus RNA 2 NC_005095
999
Penaeus vannamei nodavirus RNA 2 NC_014977
612
Nodamura virus RNA 2 NC_002691
Alphanodavirus
560
Boolarra virus RNA 2 NC_004145
587
Black beetle virus RNA 2 NC_002037
664
Flock house virus RNA 2 NC_004144
0.5

Figure 5. Maximum *Picornavirales* likelihood phylogenetic tree based on the partial polymerase protein, including 86 species and *Tetranychus urticae* associated picornaviruses (in bold). The alignment of 325 amino acids in length was produced using MUSCLE 3.7 (16 iterations) and hand ungapped. The tree was mid-point rooted. Bootstrap values > 50% are indicated at each node. Scale bars correspond to amino acid substitutions per site. Families of the *Picornavirales* order and genera of the *Dicistroviridae* family are indicated in brackets. Genomic organization of *Tetranychus urticae* associated picornaviruses is also indicated. Grey arrows and rectangles: predicted open reading frames (ORF), Light grey: putative non-structural protein; dark grey: putative structural protein. Arrow: complete ORF; Rectangle: truncated ORF.

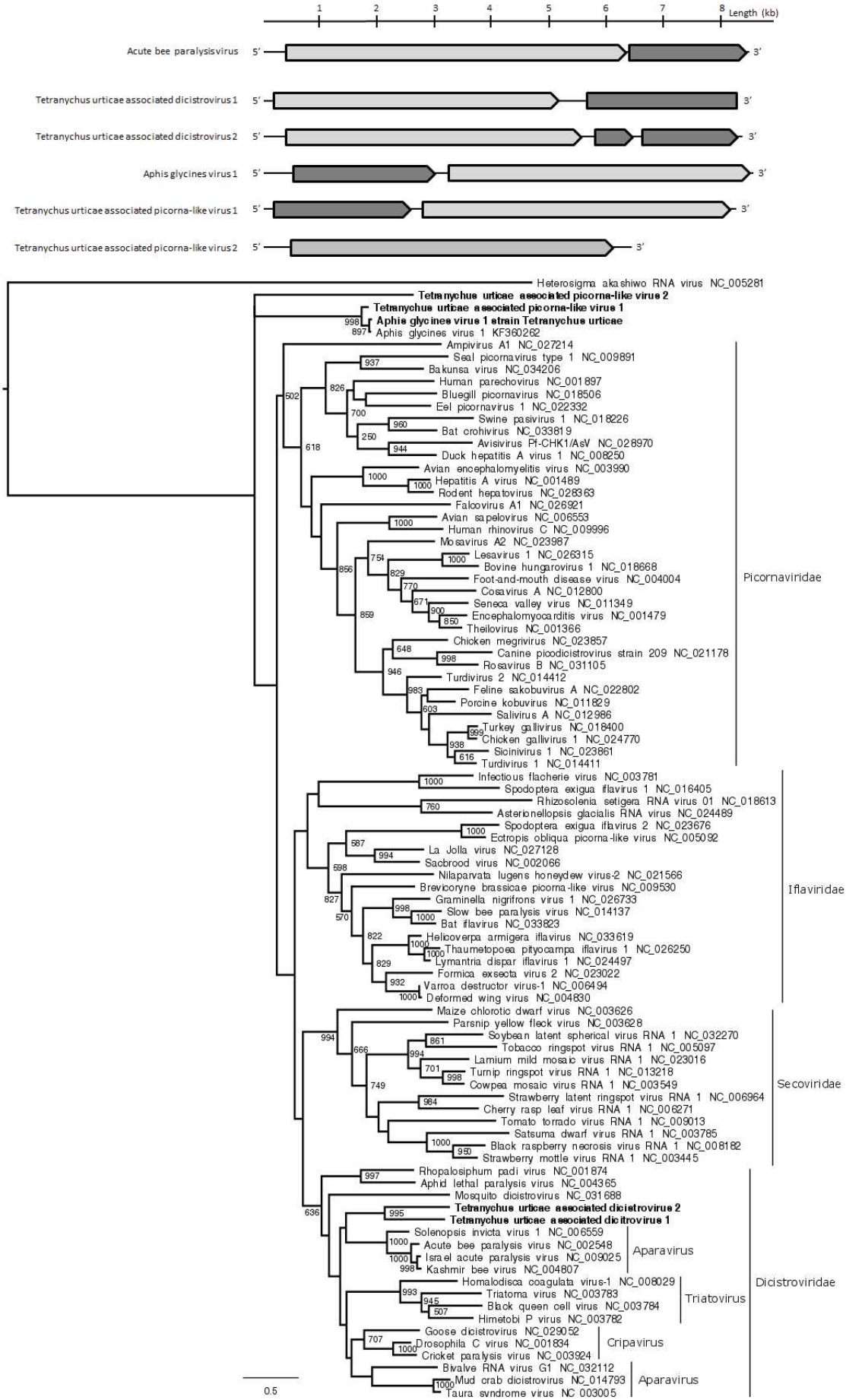
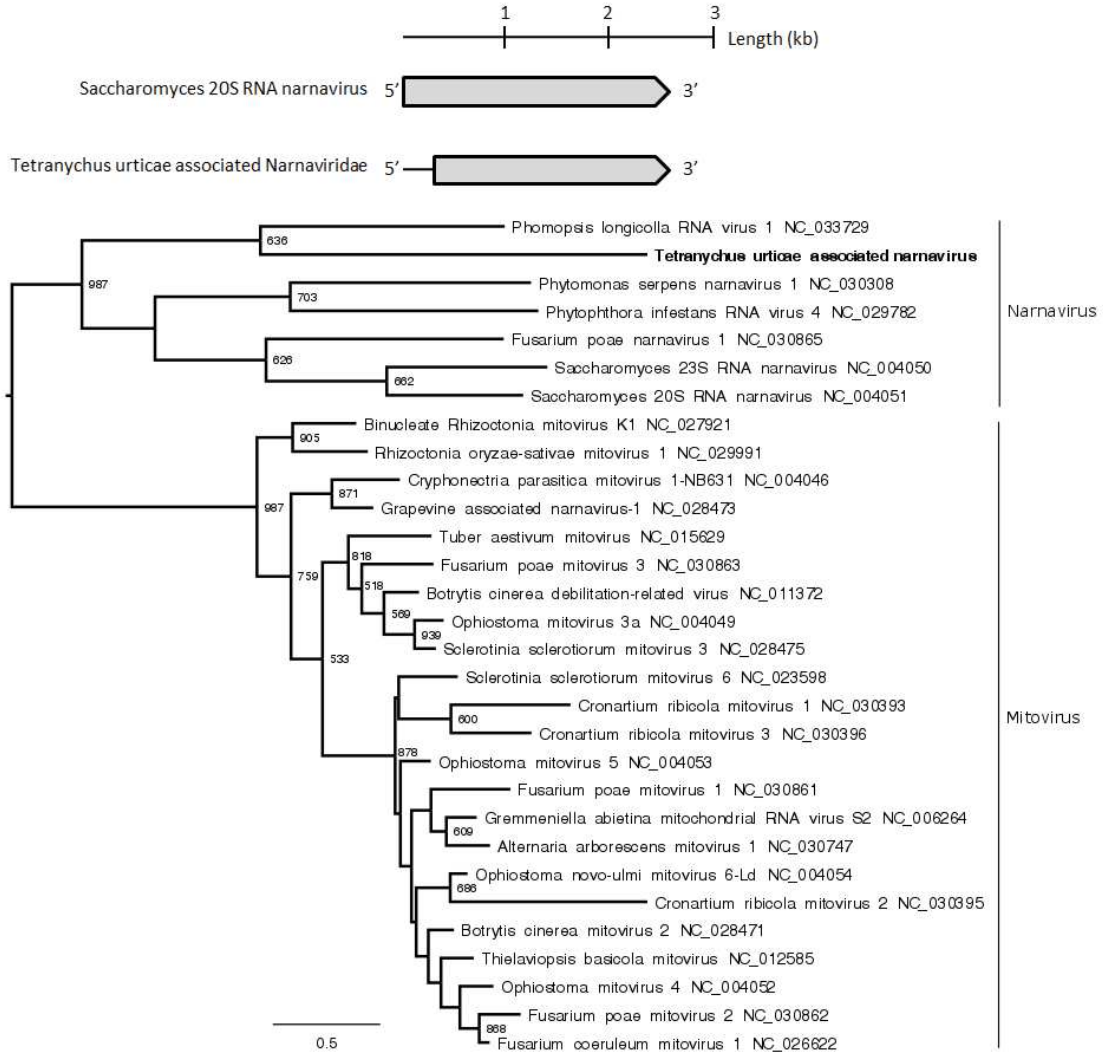

1
2
3
4
5
6
7
8
Length (kb)
Acute bee paralysisvirus
Tetranychus urticae associated dicistrovirus 1
Tetranychus urticae associated dicistrovirus 2
Aphis glycines virus 1
Tetranychus urticae associated picorna-like virus 1
Tetranychus urticae associated picorna-like virus 2
5'
3'
Heterosigma akashiwo RNA virus NC_005281
Tetranychus urticae associated picorna-like virus 2
Tetranychus urticae associated picorna-like virus 1
Aphis glycines virus 1 strain Tetranychus urticae
Aphis glycines virus 1 KF360262
Ampivirus A1 NC_027214
Seal picornavirus type 1 NC_009891
Bakunsa virus NC_034206
Human parechovirus NC_001897
Bluegill picornavirus NC_018506
Eel picornavirus 1 NC_022332
Swine pasivirus 1 NC_018226
Bat crohivirus NC_033819
Avisivirus Pf-CHK1/AsV NC_028970
Duck hepatitis A virus 1 NC_008250
Avian encephalomyelitis virus NC_003990
Hepatitis A virus NC_001489
Rodent hepatovirus NC_028363
Falcovirus A1 NC_026921
Avian sapelovirus NC_006553
Human rhinovirus C NC_009996
Mosavirus A2 NC_023987
Lesavirus 1 NC_026315
Bovine hungarovirus 1 NC_018668
Foot-and-mouth disease virus NC_004004
Cosavirus A NC_012800
Seneca valley virus NC_011349
Encephalomyocarditis virus NC_001479
Theilovirus NC_001366
Chicken megrivirus NC_023857
Canine picodicistrovirus strain 209 NC_021178
Rosavirus B NC_031105
Turdivirus 2 NC_014412
Feline sakobuvirus A NC_022802
Porcine kobuvirus NC_011829
Salivirus A NC_012986
Turkey gallivirus NC_018400
Chicken gallivirus 1 NC_024770
Sicinivirus 1 NC_023861
Turdivirus 1 NC_014411
Picornaviridae
Infectious flacherie virus NC_003781
Spodoptera exigua iflavirus 1 NC_016405
Rhizosolenia setigera RNA virus 01 NC_018613
Asterionellopsis glacialis RNA virus NC_024489
Spodoptera exigua iflavirus 2 NC_023676
Ectropis obliqua picorna-like virus NC_005092
La Jolla virus NC_027128
Sacbrood virus NC_002066
Nilaparvata lugens honeydew virus-2 NC_021566
Brevicoryne brassicae picorna-like virus NC_009530
Graminella nigrifrons virus 1 NC_026733
Slow bee paralysis virus NC_014137
Bat iflavirus NC_033823
Helicoverpa armigera iflavirus NC_033619
Thaumetopoea pityocampa iflavirus 1 NC_026250
Lymantria dispar iflavirus 1 NC_024497
Formica exsecta virus 2 NC_023022
Varroa destructor virus-1 NC_006494
Deformed wing virus NC_004830
Iflaviridae
Maize chlorotic dwarf virus NC_003626
Parsnip yellow fleck virus NC_003628
Soybean latent spherical virus RNA 1 NC_032270
Tobacco ringspot virus RNA 1 NC_005097
Lamium mild mosaic virus RNA 1 NC_023016
Turnip ringspot virus RNA 1 NC_013218
Cowpea mosaic virus RNA 1 NC_003549
Strawberry latent ringspot virus RNA 1 NC_006964
Cherry rasp leaf virus RNA 1 NC_006271
Tomato torrado virus RNA 1 NC_009013
Satsuma dwarf virus RNA 1 NC_003785
Black raspberry necrosis virus RNA 1 NC_008182
Strawberry mottle virus RNA 1 NC_003445
Secoviridae
Rhopalosiphum padi virus NC_001874
Aphid lethal paralysis virus NC_004365
Mosquito dicistrovirus NC_031688
Tetranychus urticae associated dicistrovirus 2
Tetranychus urticae associated dicitrovirus 1
Solenopsis invicta virus 1 NC_006559
Acute bee paralysis virus NC_002548
Israel acute paralysis virus NC_009025
Kashmir bee virus NC_004807
Aparavirus
Homalodisca coagulata virus-1 NC_008029
Triatoma virus NC_003783
Black queen cell virus NC_003784
Himetobi P virus NC_003782
Triatovirus
Goose dicistrovirus NC_029052
Drosophila C virus NC_001834
Cricket paralysis virus NC_003924
Cripavirus
Bivalve RNA virus G1 NC_032112
Mud crab dicistrovirus NC_014793
Taura syndrome virus NC_003005
Aparavirus
Dicistroviridae
0.5

Figure 6. Maximum *Narnaviridae* likelihood phylogenetic tree based on the partial polymerase protein, including 29 narnavirus species and *Tetranychus urticae associated narnavirus* (in bold). The alignment of 127 amino acids in length was produced using MUSCLE 3.7 (16 iterations) and hand ungapped. The tree was mid-point rooted. Bootstrap values > 50% are indicated at each node. Scale bars correspond to amino acid substitutions per site. Genera of the *Narnaviridae* family are indicated in brackets. Genomic organization of *Tetranychus urticae associated narnavirus* is also indicated. Grey arrows and rectangles: predicted open reading frames (ORF), Light grey: putative non-structural protein; dark grey: putative structural protein. Arrow: complete ORF; Rectangle: truncated ORF.



Supplementary figures

Table S1: Detailed information about sequences found in found in *Tetranychus urticae* genomic and/or transcriptomic databases (Refseq_genomic, EST and TSA) that display homologies with viral species discovered in the viromes generated in this study.

Viral species	Contig length (nt)	Database	Best Hit Description	Query cover (%)	E-value	Identity (%)	Accession Number
Tetranychus urticae associated Picorna-like virus 1	8590	EST	CCIO4537.b1 CCIO Tetranychus urticae London embryo (H) Tetranychus urticae cDNA clone CCIO4537 5', mRNA sequence	9%	0.0	98%	GW002029.1
			CCZA2763.b1 CCZA Tetranychus urticae London larvae normalized Tetranychus urticae cDNA clone CCZA2763 5', mRNA sequence	9%	0.0	98%	GW010081.1
			CCZB3128.b1 CCZB Tetranychus urticae London adult normalized Tetranychus urticae cDNA clone CCZB3128 5', mRNA sequence	9%	0.0	97%	GW039799.1
			CCIO1765.b1 CCIO Tetranychus urticae London embryo (H) Tetranychus urticae cDNA clone CCIO1765 5', mRNA sequence	9%	0.0	97%	GT997100.1
			CCIO5452.b1 CCIO Tetranychus urticae London embryo (H) Tetranychus urticae cDNA clone CCIO5452 5', mRNA sequence	9%	0.0	98%	GW003803.1
			CCZA15068.b1 CCZA Tetranychus urticae London larvae normalized Tetranychus urticae cDNA clone CCZA15068 5', mRNA sequence	9%	0.0	98%	GW031879.1
			CCIN6513.b1 CCIN Tetranychus urticae London embryo (L) Tetranychus urticae cDNA clone CCIN6513 5', mRNA sequence	8%	0.0	98%	GT993695.1
			CCIO4391.b1 CCIO Tetranychus urticae London embryo (H) Tetranychus urticae cDNA clone CCIO4391 5', mRNA sequence	9%	0.0	96%	GW001762.1
			CCZA13690.b1 CCZA Tetranychus urticae London larvae normalized Tetranychus urticae cDNA clone CCZA13690 5', mRNA sequence	8%	0.0	98%	GW029425.1
			CCZA8138.b1 CCZA Tetranychus urticae London larvae normalized Tetranychus urticae cDNA clone CCZA8138 5', mRNA sequence	9%	0.0	97%	GW020009.1
			CCZA13690.g1 CCZA Tetranychus urticae London larvae normalized Tetranychus urticae cDNA clone CCZA13690 3', mRNA sequence	8%	0.0	97%	GW029426.1
			CCZA10269.b1 CCZA Tetranychus urticae London larvae normalized Tetranychus urticae cDNA clone CCZA10269 5', mRNA sequence	8%	0.0	98%	GW022207.1
			CCZB12088.b1 CCZB Tetranychus urticae London adult normalized Tetranychus urticae cDNA clone CCZB12088 5', mRNA sequence	8%	0.0	98%	GW056395.1
			CCIO3913.b1 CCIO Tetranychus urticae London embryo (H) Tetranychus urticae cDNA clone CCIO3913 5', mRNA sequence	8%	0.0	98%	GW000846.1
			CCZA2557.b1 CCZA Tetranychus urticae London larvae normalized Tetranychus urticae cDNA clone CCZA2557 5', mRNA sequence	8%	0.0	98%	GW009699.1
			CCIN5995.b1 CCIN Tetranychus urticae London embryo (L) Tetranychus urticae cDNA clone CCIN5995 5', mRNA sequence	9%	0.0	96%	GT992614.1
			CCZA15700.b1 CCZA Tetranychus urticae London larvae normalized Tetranychus urticae cDNA clone CCZA15700 5', mRNA sequence	9%	0.0	96%	GW033019.1
			CCZA8418.b1 CCZA Tetranychus urticae London larvae normalized Tetranychus urticae cDNA clone CCZA8418 5', mRNA sequence	8%	0.0	97%	GW020536.1
			CCZA5505.b1 CCZA Tetranychus urticae London larvae normalized Tetranychus urticae cDNA clone CCZA5505 5', mRNA sequence	8%	0.0	97%	GW015125.1
			CCZB6366.b1 CCZB Tetranychus urticae London adult normalized Tetranychus urticae cDNA clone CCZB6366 5', mRNA sequence	8%	0.0	98%	GW046163.1
			CCZA8282.b1 CCZA Tetranychus urticae London larvae normalized Tetranychus urticae cDNA clone CCZA8282 5', mRNA sequence	9%	0.0	97%	GW020275.1
			CCIO4292.b1 CCIO Tetranychus urticae London embryo (H) Tetranychus urticae cDNA clone CCIO4292 5', mRNA sequence	8%	0.0	96%	GW001574.1
			CCIO3041.b1 CCIO Tetranychus urticae London embryo (H) Tetranychus urticae cDNA clone CCIO3041 5', mRNA sequence	9%	0.0	96%	GT999313.1
			CCZA7044.b1 CCZA Tetranychus urticae London larvae normalized Tetranychus urticae cDNA clone CCZA7044 5', mRNA sequence	9%	0.0	95%	GW017992.1
			CCIN4631.b1 CCIN Tetranychus urticae London embryo (L) Tetranychus urticae cDNA clone CCIN4631 5', mRNA sequence	9%	0.0	96%	GT989770.1
			CCZA12238.b1 CCZA Tetranychus urticae London larvae normalized Tetranychus urticae cDNA clone CCZA12238 5', mRNA sequence	9%	0.0	96%	GW026787.1
			CCZB9114.g1 CCZB Tetranychus urticae London adult normalized Tetranychus urticae cDNA clone CCZB9114 3', mRNA sequence	9%	0.0	95%	GW051701.1
			CCZA14541.b1 CCZA Tetranychus urticae London larvae normalized Tetranychus urticae cDNA clone CCZA14541 5', mRNA sequence	8%	0.0	98%	GW030934.1
			CCZB9114.b1 CCZB Tetranychus urticae London adult normalized Tetranychus urticae cDNA clone CCZB9114 5', mRNA sequence	8%	0.0	98%	GW051700.1
			CCZA8138.g1 CCZA Tetranychus urticae London larvae normalized Tetranychus urticae cDNA clone CCZA8138 3', mRNA sequence	8%	0.0	97%	GW020010.1
			CCZA15068.g1 CCZA Tetranychus urticae London larvae normalized Tetranychus urticae cDNA clone CCZA15068 3', mRNA sequence	8%	0.0	97%	GW031880.1
			CCZA16505.b1 CCZA Tetranychus urticae London larvae normalized Tetranychus urticae cDNA clone CCZA16505 5', mRNA sequence	8%	0.0	96%	GW034462.1
			CCZA13023.b1 CCZA Tetranychus urticae London larvae normalized Tetranychus urticae cDNA clone CCZA13023 5', mRNA sequence	8%	0.0	97%	GW028229.1
			CCZA8861.b1 CCZA Tetranychus urticae London larvae normalized Tetranychus urticae cDNA clone CCZA8861 5', mRNA sequence	8%	0.0	97%	GW021341.1
			CCZA7338.g1 CCZA Tetranychus urticae London larvae normalized Tetranychus urticae cDNA clone CCZA7338 3', mRNA sequence	8%	0.0	98%	GW018532.1
			CCZA13376.b1 CCZA Tetranychus urticae London larvae normalized Tetranychus urticae cDNA clone CCZA13376 5', mRNA sequence	8%	0.0	96%	GW028854.1
			CCZB11530.b1 CCZB Tetranychus urticae London adult normalized Tetranychus urticae cDNA clone CCZB11530 5', mRNA sequence	8%	0.0	96%	GW055363.1
			CCZA15572.b1 CCZA Tetranychus urticae London larvae normalized Tetranychus urticae cDNA clone CCZA15572 5', mRNA sequence	8%	0.0	96%	GW012788.1
			CCZA12709.b1 CCZA Tetranychus urticae London larvae normalized Tetranychus urticae cDNA clone CCZA12709 5', mRNA sequence	8%	0.0	96%	GW027651.1
			CCZB13861.b1 CCZB Tetranychus urticae London adult normalized Tetranychus urticae cDNA clone CCZB13861 5', mRNA sequence	8%	0.0	96%	GW059560.1
			CCZA8695.b1 CCZA Tetranychus urticae London larvae normalized Tetranychus urticae cDNA clone CCZA8695 5', mRNA sequence	8%	0.0	95%	GW021039.1
			CCZB5452.g1 CCZB Tetranychus urticae London adult normalized Tetranychus urticae cDNA clone CCZB5452 3', mRNA sequence	8%	0.0	95%	GW044334.1
			CCZB11993.b1 CCZB Tetranychus urticae London adult normalized Tetranychus urticae cDNA clone CCZB11993 5', mRNA sequence	8%	0.0	96%	GW017697.1
			CCZA5068.b1 CCZA Tetranychus urticae London larvae normalized Tetranychus urticae cDNA clone CCZA5068 5', mRNA sequence	8%	0.0	96%	GW014308.1
			CCZA16090.b1 CCZA Tetranychus urticae London larvae normalized Tetranychus urticae cDNA clone CCZA16090 5', mRNA sequence	8%	0.0	98%	GW031717.1
			CCZA9053.b1 CCZA Tetranychus urticae London larvae normalized Tetranychus urticae cDNA clone CCZA9053 5', mRNA sequence	8%	0.0	96%	GW021688.1
			CCZA4808.b1 CCZA Tetranychus urticae London larvae normalized Tetranychus urticae cDNA clone CCZA4808 5', mRNA sequence	8%	0.0	96%	GW013817.1
			CCZA10100.b1 CCZA Tetranychus urticae London larvae normalized Tetranychus urticae cDNA clone CCZA10100 5', mRNA sequence	8%	0.0	96%	GW022900.1
			CCZA7338.b1 CCZA Tetranychus urticae London larvae normalized Tetranychus urticae cDNA clone CCZA7338 5', mRNA sequence	8%	0.0	95%	GW018531.1
			CCZB14383.b1 CCZB Tetranychus urticae London adult normalized Tetranychus urticae cDNA clone CCZB14383 5', mRNA sequence	8%	0.0	96%	GW060484.1
			CCZA16874.b1 CCZA Tetranychus urticae London larvae normalized Tetranychus urticae cDNA clone CCZA16874 5', mRNA sequence	7%	0.0	97%	GW035115.1
			CCZB4341.b1 CCZB Tetranychus urticae London adult normalized Tetranychus urticae cDNA clone CCZB4341 5', mRNA sequence	8%	0.0	96%	GW042112.1
			CCZA7011.b1 CCZA Tetranychus urticae London larvae normalized Tetranychus urticae cDNA clone CCZA7011 5', mRNA sequence	8%	0.0	95%	GW017931.1
			CCZA5091.b1 CCZA Tetranychus urticae London larvae normalized Tetranychus urticae cDNA clone CCZA5091 5', mRNA sequence	7%	0.0	96%	GW014352.1
			CCZA11727.b1 CCZA Tetranychus urticae London larvae normalized Tetranychus urticae cDNA clone CCZA11727 5', mRNA sequence	8%	0.0	95%	GW025852.1
			CCZB5452.b1 CCZB Tetranychus urticae London adult normalized Tetranychus urticae cDNA clone CCZB5452 5', mRNA sequence	8%	0.0	95%	GW044333.1
			CCIO2762.b1 CCIO Tetranychus urticae London embryo (H) Tetranychus urticae cDNA clone CCIO2762 5', mRNA sequence	7%	0.0	95%	GT998825.1
			CCZA1698.b1 CCZA Tetranychus urticae London larvae normalized Tetranychus urticae cDNA clone CCZA1698 5', mRNA sequence	6%	0.0	99%	GW008097.1
			CCZA1646.b1 CCZA Tetranychus urticae London larvae normalized Tetranychus urticae cDNA clone CCZA1646 5', mRNA sequence	6%	0.0	95%	GW008003.1
			CCZA2557.g1 CCZA Tetranychus urticae London larvae normalized Tetranychus urticae cDNA clone CCZA2557 3', mRNA sequence	5%	0.0	99%	GW009700.1
			CCZA12399.b1 CCZA Tetranychus urticae London larvae normalized Tetranychus urticae cDNA clone CCZA12399 5', mRNA sequence	5%	0.0	99%	GW027081.1
			CCZA8418.g1 CCZA Tetranychus urticae London larvae normalized Tetranychus urticae cDNA clone CCZA8418 3', mRNA sequence	5%	0.0	99%	GW020527.1
			CCZA5505.g1 CCZA Tetranychus urticae London larvae normalized Tetranychus urticae cDNA clone CCZA5505 3', mRNA sequence	5%	0.0	99%	GW015126.1
			CCIO1745.g1 CCIO Tetranychus urticae London embryo (H) Tetranychus urticae cDNA clone CCIO1745 3', mRNA sequence	5%	0.0	99%	GT997101.1
			CCZB11993.g1 CCZB Tetranychus urticae London adult normalized Tetranychus urticae cDNA clone CCZB11993 3', mRNA sequence	5%	0.0	99%	GW017698.1
			CCZA5091.g1 CCZA Tetranychus urticae London larvae normalized Tetranychus urticae cDNA clone CCZA5091 3', mRNA sequence	5%	0.0	99%	GW014353.1
			CCZB3128.g1 CCZB Tetranychus urticae London adult normalized Tetranychus urticae cDNA clone CCZB3128 3', mRNA sequence	5%	0.0	98%	GW039800.1
			CCZA1698.g1 CCZA Tetranychus urticae London larvae normalized Tetranychus urticae cDNA clone CCZA1698 3', mRNA sequence	5%	0.0	98%	GW008098.1
			CCZA8695.g1 CCZA Tetranychus urticae London larvae normalized Tetranychus urticae cDNA clone CCZA8695 3', mRNA sequence	5%	0.0	99%	GW021040.1
			CCIN4631.g1 CCIN Tetranychus urticae London embryo (L) Tetranychus urticae cDNA clone CCIN4631 3', mRNA sequence	5%	0.0	99%	GT989771.1
			CCZB4341.g1 CCZB Tetranychus urticae London adult normalized Tetranychus urticae cDNA clone CCZB4341 3', mRNA sequence	5%	0.0	99%	GW042113.1
			CCZA12238.g1 CCZA Tetranychus urticae London larvae normalized Tetranychus urticae cDNA clone CCZA12238 3', mRNA sequence	5%	0.0	99%	GW026788.1
			CCZA8861.g1 CCZA Tetranychus urticae London larvae normalized Tetranychus urticae cDNA clone CCZA8861 3', mRNA sequence	5%	0.0	99%	GW021342.1
			CCZB12088.g1 CCZB Tetranychus urticae London adult normalized Tetranychus urticae cDNA clone CCZB12088 3', mRNA sequence	5%	0.0	98%	GW056396.1
			CCZB6366.g1 CCZB Tetranychus urticae London adult normalized Tetranychus urticae cDNA clone CCZB6366 3', mRNA sequence	5%	0.0	97%	GW046164.1
			CCIO4391.g1 CCIO Tetranychus urticae London embryo (H) Tetranychus urticae cDNA clone CCIO4391 3', mRNA sequence	5%	0.0	99%	GW001763.1
			CCIO3913.g1 CCIO Tetranychus urticae London embryo (H) Tetranychus urticae cDNA clone CCIO3913 3', mRNA sequence	5%	0.0	99%	GW000847.1
			CCZA15572.g1 CCZA Tetranychus urticae London larvae normalized Tetranychus urticae cDNA clone CCZA15572 3', mRNA sequence	5%	0.0	99%	GW032789.1
			CCZB13861.g1 CCZB Tetranychus urticae London adult normalized Tetranychus urticae cDNA clone CCZB13861 3', mRNA sequence	5%	0.0	99%	GW059561.1
			CCZA7011.g1 CCZA Tetranychus urticae London larvae normalized Tetranychus urticae cDNA clone CCZA7011 3', mRNA sequence	5%	0.0	99%	GW017932.1
			CCZA4808.g1 CCZA Tetranychus urticae London larvae normalized Tetranychus urticae cDNA clone CCZA4808 3', mRNA sequence	5%	0.0	98%	GW013818.1
			CCZA9053.g1 CCZA Tetranychus urticae London larvae normalized Tetranychus urticae cDNA clone CCZA9053 3', mRNA sequence	4%	0.0	98%	GW021689.1
			CCZA13023.g1 CCZA Tetranychus urticae London larvae normalized Tetranychus urticae cDNA clone CCZA13023 3', mRNA sequence	4%	0.0	99%	GW028230.1
			CCIO4292.g1 CCIO Tetranychus urticae London embryo (H) Tetranychus urticae cDNA clone CCIO4292 3', mRNA sequence	4%	0.0	99%	GW001575.1
			CCZA7044.g1 CCZA Tetranychus urticae London larvae normalized Tetranychus urticae cDNA clone CCZA7044 3', mRNA sequence	4%	0.0	97%	GW017993.1
			CCZB14383.g1 CCZB Tetranychus urticae London adult normalized Tetranychus urticae cDNA clone CCZB14383 3', mRNA sequence	4%	0.0	99%	GW060485.1
			CCZA10100.g1 CCZA Tetranychus urticae London larvae normalized Tetranychus urticae cDNA clone CCZA10100 3', mRNA sequence	4%	0.0	99%	GW022901.1
			CCZA6904.b1 CCZA Tetranychus urticae London larvae normalized Tetranychus urticae cDNA clone CCZA6904 5', mRNA sequence	4%	0.0	95%	GW017735.1
			CCZA10269.g1 CCZA Tetranychus urticae London larvae normalized Tetranychus urticae cDNA clone CCZA10269 3', mRNA sequence	4%	0.0	99%	GW022208.1
			CCZA16874.g1 CCZA Tetranychus urticae London larvae normalized Tetranychus urticae cDNA clone CCZA16874 3', mRNA sequence	4%	0.0	99%	GW035116.1
			CCIO5452.g1 CCIO Tetranychus urticae London embryo (H) Tetranychus urticae cDNA clone CCIO5452 3', mRNA sequence	4%	0.0	98%	GW003804.1
			CCZA12399.g1 CCZA Tetranychus urticae London larvae normalized Tetranychus urticae cDNA clone CCZA12399 3', mRNA sequence	4%	3,00E-178	99%	GW027082.1
			CCZA11727.g1 CCZA Tetranychus urticae London larvae normalized Tetranychus urticae cDNA clone CCZA11727 3', mRNA sequence	4%	1,00E-165	99%	GW025853.1
			CCZA16505.g1 CCZA Tetranychus urticae London larvae normalized Tetranychus urticae cDNA clone CCZA16505 3', mRNA sequence	3%	2,00E-156	98%	GW034463.1
			CCZA8834.g1 CCZA Tetranychus urticae London larvae normalized Tetranychus urticae cDNA clone CCZA8834 3', mRNA sequence	3%	7,00E-155	99%	GW021292.1
			CCZA8834.b1 CCZA Tetranychus urticae London larvae normalized Tetranychus urticae cDNA clone CCZA8834 5', mRNA sequence	3%	7,00E-155	99%	GW021291.1
			CCZB2616.g1 CCZB Tetranychus urticae London adult normalized Tetranychus urticae cDNA clone CCZB2616 3', mRNA sequence	3%	1,00E-126	99%	GW038844.1
			CCZB2616.b1 CCZB Tetranychus urticae London adult normalized Tetranychus urticae cDNA clone CCZB2616 5', mRNA sequence	3%	1,00E-126	99%	GW038843.1
			CCZA8315.b1 CCZA Tetranychus urticae London larvae normalized Tetranychus urticae cDNA clone CCZA8315 5', mRNA sequence	2%	6,00E-124	100%	GW020337.1
			CCZA14541.g1 CCZA Tetranychus urticae London larvae normalized Tetranychus urticae cDNA clone CCZA14541 3', mRNA sequence	2%	4,00E-120	99%	GW030935.1
			CCIO3041.g1 CCIO Tetranychus urticae London embryo (H) Tetranychus urticae cDNA clone CCIO3041 3', mRNA sequence	2%	3,00E-109	98%	GT999314.1
			CCIN5995.g1 CCIN Tetranychus urticae London embryo (L) Tetranychus urticae cDNA clone CCIN5995 3', mRNA sequence	2%	9,00E-97	99%	GT992615.1
			CCZA13376.g1 CCZA Tetranychus urticae London larvae normalized Tetranychus urticae cDNA clone CCZA13376 3', mRNA sequence	2%	8,00E-91	100%	GW028855.1
			CCZA13119.g1 CCZA Tetranychus urticae London larvae normalized Tetranychus urticae cDNA clone CCZA13119 3', mRNA sequence	1%	2,00E-41	99%	GW028395.1
			CCZA8843.b1 CCZA Tetranychus urticae London larvae normalized Tetranychus urticae cDNA clone CCZA8843 5', mRNA sequence	0%	1,00E-12	98%	GW017620.1
Tetranychus urticae associated Picorna-like virus 2	8151	TSA	Tetranychus urticae Tu-Houten1_isotig05055 mRNA sequence	17%	0.0	98%	JR695244.1
			Tetranychus urticae Tu-DeLier1_isotig00010 mRNA sequence	14%	0.0	98%	JR673760.1
			Tetranychus urticae Tu-DeLier1_isotig00009 mRNA sequence	14%	0.0	98%	JR673759.1
			Tetranychus urticae Tu-DeLier1_isotig00008 mRNA sequence	14%	0.0	98%	JR673758.1
			Tetranychus urticae Tu-DeLier1_isotig00007 mRNA sequence	14%	0.0	98%	JR673757.1
			Tetranychus urticae Tu-DeLier1_isotig00006 mRNA sequence	14%	0.0	98%	JR673756.1
			Tetranychus urticae Tu-DeLier1_isotig00005 mRNA sequence	14%	0.0	98%	JR673755.1
			Tetranychus urticae Tu-DeLier1_isotig00004 mRNA sequence	14%	0.0	98%	JR673754.1
			Tetranychus urticae Tu-DeLier1_isotig00003 mRNA sequence	14%	0.0	98%	JR673753.1
			Tetranychus urticae Tu-DeLier1_isotig00002 mRNA sequence	14%	0.0	98%	JR673752.1
			Tetranychus urticae Tu-DeLier1_isotig00001 mRNA sequence	14%	0.0	98%	JR673751.1
			Tetranychus urticae Tu-Santpoort1_contig01903 mRNA sequence	7%	0.0	98%	JR685191.1
			Tetranychus urticae Tu-DeLier1_isotig06309 mRNA sequence	7%	0.0	98%	JR680060.1
			Tetranychus urticae Tu-Houten1_isotig13217 mRNA sequence	5%	0.0	98%	JR703406.1
Tetranychus urticae associated Nodaviridae Segment B 1	1661	EST	CCZA8843.b1 CCZA Tetranychus urticae London larvae normalized Tetranychus urticae cDNA clone CCZA8843 5', mRNA sequence	3%	1,00E-12	98%	GW017620.1

Article de recherche 4

Diversity and composition of arthropod pests' viral communities, and insights about pest viruses distribution in arthropod communities

Sarah François[1,2], Maximilien Kulikowski[1], Marie Frayssinet[2], Denis Filloux[3], Emmanuel Fernandez[3], Philippe Roumagnac[3], Rémy Froissart[4], Mylène Ogliastro[2]

[1] Université de Montpellier, UMR 1333 DGIMI « Diversité, Génomes et Interactions Microorganismes-Insectes », place Eugène-Bataillon, 34095 Montpellier cedex 5, France

[2] Laboratoire « Diversité, Génomes et Interactions Microorganismes Insectes » (DGIMI) UMR 1333, INRA, Université Montpellier, 34095 Montpellier, France.

[3] Laboratoire « Biologie et Génétique des Interactions Plante-Parasite » UMR BGPI, CIRAD-INRA-SupAgro, Campus International de Montferrier-Baillarguet, Montpellier Cedex-5, France.

[4] Laboratoire « Maladies Infectieuses et Vecteurs : Écologie, Génétique, Évolution et Contrôle » (MIVEGEC), UMR 5290, CNRS-IRD-UM, 911 avenue Agropolis, 34394, Montpellier, France.

En preparation.

Abstract

Viral metagenomics is a powerful tool to decipher virus diversity and prevalence in agroecosystems, which may help to improve their functioning and their management. Toward that goal, we need to bring ecological concepts into virology, first establishing an inventory of viruses circulating between plants and insects. We defined prototypic agroecosystems in Montpellier area, made of two adjacent ecosystems, alfalfa field and grassland. In these fields and localities, we sampled plants and three invasive pest species, cotton bollworm (*Helicoverpa armigera*) and alfalfa weevil (the curculionid *Hypera postica*) larvae and pea aphids (*Acyrthosiphon pisum*). The viromes of about 4000 individuals, grouped in 169 samples, were obtained by viral metagenomics. We discovered a high diversity of new potential arthropod and plant virus species. Arthropod viruses' composition differs completely between pest species, and plants viruses were shared between pest and plants. Finally, we analyzed the distribution of thirteen abundant viral species found in crop pests' viromes in arthropod communities using PCR approach. Sampled arthropods represent twelve orders grouping about 3500 individuals. Five virus contigs found in pest viromes were associated with other arthropod species, and were mainly present in predators. Our results showed that coupling viral metagenomics with PCR give insights into the viral diversity and distribution in arthropods.

Key Words

Crop pests, Arthropod communities, Viral metagenomics, Viral discovery, Viral ecology

Introduction

Agroecosystems, e.g. agricultural ecosystems, represent about 10% of emerging lands. They are mainly constituted of fields and grasslands which differ in term of plant diversity, and therefore in arthropod diversity and abundance (Lewinsohn and Roslin, 2008; Nyman *et al.*, 2012; Siemann *et al.*, 1998). There is an increasingly evidence that the level of internal regulation of function in agroecosystems is largely dependent on the level of plant and animal biodiversity present. Biodiversity performs a variety of ecological services, including for example suppression of undesirable organisms and thus can enhance the production of food (Altieri, 1999). But over the last decades the rural areas have undergone habitat homogenization and fragmentation. The development of agricultural production thus resulted in a diminution of plant species and plant genetic diversity and in increasing in the densities and surfaces of field crops (Planning *et al.*, 2002).

These factors are responsible for the presence of arthropod pests in agroecosystems. Arthropods are the most abundant and diverse animals in agroecosystems. Less than 0.5% of arthropod species are considered as pests, however, some of them are a serious threat to crops (http://www.fao.org/). Indeed, they are responsible for the loss of almost 1/15 of the total annual agricultural production, due to their herbivory. Moreover, the role of some of them in the propagation of phytoviruses also contributes to the reduction of agricultural yields. For example, the Tomato spotted wilt virus, transmitted by thrips, was responsible for yields reduction estimated at more than $ 1 billion in the 1990s (Scholthof *et al.*, 2011).

The regulation of crop pest populations is mainly achieved through the use of chemicals. However, the use of pesticides poses many problems, such as loss of biodiversity, impact on human health and the selection of resistant arthropod pests populations (Tilman *et al.*, 2002). Therefore, there is an urgent need for sustainable alternatives to pesticides. Biological control, i.e. the use of auxiliary organisms to prevent or reduce damage caused by pests, has gained in interest in recent years as it is less dangerous to human health and more selective than chemical insecticides (Lacey, Frutos, Kaya, & Vail, 2001). Entomopathogenic viruses, that naturally occurring in arthropod populations, represent largely unexplored biological control resources. Indeed, viruses are known to play a major role in the regulation of their host populations, as shown by the impacts of some of the emerging viruses on humans, crops and livestock (Jones *et al.*, 2008; Woolhouse and Gowtage-sequeria, 2005),

which can have up to fundamental consequences in biochemical processes on a global scale (Suttle, 2007; Yoon *et al.*, 2011). However, only few viruses belonging to the *Baculoviridae* family are currently used in biological control (Lacey *et al.*, 2015). However, baculoviruses already show their limitations as resistances have appeared in target insect populations. All these facts highlight the need to explore new pathogenic viral resources for arthropod pests regulation.

For all the reasons mentioned above, the characterization of viral communities associated with agroecosystems, and particularly with arthropod pests, is important in both terms of theoretical and applied outcomes in cultivated plants health. Viral metagenomics that represents without molecular a priori viral enrichment technics coupled to high throughput sequencing is a powerful tool for exploring viral diversity. The use of viral metagenomics allowed to gain exponential knowledges of viral communities diversity found in a high varieties of hosts and ecosystems (Labonté and Suttle, 2013; Lecuit and Eloit, 2013; Mohiuddin and Schellhorn, 2015; Rosario and Breitbart, 2011; Roux *et al.*, 2016; Zablocki *et al.*, 2014). A first step toward the comprehension of viral importance in agroecosystems is to establish an inventory of viruses circulating between plants and insects. However, to our knowledge, the only published arthropod pests viromes are those of whiteflies (Bemisia tabaci), and the authors of these study focused on phytoviruses present in this insect species (Ng *et al.*, 2011; Rosario *et al.*, 2015, 2014).

To gain insights into diversity and composition of agroecosystems viral communities, we defined prototypic agroecosystems in Montpellier area, made of two adjacent ecosystems: alfalfa field and grassland. Alfalfa (*Medicago sativa*), cultivated as forage to livestock, represents the most cultivated legume in the word, with an annual production of 454 million of tons/year (FAO, 2002). Alfalfa is attacked by many pests, as alfalfa weevils and caterpillars that cause direct damage to the foliage. Some of them, as pea aphid (as *Acyrthosiphon pisum*), can be vectors of plant viruses (Frame *et al.*, 1998), as the alfalfa mosaic virus that has caused yield losses of 24% to 67% (Forster *et al.*, 1997).

In these ecosystems, we sampled three invasive pest species displaying a worldwide geographical distribution: the cotton bollworm *Helicoverpa armigera* (Lepidoptera, Noctuidae) (http://www.cabi.org/isc/datasheet/26757), the alfalfa weevil *Hypera postica* (Coleoptera, Curculionidae) (http://www.cabi.org/isc/datasheet/28335) and the pea aphid *Acyrthosiphon pisum* (Hemiptera, Aphididae) (http://www.cabi.org/isc/datasheet/3147). We

also sampled plants: Alfalfa (*Medicago sativa*) and grassland plants. The viromes of about 4000 individuals, grouped in 169 samples, were obtained by viral metagenomics.

Finally, to analyze the circulation of arthropod pest associated viruses, PCR screening was performed in arthropod communities that comprise twelve orders grouping about 3500 individuals. Coupling viral metagenomics with PCR gave us insights into viral diversity and distribution in arthropods.

Material and Methods

Arthropod pests, predators and plants sampling

Arthropod species and plant species were collected from two adjacent ecosystems that were not treated with pesticides: alfalfa fields and grasslands. The same sampling effort was carried out for both ecosystems. *Helicoverpa armigera* larvae individuals (cotton bollworm, Lepidoptera) were sampled between 2014 and 2016 in 5 different localities in Southern France (Prades le Lez, Mauguio, Lattes, St Martin de Londres and Candillargues) and in Spain (Sevilla and Cordoba). In April 2015, *Hypera postica* (alfalfa weevil, Coleoptera) larvae and *Acyrthosiphon pisum* (aphid, Hemiptera) individuals were collected in Prades le Lez. Arthropods were collected using nets on a surface of $100m^2$/samples. Alfalfa (*Medicago sativa, Fabales*) and grassland plants were randomly collected in each ecosystem between 2015 and 2016 **(Fig. 1 - Tab. 1. – SM1)**. After collection, samples were maintained in ice and transferred to the lab within 5h and stored at -80°C without addition of any preservative solutions.

Virome preparation and sequencing

Viromes were obtained by the method described by Candresse *et al.* (Candresse *et al.*, 2014). Briefly, one gram of insect or plant material were processed using a virion-associated nucleic acids (VANA) based metagenomics approach to screen for the presence of DNA and RNA viruses. After material grounding, supernatant were then filtered through a 0.45 µm filter. The filtrate was then centrifuged at 140000g for 2.5 hrs to concentrate viral particles. The resulting pellet was suspended and nonencapsidated nucleic acids were eliminated by

adding DNase and RNase incubation at 37°C for 1.5 hrs. Total nucleic acids were extracted. After reverse transcription and creation of the complementary strand, random PCR amplification was carried out. Tagging steps occurred during RT and PCR steps. Amplified and tagged DNA products produced by the VANA approach were sequenced using an Illumina platfom (MiSeq sequencing: 2x300 bp and 2x250 bp paired-end sequencing with V3 chemistry, Beckman Coulter Genomics, USA).

Viromes cleaning, de novo assembly, mapping, and taxonomic assignment

Raw reads were first demultiplexed using a agrep (Wu and Manber, 1992). Illumina adaptors were removed using Cutadapt 1.9 (Martin, 2011) and they were filtered for quality (q30, elimination of reads < 15 nt length) and. The cleaned reads were subjected to *de novo* assembly using SPAdes (different kmer sizes: 21,33,55,77,125) (Bankevich *et al.*, 2012) followed with CAP3 with default parameters (Huang, 1999). Mapping step of cleaned reads was performed on contigs produced by *de novo* assembly using Bowtie 2.1.0 (with local and very sensitive options on) (Langmead, 2010; Toland *et al.*, 2013). All contigs and unassembled reads were then subjected to BLASTx searches against the non-redundant GenBank protein sequences database for taxonomic attribution (e-value cutoff of < 10^{-3}; Altschul *et al.*, 1990). Putative viral genomes ORF were obtained using ORF finder on Geneious 1.7 (Kearse *et al.*, 2012). Viral contigs were classified as viral operational taxonomic units (vOTU). In order to remove inter-sample and laboratory contamination, we focused on the most abundant vOTU by using an arbitrary abundance cutoff of <0.1% that was applied for each vOTU in all samples. Each viral species found below this threshold was discarded from further analyses. Moreover, samples containing <5000 reads, resulting from a failed nucleic acid extraction, were discarded for further analyses. To assess whether vOTUs represent novel or already described viruses, their full-length proteins were aligned and compared with their closest relative viral proteins (found in GenBank database) using MUSCLE 3.7 (16 iterations) according to the species demarcation thresholds recommended within the online (9th or 10th) Reports of the ICTV (https://talk.ictvonline.org/ictv-reports/ictv_online_report/).

Arthropods communities sampling and viral screening

To analyze the prevalence of viruses found in arthropod pests in the ecosystems, arthropods communities were also sampled in the same fields and grasslands than those where crop species were sampled. The sampling was carried out in 2016 over June-November. The sampling was carried out with net on a surface of 300 m^2. The harvested arthropods were pooled by species, a sample being composed of individuals of the same species collected at a given date, location and ecosystem (**Fig.1**). We next screened these pools for the presence of thirteen abundant viral species found in crop pests' viromes using PCR approach (**Tab. 2**). PCR primers were tested with success on arthropod pest samples.

Samples were washed with MilliQ water before adding of 200 – 500 µl of PBS buffer. For steel balls of 2-3 mm in diameter were placed in each samples, and they were grounded for 3 min using Tissue Lyser (Qiagen). RNA extraction was performed on grounded samples using RNeasy Mini Kit (Qiagen) floowing manufacturer instructions. Reverse transcription step was carried out using SuperScript III Reverse Transcriptase kit (Invitrogen) according to manufacturer's instructions. PCR mix was prepared using Hot Star Master mix (Qiagen) using manufacturers protocol. PCR was performed with the following cycling conditions: 95°C for 2 min, 35 cycles of 95°C for 2 min, 55°C for 1 min, 72°C for 2 min and an additional final extension for 5 min at 72°C. Visualization, purification and sequencing of the PCR products has been performed as previously described.

Identification of arthropod species

Arthropod species were identified visually and by PCR barcoding using the COI gene primers Uni-MinibarF1: 5'-TCC ACT AAT CAC AAR GAT ATT GGT AC-3' and Uni-MinibarR1: 5'- GAA AAT CAT AAT GAA GGC ATG AGC-3' (Meusnier *et al.*, 2008). Insect DNA was extracted from 1 to 10 grinded individuals using Wizard DNA purification kit (Promega). The amplification was performed by PCR using the HotStarTaq Master Mix Kit (Qiagen) according to the manufacturer's protocol. PCR products were purified using Wizard cleanup kit (Promega) according to the manufacturer's protocol. The following cycling conditions were used: one cycle of 95°C for 2 min, 5 cycles of 95°C for 1 min, 46°C for 1 min, 72°C for 30 sec and 35 cycles of 95°C for 1 min, 53°C for 1 min, 72°C for 30 sec. An additional final extension for 5 min at 72°C was then performed. The yield of the PCR

products was verified by migration of 8µl of PCR products to agarose gel and visualized by staining with ethidium bromide. PCR products were cleaned up using Wizard SV Gel and PCR Clean-Up kit (Promega) according to the manufacturer's protocol. The cleaned PCR products were sequenced using Sanger's method (Beckman Coulter Genomics, France). After quality cleaning, taxonomic assignment was performed by comparing sequences against BOLD database by BLASTn (Ratnasingham and Herbert, 2007).

Statistical analyses

Statistical analyses were performed using R 3.3.3. Viral diversity accumulation curves of *Helicoverpa armigera* viromes were made using the Vegan package (Oksanen, 2016). Differences in viromes diversity and composition of arthropod pest species and plants were analyzed using Venn diagrams. The impact of the spatio-temporality, of arthropod taxonomy and of the number of individuals per sample on viral presence and abundance was analyzed by global linearized models (GLM).

Database screening

All the near complete viral genomes were used as queries to perform BLASTn searches within the *Helicoverpa armigera*, *Acyrthosiphon pisum*, and *Medicago sativa* RefSeq (http://www.ncbi.nlm.nih.gov/refseq/) and WGS genomic databases (https://www.ncbi.nlm.nih.gov/genbank/wgs/), as well as in EST (http://www.ncbi.nlm.nih.gov/nucest/) and TSA (http://www.ncbi.nlm.nih.gov/genbank/tsa) transcriptomic databases with a minimum percentage of nucleotides similarity cutoff of ≥95% and an e-value cutoff of $<10^{-10}$.

Phylogenetic analyses

The putative amino acid sequences of viral polymerase proteins viral were used for phylogenetic analyses. All ORFs were translated *in silico* using ORF finder (cut off >300bp) on Geneious 1.7 (Kearse *et al.*, 2012). They were aligned with the corresponding polymerase fragments of relative viruses deposited on Genbank nr database using MUSCLE 3.7 (16 iterations) with default settings (Edgar, 2004). Aligned sequences were edited by Gblocks

0.91 with default setting to eliminate poor aligned regions (Talavera and Castresana, 2007), or were manually edited to remove GAPs (*Parvoviridae* and *Picornavirales* alignments). Maximum likelihood phylogenetic trees were produced from these alignments using PhyML 3.1 (Dereeper *et al.*, 2008; Guindon *et al.*, 2010) with substitution models chosen as the best-fit using Prottest 2.4 (Abascal *et al.*, 2005). One thousand bootstrap replicates were used to test the support of branches. Trees were visualized with FigTree 1.4 (http://tree.bio.ed.ac.uk/software /figtree/). Outgroups were used when possible, otherwise trees were mid-point rooted.

Results

Viromes overview

After reads quality cleaning and low quality samples removal, we obtained cleaned viromes of three arthropod pests and plants. *Helicoverpa armigera* cleaned viromes were obtained for 132 samples (114 field samples - 18 grassland samples) comprising 1590 individuals. A total of 9 318 409 cleaned reads were obtained from *H. armigera* viromes (representing on average 70 600 reads/sample). *Hypera postica* viromes were obtained for 14 samples (11 field samples - 3 grassland samples), comprising about 1400 individuals. A total of 2 928 803 cleaned reads were obtained from these viromes (~210 000 reads/sample). *Acyrthosiphon pisum* viromes contained 4 samples (2 field samples- 2 grassland samples) of about 400 individuals. 1 450 147 cleaned reads were obtained (~362 536 reads/sample). *Medicago sativa* (alfalfa) and grassland plants viromes were obtained from 530 individuals (10 field samples - 9 grassland samples). 1 423 462 cleaned reads were obtained from plant viromes (~75 000 reads/sample) (**Tab. 1 - SM1)**.

H. armigera viromes are dominated by arthropod virus sequences (5% of reads), as well as *H. postica* viromes (72%) (**Fig. 2**). *A. pisum* viromes are dominated by one bacteriophage species: *Acyrthosiphon pisum associated bacteriophage* (NC_000935) which infects its secondary endosymbiont *Candidatus Hamiltonnella defensa* (56%). *A. pisum* viromes also contained arthropod virus sequences (9%). Arthropod viromes finally contained a significant proportion of plant virus sequences: 5,8% for *H. postica* and 0,6% for *H. armigera*. *A. pisum* viromes do not contained plant virus sequences. Finally, plant viromes

were dominated by plant virus sequences (12,5% of reads), and traces of arthropod viruses and bacteriophages (0,1% and 0,22%, respectively) (**Fig. 2**).

Viral discovery

Based on species demarcation threshold recommended by the ICTV on the most complete virus coding sequences obtained by *de novo* assembly; a total of 21 new arthropod virus species and of 6 new plant virus species could have been discovered in the viromes of the three crop pest species (*H. armigera*: 10 arthropod viruses and 2 plant viruses; *H. postica*: 6 arthropod viruses and 4 plant viruses and *A. pisum*: 5 arthropod viruses). Finally, alfalfa viromes contained contigs that could represent 2 new plant virus species **(Tab. 3)**. Contigs displaying sufficient length were placed in phylogenetic trees based on polymerase proteins using maximum likelihood method and bootstraps as branch support.

H. armigera associated Rhabdoviridae contig is clustered with *Spodoptera frugiperda rhabdovirus* (NC_025382) with which it shared about 90% of amino acid similarities on its polymerase protein. This contig is placed in a clade grouping other arthropod associated unclassified *Rhabdoviridae* that could represent a new genus-level lineage (**Fig. 3**). *H. armigera associated Phenuiviridae* contig belongs to an unassigned clade (**Fig 4**). *H. armigera associated Nodaviridae* contigs are clustered in the *Alphanodavirus* genus (**SM2**). *H. armigera associated Reoviridae* contig belongs to an unclassified clade within the *Spinareovirinae* subfamily **(SM3)**. *A. pisum associated Parvoviridae* contig is clustered in the *Ambidensovirus* genus with a densovirus infecting an aphid (*Myzus persicae densovirus*) (**SM4**). *A. pisum associated Carmotetraviridae* contig is grouped with the *Providence virus* (NC_014126) which is the only one described *Carmotetraviridae* species. *H. postica* and *H. armigera associated Alphatetraviridae* contigs could belong to new genus-level lineage species within the *Alphatetraviridae* family (**Fig 5**). Sinaivirus tree shows that *H. armigera associated sinaivirus* contig is clustered into this genus while the *H. postica associated sinaivirus* contig is basal to this group, so it could represent a new genus-level lineage (**Fig. 6**). The *Picornavirales* order tree grouped 5 new species-level lineages associated with *H. postica* and *H. armigera*, all are clustered within the *Iflavirus* genus (**Fig. 7**).

Concerning plant viruses, contigs that potentially belong to five new *Sobemovirus* species-level lineages were found in our viromes. They are clustered in a monophyletic clade separated from the other sobemoviruses (**Fig. 8**). Concerning the *Tymoviridae* family, *M.*

sativa associated Tymoviridae contig is clustered within the *Marafivirus* genus (**Fig. 9**). *H. postica associated Alphaflexiviridae* contig is clustered within the genus *Sclerodarnavirus* genus (**SM5**). The two *M. sativa* associated *Partitiviridae* contigs coding for polymerase proteins are clustered within the *Alphapartitivirus* genus (**SM6**). Overall, 80% of the contigs placed in phylogenetic trees could represent new species-level lineages. Moreover, phylogenetic analyses show that 5 of these contigs could represent viruses belonging to undescribed genus-level lineages.

Diversity and composition of crop pests viromes

H. armigera viromes contained contigs belonging to 15 arthropod families and unclassified genera. They belong to dsDNA viruses: *Baculoviridae* (*H. armigera granulovirus* (NC_010240.1) and *Xestia c nigrum granulovirus* (NC_002331.1), *Nudiviridae* and *Polydnaviridae* (*Hyposoter didimator ichnovirus* and fragments of another species-level lineage). They also contained ssDNA viruses: *Bidnaviridae*, *Parvoviridae*; dsRNA viruses: *Reoviridae*, ss+RNA viruses *Flaviviridae*, *Nodaviridae* (2 potential species-level lineages), *Tetraviridae*, *Dicistroviridae* (*Aphid lethal paralysis virus*), *Iflaviridae* (2 potential species-level lineages), *Negevirus* (1 potential species-level lineage), *Sinaivirus* (1 potential species-level lineage) and ss-RNA viruses: *Phenuiviridae* (1 potential species-level lineage), *Rhabdoviridae* (1 potential species-level lineage) (**SM7**).

The three most prevalent arthropod virus families present in *H. armigera* viromes are *Phenuiviridae* (present in 51,5% of samples), *Baculoviridae* (present in 36,4% of samples), and *Rhabdoviridae* (present in 19,7% of samples) (**Fig.10A**). The influence of the spatio-temporal factors and of the number of individuals sampled on viral species presence was tested by GLM models. Statistical analyses show that *Phenuiviridae* are more present in crops than in grasslands (Chisq test, $p=9,7x10^{-5}$) and that there is a correlation between the presence of *H. armigera associated Phenuiviridae* and *H. armigera associated Rhabdoviridae* species (Chisq test, $p=8.4x10^{-5}$). The presence of one of these viruses increased by a factor two the probability that the other is present in the same sample (IC 95%: relative risk between 1,5 - 2,5). *H. armigera* viromes contained fragments of dsDNA bacteriophages: *Podoviridae* (18% of samples), *Siphoviridae* (16%) and *Myoviridae* (2%); and ssDNA bacteriophages *Microviridae* (3%). *H. armigera* viromes finally comprise contigs belonging to 8 phytovirus families: ssDNA: *Gemicirculavirus*, dsRNA: *Amalgaviridae* and *Partitiviridae*, and ss+RNA:

Alfaflexiviridae, *Bromoviridae*, *Tymoviridae*, *Virgaviridae* and *Sobemovirus*. They finally contained one fungus virus family: *Totiviridae* (dsRNA). The three most prevalent families of plant or fungus viruses in *H. armigera* viromes are *Partitiviridae* (37,1% of samples), *Totiviridae* (17,4% of samples) and *Sobemovirus* (9% of samples) (**SM7**). Statistical analyses show that the sampling date and locality have a statistical influence on the presence of *Partitiviridae* and *Totiviridae* in *H. armigera* viromes (Chisq test, p<0,0009).

Accumulation curve made on *H. armigera* viromes reaches an asymptote, so the viral family diversity present in this arthropod species was correctly represented (**Fig. 11**) Statistical analyses show that the ecosystems sampled have an influence on arthropod virus richness in *H. armigera* viromes: there are less arthropod viruses on grasslands than in crops (F test, p=0,003). This result can be correlated to the fact that, for the same sampling effort between crop and grassland ecosystem, there are less *H. armigera* individuals in grasslands (55 individuals) than in crops (1535 individuals) (F test, p=0,002). The model finally also show that increase the number of individuals sampled increased also the number of plant viruses detected (F test, p=0,007) and the probability of detecting *Phenuiviridae* (F test, p=0,0002) in *H. armigera viromes*.

H. postica viromes contained contigs that could belong to 6 arthropod virus species, all are ss+RNA viruses belonging to *Alphatetraviridae*, *Iflaviridae* (3 potential species), unclassified *Picornavirales* (2 potential species), and *Sinaivirus*. The three most prevalent arthropod viruses species found in *H. postica*, present in all samples, are *H. postica associated Iflaviridae 1* (average read abundance of 36%), *H. postica associated Iflaviridae 3* (~28%), *H. postica associated Iflaviridae 2* (~7,6%). They represent the core virome of *H. postica* (**Fig. 10B – SM8**). *H. postica* viromes contained contigs that could belong to 4 plant virus species-level lineages. The three most prevalent plant viruses families found in *H. postica* viromes are *Partitiviridae* (present in all samples – average read abundance of 2,8%), *Luteoviridae* (present in 92% of samples – average read abundance of 0,45%) and *Alphaflexiviridae* (present in 79% of samples – average read abundance of 1,9%) (**SM8**).

A. pisum viromes contained contigs that could represent 7 arthropod virus species: ssDNA viruses belonging to *Parvoviridae* family (1 species), and ss+RNA viruses: *Carmotetraviridae* (1 species), *Ilfaviridae* (*H. postica associated Iflaviridae 3* present in one sample at low abundance), *Negevirus* (2 species), and two unclassified viruses: *Wuhan aphid*

virus 2 and one other unclassified virus. The three most prevalent arthropod viruses species found in *A. pisum* viromes are *A. pisum associated Parvoviridae* (present in ¾ samples – average read abundance of 7,5%), *A. pisum associated unclassified ss+RNA virus* (present in ¾ samples – average read abundance of 2,9%) and *A. pisum associated Negevirus 1* (present in all samples – average read abundance of 1,4%) (**Fig. 10C – SM8**). Statistical analyses show that ecosystems sampled don't have a statistical influence on virus presence nor on abundance in *H. postica* and *A. pisum* viromes, but these results should be taken with caution because of the low number of samples.

Alfalfa and grassland plants viromes contained contigs belonging to 10 families of plant or fungus infecting viruses: ssDNA viruses (*Geminiviridae*), dsRNA viruses (*Amalgaviridae* (1 potential species), *Partitiviridae* (2 potential species), *Totiviridae*, and ss+RNA viruses (*Alphaflexiviridae* (1 potential species), *Bromoviridae* (*Alfalfa mosaic virus*), *Luteoviridae* (*Bean leafroll virus*), *Potyviridae* (1 potential species), *Secoviridae* (1 potential species), and *Tymoviridae* (1 potential species)). The most prevalent viral families found in plant viromes are *Amalgaviridae* (present in 16/19 samples – average read abundance of 2,8%), *Partitiviridae* (present in 14/19 samples – average read abundance of 5,4%), *Totiviridae* (present in 10/19 samples – average read abundance of 1,4%) and *Bromoviridae* (present in 10/19 samples – average read abundance of 2,2%) (**Fig. 10D – SM9**). Plants viromes also contain traces of arthropod viruses (between 0% and 0,32% of read abundance per plant sample) and traces of bacteriophages (between 0% and 2,75%) (**SM9**). Concerning the 19 plant viromes, spatio-temporal factors don't show a statistical influence on virus diversity and abundance. But these results should be taken with caution because of the small number of samples.

Arthropod virus virome composition differs totally between the three crop pest species. *H. armigera* viromes are dominated – in terms of prevalence - by ss-RNA viruses (*Phenuiviridae and Rhabdoviridae*) and by dsDNA viruses (*Baculoviridae*) while *H. postica* viromes are dominated by ss+RNA viruses (*Picornavirales*) and those of *A. pisum* by the *Acyrthosiphon pisum associated bacteriophage* (NC_000935) (dsDNA virus). The arthropod viruses found in common between two pests are present a very low read abundance and were identified in only one sample, this could reflect inter sample contamination rather than

infecting viruses (**Fig. 12A**). These results are confirmed by GLM models analyses showing that the presence of arthropod viruses (Chisq test, p=0,0004) and of *Acyrthosiphon pisum associated bacteriophage* (Chisq test, p=0,0006) are statistically explained by the host taxonomy.

Some plant viruses are shared between the crop pests and plants. Sequences from five plant virus contigs are shared between plants and two crop pest species, they represent the most abundant plant viruses identified in viromes (e.g. *H. postica associated Alphaflexiviridae*, *M. sativa associated Partitiviridae 1*, *M. sativa associated Tymoviridae*, *M. sativa associated Amalgaviridae* and *Alfalfa mosaic virus*). Four viruses are shared between *H. armigera* and plants and 1 virus contig is shared between *H. postica* and plants (**Fig. 12B**). Finally, 3 plant viruses were only found in pest viromes while two plant viruses were only found in plant viromes.

Worldwide distribution of viruses found in crop pests and in alfalfa viromes

To obtain insights into the geographical distribution of the viruses found in this study, they were searched in genomic and transcriptomic databases with a threshold of ≥95% of nucleotide identity. Five of eight aphid virus contigs were found in the transcriptome of *A. pisum*: *Acyrthosiphon pisum bacteriophage*, *Aphid lethal paralysis virus*, *Acyrthosiphon pisum associated Negevirus 1*, *Wuhan aphid virus 2* and *Acyrthosiphon pisum associated unclassified ss+RNA virus*. *A. pisum* transcriptome came from an English strain (Cambridge). The *A. pisum* symbiont bacteriophage was also found in its genome derived from an USA strain, which suggests that this bacteriophage may have a wide geographical distribution (**Tab. 4**, **SM10)**. These findings ask questions about the impact of these viruses on aphid fitness as infected individuals come from laboratory rearing. Moreover, the two aphid viruses that belong to already described species were described in USA (*Aphid lethal paralysis virus*) or in China (*Wuhan aphid virus 2*) (Liu *et al.*, 2014; Shi *et al.*, 2016).

The *Baculoviridae* species found in *H. armigera* viromes were described in Japan (Hayakawa *et al.*, 1999).

Finally, two alfalfa associated viruses were also found in alfalfa transcriptomes: *Medicago sativa associated Amalgaviridae*, and *Medicago sativa associated Partitiviridae 2*.

These transcriptomes came from studies made in China (TSA) and in USA (EST) (Aziz *et al.*, 2005).

Thus, all these viruses could also have a worldwide distribution.

Distribution of crop pest viruses in arthropod communities

Arthropod communities' samples represent 295 pools of individuals containing sufficient RNA amount for RT-PCR. These sampled comprise 12 orders containing a total of 3492 individuals. To be sure that there was the same sampling effort between communities sampled, we tested the influence of the date, the ecosystem, the locality and the taxonomy on the number of sampled individuals. Globally, the spatio-temporal factors don't show significant influence on the number of individuals sampled, showing that there was the same sampling effort between the communities sampled. Only the taxonomy had a significant influence on the number of individuals, with four orders grouping about 80% of individuals: Hemiptera (23%), Hymenoptera (23%), Coleoptera (20%) and Orthoptera (13%) (**Tab. 5**) (F test, $p=6,2x10^{-12}$). The principal factor structuring arthropod communities is the ecosystem from which they came: there are more individuals sampled in fields than in grasslands in 3/13 orders (Hemiptera, Lepidoptera and Arenae) (F test, $p<0,001$). On the other side, sampling date played a significant role in the presence of 2/13 orders: there are more Hemiptera and Arenae sampled in summer than in autumn (F test, $p<0,003$). Finally, Hemiptera seemed to be more abundant in one locality than in other localities (F test, $p=0,004$).

These communities were screened for the presence of contigs that could belong to 13 arthropod virus species that are abundant in arthropod pest species: 6 *H. armigera* associated virus species, 4 *A. pisum* associated virus species, 2 *H. postica* associated virus species and *Aphid lethal paralysis virus* (NC_004365) than was found in a virome of aphid predator *Coccinella septempunctata* (Coleoptera, *Coccinellidae*) (**Tab. 2**).

Eight virus species were not found in arthropod communities samples: all *A. pisum* associated viruses, *H. postica associated Iflaviridae 3*, *H. armigera* associated reoviruses and *Xestia c-nigrum granulovirus* (NC_002331).

Three viruses were found in low prevalence in arthropod communities. *H. postica associated Iflaviridae 1* (2,4% of positive pools) and *H. armigera associated Alphatetraviridae* (1,7% of positive pools) are only present in predators (as spiders, daddy-long-legs, ladybugs and ants). *H. armigera associated Nodaviridae 2* (4,1% of positive pools) are mainly present in predators but also in crickets pools (**Tab. 5**). These results could rather represent predator diet contamination rather than infection.

Virus absence - or presence in few samples - could reflect host low abundance in arthropod communities: for example *A. pisum* were absent in sampled communities and none of its viruses were found in these communities. They could also represent viruses that are found in low frequencies in their hosts (rare viruses), although *H. postica associated Iflaviridae 3* and *Xestia c-nigrum granulovirus* (NC_002331) were highly present in crop pest viromes.

Finally, two viruses are prevalent in arthropod communities - particularly in predators. *Aphid lethal paralysis virus* (NC_004365) (10,8% of positive pools) is particularly present in ladybugs. This virus has been shown to have a wide host range as it can infect leafhoppers (Liu *et al.*, 2014). This virus is also present in low abundance in one plant virome. And *H. armigera associated Iflaviridae 2* (21,4% of positive pools) is present in Diptera and Lepidoptera (**Tab. 5**). Statistical analyses show these 2 virus species are statistically more present in fields than in grasslands (F test, $p<0,0007$). This result could reflect differences in host abundance between to the ecosystem (3% of *H. armigera* individuals were sampled from grasslands). Moreover, *H. armigera Iflaviridae* shows a prevalence peak in summer (Chisq test, $p=2,4x10^{-6}$).

To conclude, the 5/13 pest viruses present in arthropod communities are present in predators (particularly in spiders, daddy-long-legs and ladybugs), potentially indicating viral accumulation by trophic network.

Discussion

Viruses are among the most abundant and diverse biological entities on Earth (Suttle, 2005). With the expansion of viral metagenomics, we gained exponential knowledges about viral diversity (Rosario and Breitbart, 2011). However viral diversity associated with agroecosystems, and particularly with arthropods, is poorly documented. This can be explained by the lack of viral diversity studies on non-bloodsucking arthropods compared to huge number of viral diversity studies on Human, economically important vertebrates and cultivated plants (Junglen and Drosten, 2013). However, gaining knowledges on viral communities present in agroecosystems may help to improve their functioning and their management. Toward that goal, we need to bring ecological concepts into virology, first establishing an inventory of viruses circulating in plants and insects communities. To develop such approach, we sampled three pest species displaying a worldwide distribution: the cotton bollworm (*Helicoverpa armigera*), larvae of the alfalfa weevil (*Hypera postica*) and aphids (*Acyrthosiphon pisum*), and plants, in alfalfa fields and grasslands in the Montpellier area. The viromes of about 4000 individuals, grouped in 169 samples, were obtained by viral metagenomics.

Crop pests and alfalfa viromes contained contigs representing already described viruses, and their presence is congruent with literature that already reported these viruses in the same hosts. But our viromes could contain a majority of new virus-species lineages (80% of contigs). These results are congruent with those obtained from other arthropod viral diversity studies (Li *et al.*, 2015; Shi *et al.*, 2015; Tokarz *et al.*, 2014; Webster *et al.*, 2015). Moreover, phylogenetic analyses show that some of these viruses could belong to undescribed genus-level lineages. These results reflect our lack of knowledge about arthropod viral diversity. The high viral diversity found in arthropods is consistent to the fact that, taken into account the co-evolution between some of virus and the arthropods, the arthropod viral diversity may overcome the vertebrate viral diversity (Li *et al.*, 2015; Marklewitz *et al.*, 2015; Shi *et al.*, 2016).

The composition of arthropod virus communities differs totally between the three crop pests. *H. armigera* viromes are dominated – in terms of read abundance - by ss-RNA viruses (*Phenuiviridae and Rhabdoviridae*) and by dsDNA viruses (*Baculoviridae*) while *H. postica* viromes are dominated by ss+RNA viruses (*Picornavirales*) and those of *A. pisum* by the

Acyrthosiphon pisum associated bacteriophage (NC_000935) (dsDNA virus). It is interesting to note that there was no effect of the read number per sample on the viral diversity; however, reads number has significant influence in viral genomes reconstitution. For example, *H. postica* and *A. pisum* viromes contained a higher sequencing depth than *H. armigera* viromes (3 to 5 fold); and are dominated by small viruses (<10 kb in length) unlike *H. armigera* viromes in which baculoviruses (80-180 kb in length) are highly present. Consequently, *H. armigera* viral genomes are more scattered than those of the other two crop pest species.

In our viral diversity analyses, viruses that were present in low abundance (<0.1%) were not taken into account. These restrictions have for consequence to greatly reduce the overestimation of viral diversity and laboratory contaminations, at the expense of not taking into account rare viruses that could be present in samples. Moreover, our sampling was made on regional or local scales. So our sampling is not representative of the whole viral diversity found in the sampled hosts, and complementary studies are needed draw a general conclusion about the diversity and composition of these crop pest viral communities. However, it is to notice that 20% of virus species were found in crop pest and alfalfa genomes and transcriptomes or belong to species already described in England, China and USA and Japan. Thus, these viruses seemed to display a large geographical distribution and could thus constitute crop pests and alfalfa core viromes.

To better apprehend viral circulation in agroecosystems, plants and arthropod communities were sampled. Viromes of arthropod pests and plants were obtained, and the presence of some pest associated viruses was tested on the other members of arthropod communities. The results seemed to show viral accumulation by trophic network. Indeed, plants viromes contained a majority of phytoviruses and traces of arthropod viruses that could be due to contamination of plant surface by infected arthropod feces (François *et al.*, 2014). On the other side, arthropod pests contained significant proportion of plant viruses, while there was arthropod pests associated viruses in arthropod predators that indicates that they could be used as monitors for virus surveillance in agroecosystems, as it was tested with success in other viral metagenomics studies (Dayaram *et al.*, 2014, 2013; Ng *et al.*, 2011; Rosario *et al.*, 2015).

One of the main caveats of this study, as well as other viral metagenomic studies, is an uncertainty about the newly discovered viruses host range. As this study is only based on virus sequences presence, even if some of them are very abundant in the viromes, it is not a

proof that the organisms from which the viromes were prepared are their hosts (Mokili *et al.*, 2012). Further studies are needed to identify the hosts, as well as the virulence, of these viruses.

To conclude, viral metagenomics represent a without molecular a priori approach allowing viral discovery; while PCR approach allows the screening of known viral taxa in a high number of samples. The combination of these two approaches gave us insights into the diversity and composition of agroecosystems viral communities. That may help to improve agroecosystems functioning and management.

Acknowledgments: We are particularly grateful to the Conseil General de l'Hérault and to farmers for providing us the opportunity to collect insects and plants in the Domaine de Restinclières, Mauguio, Lattes, St. Martin de Londres, and Candillargues. We warmly thank Armelle Coeur d'Acier for her help on the collect and on the identification of aphid species. S. F. is a doctoral fellow from the University of Montpellier and was supported by a scholarship from Institut National de la Recherche Agronomique (INRA).

Author Contributions: Data Acquisition (S.F., M.K., M.F.); Analysis and interpretation of data (S.F., M.K., M.O. and R.F.); Manuscript preparation (S.F., D.F., E.F., P.R., R.F. and M.O.); Study supervision (S.F., R.F. and M.O.).

Additional Information: The authors declare no competing financial interests.

Bibliography

Abascal, F., Zardoya, R., Posada, D., 2005. ProtTest: selection of best-fit models of protein evolution. Bioinformatics 21, 2104–5. doi:10.1093/bioinformatics/bti263

Altieri, M.A., 1999. The ecological role of biodiversity in agroecosystems. Agric. Ecosyst. Environ. 74, 19–31.

Altschul, S., Gish, W., Miller, W., Myers, E., Lipman, D., 1990. Basic local alignment search tool. J. Mol. Biol. 215, 403–410.

Aziz, N., Paiva, N.L., May, G.D., Dixon, R.A., 2005. Transcriptome analysis of alfalfa glandular trichomes. Planta 221, 28–38. doi:10.1007/s00425-004-1424-1

Bankevich, A., Nurk, S., Antipov, D., Gurevich, A.A., Dvorkin, M., Kulikov, A.S., Lesin, V.M., Nikolenko, S.I., Pham, S., Prjibelski, A.D., Pyshkin, A. V, Sirotkin, A. V, Vyahhi, N., Tesler, G., Alekseyev, M.A., Pevzner, P.A., 2012. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. J. Comput. Biol. 19, 455–77. doi:10.1089/cmb.2012.0021

BARCODING BOLD : The Barcode of Life Data System, 2007. 355–364. doi:10.1111/j.1471-8286.2006.01678.x

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. doi:10.1371/journal.pone.0102945

Dayaram, A., Galatowitsch, M., Harding, J.S., Argüello-astorga, G.R., Varsani, A., 2014. Infection , Genetics and Evolution Novel circular DNA viruses identified in Procordulia grayi and Xanthocnemis zealandica larvae using metagenomic approaches. Infect. Genet. Evol. 22, 134–141. doi:10.1016/j.meegid.2014.01.013

Dayaram, A., Potter, K.A., Moline, A.B., Rosenstein, D.D., Marinov, M., Thomas, J.E., Breitbart, M., Rosario, K., Argu, G.R., 2013. High global diversity of cycloviruses amongst dragonflies. J. Gen. Virol. 94, 1827–1840. doi:10.1099/vir.0.052654-0

Dereeper, a, Guignon, V., Blanc, G., Audic, S., Buffet, S., Chevenet, F., Dufayard, J.-F., Guindon, S., Lefort, V., Lescot, M., Claverie, J.-M., Gascuel, O., 2008. Phylogeny.fr: robust phylogenetic analysis for the non-specialist. Nucleic Acids Res. 36, W465-9. doi:10.1093/nar/gkn180

Edgar, R.C., 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 32, 1792–7. doi:10.1093/nar/gkh340

Forster, R., Beck, A., Lough, T., 1997. Engineering for resistance to virus diseases. pp. 291–315.

Frame, J., Charlton, J.F.., Laidlaw, A.., 1998. Temperate Forage Legumes., CAB Intern. ed.

François, S., Bernardo, P., Filloux, D., Roumagnac, P., Yaverkovski, N., Froissart, R., Ogliastro, M., 2014. A Novel Itera-Like Densovirus Isolated by Viral Metagenomics from the Sea Barley Hordeum marinum 2, 13–14. doi:10.1128/genomeA.01196-14.Copyright

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, 307–21. doi:10.1093/sysbio/syq010

Hayakawa, T., Ko, R., Okano, K., Seong, S., Goto, C., Maeda, S., 1999. Sequence Analysis of the Xestia c-nigrum Granulovirus Genome. Virology 262, 277–297.

Huang, X., 1999. CAP3: A DNA Sequence Assembly Program. Genome Res. 9, 868–877. doi:10.1101/gr.9.9.868

Jones, K.E., Patel, N.G., Levy, M.A., Storeygard, A., Balk, D., Gittleman, J.L., Daszak, P., 2008. Global trends in emerging infectious diseases. Nature 451, 990–993. doi:10.1038/nature06536

Junglen, S., Drosten, C., 2013. Virus discovery and recent insights into virus diversity in arthropods. Curr. Opin. Microbiol. 16, 507–513.

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 analysis of sequence data. Bioinformatics 28, 1647–9. doi:10.1093/bioinformatics/bts199

Labonté, J.M., Suttle, C. a, 2013. Previously unknown and highly divergent ssDNA viruses populate the oceans. ISME J. 7, 2169–2177. doi:10.1038/ismej.2013.110

Lacey, L.., Frutos, R., Kaya, H.., Vail, P., 2001. Insect Pathogens as Biological Control Agents: Do They Have a Future? Biol. Control 21, 230–248. doi:10.1006/bcon.2001.0938

Lacey, L.A., Grzywacz, D., Shapiro-ilan, D.I., Frutos, R., Brownbridge, M., Goettel, M.S., 2015. Insect pathogens as biological control agents : Back to the future. J. Invertebr. Pathol. 132, 1–41. doi:10.1016/j.jip.2015.07.009

Langmead, B., 2010. Aligning short sequencing reads with Bowtie. Curr Protoc Bioinforma. 11.

Lecuit, M., Eloit, M., 2013. The human virome : new tools and concepts. Trends Microbiol. 21, 510–515.

Lewinsohn, T.M., Roslin, T., 2008. Four ways towards tropical herbivore megadiversity. Ecol. Lett. 11, 398–416. doi:10.1111/j.1461-0248.2008.01155.x

Li, C.-X., Shi, M., Tian, J.-H., Lin, X.-D., Kang, Y.-J., Chen, L.-J., Qin, X.-C., Xu, J., Holmes, E.C., Zhang, Y.-Z., 2015. Unprecedented genomic diversity of RNA viruses in

arthropods reveals the ancestry of negative-sense RNA viruses. Elife 4, 1–26. doi:10.7554/eLife.05378

Liu, S., Vijayendran, D., Carrillo-Tripp, J., Miller, W.A., Bonning, B.C., 2014. Analysis of new aphid lethal paralysis virus (ALPV) isolates suggests evolution of two ALPV species. J. Gen. Virol. 95, 2809–19. doi:10.1099/vir.0.069765-0

Marklewitz, M., Zirkel, F., Kurth, A., Drosten, C., Junglen, S., 2015. Evolutionary and phenotypic analysis of live virus isolates suggests arthropod origin of a pathogenic RNA virus family. Proc. Natl. Acad. Sci. U. S. A. 112, 7536–41. doi:10.1073/pnas.1502036112

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

Meusnier, I., Singer, G. a C., Landry, J.-F., Hickey, D. a, Hebert, P.D.N., Hajibabaei, M., 2008. A universal DNA mini-barcode for biodiversity analysis. BMC Genomics 9, 214. doi:10.1186/1471-2164-9-214

Mohiuddin, M., Schellhorn, H.E., 2015. Spatial and temporal dynamics of virus occurrence in two freshwater lakes captured through metagenomic analysis. Front. Microbiol. 6, 960. doi:10.3389/fmicb.2015.00960

Mokili, J.L., Rohwer, F., Dutilh, B.E., 2012. Metagenomics and future perspectives in virus discovery. Curr. Opin. Virol. 2, 63–77. doi:10.1016/j.coviro.2011.12.004

Ng, T.F.F., Duffy, S., Polston, J.E., Bixby, E., Vallad, G.E., Breitbart, M., 2011. Exploring the diversity of plant DNA viruses and their satellites using vector-enabled metagenomics on whiteflies. PLoS One 6, e19050. doi:10.1371/journal.pone.0019050

Nyman, T., Linder, H.P., Peña, C., Malm, T., Wahlberg, N., 2012. Climate-driven diversity dynamics in plants and plant-feeding insects. Ecol. Lett. 15, 889–98. doi:10.1111/j.1461-0248.2012.01782.x

Oksanen, J., 2016. vegan: Community Ecology Package. Ordination methods, diversity analysis and other functions for community and vegetation ecologists. Version 2.4-1. URL https://CRAN.R-project.org/package=vegan.

Planning, U., Jongman, R.H.G., Biodiversity, E., Network, O., Of, D., 2002. Homogenisation and fragmentation of the European landscape : Ecological consequences and solutions Homogenisation and fragmentation of the European landscape : ecological consequences and solutions. doi:10.1016/S0169-2046(01)00222-5

Rosario, K., Breitbart, M., 2011. Exploring the viral world through metagenomics. Curr. Opin. Virol. 1, 289–297. doi:10.1016/j.coviro.2011.06.004

Rosario, K., Capobianco, H., Fei, T., Ng, 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, e886748. doi:10.1371/journal.pone.0086748

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. Viruses 7, 5553–5570. doi:10.3390/v7102895

Roux, S., Enault, F., Ravet, V., Colombet, J., Bettarel, Y., Auguet, J.C., Bouvier, T., Lucas-Staat, S., Vellet, A., Prangishivili, D., Forterre, P., Debroas, D., Sime-Ngando, T., 2016. Analysis of metagenomic data reveals common features of halophilic viral communities across continents. Environ. Microbiol. 18, 889–903. doi:10.1111/1462-2920.13084

Scholthof, K.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. Mol. Plant Pathol. 12, 938–954. doi:10.1111/J.1364-3703.2011.00752.X

Shi, C., Liu, Y., Hu, X., Xiong, J., Zhang, B., Yuan, Z., 2015. A metagenomic survey of viral abundance and diversity in mosquitoes from Hubei province. PLoS One 10, e0129845. doi:10.1371/journal.pone.0129845

Shi, M., Lin, X., Vasilakis, N., Tian, J., Li, C., Chen, L., Eastwood, G., Diao, X., 2016. Divergent Viruses Discovered in Arthropods and Vertebrates Revise the Evolutionary History of the Flaviviridae and Related Viruses. J. Virol. 90, 659–669. doi:10.1128/JVI.02036-15.Editor

Siemann, E., Tilman, D., Haarstad, J., Ritchie, M., 1998. Experimental tests of the dependence of arthropod diversity on plant diversity. Am Nat 152, 738–750.

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

Suttle, C. a, 2005. Viruses in the sea. Nature 437, 356–61. doi:10.1038/nature04160

Talavera, G., Castresana, J., 2007. Improvement of phylogenies after removing divergent and ambiguously aligned blocks from protein sequence alignments. Syst. Biol. 56, 564–77. doi:10.1080/10635150701472164

Tilman, D., Cassman, K.G., Matson, P.A., Naylor, R., Polasky, S., 2002. Agricultural sustainability and intensive production practices. Nature 418, 671–677.

Tokarz, R., Williams, S.H., Sameroff, S., Sanchez Leon, M., Jain, K., Lipkin, W.I., 2014. Virome analysis of Amblyomma americanum, Dermacentor variabilis, and Ixodes scapularis ticks reveals novel highly divergent vertebrate and invertebrate viruses. J. Virol. 88, 11480–11492. doi:10.1128/JVI.01858-14

Toland, A.E., Çatalyürek, Ü. V, Hatem, A., Bozda, D., 2013. Benchmarking short sequence mapping tools. BMC Bioinformatics 7, 184.

Webster, C.L., Waldron, F.M., Robertson, S., Crowson, D., Ferrari, G., Quintana, J.F., Brouqui, J.-M., Bayne, E.H., Longdon, B., Buck, A.H., Lazzaro, B.P., Akorli, J., Haddrill, P.R., Obbard, D.J., 2015. The Discovery, Distribution, and Evolution of

Viruses Associated with Drosophila melanogaster. PLoS Biol. 13, e1002210. doi:10.1371/journal.pbio.1002210

Woolhouse, M.E.J., Gowtage-sequeria, S., 2005. Host Range and Emerging and Reemerging Pathogens. Emerg. Infect. Dis. 11, 1842–1847.

Wu, S., Manber, U., 1992. A fast approximate pattern-matching tool. Usenix Winter 1992 Tech. Conf.

Yoon, H.S., Price, D.C., Stepanauskas, R., Rajah, V.D., Sieracki, M.E., Wilson, W.H., Yang, E.C., Duffy, S., Bhattacharya, D., 2011. Single-cell genomics reveals organismal interactions in uncultivated marine protists. Science. 332, 714–717.

Zablocki, O., van Zyl, L., Adriaenssens, E.M., Rubagotti, E., Tuffin, M., Cary, S.C., Cowan, D., 2014. High-level diversity of tailed phages, eukaryote-associated viruses, and virophage-like elements in the metaviromes of antarctic soils. Appl. Environ. Microbiol. 80, 6888–6897. doi:10.1128/AEM.01525-14

Figures

Figure 1: Summary of sampling places. Silhouettes represent the arthropod sampled.

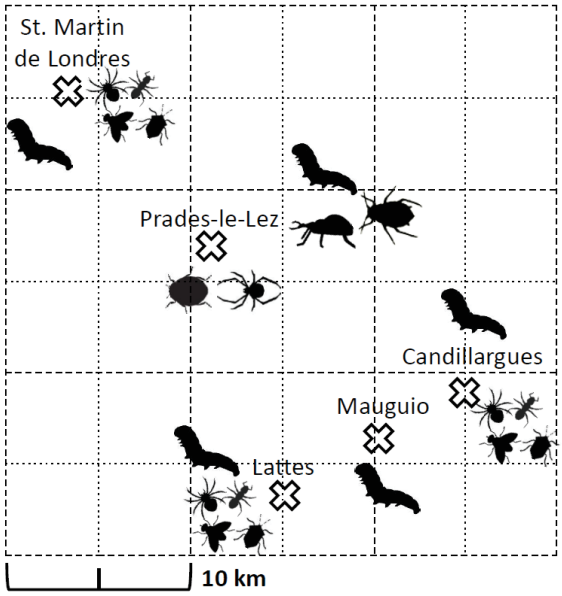


Figure 2: Percentage of arthropod viruses (blue), plant viruses (green) and bacteriophages (red) found in crop pests and in predators viromes.

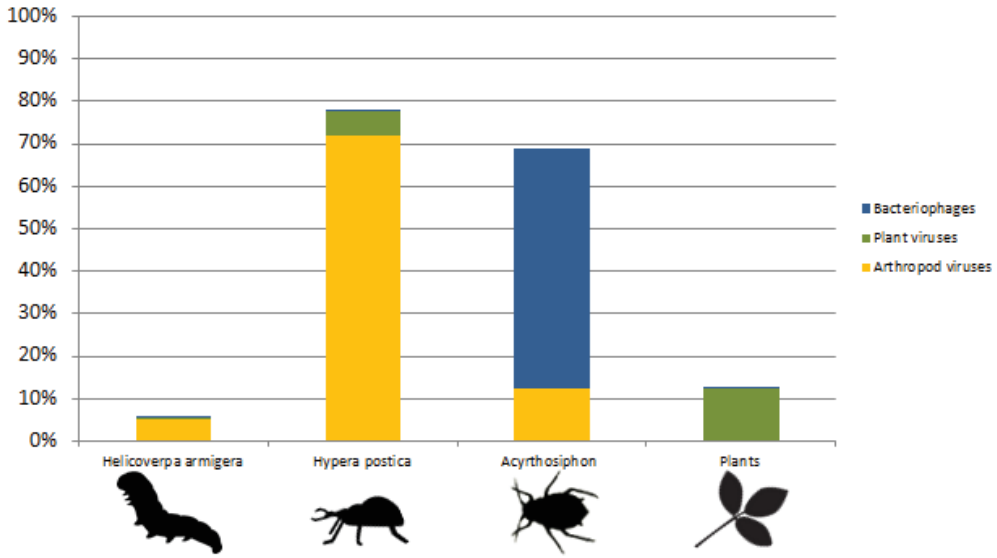


Figure 3: Maximum *Rhabdoviridae* likelihood phylogenetic tree based on partial polymerase protein, including 81 rhabdovirus species and *Helicoverpa armigera associated Rhabdoviridae* (in red). The alignment of 148 amino acids in length was produced using MUSCLE 3.7 (16 iterations) and was hand ungapped. The tree was midpoint rooted. Bootstrap values ≥50% are indicated at each node. Scale bars correspond to amino acid substitutions per site. Genera of the *Rhabdoviridae* family are indicated in brackets.

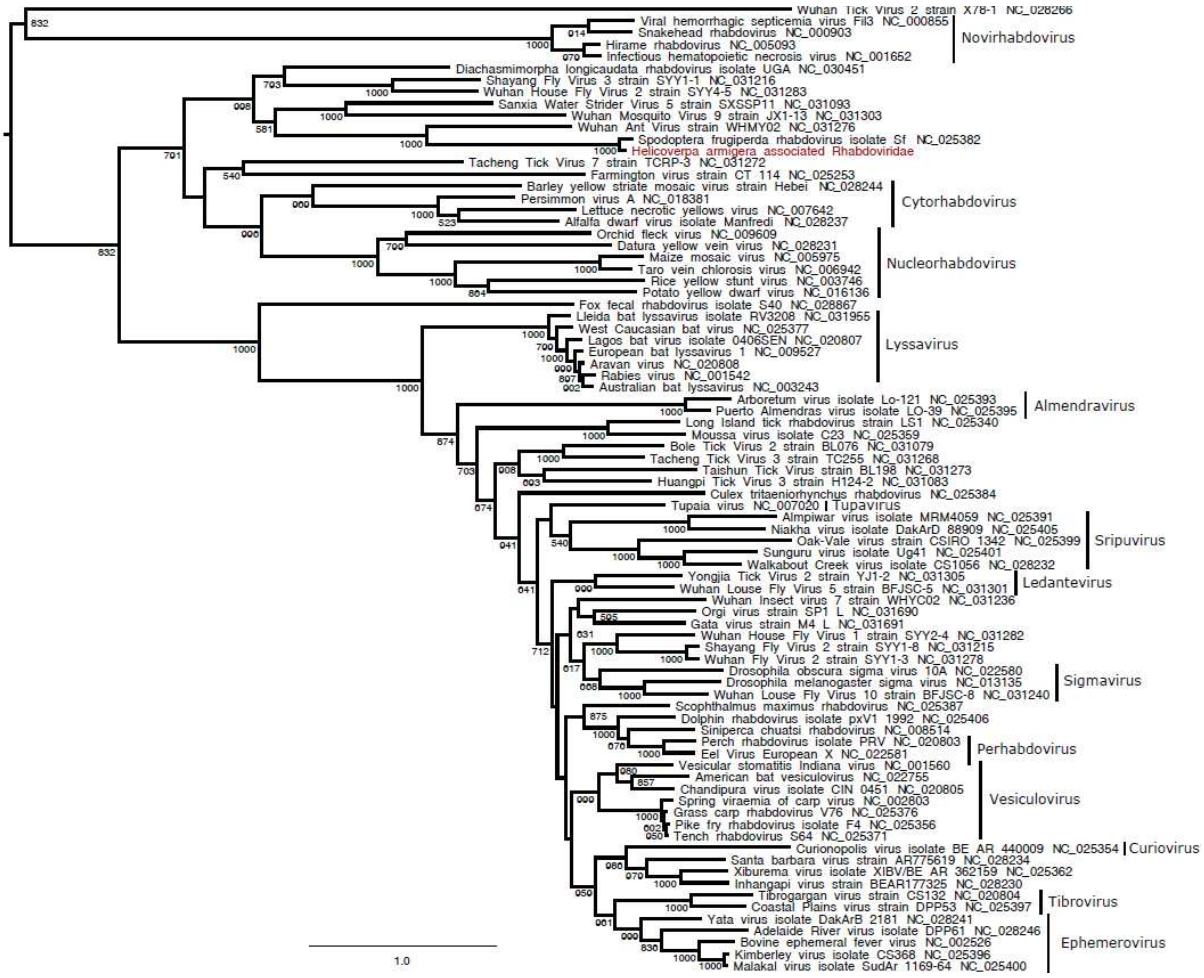


Figure 4: Maximum *Phenuiviridae* likelihood phylogenetic tree based on partial polymerase protein, including 51 phenuivirus species and *Helicoverpa armigera associated Phenuiviridae* (in red). The alignment of 351 amino acids in length was produced using MUSCLE 3.7 (16 iterations) and was hand ungapped. The tree was midpoint rooted. Bootstrap values ≥50% are indicated at each node. Scale bars correspond to amino acid substitutions per site. Genera of the *Phenuiviridae* family are indicated in brackets.

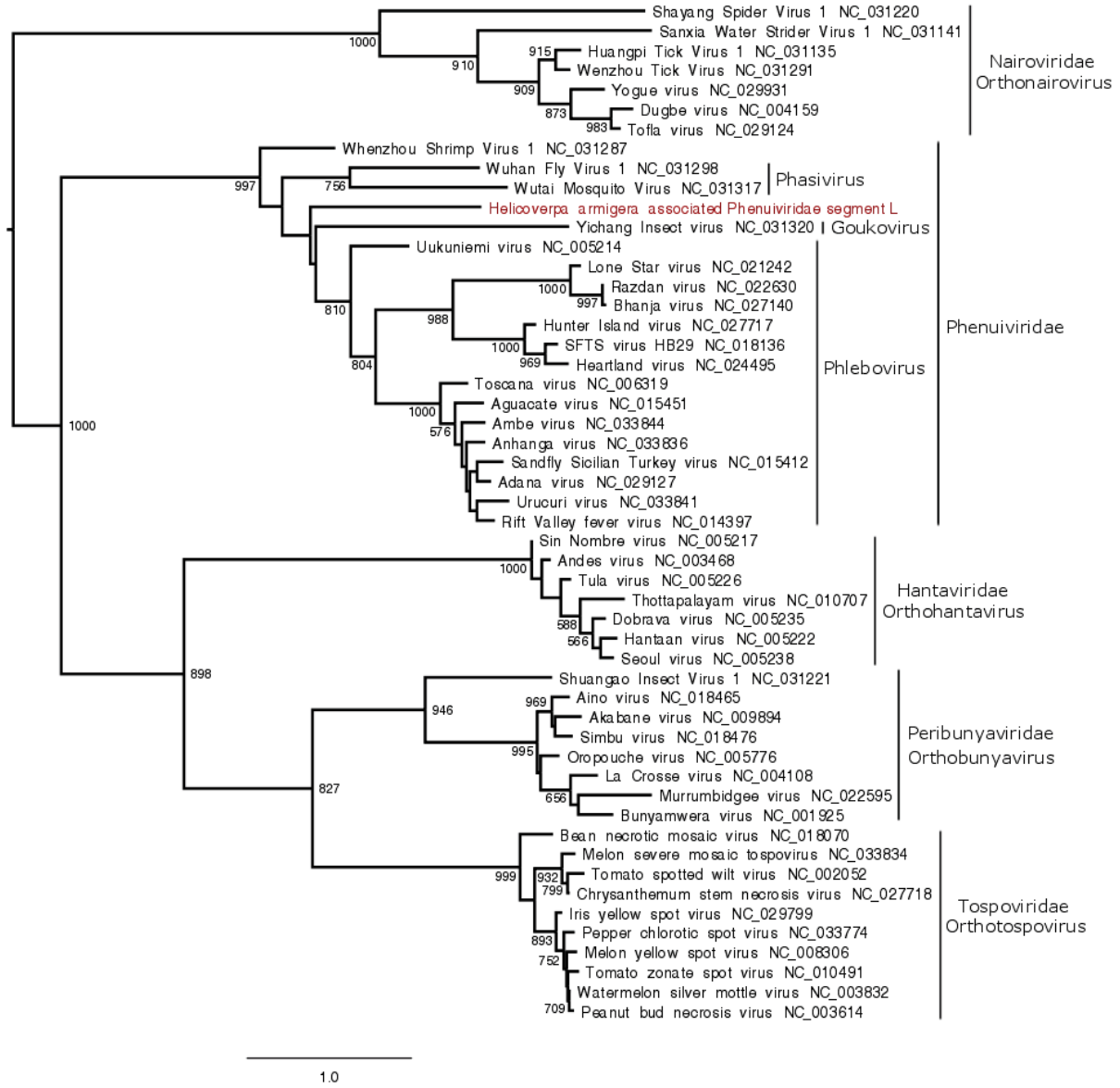


Figure 5: Maximum *Tetraviridae* likelihood phylogenetic tree based on partial polymerase protein, including 9 tetravirus species and *Helicoverpa armigera associated Alphatetraviridae*, *Hypera postica Alphatetraviridae* and *Acyrthosiphon pisum Carmotetraviridae* (in red). The alignment of 400 amino acids in length was produced using MUSCLE 3.7 (16 iterations) and was hand ungapped. The tree was midpoint rooted. Bootstrap values ≥50% are indicated at each node. Scale bars correspond to amino acid substitutions per site. Genera of the *Tetraviridae* are indicated in brackets.

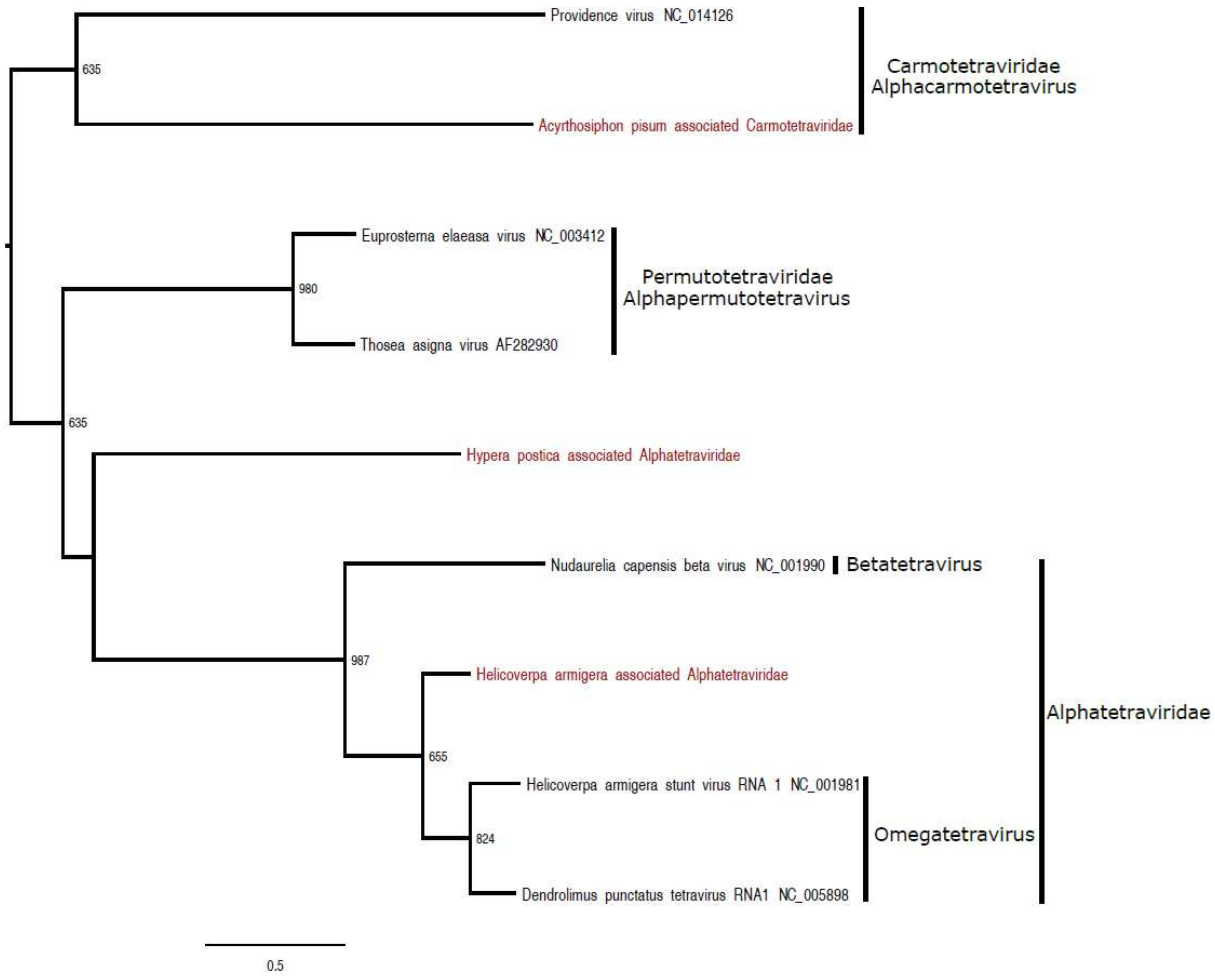


Figure 6: Maximum *Sinaivirus* likelihood phylogenetic tree based on partial polymerase protein, including 11 sinaivirus species and *Helicoverpa armigera associated Sinaivirus*, *Hypera postica associated Sinaivirus* (in red). The alignment of 535 amino acids in length was produced using MUSCLE 3.7 (16 iterations) and was hand ungapped. The tree was midpoint rooted. Bootstrap values ≥50% are indicated at each node. Scale bars correspond to amino acid substitutions per site.

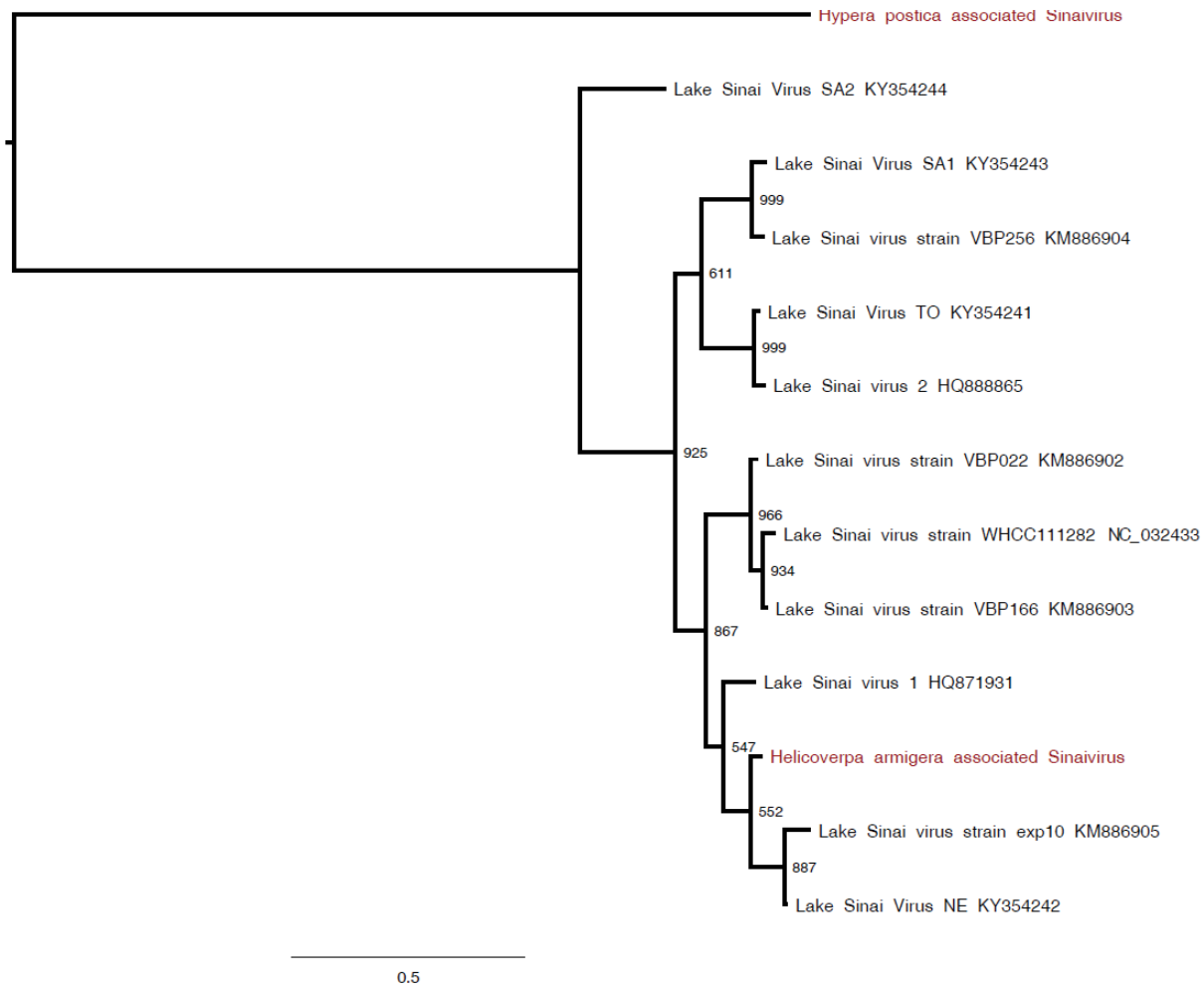


Figure 7: Maximum *Picornavirales* likelihood phylogenetic tree based on partial polymerase protein, including 71 picornavirales species and 5 viromes picornavirales (in red). The alignment of 494 amino acids in length was produced using MUSCLE 3.7 (16 iterations) and was hand ungapped. The tree was midpoint rooted. Bootstrap values ≥50% are indicated at each node. Scale bars correspond to amino acid substitutions per site. Families of the *Picornavirales* order are indicated in brackets.

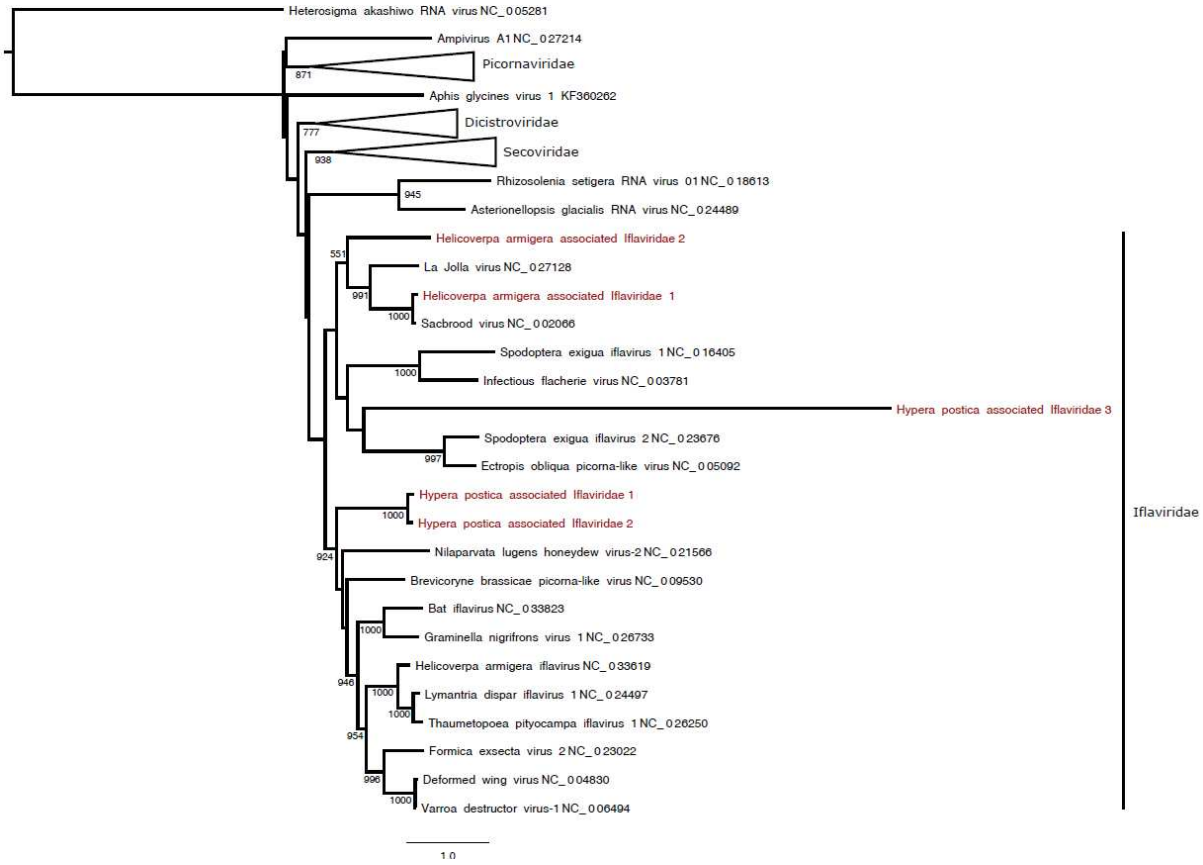


Figure 8: Maximum *Sobemovirus* likelihood phylogenetic tree based on partial polymerase protein, including 18 sobemovirus species and 5 viromes sobemoviruses (in red). The alignment of 127 amino acids in length was produced using MUSCLE 3.7 (16 iterations) and was hand ungapped. The tree was midpoint rooted. Bootstrap values ≥50% are indicated at each node. Scale bars correspond to amino acid substitutions per site.

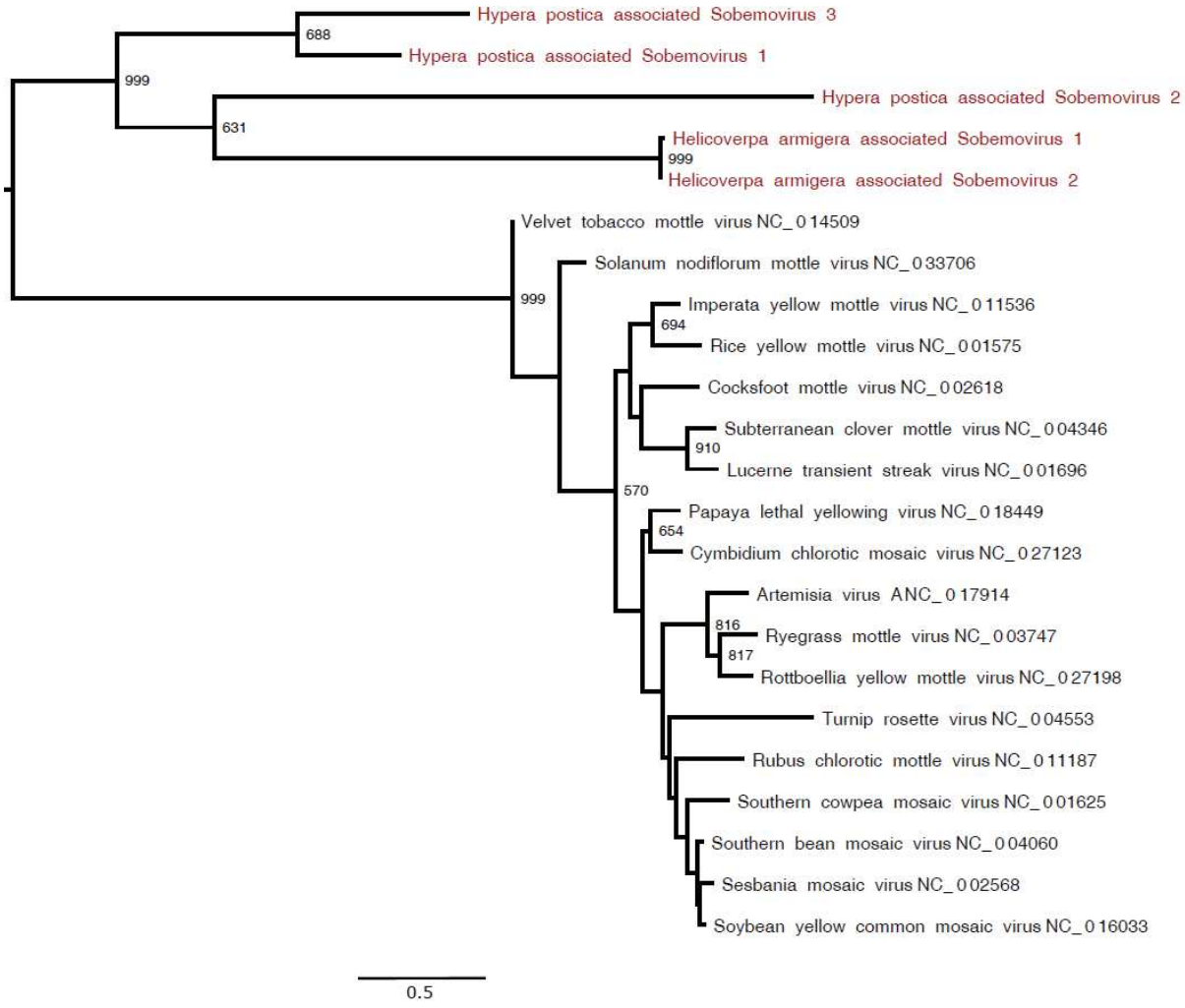


Figure 9: Maximum *Tymoviridae* likelihood phylogenetic tree based on partial polymerase protein, including 37 tymovirus species and 1 viromes tymovirus (in red). The alignment of 727 amino acids in length was produced using MUSCLE 3.7 (16 iterations) and was hand ungapped. The tree was midpoint rooted. Bootstrap values ≥50% are indicated at each node. Scale bars correspond to amino acid substitutions per site. Genera of the *Tymoviridae* family are indicated in brackets.

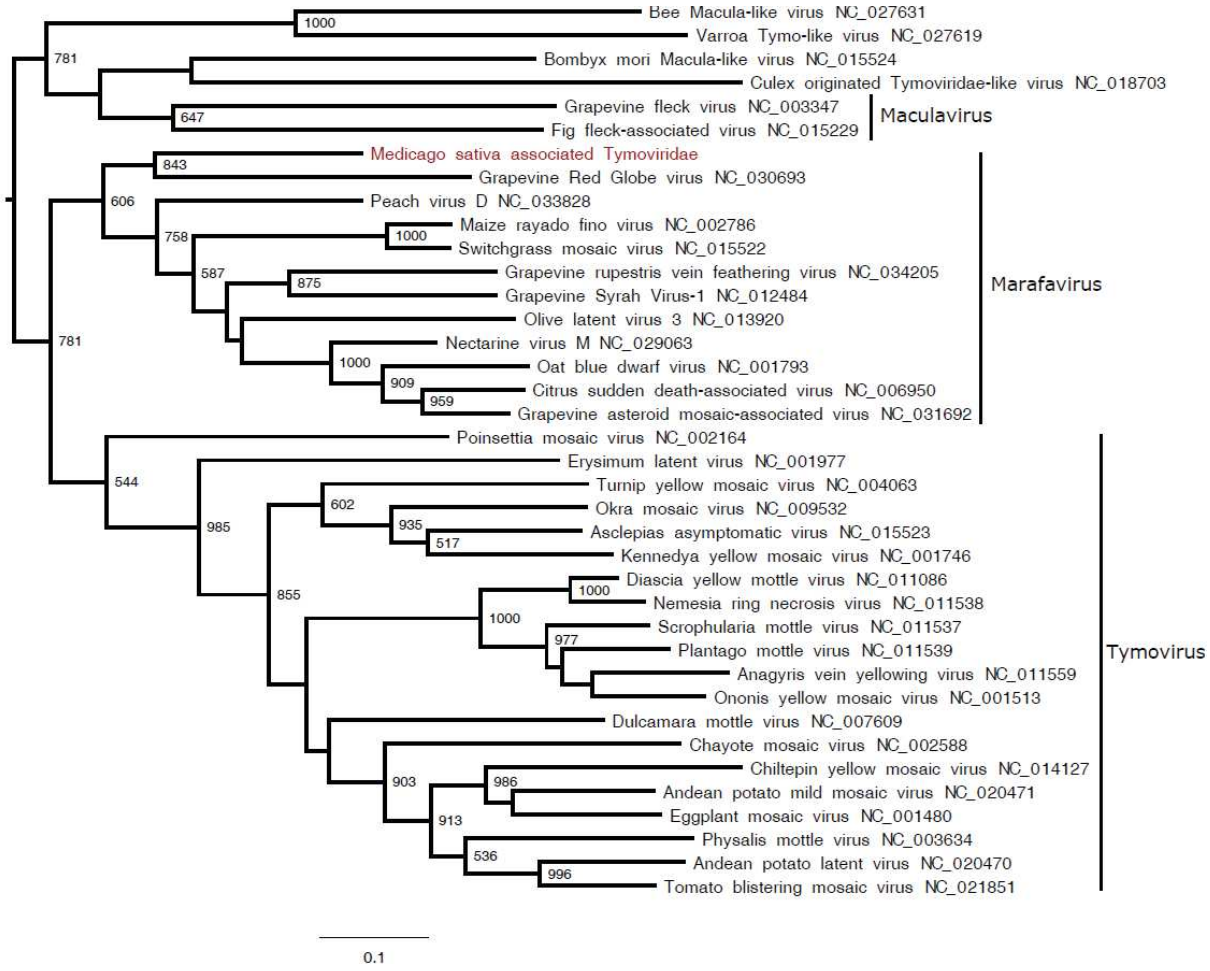


Figure 10: Percentage of samples for which viral families presence was found for *Helicoverpa armigera* viromes (A), *Hypera postica* viromes (B), *Acyrthosiphon pisum* viromes (C), and plants viromes (D).

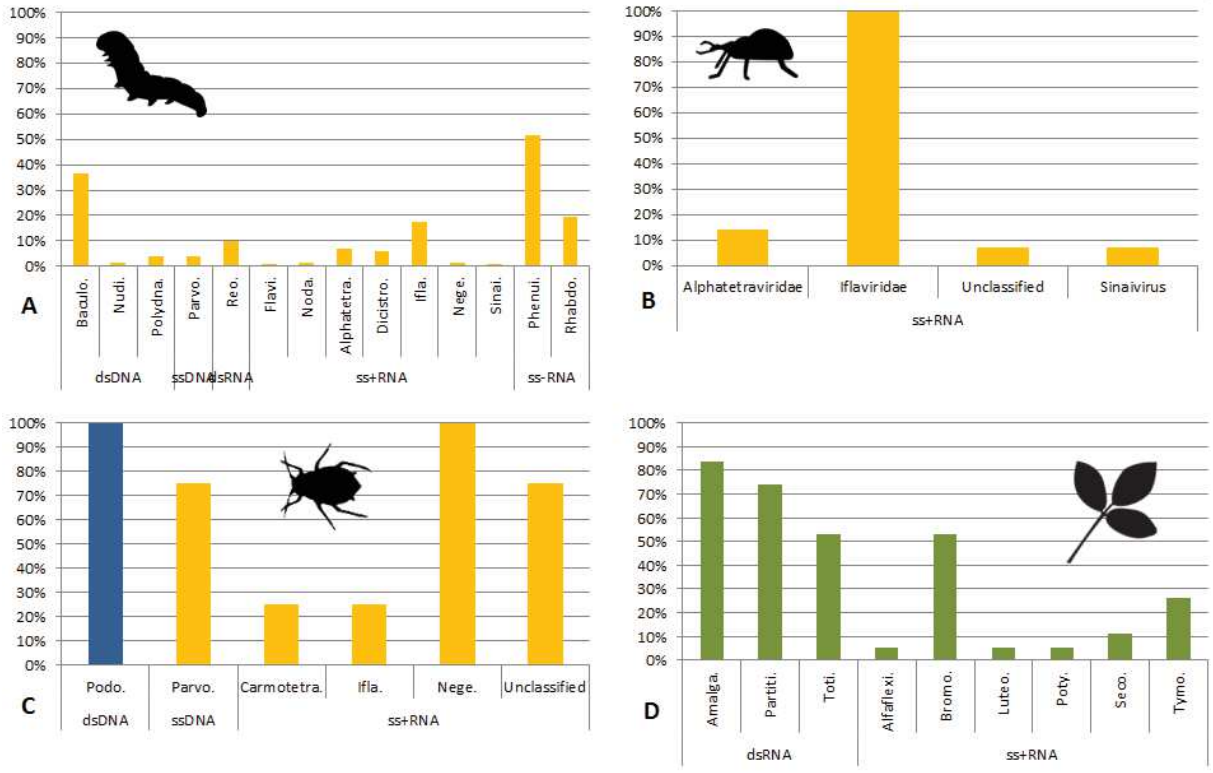


Figure 11: Accumulation curves of viral families in *Helicoverpa armigera* viromes. This accumulation curve is based on viral families that were represented by >0.1% of total reads.

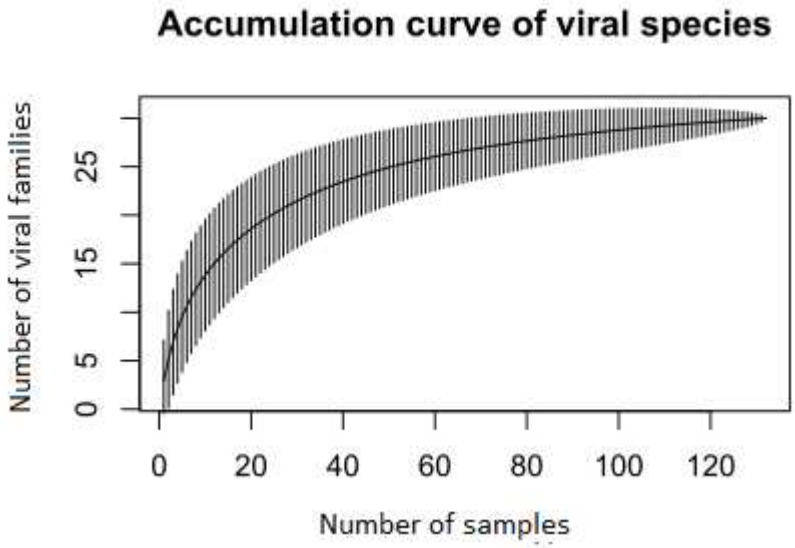


Figure 12: Venn diagrams on arthropod viruses shared between crop pest species (A) and of plant viruses shared between crop pests species and plants (B).

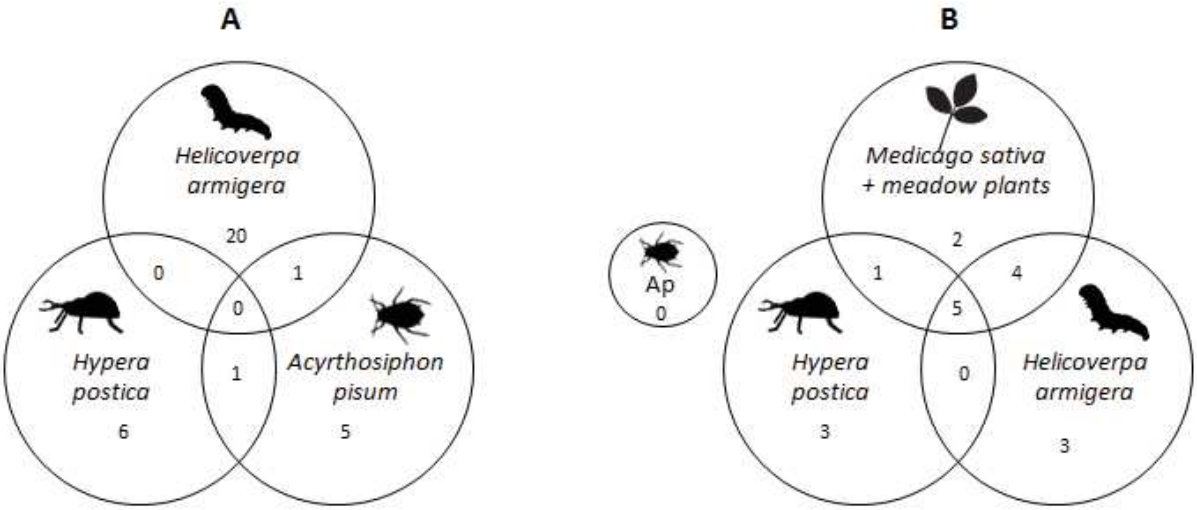


Table 1: Summary of arthropods and plants sampled in alfalfa fields and in grasslands.

Host	Number of samples	Number of individuals	Number of cleaned reads
Helicoverpa armigera	132	1 590	9 318 409
Hypera postica	14	≈ 1400	2 928 803
Acyrthosiphon pisum	4	≈ 400	1 450 950
Medicago sativa + grassland plants	19	530	1 423 462
Total	**169**	**≈ 4 000**	**15 121 624**

Table 2: Summary of crop pests associated viruses searched in arthropods communities, including PCR primers sequences.

Viral species	Foward Primer	Reverse Primer
Xestia c-nigrum granulovirus - Baculoviridae (NC_002331)	ACCCGACTGCATGTTCTGGACT	TCGCGACTCCAACTCGTCGTAT
Helicoverpa armigera associated Reoviridae 1	GCGCAGCTCTGGATTCAGTACG	CCATGCCTTTCACATCCGCACA
Helicoverpa armigera associated Reoviridae 2	CTGGTAAGGCAGAGCCAGGAGT	TTGATCCGCCGTTGGAAACCAC
Helicoverpa armigera associated Iflaviridae 2	ACAGGTAAGACAGAGCTCGCCC	GTTAGCTGCACGCATAGTGGCA
Helicoverpa armigera associated Nodaviridae 2	GCGACCCCAAGTTGACACGAAA	CGCGCGTGGTGAATCTTATGGT
Helicoverpa armigera associated Alphatetraviridae	GATGGACAGCGGTAAAGTCGGC	GAGCGATGCGTACAGGGTCAAC
Hypera postica associated Iflaviridae 1	GCTGGCTTTTCAGACGGCTCTA	TGGATTACCGCTAGGCATCCCA
Hypera postica associated Iflaviridae 3	GCCACTGAGGGAAGATTGCACA	TGCAGGGACTAGGGTCTCAAGG
Acyrthosiphon pisum bacteriophage APSE-1 (NC_000935)	GCCTGGTTCAAAGACACACGCT	CGTAAAAATTGCCGTGCCCGTG
Acyrthosiphon pisum associated Parvoviridae	TACATGCAGCTCCAAATGCGGG	GGGCCTTGTAAAGGAGCATCGC
Acyrthosiphon pisum associated Negevirus	TTTGATGTCGCGACCGTTGACC	CGCGCCAGATAACATGGCTTCA
Acyrthosiphon pisum associated Unclassified ss+ RNA virus	AGCTTCACTTCCAGACCGAGGG	CGGACCTGGGGCTTATGTTCCT
Aphid lethal paralysis virus (NC_004365)	GAGCACAGGCCCCGGTATATGA	AGAGCAGTCTTCACGCTTCCCA

Table 3: Summary of viral contigs found in arthropod pests and plants viromes.

	Viral taxonomy			Contig name	Length	Protein	Closest organism	Accession Number	Protein identity (%)
	Baltimore classification	Family	Genera						
Arthropod viruses	ss DNA	Parvoviridae	Ambidensovirus	Acyrthosiphon pisum associated Parvoviridae	3369	Polymerase (NS1)	Acyrthosiphon pisum	XP_003246947	62.5 %
						Capsid (VP)	Acyrthosiphon pisum	XP_016664130	52.0%
	ds RNA	Reoviridae	Spinareovirinae - Unclassified	Helicoverpa associated Reoviridae 1	4257	Polymerase	Operophtera brumata reovirus	YP_392501	68.7%
		Nodaviridae	Alphanodavirus	Helicoverpa associated Nodaviridae 1 segment 1	3109	Polymerase	Shuangao insect virus 11	YP_009337806	58.8%
				Helicoverpa associated Nodaviridae 2 segment 1	1808	Polymerase	Lutzomyia nodavirus	AKP18615	57.0%
				Helicoverpa associated Nodaviridae 2 segment 2	2300	Capsid	Wuhan nodavirus	ABB71128	67.6%
		Carmotetraviridae	Alphacarmotetravirus	Acyrthosiphon pisum associated Carmotetraviridae	6221	Polymerase	Providence virus	YP_003620397	14.0%
						Capsid	Lake Sinai virus 2	AEH26188	28.1%
		Permutotetraviridae	Unclassified	Hypera postica associated Alphatetraviridae	4784	Polymerase	Daeseongdong virus 2	YP_009182196	37.0%
						Capsid	Hubei permutotetra-like virus 11	YP_009337277	40.8%
				Helicoverpa armigera associated Alphatetraviridae	2517	Polymerase	Helicoverpa armigera stunt virus	AR024698	45.0%
				Helicoverpa armigera associated Iflaviridae 1	8773	Polyprotein	Sacbrood virus	AST13087	83.3%
				Helicoverpa armigera associated Iflaviridae 2	7659	Polyprotein	Hubei picorna-like virus 34	YP_009337152	53.0%
				Hypera postica associated Iflaviridae 1	8672	Polyprotein	Hubei picorna-like virus 31	YP_009330055	38.3%
				Hypera postica associated Iflaviridae 2	7771	Polyprotein	Hubei picorna-like virus 31	YP_009330055	41.3%
				Hypera postica associated Iflaviridae 3	12260	Polyprotein	Wuhan house centipede virus 3	YP_009342325	19.5%
		Unclassified Picornavirales	Unclassified	Hypera postica associated unclassified Picornavirales	4739	Putative protein	Hubei picorna-like virus 43	YP_009337372	22.9%
	ss+ RNA	Unclassified	Negevirus	Acyrthosiphon pisum associated Negevirus 1	9646	Polymerase	Wuhan insect virus 8	YP_009344994	72.8%
						Putative protein 2	Wuhan insect virus 8	YP_009344995	56.6%
						Putative protein 3	Wuhan insect virus 8	YP_009344996	65.6%
				Acyrthosiphon pisum associated Negevirus 2	9519	Polymerase	Wuhan house centipede virus 1	YP_009342435	49.6%
						Putative protein 2	Wuhan house centipede virus 1	YP_009342436	39.8%
						Putative protein 3	Wuhan house centipede virus 1	YP_009342437	51.0%
				Helicoverpa armifera associated Negevirus	3650	Polymerase	Negev virus	BAR91509	45.3%
						Putative protein 2	Castlerea virus	AQZ55406	21.7%
						Putative protein 3	Loreto virus	YP_009351837	25.5%
						ORF1	Lake Sinai virus 2	AS583298	28.7%
		Unclassified	Sinaivirus	Hypera postica associated Sinaivirus	5667	Polymerase	Lake Sinai virus	AJR19138	39.7%
						Capsid	Lake Sinai Virus TO	YP_009388503	26.9%
						ORF1	Lake Sinai Virus NE	YP_009388494	97.2%
				Helicoverpa armigera associated Sinaivirus	5724	Polymerase	Lake Sinai Virus NE	YP_009388495	92.9
						Capsid	Lake Sinai virus 1	AS583258	91.9%
				Acyrthosiphon pisum associated Wuhan aphid virus 2 segment 3 variant	2410	NS3 protein	Wuhan aphid virus 2	YP_009179379	95.5 %
		Unclassified	Unclassified	Acyrthosiphon pisum associated Wuhan aphid virus 2 segment 4 variant	2601	Capsid	Wuhan aphid virus 2	YP_009179386	92.0%
						Putative protein	Wuhan aphid virus 2	YP_009179387	97.2%
				Acyrthosiphon pisum associated Unclassified ss+RNA virus Segment 1	3625	Putative protein	Wuhan insect virus 21	YP_009342466	79.3%
						Putative protein 2	Wuhan insect virus 21	YP_009342467	89.3%
				Acyrthosiphon pisum associated Unclassified ss+RNA virus Segment 2	2097	Capsid protein	Wuhan insect virus 21	YP_009342453	13.2%
		Phenuiviridae	Unclassified	Helicoverpa armigera associated Phenuiviridae Segment G	5214	Glycoprotein	Hubei lepidoptera virus 1	YP_009330257	19.1%
				Helicoverpa armigera associated Phenuiviridae Segment L	7825	Polymerase protein	Hubei lepidoptera virus 1	YP_009330283	29.8%
	ss- RNA	Rhabdoviridae	Unclassified	Helicoverpa armigera associated Rhabdoviridae	12524	Polymerase protein (L)	Spodoptera frugiperda rhabdovirus	YP_009094312	89.3%
						Protein N	Spodoptera frugiperda rhabdovirus	ASU89572	84.1%
						Protein P	Spodoptera frugiperda	ANJ26482	82.8%
						Protein M	Spodoptera frugiperda rhabdovirus	YP_009094309	85.9%
						Protein G	Spodoptera frugiperda rhabdovirus	YP_009094310	79.2%
Plant viruses	ds RNA	Amalgaviridae	Amalgavirus	Medicago sativa associated Amalgaviridae	3375	Fusion protein	Trifolium subterraneum	GAU16045	28.3%
						Capsid protein	Trifolium subterraneum	GAU16045	25.5%
						Fusion protein	Spinach amalgavirus 1	YP_009388304	52.70%
		Partitiviridae	Alphapartitivirus	Medicago sativa associated Partitiviridae 1 Segment 1	1853	Polymerase protein	Pear alphapartitivirus	BBA66577	85.2%
				Medicago sativa associated Partitiviridae 1 Segment 2	1663	Capsid protein	Medicago sativa alphapartitivirus 1	AT100051	99.8%
				Medicago sativa associated Partitiviridae 2 Segment 1	1766	Polymerase protein	Medicago sativa alphapartitivirus 1	AT100050	95.9%
				Medicago sativa associated Partitiviridae 2 Segment 2	1185	Capsid protein	Pyrus pyrifolia cryptic virus	BBA20646	16.7%
		Alphaflexiviridae	Scledornavirus	Hypera postica associated Alphaflexiviridae	7642	Polymerase protein	Garlic virus A	AGC09135	33.6%
	ss+ RNA	Bromoviridae	Alfamovirus	Alfalfa mosaic virus Bromoviridae RNA1 (NC 001495)	3644	100%			
				Alfalfa mosaic virus Bromoviridae RNA2 (NC 002024)	2593				
				Alfalfa mosaic virus Bromoviridae RNA3 (NC 002025)	2037				
		Luteoviridae	Lutevirus	Bean leafroll virus Luteoviridae (NC 003369)	5964	100%			
		Tymoviridae	Marafavirus	Medicago sativa associated Tymoviridae	6407	Polymerase protein	Medicago sativa marafivirus 1	AT100054	34.5 %
		Unclassified	Sobemovirus	Hypera postica associated Sobemovirus 1	3245	Polymerase protein	Hubei diptera virus 13	YP_009337911	41.3%
						Capsid protein	Hubei sobemo-like virus 20	YP_009330032	22.2%
				Hypera postica associated Sobemovirus 2	2842	Polymerase protein	Wuhan house centipede virus 4	YP_009342451	55.1%
						Capsid protein	Wuhan house centipede virus 4	YP_009342450	39.3%
				Hypera postica associated Sobemovirus 3	2151	Polymerase protein	Hubei sobemo-like virus 9	YP_009329982	12.9%
						Capsid protein	Hubei sobemo-like virus 29	YP_009330084	48.8%
				Helicoverpa armigera associated Sobemovirus 1	1890	Putative protein 1	Hubei sobemo-like virus 8	YP_009330087	28.6%
				Helicoverpa armigera associated Sobemovirus 2	2682	Putative protein 2	Sanxia water strider virus 10	YP_009337843	38.8%
				Acyrthosiphon pisum bacteriophage (NC 000935)	36524	Putative protein	Hubei sobemo-like virus 48	APG75767	35.0%
Bacteriophages	ds DNA	Podoviridae	Unclassified			100%			

Table 4: Summary of the number of sequences found in putative host genomic and/or transcriptomic databases (Refseq_genomic, WGS, EST and TSA) that display significant homologies with viral species discovered in the viromes generated for this study.

Host	Viral species	Database	Number of sequences
Acyrthosiphon pisum	Acyrthosiphon pisum bacteriophage APSE-1	refseq_genomics	2
		EST	2
	Aphid lethal paralysis virus	EST	25
	Acyrthosiphon pisum associated Negevirus 1	EST	39
	Wuhan aphid virus 2	EST	4
	Acyrthosiphon pisum associated Unclassified ss+RNA virus	EST	4
Medicago sativa	Medicago sativa associated Amalgaviridae	TSA	1
	Medicago sativa associated Partitiviridae 2 Segment 1	EST	30
		TSA	3

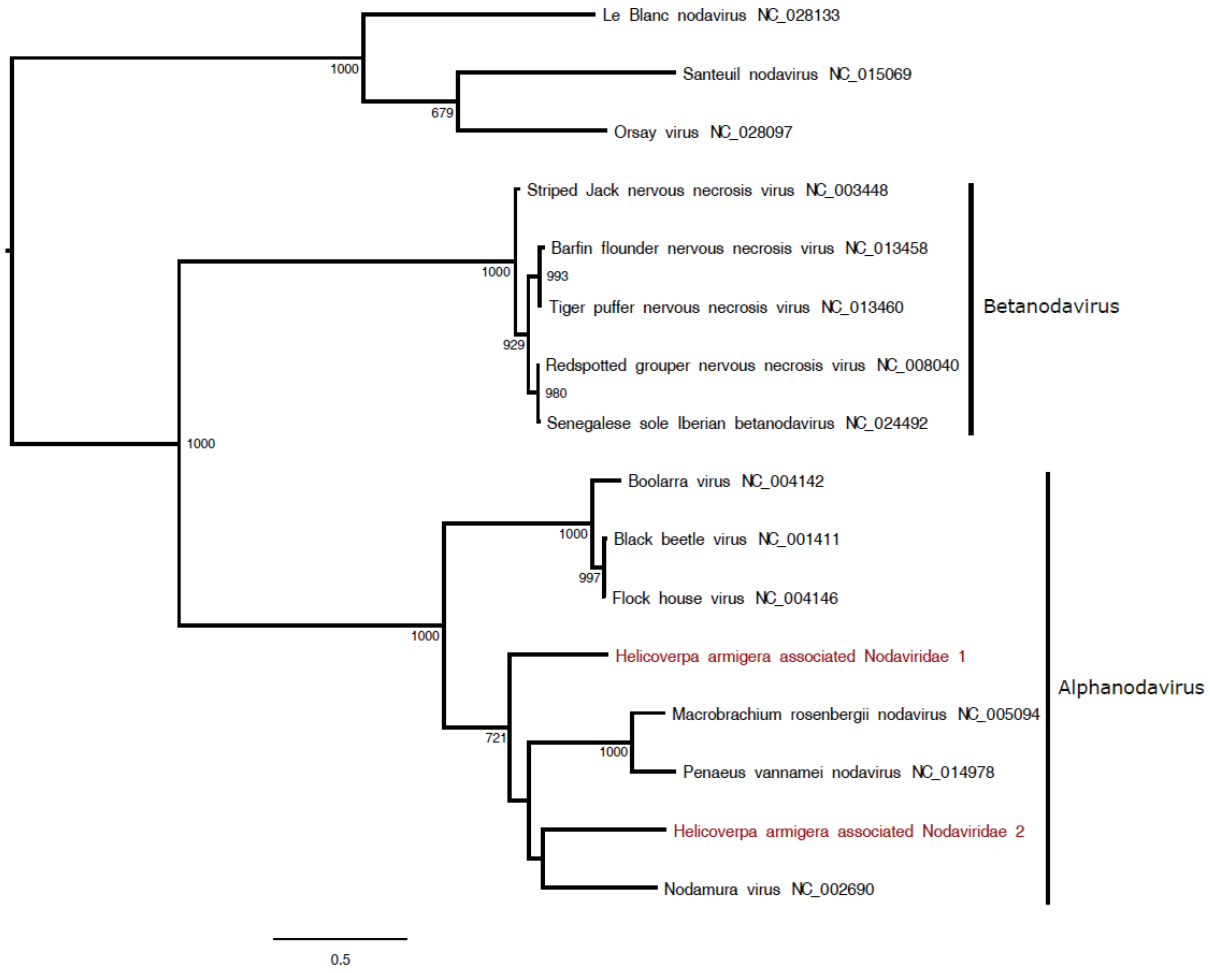
Table 5: Summary of the five crop pests associated viruses that were found in arthropods and molluscs viromes, and details about the composition of arthropods and molluscs communities.

	Host taxonomy	Pools (%)	Individuals (%)	Helicoverpa armigera associated Iflaviridae 2	Helicoverpa armigera associated Nodaviridae 2	Helicoverpa armigera associated Alphatetraviridae	Hypera postica associated Iflaviridae 1	Aphid lethal paralysis virus
Hexapoda	Blattodea	0%	0%	0%	0%	0%	0%	0%
	Coleoptera	16%	20%	15%	0%	0%	6%	17%
	Dermaptera	2%	1%	0%	0%	0%	0%	0%
	Diptera	5%	6%	40%	0%	0%	0%	0%
	Hemiptera	20%	23%	22%	2%	0%	2%	15%
	Hymenoptera	5%	23%	33%	13%	0%	7%	33%
	Lepidoptera	13%	4%	18%	0%	0%	0%	0%
	Mantoptera	1%	0%	33%	0%	0%	0%	0%
	Neuroptera	0%	0%	0%	0%	0%	0%	0%
	Orthoptera	20%	13%	10%	2%	0%	0%	3%
Chelicerata	Araneae	12%	7%	41%	9%	9%	3%	15%
	Opiliones	4%	3%	33%	42%	17%	8%	25%
	Total	100%	100%	21,4%	4,1%	1,7%	2,4%	10,8%

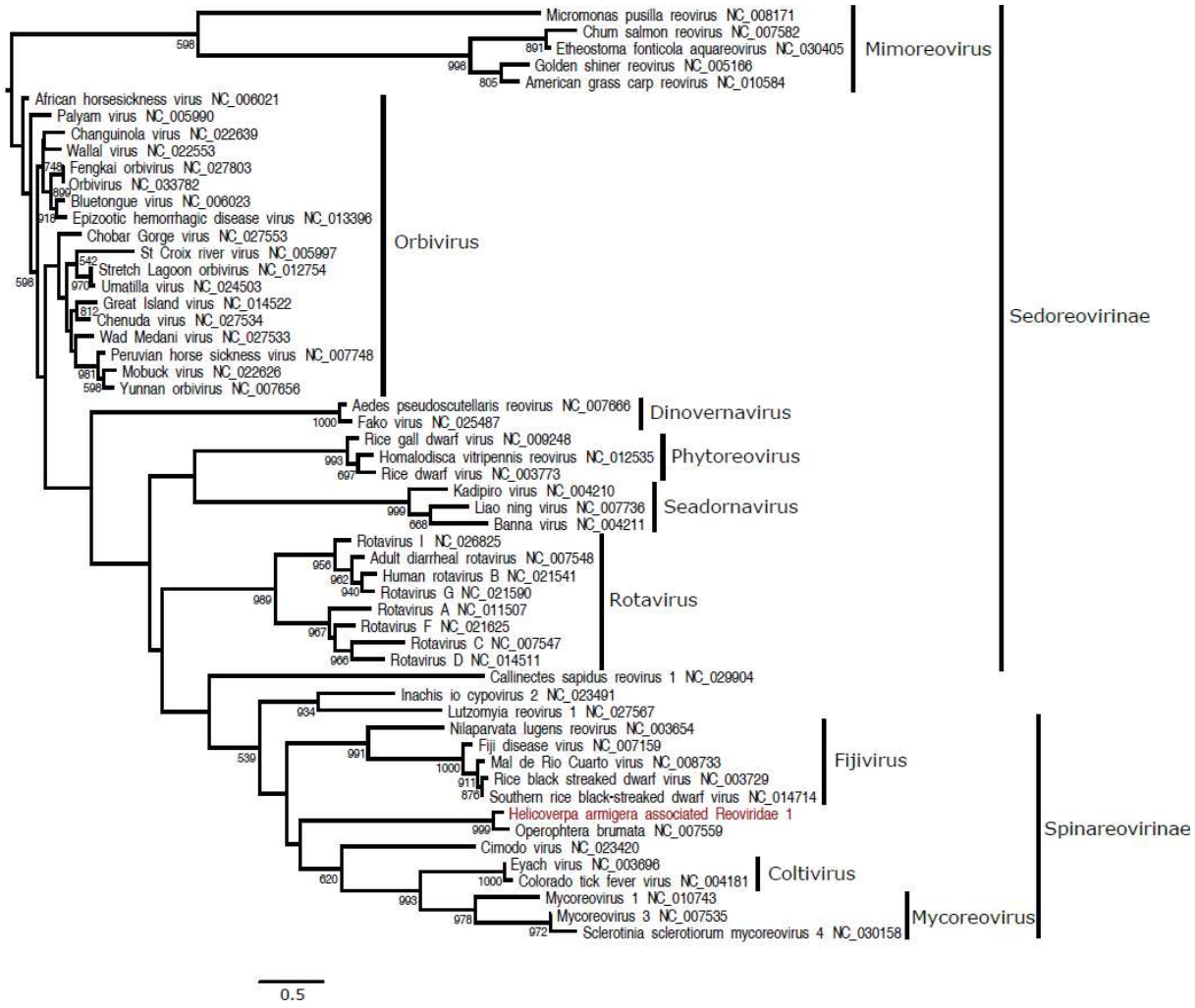
SM1: Detailed information about crop pests and plants samples.

	Sample name	Date	Locality	Ecosystem	Number of individuals	Number of cleaned reads
Hypera postica	*Hypera postica* 1	15/04/2015	Prades le Lez	Crop	> 100	165 836
	Hypera postica 2				> 100	144 967
	Hypera postica 3				> 100	214 565
	Hypera postica 4				> 100	444 691
	Hypera postica 5				> 100	115 578
	Hypera postica 6				> 100	107 610
	Hypera postica 7				> 100	162 564
	Hypera postica 8				> 100	330 252
	Hypera postica 9				> 100	275 280
	Hypera postica 10				> 100	275 503
	Hypera postica 11				> 100	128 579
	Hypera postica 12			Meadow	> 100	178 975
	Hypera postica 13				> 100	130 088
	Hypera postica 14				> 100	254 130
Acyrthosiphon pisum	*Acyrthosiphon pisum* 1			Crop	> 100	224 406
	Acyrthosiphon pisum 2				> 100	272 076
	Acyrthosiphon pisum 3			Meadow	> 100	183 805
	Acyrthosiphon pisum 4				> 100	569 886
			Total	23	> 1000	4 378 950
Helicoverpa armigera	SWP B7	01/07/2014	Sevilla	Crop	5	598 070
	SWP 93				16	423 786
	CWP 93		Cordoba		5	387 799
	MS 1	01/07/2014	Mauguio		6	482 485
	MS 2				6	307 440
	MN5 1	01/09/2014			10	245 479
	MN5 2				30	747 521
	MN5 1	01/10/2014			100	347 268
	R1 2014	01/10/2014	Prades le Lez		3	43 166
	R2 2014				3	93 464
	R3 2014				4	91 548
	R4 2014				5	80 243
	R5 2014				4	165 232
	M1L1.2	06/08/2015	Mauguio		13	43 431
	M1L2.1	10/08/2015			26	9 982
	M1L3.1	12/08/2015			26	31 387
	M1L3.2				9	39 938
	M1L3.4				6	46 189
	M1L4.1	13/08/2015			8	24 507
	M1L4.2				6	45 107
	M1L4.3				1	9 103
	M2L1.1	05/10/2015			19	32 626
	M2L1.2				16	42 293
	M2L1.3				9	82 330
	M2L2.1	07/10/2015			4	14 611
	M2L3.1	14/10/2015			4	18 623
	M2L3.2				5	11 484
	M2L3.3				5	61 709
	S1L1.1 2015	25/08/2015	St Martin de Londres		8	144 131
	S1L1.2 2015				11	92 883
	S1L1.3 2015				11	83 678
	S1L2.2 2015	26/08/2015			13	76 360
	S1L3.2 2015	27/08/2015			12	50 483
	S1L3.3 2015				18	23 855
	S2L1.1 2015	14/10/2015			5	48 309
	S2L1.3 2015				1	117 307
	S2L2.1 2015	22/10/2015			4	41 380
	S2L3.2 2015				15	16 323
	R1L1.1	31/08/2015	Prades le Lez		2	7 889
	R2L2.1	06/11/2015			2	36 213
	R2L2.2				3	28 514
	R2L3.1	10/11/2015			2	25 841
	R1P1.1	31/08/2015		Meadow	1	22 085
	R1P3.1	01/09/2015			2	47 911
	R1P3.2				1	63 585
	L0L1.1	16/06/2016	Lattes	Crop	6	44 146
	L0L1.2				10	43 481
	L0L1.3				15	20 959
	L1 L1.1	19/07/2016			31	16 924
	L1 L1.2				34	61 358
	L1 L1.3				50	42 759
	L1 L2.1	20/07/2016			25	60 063
	L1 L2.2				38	81 786
	L1 L2.3				16	12 982
	L1 L3.1	21/07/2016			24	61 137
	L2 L1.1	08/09/2016			2	55 899
	L2 L1.2				6	44 818
	L2 L1.3				7	9 377
	L2 L2.1	09/09/2016			4	45 736
	L2 L2.2				4	18 003
	L2 L2.3				3	42 071
	L2 L3.1	20/09/2016			5	47 511
	L3 L1.1	28/10/2016			21	39 779
	L3 L1.2				23	55 092
	L3 L1.3				25	23 693
	L3 L2.1	31/10/2016			14	21 926
	L3 L2.2				10	29 406
	L3 L2.3				9	30 851
	L3 L3.1	28/10/2016			11	60 329
	L3 L3.2				6	129 633
	L3 L3.3				16	43 127
	L0P1.1	16/06/2016		Meadow	2	36 323
	L0P1.2				4	40 952
	L1 P1.1	19/07/2016			1	34 275
	L1 P1.2				5	20 815
	L1 P1.3				7	72 112
	L1 P2.1	20/07/2016			4	30 729
	L1 P2.2				2	29 435
	L2 P1.1	08/09/2016			2	7 557
	L2 P2.2	16/09/2016			3	88 966
	L2 P2.3				7	43 379
	L2 P3.1	20/09/2016			4	71 646
	L3 P3.1	28/10/2016			1	15 735
	S1 L2.1	28/07/2016	St Martin de Londres	Crop	10	112 280
	S1 L2.2				8	32 174
	S1 L2.4.5				2	57 048
	S1 L2.6				1	36 128
	S1 L3.1	02/08/2016			2	33 508
	S1 L3.2				2	38 839
	S1 L3.3				6	22 450
	S1 L4.1	04/08/2016			2	10 846
	S2 L1.1	03/10/2016			19	45 510
	S2 L1.2				20	15 681
	S2 L1.3				19	102 971
	S2 L2.2	05/10/2016			15	28 202
	S2 L2.3				22	96 557
	S2 L3.1	06/10/2016			12	10 291
	S3 L1.3	21/10/2016			18	35 697
	S3 L2.1				21	79 205
	S3 L2.2				8	81 704
	S3 L2.3				7	51 680
	S3 L3.1	27/10/2016			7	25 431
	S3 L3.2				1	169 661
	S2 P3.1	06/10/2016		Meadow	1	45 905
	S3 P3.1	27/10/2016			3	53 216
	C0 L1.1	21/06/2016	Candillargues	Crop	1	27 798
	C0 L1.2				6	47 524
	C1 L1.1	19/07/2016			12	10 050
	C1 L2.1	20/07/2016			10	6 146
	C1 L2.2				11	6 044
	C1 L2.3				18	39 117
	C1 L3.1	26/07/2016			2	43 481
	C1 L3.2				4	17 444
	C1 L3.3				4	15 638
	C2 L1.2	22/09/2016			17	36 096
	C2 L1.3				12	73 394
	C2 L2.1	23/09/2016			21	44 584
	C2 L2.2				11	44 705
	C2 L2.3				19	22 359
	C2 L3.1	27/09/2016			10	13 109
	C2 L3.2				3	33 320
	C2 L3.3				9	39 822
	C3 L1.1	18/10/2016			36	45 953
	C3 L1.2				52	118 658
	C3 L1.3				59	29 785
	C3 L2.1	19/10/2016			21	80 788
	C3 L2.2				34	59 372
	C3 L2.3				25	34 913
	C3 L3.1				22	27 799
	C3 L3.2				14	51 651
	C3 L3.3				22	20 072
	C3 P3.1			Meadow	5	55 913
			Total	132	5 590	9 318 409
Plants	R1L	15/04/2015	Prades le Lez	Crop	10	160 833
	R1P	15/04/2015		Meadow	10	267 375
	L1L	21/07/2016	Lattes	Crop	30	80 522
	L2L	20/09/2016			30	56 625
	L3L	31/10/2016			30	80 158
	L1P	21/07/2016		Meadow	30	30 687
	L2P	20/09/2016			30	51 786
	L3P	31/10/2016			90	34 108
	S1L	04/08/2016	St Martin de Londres	Crop	30	73 329
	S2L	06/10/2016			30	104 789
	S3L	27/10/2016			30	194 795
	S1P	04/08/2016		Meadow	30	64 012
	S2P	06/10/2016			30	24 152
	S3P	27/10/2016			30	24 486
	C1L	26/07/2016	Candillargues	Crop	30	10 366
	C2L	27/09/2016			30	47 338
	C3L	19/10/2016			30	9 940
	C1P	26/07/2016		Meadow	30	75 772
	C2P	27/09/2016			30	32 283
			Total	19	530	1 423 462

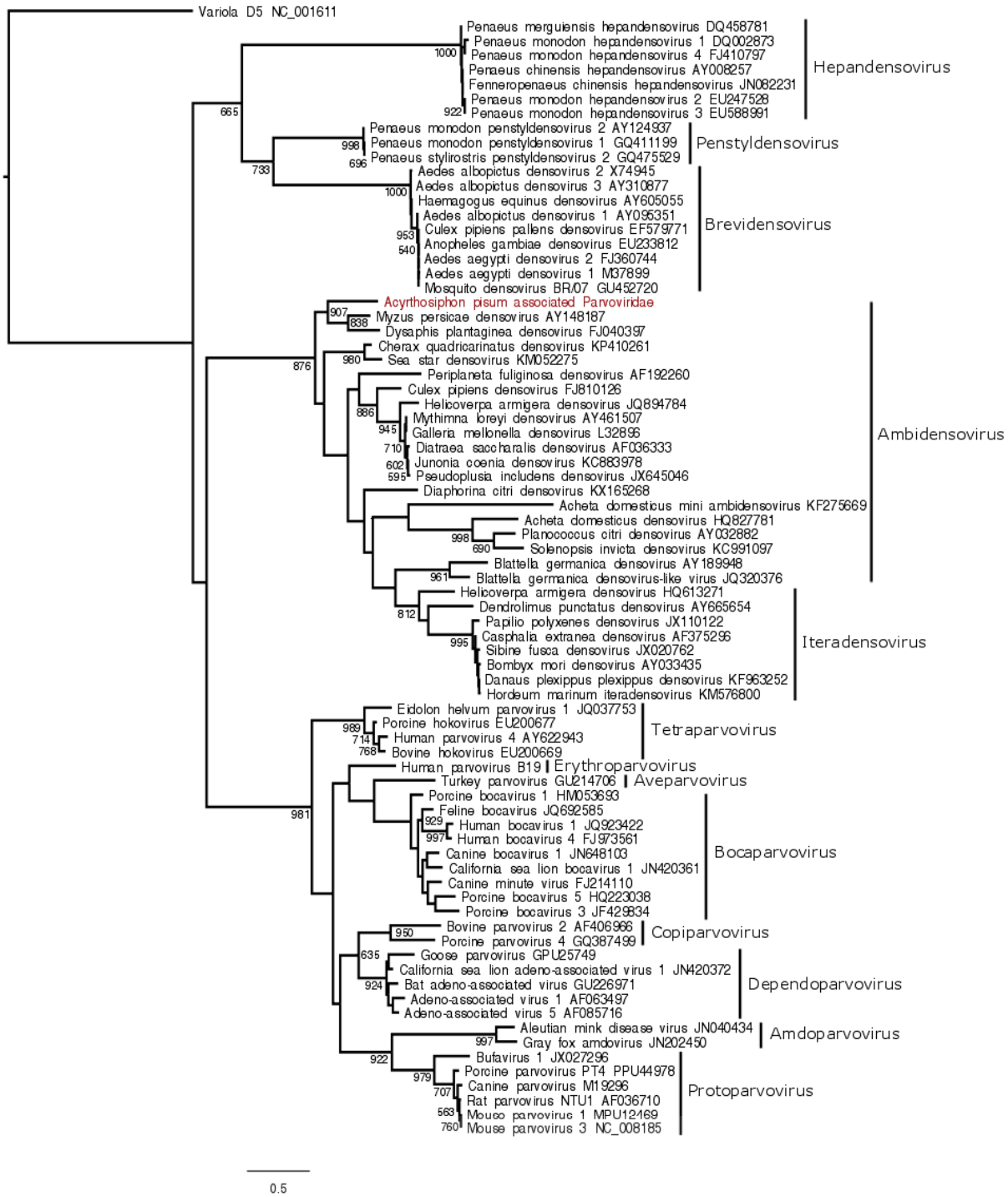
SM2: Maximum *Nodaviridae* likelihood phylogenetic tree based on partial polymerase protein, including 14 nodavirus species and 2 *Helicoverpa armigera* associated nodaviruses (in red). The alignment of 762 amino acids in length was produced using MUSCLE 3.7 (16 iterations) and was hand ungapped. The tree was midpoint rooted. Bootstrap values ≥50% are indicated at each node. Scale bars correspond to amino acid substitutions per site. Genera of the *Nodaviridae* family are indicated in brackets.



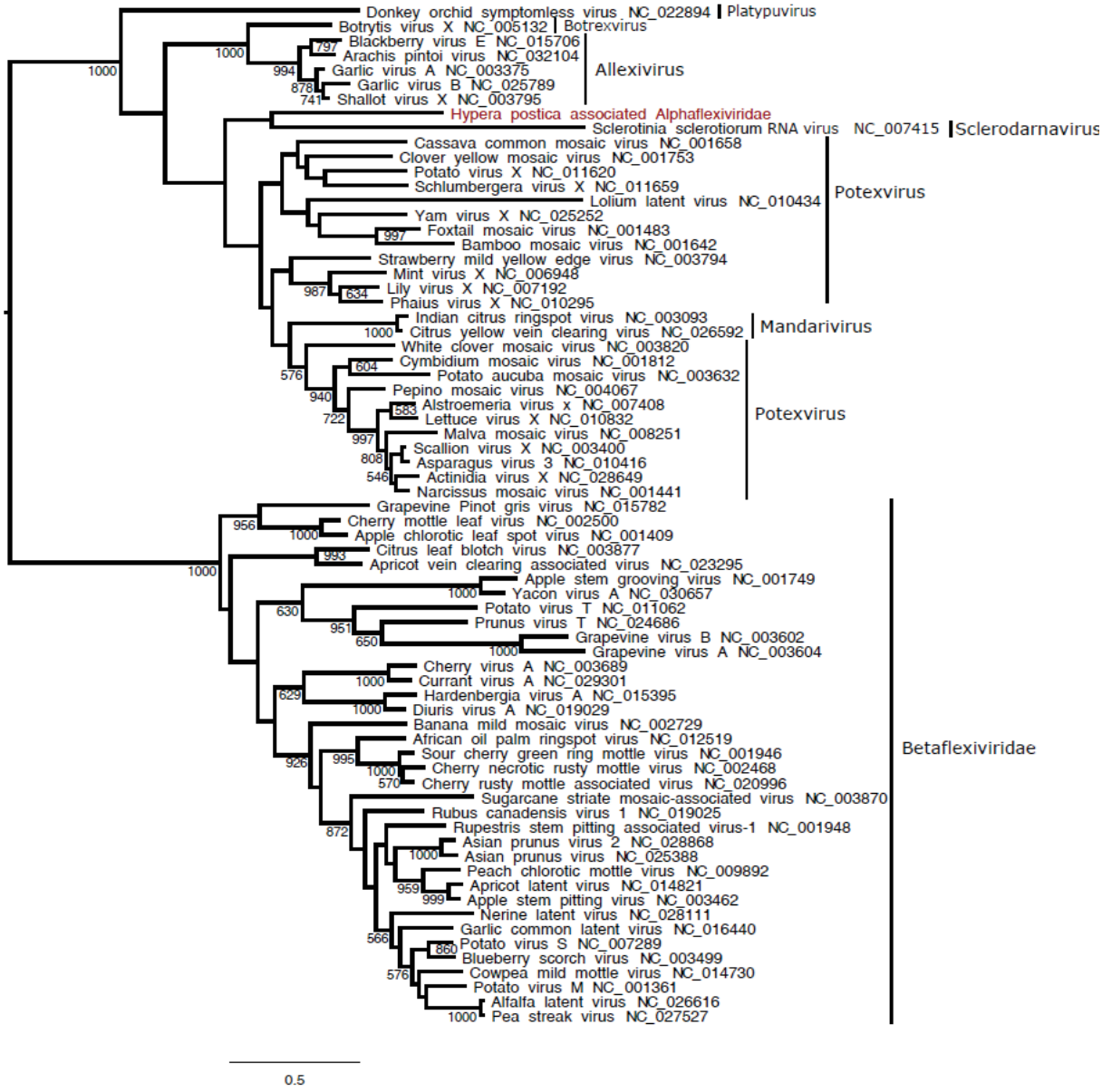
SM3: Maximum *Reoviridae* likelihood phylogenetic tree based on partial polymerase protein, including 54 nodavirus species and *Helicoverpa armigera associated Reoviridae* (in red). The alignment of 287 amino acids in length was produced using MUSCLE 3.7 (16 iterations) and was hand ungapped. The tree was midpoint rooted. Bootstrap values ≥50% are indicated at each node. Scale bars correspond to amino acid substitutions per site. Genera of the *Reoviridae* family are indicated in brackets.



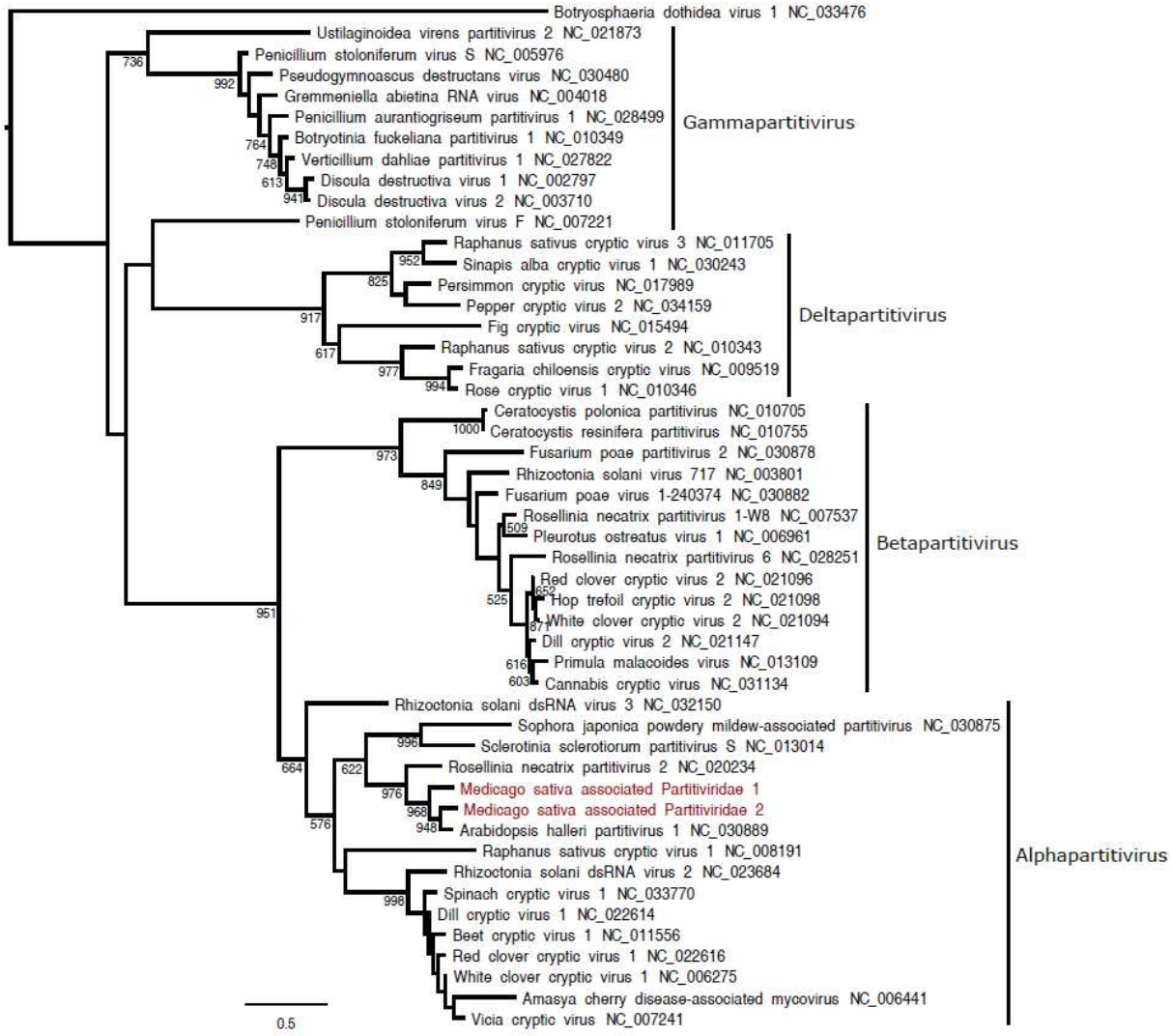
SM4: Maximum *Parvoviridae* likelihood phylogenetic tree based on partial polymerase protein, including 77 parvovirus species and *Acyrthosiphon pisum associated Parvoviridae* (in red). The alignment of 158 amino acids in length was produced using MUSCLE 3.7 (16 iterations) and was hand ungapped. The tree was rooted using variola D5 protein. Bootstrap values ≥50% are indicated at each node. Scale bars correspond to amino acid substitutions per site. Genera of the *Parvoviridae* family are indicated in brackets.



SM5: Maximum flexiviruses likelihood phylogenetic tree based on partial polymerase protein, including 69 flexivirus species and *Hypera postica associated Alphaflexiviridae* (in red). The alignment of 337 amino acids in length was produced using MUSCLE 3.7 (16 iterations) and was hand ungapped. The tree was midpoint rooted. Bootstrap values ≥50% are indicated at each node. Scale bars correspond to amino acid substitutions per site. Genera of the flexiviruses are indicated in brackets.



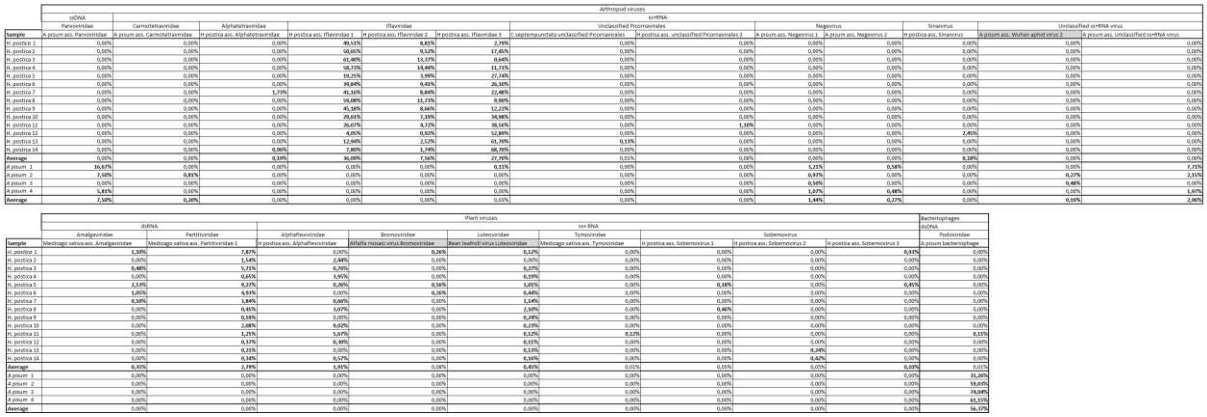
SM6: Maximum *Partitiviridae* likelihood phylogenetic tree based on partial polymerase protein, including 47 partitivirus species and *Medicago sativa* associated partitiviruses (in red). The alignment of 173 amino acids in length was produced using MUSCLE 3.7 (16 iterations) and was hand ungapped. The tree was midpoint rooted. Bootstrap values ≥50% are indicated at each node. Scale bars correspond to amino acid substitutions per site. Genera of the *Partitiviridae* family are indicated in brackets.



SM7: Detailed information about the abundance of viral families (in terms of read number) found in *Helicoverpa armigera* viromes.

	Arthropod viruses															Plant viruses											Bacteriophages			
	dsDNA			ssDNA		dsRNA	ss+RNA							ss-RNA		ssDNA		dsRNA			ss+RNA						dsDNA			ssDNA
Sample	Baculo.	Nudi.	Polydna.	Bidna.	Parvo.	Reo.	Flavi.	Noda.	Alphatetra.	Dicistro.	Ifla.	Negevirus	Sinaivirus	Phenui.	Rhabdo.	Gemini.	Circulavirus	Amalga.	Partiti.	Toti.	Alfaflexi.	Bromo.	Tymo.	Virga.	Sobemovirus	**Satellites**	Myo.	Podo.	Sipho.	Micro.
SWP 87	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	71,83%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,28%	0,00%
SWP 93	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	11,81%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,12%	0,00%	0,00%
CWP 93	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%
MS 1	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%
MS 2	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,18%	0,00%	0,00%	0,00%
MNS 1	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	1,30%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%
MNS 2	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,55%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%
MNS 3	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	2,21%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%
R1 2014	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%
R2 2014	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	10,44%	1,64%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%
R3 2014	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	4,90%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%
R4 2014	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	14,24%	0,28%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,21%	0,00%	0,00%	0,00%	0,00%	0,00%
R5 2014	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	2,71%	0,00%	1,85%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%
M1L1.2	0,00%	0,13%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%
M1L2.1	0,28%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%
M1L3.1	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,76%	0,00%	0,00%	0,00%	0,00%	0,35%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%
M1L3.2	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,15%	0,00%	0,00%	0,00%	0,00%	0,45%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%
M1L3.4	0,28%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,67%	0,00%	0,00%	0,00%	0,00%	0,00%
M1L4.1	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%
M1L4.2	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,22%	0,19%	0,00%	0,00%	0,00%	0,11%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%
M1L4.3	0,41%	0,00%	0,82%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	1,04%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%
M2L1.1	20,68%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,20%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,46%	0,00%
M2L1.2	15,94%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	1,04%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%
M2L1.3	31,33%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%
M2L2.1	33,63%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,20%	0,00%	0,00%
M2L3.1	31,12%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,27%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%
M2L3.2	13,89%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,89%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	3,13%	0,00%
M2L3.3	23,17%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,21%	0,00%	0,00%
S1L1.1 2015	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	2,11%	0,00%	0,00%	0,00%	0,00%	0,23%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%
S1L1.2 2015	0,72%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	2,12%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%
S1L1.3 2015	0,11%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%
S1L2.2 2015	0,00%	0,00%	0,11%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	2,08%	0,35%	0,00%	0,00%	0,00%	0,21%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,29%	0,00%	0,00%
S1L3.2 2015	0,11%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,61%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%
S1L3.3 2015	0,12%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%
S2L1.1 2015	14,01%	0,00%	0,00%	0,00%	0,00%	1,01%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,28%	0,12%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%
S2L1.3 2015	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,33%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%
S2L2.1 2015	32,75%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%
S2L3.2 2015	28,05%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,50%	0,00%	0,00%	0,00%	0,13%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%
R1L1.1	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%
R2L2.1	0,00%	0,00%	0,12%	0,00%	0,00%	2,93%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%
R2L2.2	0,00%	0,00%	0,31%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,12%	0,00%
R2L3.1	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,43%	0,00%	0,00%
R1P1.1	0,77%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%
R1P3.1	0,74%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%
R1P3.2	0,42%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%
L0L1.1	0,69%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	16,87%	0,00%	0,00%	0,00%	0,00%	0,17%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,15%	0,00%	0,00%
L0L1.2	0,73%	0,11%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,73%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%
L0L1.3	0,62%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,53%	0,00%	0,00%	0,00%	0,00%	0,19%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	4,50%	0,00%	0,00%
L1 L1.1	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,12%	0,00%	0,00%	0,00%	0,00%	0,14%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%
L1 L1.2	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,29%	0,00%	0,00%
L1 L1.3	0,10%	0,00%	0,00%	0,00%	0,16%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	2,62%	0,00%	0,00%	1,34%	0,00%	0,00%	0,18%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%
L1 L2.1	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	27,29%	1,51%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%
L1 L2.2	0,20%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,36%	0,00%	0,14%	0,00%	0,00%	0,41%	0,00%	0,00%	0,00%	0,00%	0,00%	0,11%	0,00%	0,00%	0,00%	0,00%	0,00%
L1 L2.3	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%
L1 L3.1	0,48%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,21%	0,00%	0,00%	0,00%	0,00%	0,72%	0,00%	0,15%	0,00%	0,17%	0,40%	0,00%	0,00%	0,22%	0,00%	0,00%	0,00%	0,00%	0,00%	0,16%	0,00%	0,19%
L2 L1.1	0,16%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%
L2 L1.2	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,29%	0,00%	0,00%	0,34%	0,93%	0,00%	0,00%	0,00%	0,25%	0,15%	0,00%	0,39%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,33%	0,00%
L2 L1.3	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,21%	0,00%	0,00%	0,53%	0,00%	0,00%	0,00%	0,00%	0,00%	0,18%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%
L2 L2.1	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,39%	0,00%	0,00%	5,14%	0,00%	0,00%	0,00%	0,00%	0,00%	0,12%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%
L2 L2.2	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,13%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	1,19%	0,00%	0,00%	0,44%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%
L2 L2.3	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,32%	0,00%	0,00%	7,43%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%
L2 L3.1	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,19%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,29%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%
L3 L1.1	0,16%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,77%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,11%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%
L3 L1.2	0,27%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,29%	0,00%	0,00%	0,00%	0,00%	0,33%	0,11%	0,00%	0,32%	0,00%	0,21%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%
L3 L1.3	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,37%	0,00%	0,00%	0,00%	0,00%	1,73%	0,15%	0,00%	0,00%	0,00%	0,19%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%
L3 L2.1	0,37%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,10%	0,00%	0,00%	0,00%	0,00%	0,91%	0,17%	0,00%	0,62%	0,00%	1,56%	0,00%	0,00%	0,00%	0,00%	0,00%	0,10%	0,00%	0,00%	0,00%	0,00%	0,10%
L3 L2.2	11,14%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,22%	0,00%	0,00%	0,00%	0,00%	1,87%	0,62%	0,00%	0,83%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,12%	0,00%	0,00%	0,00%	0,00%	0,00%
L3 L2.3	0,18%	0,00%	0,00%	0,00%	0,00%	0,62%	0,00%	0,00%	1,33%	0,00%	0,00%	0,00%	0,00%	0,26%	0,00%	0,00%	0,00%	0,00%	0,12%	0,00%	0,00%	0,00%	0,00%	0,00%	2,85%	0,00%	0,00%	0,00%	0,00%	0,00%
L3 L3.1	0,00%	0,00%	0,00%	0,00%	0,31%	0,46%	0,00%	0,00%	6,57%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,11%	0,00%	3,27%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,12%	0,00%	0,00%
L3 L3.2	0,00%	0,00%	0,00%	0,44%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,11%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%
L3 L3.3	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	1,51%	0,00%	0,00%	0,00%	0,00%	0,49%	0,51%	0,00%	0,10%	0,00%	0,42%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%
L0P1.1	0,13%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%
L0P1.2	0,49%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%
L1 P1.1	0,60%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,28%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%
L1 P1.2	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	1,26%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,29%	0,00%
L1 P1.3	0,18%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,76%	0,30%	0,00%	0,00%	0,00%	0,81%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,10%	0,00%
L1 P2.1	0,20%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,10%	0,00%
L1 P2.2	0,44%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%
L2 P1.1	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,20%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%
L2 P2.2	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,11%	0,00%
L2 P2.3	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,58%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%
L2 P3.1	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,22%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,56%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%
L3 P3.1	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,17%	0,00%
S1 L2.1	0,24%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,13%	0,00%	1,30%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%
S1 L2.2	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%
S1 L2.4.5	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	1,64%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%
S1 L2.6	0,00%	0,00%	0,21%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,12%	0,00%
S1 L3.1	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,11%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%
S1 L3.2	0,13%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,22%	0,00%	0,00%	1,98%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,18%	0,00%
S1 L3.3	0,00%	0,00%	0,00%	0,00%	0,00%	0,27%	0,00%	0,00%	0,00%	0,00%	0,24%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,25%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,14%	0,00%
S1 L4.1	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%
S2 L1.1	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	10,88%	0,15%	0,00%	0,00%	0,00%	0,99%	0,28%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%
S2 L1.2	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,17%	0,00%	0,00%	0,00%	0,00%	0,13%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,11%	0,27%	0,00%
S2 L1.3	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,13%	0,00%	0,00%	0,00%	0,00%	0,15%	0,21%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%
S2 L2.2	0,16%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	3,61%	0,57%	0,00%	0,32%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%
S2 L2.3	0,00%	0,00%	0,00%	0,00%	0,10%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,22%	0,00%	0,00%	0,00%	0,00%	0,22%	0,00%	0,00%	0,15%	0,00%	0,00%	0,00%	0,00%	0,00%	0,10%	0,00%	0,00%
S2 L3.1	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	2,76%	0,34%	0,00%	0,00%	0,00%	0,00%	0,10%	0,00%	0,00%	0,30%	0,00%	0,00%
S3 L1.3	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,97%	0,18%	0,00%	0,00%	0,12%	3,19%	0,14%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%
S3 L2.1	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	1,30%	0,12%	0,00%	0,00%	0,00%	0,61%	0,45%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,21%	0,00%
S3 L2.2	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,13%	0,43%	0,00%	0,00%	0,20%	0,79%	0,54%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,10%	0,00%
S3 L2.3	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,52%	1,43%	0,00%	0,00%	0,00%	4,22%	0,21%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%
S3 L3.1	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,38%	0,00%	0,18%	0,00%	1,12%	0,00%	0,00%	0,00%	0,00%	0,00%	0,62%	0,00%	0,00%	0,00%	0,17%	0,00%
S3 L3.2	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	11,96%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,14%	0,00%	0,00%	0,00%
S2 P3.1	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,15%	0,00%
S3 P3.1	1,13%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	1,44%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%
C0 L1.1	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	4,42%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	3,57%	0,35%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%
C0 L1.2	0,00%	0,00%	0,00%	0,00%	0,00%	5,55%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,36%	0,00%	0,00%	0,00%	0,00%	0,00%	1,15%	0,39%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%
C1 L1.1	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	4,47%	0,00%	0,00%	1,62%	0,15%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%
C1 L2.1	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,16%	2,89%	0,00%	0,00%	0,18%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,39%	0,00%	0,00%
C1 L2.2	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,30%	4,32%	0,00%	0,00%	0,63%	0,78%	0,00%	0,00%	0,00%	0,00%	0,00%	0,18%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,17%	0,00%	0,55%
C1 L2.3	0,00%	0,00%	0,00%	0,00%	0,00%	1,62%	0,00%	0,00%	0,00%	0,00%	0,31%	0,00%	0,00%	0,41%	0,00%	0,00%	0,00%	0,00%	0,00%	0,25%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%
C1 L3.1	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,11%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,15%	0,00%
C1 L3.2	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	1,39%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%
C1 L3.3	0,38%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	5,54%	0,00%	0,38%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	2,22%	0,00%	0,00%	0,00%	0,00%	0,00%
C2 L1.2	0,24%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	3,76%	1,32%	0,00%	0,00%	0,49%	0,00%	0,00%	0,00%	0,00%	0,31%	0,00%	0,00%	0,00%	0,10%	0,00%	0,00%	0,00%	0,00%	1,96%	0,00%	0,00%
C2 L1.3	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	2,80%	1,25%	0,00%	0,00%	0,38%	0,00%	0,19%	0,00%	0,12%	0,12%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	2,21%	0,00%	0,00%
C2 L2.1	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,30%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,45%	8,62%	0,00%	0,00%
C2 L2.2	0,00%	0,00%	0,00%	0,00%	0,00%	7,64%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,28%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,32%	0,00%	0,00%
C2 L2.3	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	3,20%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,93%	0,00%	0,00%
C2 L3.1	0,14%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,21%	0,00%	0,00%	0,00%	0,14%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%
C2 L3.2	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,21%	0,00%	0,00%	0,00%	0,65%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,91%	0,00%	0,00%
C2 L3.3	0,24%	0,00%	0,00%	0,00%	0,17%	2,86%	0,00%	0,00%	0,00%	0,95%	0,58%	0,00%	0,00%	0,00%	0,34%	0,00%	0,00%	0,00%	0,00%	2,25%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	13,67%	0,00%	0,00%
C3 L1.1	0,10%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,44%	1,02%	0,00%	0,70%	0,52%	0,00%	0,00%	0,29%	0,61%	0,00%	0,00%	0,00%	0,00%	0,27%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%
C3 L1.2	0,00%	0,00%	0,00%	0,00%	0,00%	2,25%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	1,08%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%
C3 L1.3	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	23,41%	0,00%	0,00%	0,00%	0,00%	0,22%	0,00%	0,00%	0,00%	3,10%	0,00%	0,65%	0,00%	0,00%	0,00%	0,00%	0,00%
C3 L2.1	0,00%	0,00%	0,00%	0,00%	0,00%	2,29%	0,00%	0,00%	0,00%	0,00%	0,10%	0,00%	0,00%	0,59%	0,12%	0,00%	0,00%	0,00%	0,16%	0,72%	0,00%	0,00%	0,12%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%
C3 L2.2	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,22%	0,00%	0,00%	28,23%	0,91%	0,00%	0,00%	0,00%	0,18%	0,29%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,12%	0,00%
C3 L2.3	0,00%	0,00%	0,00%	0,00%	0,00%	3,26%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,27%	0,00%	0,00%	0,00%	0,00%	0,13%	2,09%	0,00%	0,00%	0,46%	0,00%	0,44%	0,00%	0,00%	0,00%	0,00%	0,00%
C3 L3.1	0,14%	0,00%	0,00%	0,00%	0,00%	0,19%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,12%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,23%	0,00%	0,00%	0,00%	0,00%	0,00%
C3 L3.2	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	3,83%	0,00%	0,00%	0,00%	0,25%	0,26%	0,44%	0,25%	0,00%	0,49%	0,00%	0,00%	0,00%	0,00%	0,71%	0,60%	1,53%
C3 L3.3	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,89%	0,00%	0,00%	0,00%	0,00%	0,28%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,30%	0,00%	0,00%
C3 P3.1	0,00%	0,00%	0,00%	0,00%	0,76%	0,00%	0,17%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%

SM8: Detailed information about the abundance of viral families (in terms of read number) found in *Hypera postica*, *Acyrthosiphon pisum*, *Coccinella septempunctata* and *Synaema globosum* viromes.



SM9: Detailed information about the abundance of viral families (in terms of read number) found in plants viromes.

Sample	Arthropod viruses				Plant viruses										Bacteriophages	
	dsRNA	ss+RNA		ss-RNA	ssDNA	dsRNA			ss+RNA					Satellites	dsDNA	
	Reo.	Dicistro.	Unclassified Picornavirales	Phenui.	Geminicirculavirus	Amalga.	Partiti.	Toti.	Alfaflexi.	Bromo.	Luteo.	Poty.	Seco.	Tymo.	Myo.	Sipho.
R1L	0,00%	0,00%	0,00%	0,00%	0,00%	**9,88%**	**23,32%**	0,00%	0,00%	**0,22%**	**1,68%**	0,00%	0,00%	0,00%	0,00%	0,00%
R1P	0,00%	0,00%	0,00%	0,00%	0,00%	**20,89%**	**2,53%**	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	**1,40%**	0,00%	0,00%
L1L	0,00%	0,00%	0,00%	0,00%	0,00%	**2,21%**	**4,40%**	**0,30%**	0,00%	**6,71%**	0,00%	0,00%	0,00%	**0,26%**	0,00%	0,00%
L2L	0,00%	0,00%	**0,32%**	0,00%	**0,72%**	0,00%	**19,09%**	**0,40%**	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%
L3L	0,00%	0,00%	0,00%	0,00%	0,00%	**0,92%**	**17,28%**	0,00%	0,00%	**0,25%**	0,00%	0,00%	0,00%	**0,44%**	0,00%	0,00%
L1P	**0,12%**	0,00%	0,00%	0,00%	0,00%	**0,91%**	**3,60%**	**0,55%**	0,00%	0,00%	0,00%	**0,24%**	0,00%	**0,50%**	0,00%	0,00%
L2P	0,00%	0,00%	0,00%	0,00%	0,00%	**0,60%**	**0,26%**	**6,41%**	0,00%	**1,26%**	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%
L3P	0,00%	0,00%	0,00%	0,00%	0,00%	**1,36%**	0,00%	**0,30%**	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	**0,17%**	**0,25%**
S1L	0,00%	0,00%	0,00%	0,00%	0,00%	**5,43%**	**4,06%**	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%
S2L	0,00%	0,00%	0,00%	0,00%	0,00%	**1,30%**	**2,63%**	**1,37%**	0,00%	**1,20%**	0,00%	0,00%	**0,21%**	0,00%	0,00%	0,00%
S3L	0,00%	0,00%	0,00%	0,00%	0,00%	**0,22%**	**14,76%**	**0,23%**	0,00%	**3,25%**	0,00%	0,00%	**0,68%**	0,00%	0,00%	0,00%
S1P	0,00%	0,00%	0,00%	0,00%	**0,65%**	**5,95%**	**0,18%**	0,00%	0,00%	**15,98%**	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%
S2P	0,00%	0,00%	0,00%	0,00%	0,00%	**0,19%**	0,00%	**1,72%**	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	**0,30%**
S3P	0,00%	0,00%	0,00%	**0,10%**	**0,13%**	0,00%	**0,11%**	**0,17%**	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%
C1L	0,00%	0,00%	0,00%	0,00%	0,00%	**0,95%**	**7,82%**	0,00%	**0,89%**	**4,48%**	0,00%	0,00%	0,00%	**0,18%**	0,00%	0,00%
C2L	0,00%	**0,30%**	0,00%	0,00%	**3,62%**	**0,69%**	**1,65%**	0,00%	0,00%	**8,09%**	0,00%	0,00%	0,00%	**0,26%**	0,00%	0,00%
C3L	0,00%	0,00%	0,00%	0,00%	**0,12%**	**1,76%**	0,00%	**15,10%**	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	**2,75%**
C1P	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	**0,20%**
C2P	0,00%	0,00%	0,00%	0,00%	0,00%	**0,13%**	0,00%	0,00%	0,00%	**1,06%**	0,00%	0,00%	0,00%	0,00%	0,00%	**0,43%**

SM10: Detailed information about sequences found in putative host genomic and/or transcriptomic databases (Refseq_genomic, WGS, EST and TSA) that display significant homologies with viral species discovered in the viromes generated for this study.

Host	Virus name	Length (bp)	Database	Number of sequences	Query cover	E-value	Identity (%)	Accession
A. pisum	A. pisum bacteriophage NC 000935	36 524	refseq_genomics	2	2%	0.0	100%	CN761186.1
					0%	2,00E-51	100%	DV749266.1
			EST	2	2%	0.0	100%	CN761186.1
					0%	6,00E-49	100%	DV749266.1
	Aphid lethal paralysis virus variant	9 654	EST	25	4%	0.0	96%	DV749361.1
					5%	0.0	96%	DV751603.1
					4%	0.0	96%	DV749431.1
					4%	0.0	96%	DV749430.1
					4%	0.0	96%	DV750722.1
					4%	0.0	96%	DV750355.1
					4%	0.0	96%	DV748456.1
					4%	0.0	96%	DV749149.1
					3%	1,00E-150	96%	DV750549.1
					3%	1,00E-150	96%	DV749958.1
					3%	8,00E-147	96%	DV751827.1
					3%	8,00E-147	96%	DV751741.1
					3%	8,00E-147	96%	DV751084.1
					3%	8,00E-147	96%	DV751032.1
					3%	9,00E-146	96%	DV748735.1
					3%	9,00E-146	96%	DV748734.1
					5%	0.0	95%	DV751679.1
					3%	6,00E-148	95%	DV751453.1
					3%	4,00E-144	95%	DV747564.1
					3%	5,00E-143	95%	DV751165.1
					3%	2,00E-141	95%	DV748338.1
					3%	6,00E-136	95%	DV751955.1
					3%	6,00E-136	95%	DV749466.1
					3%	6,00E-136	95%	DV748115.1
					2%	1,00E-119	95%	DV749000.1
	A. pisum ass. Negevirus 1 part 1	7 582	EST	3	6%	0.0	99%	CN764043.1
					6%	0.0	99%	CN753768.1
					6%	0.0	99%	CF587975.1
	A. pisum ass. Negevirus 1 part 3	1 348	EST	36	64%	0.0	99%	CN760792.1
					60%	0.0	99%	CN758286.1
					60%	0.0	99%	CN587220.1
					59%	0.0	99%	CN759304.1
					59%	0.0	99%	CN585454.1
					58%	0.0	99%	CN757075.1
					59%	0.0	99%	CN757076.1
					58%	0.0	99%	CN586131.1
					57%	0.0	99%	CN753370.1
					56%	0.0	99%	CN760230.1
					56%	0.0	99%	CN760332.1
					56%	0.0	99%	CN758603.1
					56%	0.0	99%	CN761242.1
					56%	0.0	99%	CN760750.1
					56%	0.0	99%	CN754695.1
					56%	0.0	99%	CN760556.1
					56%	0.0	99%	CN760281.1
					54%	0.0	99%	CN758122.1
					53%	0.0	99%	CN758072.1
					51%	0.0	99%	CF588152.1
					51%	0.0	99%	CN762532.1
					49%	0.0	99%	CN760579.1
					47%	0.0	99%	CN759603.1
					46%	0.0	99%	CF587515.1
					45%	0.0	97%	CN763153.1
					41%	0.0	99%	CN582708.1
					37%	0.0	99%	CN583966.1
					26%	0.0	99%	CN757673.1
					22%	4,00E-151	99%	CN583759.1
					21%	1,00E-143	99%	CN759282.1
					21%	1,00E-143	99%	CN584074.1
					20%	7,00E-141	99%	CN586207.1
					20%	3,00E-139	99%	CN760018.1
					20%	3,00E-139	99%	CN754997.1
					20%	1,00E-131	98%	CN761487.1
					19%	2,00E-130	99%	CN753903.1
	Wuhan aphid virus 2 segment 3 variant	2 410	EST	2	34%	0.0	99%	CN758075.1
					25%	0.0	99%	CN584239.1
	Wuhan aphid virus 2 segment 4 variant	2 601	EST	2	29%	0.0	98%	CN763493.1
					29%	0.0	98%	CN755338.1
	A. pisum ass. ss+RNA virus segment 1	3 625	EST	1	10%	3,00E-177	98%	CN582481.1
	A. pisum ass. ss+RNA virus segment 2	2 097	EST	3	42%	0.0	99%	CN754325.1
					32%	0.0	99%	CN753576.1
					28%	0.0	96%	CN760896.1
M. sativa	M. sativa ass. Amalgaviridae	3 375	TSA	1	99%	0.0	99%	GAFF01077243.1
	M. sativa ass. Partitiviridae 2 segment 1	1 766	EST	30	33%	0.0	99%	CO513857.1
					33%	0.0	99%	CO511770.1
					33%	0.0	99%	CO512252.1
					33%	0.0	99%	CO512165.1
					32%	0.0	99%	CO513128.1
					32%	0.0	99%	CO512026.1
					32%	0.0	98%	CO514819.1
					31%	0.0	99%	CO512123.1
					31%	0.0	99%	CO514073.1
					30%	0.0	99%	CO513583.1
					30%	0.0	98%	CO514197.1
					28%	0.0	99%	CO513743.1
					28%	0.0	98%	CO512507.1
					27%	0.0	99%	CO515433.1
					27%	0.0	100%	CO513931.1
					28%	0.0	98%	CO516496.1
					27%	0.0	99%	CO515236.1
					26%	0.0	98%	CO512002.1
					26%	0.0	99%	CO514204.1
					26%	0.0	99%	CO515987.1
					24%	0.0	99%	CO512467.1
					22%	0.0	99%	EX525599.1
					22%	0.0	97%	CO516459.1
					17%	3,00E-158	99%	CO514372.1
					18%	4,00E-157	98%	CO515560.1
					15%	3,00E-126	97%	CO515895.1
					13%	3,00E-114	98%	CO516393.1
					11%	7,00E-103	100%	CO513780.1
					10%	2,00E-83	97%	CO511693.1
					5%	4,00E-42	100%	CO516226.1
			TSA	1	99%	0.0	99%	GAFF01144675.1
	M. sativa ass. Partitiviridae 2 segment 2	1 185	TSA	2	33%	0.0	99%	GAFF01172912.1
					63%	0.0	99%	GAFF01168043.1

Bilan et perspectives

Étant donné notre manque d'information sur la diversité des virus d'arthropodes, afin de comprendre le rôle qu'occupent ces virus dans le fonctionnement des agroécosystèmes, et en particulier ceux associés aux espèces d'arthropodes ravageurs de cultures, il est d'abord nécessaire d'établir un inventaire des virus circulants chez ces arthropodes, puis de comparer leur distribution dans différents environnements spatio-temporels occupés par leurs hôtes.

Nos études ont permis de mettre en évidence que la composition des viromes différait drastiquement entre les espèces de ravageurs étudiées. En outre, certaines séquences virales ont été retrouvées dans l'ensemble des viromes d'une même espèce de ravageurs. Il est néanmoins nécessaire de pondérer ces résultats par l'effort d'échantillonnage limité réalisé ici, allant d'un prélèvement unique à plusieurs prélèvements réalisés durant deux ans à une échelle spatiale régionale, ne permettant pas d'étudier la diversité virale inféodée aux espèces d'arthropodes testées de manière holistique (Breitbart, 2012). De plus, l'utilisation d'un filtre d'abondance – ici fixé arbitrairement - lors du traitement bioinformatique des données, afin de limiter les contaminations inter-échantillons, a pu éliminer des virus y étant faiblement abondants. Ainsi, la diversité virale contenue dans les échantillons traités n'a pas été entièrement explorée. Cependant, la présence de virus qui seraient inféodés à des espèces d'arthropodes particulières pose la question de l'impact de ces virus sur leurs hôtes, ainsi que celle de la prévalence de ces virus dans leurs populations d'hôtes à une échelle géographique plus large.

Ces deux études ont également permis de mettre en évidence une grande variété de contigs pouvant représenter des membres de nouvelles espèces – voire de nouveaux genres- de virus entomopathogènes ou phytopathogènes. Ainsi, l'étude de viromes de ravageurs de cultures permet d'améliorer notre appréhension de la diversité virale associée aux arthropodes ainsi qu'aux plantes dont ils se nourrissent.

Enfin, la distribution de virus associés aux ravageurs d'arthropodes a été étudiée à l'échelle des communautés d'arthropodes. Sur les 13 contigs viraux testés, 5 ont été mis en évidence par PCR chez différentes espèces d'arthropodes. De plus, ces virus sont presque exclusivement présents chez des espèces prédatrices, ce qui indiquerait d'avantage leur présence dans le bol alimentaire des prédateurs que leur infection par ces virus.

En effet, bien que nos études ont mis en évidence la présence de virus potentiellement entomopathogènes chez des espèces d'arthropodes données, cela ne signifie pas nécessairement que ces virus infectent ces espèces : ils peuvent être issus de contaminations du bol alimentaire par l'ingestion de matériel souillé, ou présents sur la cuticule des arthropodes testés. Il est donc nécessaire de réaliser des études complémentaires pour déterminer le potentiel infectieux de ces virus chez les ravageurs de cultures étudiés.

Quant aux virus potentiellement entomopathogènes découverts dans ces viromes, de nombreuses études restent à faire pour tester leur potentiel réel pour la lutte biologique, notamment leur spectre d'hôte et leur spécificité. En effet, répandre un virus dans l'environnement requiert des connaissances préalables sur son impact phénotypique sur les espèces ciblées ainsi que sur ses potentialités d'infection d'autres espèces d'arthropodes (Lacey *et al.*, 2015).

En conclusion, la métagénomique virale représente un ensemble d'outils puissants permettant d'étudier, de manière fondamentale, l'écologie virale dans les agroécosystèmes. Intégrée à des études fonctionnelles portant sur la virulence et le spectre d'hôte des virus découverts, elle pourrait aider à améliorer le fonctionnement et le management des agroécosystèmes.

Discussion

Résumé des travaux réalisés lors de la thèse

Les travaux réalisés au cours de cette thèse ont permis des avancées de notre connaissance de la diversité et de la distribution des communautés virales associées aux arthropodes.

Dans un premier temps, l'étude de la diversité génétique et du spectre d'hôte de la famille des *Parvoviridae*, réalisée à partir d'analyses de données issues de génomique et de transcriptomique, a permis de mettre en évidence leur présence au sein d'une très grande diversité d'animaux, ce qui indique la potentielle ubiquité de ces virus dans le règne animal (Chapitre I). De plus, cette étude a montré que ces virus sont bien plus diversifiés génétiquement que ne le laissent supposer les souches isolées jusqu'ici, ce qui a permis d'améliorer notre point de vue sur leur histoire évolutive.

Dans un second temps, nous avons mis en place un protocole de métagénomique virale, basé sur la purification de particules virales et l'amplification aléatoires des acides nucléiques, permettant la préparation ainsi que l'analyse de viromes à partir d'échantillons d'arthropodes et de plantes. Notre méthode permet de multiplexer des échantillons et de caractériser des virus appartenant à l'ensemble de la classification de Baltimore (Chapitre II). L'utilisation de ce protocole nous a permis d'analyser la diversité ainsi que la composition de communautés virales associées à quatre espèces d'arthropodes ravageurs de cultures (Chapitre III). Ces analyses ont notamment permis de mettre en évidence la présence d'un virome spécifique de chaque espèce hôte étudiée, ainsi que de mettre en évidence une grande diversité génétique de virus d'arthropodes et de plantes. Ces découvertes ont permis d'améliorer nos connaissances sur la diversité des virus circulants dans les agroécosystèmes, et pourraient avoir des applications potentielles en lutte biologique.

L'ensemble de ces résultats ont ainsi apporté des éléments de réponses à des questions fondamentales plus larges concernant l'histoire évolutive des virus ainsi que leur prévalence dans les écosystèmes. Cependant, la métagénomique virale se heurte à des défis d'ordres techniques et conceptuels.

Développement de la métagénomique virale : de nombreux défis à relever

L'important développement des disciplines « omiques » - en particulier la métagénomique - a bouleversé notre conception de la diversité et de l'impact des virus au sein des écosystèmes et de leurs hôtes. Cependant, la métagénomique virale fait face à de nombreux défis, autant techniques que conceptuels, qui limitent pour l'instant son utilisation dans le cadre de l'épidémiologie et du diagnostic.

Limites méthodologiques

Il existe plusieurs limitations d'ordre technique liées à l'usage des approches de métagénomique virale.

Outre les biais liés à l'échantillonnage, inhérents à tout type d'études, la méthode de traitement des échantillons utilisée peut entrainer des biais dans la représentativité des virus au sein des viromes. Tout d'abord, l'utilisation d'un filtre peut amener à l'élimination de virus de grande taille (Halary *et al.*, 2016). De plus, chaque technique d'amplification des acides nucléiques peut induire des biais d'amplification. Par conséquent, la métagénomique virale ne peut pas caractériser les communautés virales de façon strictement quantitative (Kim et Bae, 2011). Les analyses de viromes indiquent donc une tendance générale des différentes quantités de chaque type de virus présents initialement dans l'échantillon sous forme de virions. Il est donc nécessaire de mettre en place des protocoles robustes de préparation de viromes, et des réplications d'analyses pour permettre des analyses statistiques rigoureuses, afin de remplacer ces données qualitatives par des données quantitatives. Par exemple, on pourrait tester l'impact de différents protocoles de purification et d'amplification, à partir de communautés virales préalablement connue provenant d'échantillons diversifiés et dont la quantité de chaque type de virions a préalablement été quantifiée, afin d'évaluer si les quantités de séquences virales concordent avec la composition initiale du virome (Conceição-neto *et al.*, 2015). Enfin, la contamination des jeux de données par des virus présents dans les laboratoires d'études ou dans les kits d'extraction et/ou d'amplification peut nuire à notre appréhension de la diversité des communautés virales (Naccache *et al.*, 2013; Smuts *et al.*, 2014). Une solution contre ce type de contamination consiste notamment à ne prendre en

compte que les séquences virales les plus abondantes dans les viromes, au risque de ne pas prendre en compte certains virus rares.

En outre, bien que les technologies de séquençage à haut débit soient puissantes, elles présentent elles aussi des limitations. Il existe de nombreuses méthodes de séquençage qui génèrent des quantités de données différentes, ainsi que différents niveau d'erreurs de séquençage (Genohub-ngs-instrument-guide). Cette diversité de résultats issus du séquençage à haut débit a rendu nécessaire la création d'outils bioinformatiques spécifiques dédiés à leur analyse. Un grand nombre de pipelines, de logiciels et de bases de données dédiés à l'étude des viromes est actuellement disponible, mais souvent l'évaluation comparative de ces outils, ainsi que leur mise à jour, n'est pas effectuée (Roossinck, 2016), ce qui rend leur utilisation confuse pour un scientifique ne réalisant pas de veille bibliographique régulière.

Enfin, les biais liés à l'analyse des données par traitement bioinformatique peuvent aussi induire des biais dans la représentativité des viromes. Par exemple, les outils d'attribution taxonomique les plus utilisés reposent sur l'utilisation de bases de données incomplètes, restreignant nos connaissances des virus à des formes connues. Autre exemple, des biais bioinformatiques induisent une surestimation de la proportion de matière noire dans les viromes, ce qui entraine une surestimation de la diversité génétique qui y est présente. Enfin, des erreurs pouvant se produire lors de l'assemblage des reads en contigs peuvent mener à la création de chimères artéfactuelles (Charuvaka et Rangwala, 2011). Ces biais techniques ont des répercussions importantes, car ils ont une influence directe sur notre vision du monde viral.

Cette pléthore de méthodes de génération et d'analyse de viromes induisant différents types de biais, rend les études transversales (e.g. comparatives des différents types de viromes) difficiles à mettre en place. Il est donc nécessaire de mettre en place des protocoles permettant l'étude comparative de viromes (Lin *et al.*, 2017).

Enfin, des technologies de séquençage de plus en plus performantes en matière de quantités d'informations produites, et sont utilisées de manière de plus en plus régulière. Par exemple, le dernier instrument de séquençage de la plateforme Illumina, le NovaSeq 6000 S4, disponible depuis 2017, permet d'obtenir au maximum 10 milliards de reads d'une longueur de 150 nt, ce qui représente 3000 Go de données (Genohub-ngs-instrument-guide). En

comparaison, lors de mon doctorat, j'ai analysé près de 120 millions de séquences de 250 à 300 nt de longueur, ce qui a représenté 70 Go de données. Les grandes quantités de données produites par les approches « omiques » posent le problème de leur traitement, de leur stockage à long terme, de leur partage et de leur communication (Vayssier-Taussat *et al.*, 2014).

Limites conceptuelles

En plus d'être liée aux limitations techniques citées précédemment, la métagénomique virale se heurte à des problèmes d'ordre conceptuel concernant notamment la classification virale, l'intégration des virus au sein de l'holobionte, ainsi que les interprétations écologiques ou thérapeutiques des données.

Métagénomique virale et classification taxonomique des virus

Comme expliqué en introduction, les différentes études menées en métagénomique virale ont permis la description d'une diversité virale très importante, auparavant insoupçonnée, dans une grande diversité d'environnements et d'êtres vivants (Brum et Sullivan, 2015; Chandler *et al.*, 2015; Fierer *et al.*, 2012; Houldcroft, 2015; Whon *et al.*, 2012). Une partie des séquences virales caractérisées lors de ces études a été déposée dans des bases de données, dont le contenu a augmenté de manière exponentielle ces dernières années (**Fig. 12**).

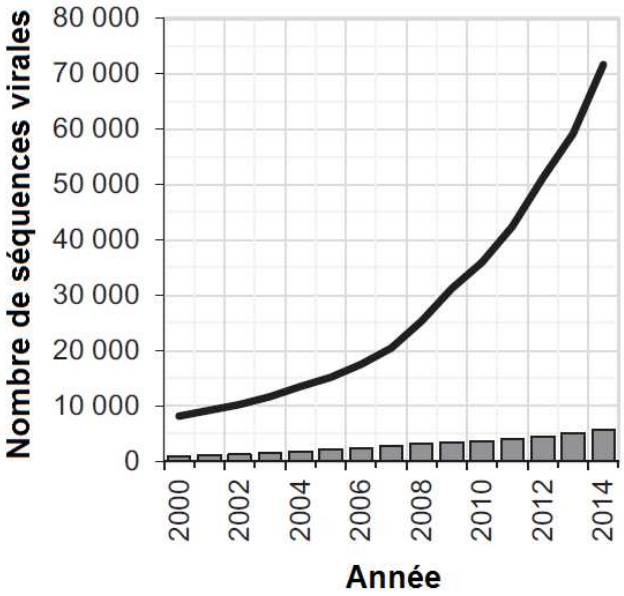


Figure 12 : Nombre de séquences virales et de viroïdes déposées dans la base de données « International Nucleotide Sequence Database Collaboration » (ligne noire), et dans la base de données RefSeq virus (GenBank) (histogramme). Les segments de virus de la grippe déposés sur la base de données « International Nucleotide Sequence Database Collaboration » n'ont pas été inclus dans ce graphique. Tirée de Brister, Ako-adjei, Bao, & Blinkova, 2015.

Afin d'intégrer les nombreux virus découverts ces dernières années au sein de la taxonomie virale (n'incluant actuellement que 4404 espèces) (ICTV, mai 2017), un assouplissement des conditions d'intégration a récemment été accepté. Elle repose à présent uniquement sur la caractérisation des séquences virales, la caractérisation phénotypique des virus devenant optionnelle (Simmonds *et al.*, 2017). Cette modification des critères d'acceptation permet de mieux appréhender la diversité des virus, notamment en prenant en compte les séquences virales possédant des similitudes très faibles avec celles actuellement présentes dans les bases de données.

Cependant, il est à noter que le concept d'espèce virale étant évolutif (i.e. soumis à une révision constante), cela pose des problèmes quant à l'attribution taxonomique des séquences virales. En effet, lors des affiliations taxonomiques, il est nécessaire d'imposer des seuils permettant de délimiter les taxas. La définition du seuil est cruciale si l'on veut interpréter la variabilité génomique des virus. Concernant les bactéries, un seuil de pourcentage de

similarité supérieur à 97% au niveau du gène codant pour l'ARN ribosomal 16S permet de délimiter des séquences appartenant à la même espèce (Thompson *et al.*, 2013). Or, l'ICTV a déterminé un seuil propre à la démarcation des espèces spécifique à chaque famille de virus. Ces changements de seuil de démarcation des familles virales sont expliqués par des caractéristiques biologiques propres aux virus appartenant aux familles concernées et au nombre de virus connus dans une famille donnée; cependant, ils compliquent l'étude de la diversité des virus à l'échelle de l'espèce. De plus, des phénomènes de recombinaison viables entre des virus ne partageant que 80% d'identité au niveau nucléotidique ont été reportés (Vuillaume *et al.*, 2011). Ces limitations illustrent le fait que la notion d'espèce pourrait être remplacée par la notion d'OTU chez les virus.

Intégration des virus à l'holobionte

Les technologies de séquençage à haut débit ont permis de changer l'échelle d'étude des virus ainsi que des microorganismes, passant d'une espèce en particulier à celui des communautés. Les nombreuses découvertes qui en résultèrent, notamment sur l'influence qu'ont certains microorganismes symbiotiques sur leurs hôtes, ont mené à la création du concept d'holobionte. L'holobionte est ici défini comme un métaorganisme, i.e. un organisme et l'ensemble des microorganismes qui lui sont associés ainsi que les virus (Guerrero *et al.*, 2013).

Au sein des différents membres de cet holobionte, il existe un continuum entre parasitisme et mutualisme obligatoire, certains virus ayant des impacts positifs sur la survie ou la reproduction de leurs hôtes de par leurs impacts négatifs sur d'autres organismes (Roossinck et Bazán, 2017). Par exemple, le bactériophage APSE-3 (*Podoviridae*), infectant la bactérie *Hamiltonella defensa*, encode une toxine qui participe à la défense de l'hôte de cette bactérie, le puceron vert du pois (*Acyrthosiphon pisum*, Hémiptère), contre le développement d'œufs de parasitoïdes (Oliver *et al.*, 2009). Autre exemple, le virus nommé D. coccinellae paralysis virus (*Iflaviridae*), stocké dans les oviductes d'une guêpe parasitoïde (*Dinocampus coccinellae*, Hyménoptère), est transmis à l'hôte de cette guêpe, la coccinelle maculée (*Coleomegilla maculata*, Coléoptère), lors de la ponte de l'œuf de la guêpe dans la coccinelle. Ce virus se réplique dans le système nerveux de la coccinelle, ce qui induit une modification du comportement de cette dernière lorsque la larve de la guêpe émerge de son corps et entame sa pupation. La coccinelle est alors paralysée par l'action du virus, et protège

la larve de la guêpe jusqu'à l'émergence de la guêpe adulte. Ensuite, la coccinelle récupère de ces symptômes, retournant à son état normal (Dheilly *et al.*, 2015).

Or, la majorité de la communauté scientifique considère encore les virus uniquement en tant qu'agents pathogènes ; leur rôle positif sur l'holobionte est donc sous-évalué (Roossinck et Bazán, 2017). De plus, les communautés de virologues et de microbiologistes sont encore cloisonnées. En conséquence, le nombre d'études incorporant l'analyse des communautés virales à celle des autres représentants de l'holobionte est extrêmement réduit, ce qui limite l'apport de connaissances concernant la place qu'occupent les virus au sein de l'holobionte. Ainsi, il est à prévoir que l'instauration de collaborations entre ces deux communautés de scientifiques permettra d'obtenir une vision holistique de l'holobionte.

De la description à la prédiction : diversité virale et écologie

La grande majorité des études menées en métagénomique virale se basent sur l'analyse d'un seul ou d'un faible nombre d'échantillons. Ces études descriptives ne permettent donc pas d'inférer des hypothèses ou des modèles explicatifs des variations de diversité et de composition des communautés virales observées entre les échantillons.

Récemment, des métaanalyses ont été menées sur des milliers de viromes dans le but de caractériser l'impact de la distance géographique ainsi que du type d'environnement sur la structuration des communautés virales (Brum *et al.*, 2015; Paez-Espino *et al.*, 2016). Ces analyses, portant en majorité sur des viromes océaniques, ont mis en évidence un impact important du type d'habitat, en comparaison de l'impact de la distance géographique, sur la structuration des communautés virales. D'autres études comparatives, menées à des échelles spatiales plus faibles, ont testé l'impact de la distance géographique ou temporelle sur la composition des communautés virales (Gustavsen *et al.*, 2014; Johannessen *et al.*, 2017).

Cependant, ce type d'étude est réalisé de manière parcellaire et est encore peu répandu en comparaison des études de phylogéographie virale menées à l'échelle intraspécifique qui permettent régulièrement de décrire la variabilité et la distribution de certains virus à l'échelle mondiale (Dudas *et al.*, 2017; Faria *et al.*, 2015; Global Consortium for H5N8 and Related Influenza Viruses., 2016). En outre, l'impact des facteurs biotiques (i.e. immunitaires, régime alimentaire, stress) sur la structuration des communautés virales est sous-étudié (Vayssier-

Taussat *et al.*, 2014). Enfin, afin de permettre le développement des métaanalyses de viromes, il est nécessaire de développer des méthodes d'analyses statistiques ainsi que de mettre en place des modèles mathématiques adaptés à leur étude comparative (Anthony *et al.*, 2015).

Métagénomique virale et diagnostic

La métagénomique virale a permis d'identifier des virus pathogènes, comme le virus de Merkel (*Polyomaviridae*) responsable d'un cancer cutané humain nommé « carcinome de Merkel » (Feng *et al.*, 2008). Le virus de Schmallenberg (*Orthobunyaviridae*), responsable de malformations et de mortalité de fœtus de bétail et transmis par des moucherons hématophages (*Culicoides sp.*, Diptère) a également été découvert par métagénomique virale (Hoffmann *et al.*, 2012).

Cependant, l'utilisation de la puissance des outils de métagénomique virale en santé humaine, animale et végétale reste essentiellement limitée – pour des raisons que nous verrons plus tard - à la surveillance de la circulation de virus pathogènes dans des animaux d'élevage (Blomström *et al.*, 2009; Munang'andu, 2016), des plantes cultivées (Idris *et al.*, 2014; Palanga *et al.*, 2016), des animaux vecteurs (Rosario *et al.*, 2016; Temmam *et al.*, 2014), des aliments (Nieuwenhuijse et Koopmans, 2017) ou d'espèces animales menacées d'extinction (Conceição-neto *et al.*, 2017). Ces études fournissent ainsi une stratégie prophylactique pour « surveiller » la circulation de virus pathogènes. Bien que l'utilisation des technologies de séquençage à haut débit permette actuellement la caractérisation de génomes viraux en près de 48h (Roossinck *et al.*, 2015), l'utilisation de la métagénomique virale dans le cadre du diagnostic est restreinte à l'étude descriptive de maladies pour lesquelles aucun agent causal n'est identifié. Ce type d'études passe par la comparaison entre les viromes d'individus asymptomatiques et d'individus malades (Fancello *et al.*, 2014; Li *et al.*, 2015; Zhang *et al.*, 2017).

Ces faits sont explicables par le fait que la plus grande limitation de la métagénomique virale repose sur le fait qu'elle permet seulement de mettre en évidence la présence de génomes viraux dans un échantillon donné, ce qui ne permet pas de déterminer leur spectre d'hôte ni si ils sont à l'origine des symptômes observés. Afin de compléter les résultats issus

de métagénomique il est actuellement nécessaire de procéder à une étape de vérification en retournant à des études de virologie classiques reposants sur l'isolation des virus et l'infection des hôtes potentiels (Soueidan *et al.*, 2014).

Pour ce faire, on revient aux postulats de Koch établis en1890 : (1) L'organisme soupçonné doit être présent dans tous les individus malades et absent des individus sains ; (2) L'organisme candidat doit être isolé et cultivé en laboratoire ; (3) L'organisme cultivé doit provoquer la maladie chez un individu sain ; (4) Le même organisme doit pouvoir être isolé de l'individu ainsi inoculé. Si toutes ces conditions sont remplies, on peut conclure que l'organisme étudié est responsable de la pathologie. Les postulats de Koch ont notamment permis d'identifier l'agent responsable de la tuberculose, et sont encore d'une grande utilité en médecine. En revanche, ils rencontrent assez vite leurs limites : (1) on ne peut actuellement pas cultiver tous les microorganismes ni tous les virus ; et (2) on commence à décrire des maladies où plusieurs microorganismes et/ou virus sont impliqués à la fois, et dans ce cas, c'est une analyse de corrélation qui pourra nous dire quels microorganismes pourraient être causatifs des maladies observées (Vayssier-Taussat *et al.*, 2014).

Une solution intermédiaire pour l'utilisation des résultats issus des disciplines « omiques » en diagnostic passe par une actualisation des postulats de Koch (Mokili *et al.*, 2012). Une de ces approches, décrite par Mokili *et al.*, repose sur la comparaison de marqueurs moléculaires, comme peuvent l'être les marqueurs métatranscriptomiques et/ou métaprotéomiques, entre individus sains et malades (**Fig. 13**). Une autre approche, décrite par Lipkin, consisterait à : (1) mettre en évidence l'agent pathogène dans les séquences ; (2) trouver le même agent dans des cellules malades, qu'il soit similaire à d'autres agents pathogènes connus, trouver des taux élevés d'anticorps indiquant une exposition récente et (3) prévenir la maladie par des médicaments, anticorps ou vaccins spécifiques (Lipkin, 2013). Le critère (3) ne s'appliquant pas aux phytopathogènes, des petits ARN interférents (siRNA) pourraient être utilisés en tant que preuve d'une réponse de défense par la plante (Roossinck, 2016).

En résumé, l'utilisation d'une approche actualisée des postulats de Koch ainsi que d'outils permettant une caractérisation non biaisée des viromes pourrait permettre la création d'un nouveau type de diagnostic des maladies infectieuses virales.

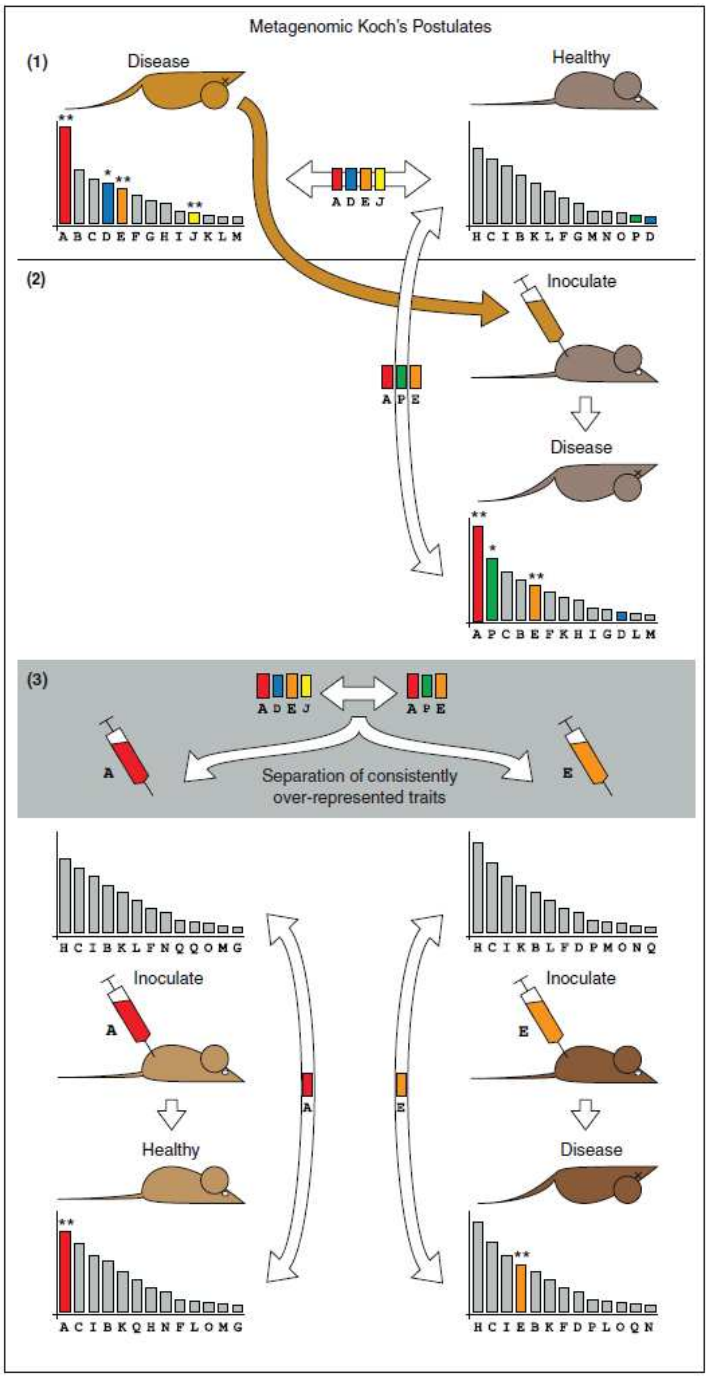


Figure 13 : Actualisation des postulats de Koch pour leur utilisation en métagénomique. La comparaison entre un animal sain et un animal malade indique une différence significative au sein des données obtenues (représentées par les histogrammes montrant par exemple l'abondance relative de reads). Pour vérifier les postulats de Koch en métagénomique, il faut (1) que des traits « omiques » du sujet malade soient significativement différents du sujet sain, (2) que l'inoculation d'échantillons provenant de l'animal malade à un animal sain induise la maladie, (3) que l'inoculation des traits purifiés dans un animal sain induise la maladie. Tirée de Mokili *et al.*, 2012.

Conclusion

La métagénomique appliquée à la virologie est une sous-discipline scientifique en plein essor qui a permis la découverte et la description d'une biodiversité virale extraordinaire et jusque-là largement sous-évaluée. Elle a entrainé un bouleversement de notre perception du rôle évolutif et écologique des virus. Nous sommes à l'ère du naturalisme moléculaire, passant par la réalisation d'inventaires génétiques d'espèces virales et microbiennes. Il en résulte que la majorité de ces études sont essentiellement descriptives. Or, ces travaux montrent surtout qu'il est extrêmement difficile de décrire l'ensemble de la diversité virale. En effet, cela demanderait un effort d'échantillonnage colossal afin de prendre en compte les virus les moins prévalents dans les hôtes ou les écosystèmes étudiés, ce qui demanderait des moyens financiers très importants. Par exemple, il a été estimé que chez les mammifères, regroupant près de 6 000 espèces, il resterait près de 320 000 espèces virales à découvrir (Anthony *et al.*, 2013). Si l'on accepte l'hypothèse que la diversité virale est proportionnelle à la diversité des hôtes, ce nombre donne une idée de la quantité de virus restant à découvrir chez les arthropodes possédant plus d'un million d'espèces. De plus, la description pure est d'un intérêt restreint pour notre compréhension holistique du rôle écologique et évolutif des virus, ainsi que pour des problématiques plus appliquées en épidémiologie et en diagnostic.

Il est à prévoir que les progrès qui seront réalisés dans l'organisation et l'analyse de la masse de données provenant du séquençage à haut débit permettront de dépasser un niveau encore majoritairement descriptif des résultats en leur donnant du sens dans le contexte des différents fronts de recherche. Les données qualitatives actuellement obtenues seront enrichies par des données quantitatives robustes basées sur les réplications d'analyses dans le temps et dans l'espace, et soumises à des analyses statistiques rigoureuses. Le développement de l'utilisation des technologies liées au séquençage haut débit, et plus particulièrement la métagénomique virale, permettraient ainsi leurs applications plus larges en évolution, en écologie, en épidémiologie, ou en diagnostic. De plus, la métagénomique virale a permis la découverte d'une importante diversité génétique qui pourrait être explorée dans le cadre de son utilisation en biotechnologie ou en médecine (par exemple, la rétrotranscriptase, utilisée en biologie moléculaire, est d'origine virale) (Rosario & Breitbart, 2011), ainsi que dans le cadre de la lutte biologique (Valles *et al.*, 2013). Enfin, ces travaux permettront de faire évoluer notre conception des virus et de leur place au sein de la biosphère. L'utilisation de ces nouveaux outils serait ainsi bénéfique pour la biologie toute entière.

Références bibliographiques

Aaskov, J., Buzacott, K., Lowry, K., Holmes, E., 2006. Long-term transmission of defective RNA viruses in humans and Aedes mosquitoes. Science. 311, 236–238.

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. 98, 11609–11614.

Altschul, S., Gish, W., Miller, W., Myers, E., Lipman, D., 1990. Basic local alignment search tool. J. Mol. Biol. 215, 403–410.

Anderson, P.K., Cunningham, A.A., Patel, N.G., Morales, F.J., Epstein, P.R., Daszak, P., 2004. Emerging infectious diseases of plants : pathogen pollution , climate change and agrotechnology drivers. Trends Ecol. Evol. 19, 535–544.

Anthony, S.J., Epstein, J.H., Murray, K.A., Navarrete-macias, I., Zambrana-torrelio, C.M., Solovyov, A., Ojeda-flores, R., Arrigo, N.C., Islam, A., Khan, A., Hosseini, P., Bogich, T.L., Mazet, J.A.K., Daszak, P., Lipkin, W., 2013. A strategy to estimate unknown viral diversity in mammals. MBio 4, 1–15.

Anthony, S.J., Islam, A., Johnson, C., Navarrete-Macias, I., Liang, E., Jain, K., Hitchens, P.L., Che, X., Soloyvov, A., Hicks, A.L., Ojeda-Flores, R., Zambrana-Torrelio, C., Ulrich, W., Rostal, M.K., Petrosov, A., Garcia, J., Haider, N., Wolfe, N., Goldstein, T., Morse, S.S., Rahman, M., Epstein, J.H., Mazet, J.K., Daszak, P., Lipkin, W.I., 2015. Non-random patterns in viral diversity. Nat. Commun. 6, 8147.

Appelt, S., Fancello, L., Bailly, L., Raoult, D., Drancourt, M., Desnues, C., 2014. Viruses in a 14th-Century Coprolite. Appl. Environ. Microbiol. 80, 2648–2655.

Arriagada, G., Gifford, R.J., 2014. Parvovirus-derived endogenous viral elements in two South American rodent genomes. J. Virol. 88, 12158–12162.

Barré-Sinoussi, F., Chermann, J., Rey, F., Nugeyre, M., Chamaret, S., Gruest, J., Dauguet, C., Axler-Blin, C., Vézinet-Brun, F., Rouzioux, C., Rozenbaum, W Montagnier, L., 1983. Isolation of a T-lymphotropic retrovirus from a patient at risk for acquired immune deficiency syndrome (AIDS). Science. 220, 868–871.

Belyi, V. a, Levine, A.J., Skalka, A.M., 2010. Sequences from ancestral single-stranded DNA viruses in vertebrate genomes: the parvoviridae and circoviridae are more than 40 to 50 million years old. J. Virol. 84, 12458–12462.

Berche, P., 2007. Une histoire des microbes, John Libbe. ed.

Bergh, O., Borsheim, K.Y., Bratbak, G., Heldal, M., 1989. High abundance of viruses found in aquatic environments. Nature 340, 467–468.

Bergoin, M., Tijssen, P., 2010. Densoviruses: A Highly Diverse Group of Arthropod Parvoviruses, in: Asgari, S., Johnson, K.N. (Eds.), Insect Virology.

Bettarel, Y., Bouvier, T., Bouvier, C., Carre, C., Desnues, A., Domaizon, I., Jacquet, S., Simengando, T., 2011. Ecological traits of planktonic viruses and prokaryotes along a full-salinity gradient. FEMS Microbiol. Ecol. 76, 360–372.

Blomström, A., Belák, S., Fossum, C., Mckillen, J., Allan, G., Wallgren, P., Berg, M., 2009. Detection of a novel porcine boca-like virus in the background of porcine circovirus type 2 induced postweaning multisystemic wasting syndrome. Virus Res. 146, 125–129.

Bovo, S., Mazzoni, G., Ribani, A., Utzeri, V.J., Bertolini, F., Schiavo, G., Fontanesi, L., 2017. A viral metagenomic approach on a non-metagenomic experiment: Mining next generation sequencing datasets from pig DNA identified several porcine parvoviruses for

a retrospective evaluation of viral infections. PLoS One 12, e0179462.

Breitbart, M., 2012. Marine Viruses: Truth or Dare. Ann. Rev. Mar. Sci. 4, 425–448.

Brister, J.R., Ako-adjei, D., Bao, Y., Blinkova, O., 2015. NCBI viral genomes resource. Nucleic Acids Res. 43, 571–577.

Bruder, K., Malki, K., Cooper, A., Sible, E., Shapiro, J.W., Watkins, S.C., Putonti, C., 2016. Freshwater Metaviromics and Bacteriophages : A Current Assessment of the State of the Art in Relation to Bioinformatic Challenges. Evol Bioinform Online 12, 25–33.

Brum, J.R., Ignacio-Espinoza, J.C., Kim, E.-H., Trubl, G., Jones, R.M., Roux, S., VerBerkmoes, N.C., Rich, V.I., Sullivan, M.B., 2016. Illuminating structural proteins in viral "dark matter" with metaproteomics. Proc. Natl. Acad. Sci. U. S. A. 113, 2436–2441.

Brum, J.R., Ignacio-espinoza, J.C., Roux, S., Doulcier, G., Acinas, S.G., Alberti, A., Chaffron, S., 2015. Ocean plankton. Patterns and ecological drivers of ocean viral communities. Science. 348, 1261498.

Brum, J.R., Sullivan, M.B., 2015. Rising to the challenge: accelerated pace of discovery transforms marine virology. Nat. Rev. Microbiol. 13, 147–159.

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.

Cantalupo, P.G., Calgua, B., Zhao, G., Hundesa, A., Wier, A.D., Katz, J.P., Grabe, M., Hendrix, R., Girones, R., Wang, D., Pipas, J., 2011. Raw Sewage Harbors Diverse Viral Populations. MBio 2, 1–11.

Chandler, J.A., Liu, R.M., Bennett, S.N., 2015. RNA shotgun metagenomic sequencing of northern California (USA) mosquitoes uncovers viruses, bacteria, and fungi. Front. Microbiol. 6:185.

Chapman, A.D., 2009. Numbers of Living Species in Australia and the World. Rep. Aust. Biol. Resour.

Charuvaka, A., Rangwala, H., 2011. Evaluation of short read metagenomic assembly. BMC Genomics 12, S8.

Chiu, C.Y., 2013. Viral pathogen discovery. Curr. Opin. Microbiol. 16, 468–78.

Claverie, J., Ogata, H., 2009. Ten good reasons not to exclude giruses from the evolutionary picture. Nat. Rev. Microbiol. 7, 615.

Clavijo, G., van Munster, M., Monsion, B., Bochet, N., Brault, V., 2016. Transcription of densovirus endogenous sequences in Myzus persicae genome. J. Gen. Virol. 97, 1000–1009.

Coffey, L.L., Page, B.L., Greninger, A.L., Herring, B.L., Russell, R.C., Doggett, S.L., Haniotis, J., Wang, C., Deng, X., Delwart, E.L., 2014. Enhanced arbovirus surveillance with deep sequencing: Identification of novel rhabdoviruses and bunyaviruses in Australian mosquitoes. Virology 448, 146–158.

Conceição-neto, N., Godinho, R., Alvares, F., Yinda, C.K., Deboutte, W., Zeller, M., Laened, L., Heylen, E., Roque, S., Petrucci-Foncesa, F., Santos, N., Van Ranst, M., Mesquita, J.R., Mattijnssens, J., 2017. Viral gut metagenomics of sympatric wild and domestic canids , and monitoring of viruses : Insights from an endangered wolf population. *Ecol*

Evol **7,** 4135–4146.

Conceição-neto, N., Zeller, M., Lefrère, H., Bruyn, P. De, Beller, L., Deboutte, W., Yinda, C., Lavigne, R., Maes, P., Van Ranst, M., Heylen, E., Matthijnssens, J., 2015. Modular approach to customise sample preparation procedures for viral metagenomics: a reproducible protocol for virome analysis. Sci. Rep. 5, 16532.

Corinaldesi, C., Tangherlini, M., Dell Anno, A., 2017. From virus isolation to metagenome generation for investigating viral diversity in deep-sea sediments. Sci. Rep. 7, 8355.

Cotmore, S.F., Agbandje-McKenna, M., Chiorini, J. a, Mukha, D. V, Pintel, D.J., Qiu, J., Soderlund-Venermo, M., Tattersall, P., Tijssen, P., Gatherer, D., Davison, A.J., 2014. The family Parvoviridae. Arch. Virol. 159, 1239–1247.

Cotmore, S.F., Tattersall, P., 2014. Parvoviruses: Small Does Not Mean Simple. Annu. Rev. Virol. 1, 517–37.

Danovaro, R., Dell'Anno, A., Corinaldesi, C., Magagnini, M., Noble, R., Tamburini, C., Weinbauer, M., 2008. Major viral impact on the functioning of benthic deep-sea ecosystems. Nature 454, 1084–1087.

Davis, B., Waldor, M.K., 2003. Filamentous phages linked to virulence of Vibrio cholerae. Curr. Opin. Microbiol. 6, 35–42.

Dayaram, A., Potter, K.A., Moline, A.B., Rosenstein, D.D., Marinov, M., Thomas, J.E., Breitbart, M., Rosario, K., Argu, G.R., 2013. High global diversity of cycloviruses amongst dragonflies. J. Gen. Virol. 94, 1827–1840.

DeBoever, C., Reid, E.G., Smith, E.N., Wang, X., Dumaop, W., Harismendy, O., Carson, D., Richman, D., Masliah, E., Frazer, K. a, 2013. Whole transcriptome sequencing enables discovery and analysis of viruses in archived primary central nervous system lymphomas. PLoS One 8, e73956.

Degnan, P.H., Ochman, H., 2011. Illumina-based analysis of microbial community diversity. ISME J. 6, 183–194.

Delwart, E., 2013. Animal virus discovery: improving animal health, understanding zoonoses, and opportunities for vaccine development. Curr. Opin. Virol. 2, 344–352.

Delwart, E.L., 2007. Viral metagenomics. Rev Med Virol. 17, 115–131.

Dheilly, N.M., Maure, F., Ravallec, M., Galinier, R., Doyon, J., Duval, D., Leger, L., Volkoff, A.-N., Missé, D., Nidelet, S., Demolombe, V., Brodeur, J., Gourbal, B., Thomas, F., Mitta, G., 2015. Who is the puppet master? Replication of a parasitic wasp-associated virus correlates with host behaviour manipulation. Proc. R. Soc. London Ser. B Biol. Sci. 282, 1–10.

Domingo-Calap, P., Cuevas, M., Sanjua, R., 2009. The fitness effects of random mutations in single-stranded DNA and RNA bacteriophages. PLoS Genet. 5, e1000742.

Dudas, G., Carvalho, L., Bedford, T., Tatem, A., Baele, G., Faria, N., Park, D., Ladner, J., Arias, A., Asogun, D., Bielejec, F., Caddy, S., Cotten, M., D'Ambrozio, J., Dellicour, S., Di Caro, A., Diclaro, J., Duraffour, S., Elmore, M., Fakoli, L., Faye, O., Gilbert, M., Gevao, S., Gire, S., Gladden-Young, A., Gnirke, A., Goba, A., Grant, D., Haagmans, B., Hiscox, J., Jah, U., Kugelman, J., Liu, D., Lu, J., Malboeuf, C., Mate, S., Matthews, D., Matranga, C., Meredith, L., Qu, J., Quick, J., Pas, S., Phan, M., Pollakis, G., Reusken, C., Sanchez-Lockhart, M., Schaffner, S., Schieffelin, J., Sealfon, R., Simon-Loriere, E., Smits, S., Stoecker, K., Thorne, L., Tobin, E., Vandi, M., Watson, S., West, K., Whitmer, S., Wiley, M., Winnicki, S., Wohl, S., Wölfel, R., Yozwiak, N., Andersen, K.,

Blyden, S., Bolay, F., Carroll, M., Dahn, B., Diallo, B., Formenty, P., Fraser, C., Gao, G., Garry, R., Goodfellow, I., Günther, S., Happi, C., Holmes, E., Kargbo, B., Keïta, S., Kellam, P., Koopmans, M., Kuhn, J., Loman, N., Magassouba, N., Naidoo, D., Nichol, S., Nyenswah, T., Palacios, G., Pybus, O., Sabeti, P., Sall, A., Ströher, U., Wurie, I., Suchard, M., Lemey, P., Rambaut, A., 2017. Virus genomes reveal factors that spread and sustained the Ebola epidemic. Nature 544, 309–315.

Duffy, S., Shackelton, L. a, Holmes, E.C., 2008. Rates of evolutionary change in viruses: patterns and determinants. Nat. Rev. Genet. 9, 267–76.

Duffy, S., Turner, P.E., Burch, C.L., 2006. Pleiotropic costs of niche expansion in the RNA bacteriophage phi 6. Genetics 172, 751–757.

Dupressoir, A., Lavialle, C., Heidmann, T., 2012. From ancestral infectious retroviruses to bona fide cellular genes: role of the captured syncytins in placentation. Placenta 33, 663–671.

Edgar, R.C., 2010. Search and clustering orders of magnitude faster than BLAST. Bioinformatics 26, 2460–2461.

Edwards, R.A., Rohwer, F., 2005. Opinion: Viral metagenomics. Nat. Rev. Microbiol. 3, 504–510.

El-Far, M., Szelei, J., Yu, Q., Fédière, G., Bergoin, M., Tijssen, P., 2012. Organization of the ambisense genome of the Helicoverpa armigera Densovirus. J. Virol. 86, 7024.

Fancello, L., Monteil, S., Popgeorgiev, N., Rivet, R., Fournier, P., Raoult, D., Desnues, C., 2014. Viral communities associated with human pericardial fluids in idiopathic pericarditis. PLoS One 9, e93367.

Fancello, L., Raoult, D., Desnues, C., 2012. Computational tools for viral metagenomics and their application in clinical research. Virology 434, 162–74.

Faria, N.R., Rambaut, A., Suchard, M.A., Baele, G., Bedford, T., Ward, M.J., Tatem, A.J., Sousa, J.D., Arinaminpathy, N., Pépin, J., Posada, D., Peeters, M., Pybus, O.G., Lemey, P., 2015. HIV epidemiology. The early spread and epidemic ignition of HIV-1 in human populations. Science. 346, 56–61.

Fei, T., Ng, F., Chen, L., Zhou, Y., Shapiro, B., Stiller, M., Heintzman, P.D., 2014. Preservation of viral genomes in 700-y-old caribou feces from a subarctic ice patch. Proc. Natl. Acad. Sci. 111, 16842–16847.

Feng, H., Shuda, M., Chang, Y., Moore, P.S., 2008. Clonal Integration of a Polyomavirus in Human Merkel Cell Carcinoma. Science. 319, 1096–1100.

Feschotte, C., Gilbert, C., 2012. Endogenous viruses: insights into viral evolution and impact on host biology. Nat. Rev. Genet. 13, 283–296.

Fierer, N., Leff, J.W., Adams, B.J., Nielsen, U.N., Thomas, S., Lauber, C.L., Owens, S., Gilbert, J.A., Wall, D.H., Caporaso, J.G., 2012. Cross-biome metagenomic analyses of soil microbial communities and their functional attributes. Proc. Natl. Acad. Sci. 109, 21390–21395.

François, S., Bernardo, P., Filloux, D., Roumagnac, P., Yaverkovski, N., Froissart, R., Ogliastro, M., 2014. A Novel Itera-Like Densovirus Isolated by Viral Metagenomics from the Sea Barley Hordeum marinum 2, 13–14.

Frey, K., Biser, T., Hamilton, T., Santos, C., Pimentel, G., Mokashi, V., Bishop-Lilly, K., 2016. Bioinformatic Characterization of Mosquito Viromes within the Eastern United States and Puerto Rico: Discovery of Novel Viruses. Evol Bioinform Online 12, 1–12.

Gall, A., 2015. Bugs full of viruses. Nat. Rev. Microbiol. 13, 253.

García-López, R., Vázquez-Castellanos, J.F., Moya, A., 2015. Fragmentation and Coverage Variation in Viral Metagenome Assemblies, and Their Effect in Diversity Calculations. Front. Bioeng. Biotechnol. 3, 141.

Ge, X., Li, Y., Yang, X., Zhang, H., Zhou, P., Zhang, Y., Shi, Z., 2012. Metagenomic analysis of viruses from bat fecal samples reveals many novel viruses in insectivorous bats in China. J. Virol. 86, 4620–4630.

Gilbert, C., Chateigner, A., Ernenwein, L., Barbe, V., Bézier, A., Herniou, E.A., Cordaux, R., 2014. Population genomics supports baculoviruses as vectors of horizontal transfer of insect transposons. Nat. Commun. 5.

Global Consortium for H5N8 and Related Influenza Viruses., 2016. Role for migratory wild birds in the global spread of avian influenza H5N8. Science. 354, 213–217.

Granberg, F., Vicente-Rubiano, M., Rubio-Guerri, C., Karlsson, O.E., Kukielka, D., Belák, S., Sánchez-Vizcaíno, J.M., 2013. Metagenomic detection of viral pathogens in Spanish honeybees: co-infection by Aphid Lethal Paralysis, Israel Acute Paralysis and Lake Sinai Viruses. PLoS One 8, e57459.

Greathead, D.., 1995. Benefits and risks of classical biological control.

Gudenkauf, B.M., Eaglesham, J.B., Aragundi, W.M., Hewson, I., 2014. Discovery of urchin-associated densoviruses (Parvoviridae) in coastal waters of the Big Island, Hawaii. J. Gen. Virol. 95, 652–658.

Guerrero, R., Margulis, L., Berlanga, M., 2013. Symbiogenesis: The holobiont as a unit of evolution. Int. Microbiol. 16, 133–143.

Gustavsen, J. a, Winget, D.M., Tian, X., Suttle, C. a, 2014. High temporal and spatial diversity in marine RNA viruses implies that they have an important role in mortality and structuring plankton communities. Front. Microbiol. 5.

Hadidi, A., Flores, R., Candresse, T., Barba, M., 2016. Next-Generation Sequencing and Genome Editing in Plant Virology 7, 1325.

Halary, S., Temmam, S., Raoult, D., Desnues, C., 2016. Viral metagenomics: are we missing the giants? Curr. Opin. Microbiol. 31, 34–43.

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. J. Virol. Methods 195, 194–204.

Hamady, M., Knight, R., 2009. Microbial community profiling for human microbiome projects : Tools , techniques , and challenges. Genome Res. 19, 1141–1152.

Han, L., Yu, D., Zhang, L., Shen, J., He, J., 2017. Genetic and functional diversity of ubiquitous DNA viruses in selected Chinese agricultural soils. Sci. Rep. 7.

Handelsmanl, J., Rondon, M.R., Goodman, R.M., Brady, S.F., Clardy, J., 1998. Molecular biological access to the chemistry of unknown soil microbes: a new frontier for natural products. Chem. Biol. 5, 245–249.

Harris, J.K., Sahl, J.W., Castoe, T.A., Wagner, B.D., Pollock, D.D., Spear, J.R., 2010. Comparison of normalization methods for construction of large, multiplex amplicon pools for next-generation sequencing. Appl. Environ. Microbiol. 76, 3863–3868.

Hayes, S., Mahony, J., Nauta, A., Van Sinderen, D., 2017. Metagenomic approaches to assess

bacteriophages in various environmental niches. Viruses 9, 1–22.

He, B., Li, Z., Yang, F., Zheng, J., Feng, Y., Guo, H., Li, Y., Wang, Y., Su, N., Zhang, F., Fan, Q., Tu, C., 2013. Virome profiling of bats from Myanmar by metagenomic analysis of tissue samples reveals more novel Mammalian viruses. PLoS One 8, e61950.

Heinz, F.X., Stiasny, K., Holzmann, H., Kundi, M., Six, W., Wenk, M., Kainz, W., Essl, A., Kunz, C., 2015. Emergence of tick-borne encephalitis in new endemic areas in Austria: 42 years of surveillance. Eurosurveillance 20, 16–19.

Hewson, I., Button, J.B., Gudenkauf, B.M., Miner, B., Newton, A.L., Gaydos, J.K., Wynne, J., Groves, C.L., Hendler, G., Murray, M., Fradkin, S., Breitbart, M., Fahsbender, E., Lafferty, K.D., Kilpatrick, a M., Miner, C.M., Raimondi, P., Lahner, L., Friedman, C.S., Daniels, S., Haulena, M., Marliave, J., Burge, C. a, Eisenlord, M.E., Harvell, C.D., 2014. Densovirus associated with sea-star wasting disease and mass mortality. Proc. Natl. Acad. Sci. U. S. A. 111, 17278–17283.

Hirsch, M.S., Aquila, R.T.D., Demeter, L.M., Hammer, S.M., Johnson, V.A., Loveday, C., Mellors, J.W., Jacobsen, D.M., Richman, D.D., 2003. Antiretroviral drug resistance testing in adults infected with human immunodeficiency virus type 1: 2003 recommendations of an International AIDS Society-USA Panel. Clin. Infect. Dis. 37, 113–128.

Hoffmann, B., Scheuch, M., Höper, D., Jungblut, R., Holsteg, M., Schirrmeier, H., Eschbaumer, M., Goller, K., Wernike, K., Fischer, M., Breithaupt, A., Mettenleiter, T., Beer, M., 2012. Novel orthobunyavirus in Cattle, Europe, 2011. Emerg. Infect. Dis. 18, 469–472.

Holmes, E.C., 2009. The origins of RNA viruses, in: The Evolution and Emergence of RNA Viruses. pp. 15–28.

Horie, M., Tomonaga, K., 2011. Non-retroviral fossils in vertebrate genomes. Viruses 3, 1836–1848.

Horner-Devine, M.C., Bohannan, B.J.M., 2006. Phylogenetic clustering and overdispersion in bacterial communities. Ecology 87, 100–108.

Horner-Devine, M.C., Carney, K.M., Bohannan, B.J.M., 2004. An ecological perspective on bacterial biodiversity. Proc. R. Soc. B Biol. Sci. 271, 113–122.

Houldcroft, C.J., 2015. Tales from the crypt and coral reef : the successes and challenges of identifying new herpesviruses using metagenomics. Front. Microbiol. 6.

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, 1219–1236.

Jeppson, L.R., Keifer, H.H., Baker, E.W., 1975. Mites Injurious to Economic Plants, Univ. Cali. ed. University of California Press.

Johannessen, T.V., Larsen, A., Bratbak, G., Edvardsen, B., Egge, E.D., Sandaa, R.-A., 2017. Seasonal Dynamics of Haptophytes and dsDNA Algal Viruses Suggest Complex Virus-Host Relationship. Viruses 9, 84.

Jones, K.E., Patel, N.G., Levy, M.A., Storeygard, A., Balk, D., Gittleman, J.L., Daszak, P., 2008. Global trends in emerging infectious diseases. Nature 451, 990–993.

Junglen, S., Drosten, C., 2013. Virus discovery and recent insights into virus diversity in arthropods. Curr. Opin. Microbiol. 16, 507–513.

Kapoor, A., Simmonds, P., Lipkin, W.I., 2010. Discovery and characterization of mammalian endogenous parvoviruses. J. Virol. 84, 12628–12635.

Kaufmann, B., Bowman, V.D., Li, Y., Szelei, J., Waddell, P.J., Tijssen, P., Rossmann, M.G., 2010. Structure of Penaeus stylirostris densovirus, a shrimp pathogen. J. Virol. 84, 11289–96.

Kew, O.M., Sutter, R.W., Gourville, E.M. De, Dowdle, W.R., Pallansch, M.A., 2005. Vaccine-derived polioviruses and the endgame strategy for global polio eradication. Annu. Rev. Microbiol. 59, 587–635.

Kim, K., Bae, J., 2011. Amplification methods bias metagenomic libraries of uncultured single-stranded and double-stranded DNA viruses. Appl. Environ. Microbiol. 77, 7663–7668.

Koonin, E. V, Dolja, V. V, Krupovic, M., 2015. Origins and evolution of viruses of eukaryotes: The ultimate modularity. Virology 479–480, 2–25.

Koonin, E. V, Wolf, Y.I., 2012. Evolution of microbes and viruses : a paradigm shift in evolutionary biology ? Front. Microbiol. 2.

Kozarewa, I., Armisen, J., Gardner, A.F., Slatko, B.E., Hendrickson, C.L., 2015. Overview of target enrichment strategies. Curr. Protoc. Mol. Biol. 112, 1–23.

Krishnamurthy, S.R., Wang, D., 2017. Origins and challenges of viral dark matter. Virus Res. 239, 136–142.

Kristensen, D.M., Mushegian, A.R., Dolja, V. V., Koonin, E. V., 2010. New dimensions of the virus world discovered through metagenomics. Trends Microbiol. 18, 11–19.

Krupovic, M., Zhi, N., Li, J., Hu, G., Koonin, E. V, Wong, S., Shevchenko, S., Zhao, K., Young, N.S., 2015. Multiple Layers of Chimerism in a Single-Stranded DNA Virus Discovered by Deep Sequencing. Genome Biol. Evol. 7, 993–1001.

Kuhn, J., Jahrling, P.B., 2010. Clarification and guidance on the proper usage of virus and virus species names. 155, 445–453.

La Scola, B., Desnues, C., Pagnier, I., Robert, C., Barrassi, L., Fournous, G., Merchat, M., Suzan-Monti, M., Forterre, P., Koonin, E., Raoult, D., 2008. The virophage as a unique parasite of the giant mimivirus. Nature 455, 100–104.

Labonté, J.M., Suttle, C. a, 2013. Previously unknown and highly divergent ssDNA viruses populate the oceans. ISME J. 7, 2169–2177.

Lacey, L.., Frutos, R., Kaya, H.., Vail, P., 2001. Insect Pathogens as Biological Control Agents: Do They Have a Future? Biol. Control 21, 230–248.

Lacey, L.A., Grzywacz, D., Shapiro-ilan, D.I., Frutos, R., Brownbridge, M., Goettel, M.S., 2015. Insect pathogens as biological control agents : Back to the future. J. Invertebr. Pathol. 132, 1–41.

Lang, A.S., Zhaxybayeva, O., Beatty, J.T., 2012. Gene transfer agents: phage-like elements of genetic exchange. Nat Rev Microbiol 10, 472–482.

Lee, J.C., Bruck, D.J., Dreves, A.J., Ioriatti, C., Vogt, H., Baufeld, P., 2011. In Focus : Spotted wing drosophila , Drosophila suzukii , across perspectives. Pest Manag Sci 67, 1349–1351.

Lefeuvre, P., Lett, J.-M., Varsani, A., Martin, D.P., 2009. Widely conserved recombination patterns among single-stranded DNA viruses. J. Virol. 83, 2697–2707.

Legendre, M., Bartoli, J., Shmakova, L., Jeudy, S., Labadie, K., Adrait, A., Lescot, M.,

POirot, O., Bertaux, L., Bruley, C., Couté, Y., Rivkina, E., Abergel, C., Claverie, J., 2014. Thirty-thousand-year-old distant relative of giant icosahedral DNA viruses with a pandoravirus morphology. Proc. Natl. Acad. Sci. 111, 4274–4279.

Legendre, M., Doutre, G., Poirot, O., Lescot, M., Arslan, D., Seltzer, V., Bertaux, L., Bruley, C., Claverie, J., Abergel, C., 2013. Pandoraviruses: amoeba viruses with genomes up to 2.5 Mb reaching that of parasitic eukaryotes. Science. 341, 281–286.

Levin, S., Sela, N., Chejanovsky, N., 2016. Two novel viruses associated with the Apis mellifera pathogenic mite Varroa destructor. Sci. Rep. 24, 6:37710.

Lewandowska, D.W., Zagordi, O., Geissberger, F., Kufner, V., Schmutz, S., Böni, J., Metzner, K.J., Trkola, A., Huber, M., 2017. Optimization and validation of sample preparation for metagenomic sequencing of viruses in clinical samples. Microbiome 5, 94.

Li, C.-X., Shi, M., Tian, J.-H., Lin, X.-D., Kang, Y.-J., Chen, L.-J., Qin, X.-C., Xu, J., Holmes, E.C., Zhang, Y.-Z., 2015. Unprecedented genomic diversity of RNA viruses in arthropods reveals the ancestry of negative-sense RNA viruses. Elife 4, 1–26.

Li, C., Hatta, M., Nidom, C.A., Muramoto, Y., Watanabe, S., Neumann, G., Kawaoka, Y., 2010. Reassortment between avian H5N1 and human H3N2 influenza viruses creates hybrid viruses with substantial virulence. Proc. Natl. Acad. Sci. 107, 4687–4692.

Li, L., Giannitti, F., Low, J., Keyes, C., Ullmann, S., Deng, X., Aleman, M., Pesavento, P.A., Pusterla, N., Delwart, E., 2015. Exploring the virome of diseased horses. J. Gen. Microbiol. 96, 2721–2733.

Li, L., Victoria, J.G., Wang, C., Jones, M., Fellers, G.M., Kunz, T.H., Delwart, E., 2010. Bat guano virome: predominance of dietary viruses from insects and plants plus novel mammalian viruses. J. Virol. 84, 6955–6965.

Li, M., Zheng, Y., Zhao, G., Fu, S., Wang, D., Wang, Z., Liang, G., 2014. Tibet Orbivirus, a novel Orbivirus species isolated from Anopheles maculatus mosquitoes in Tibet, China. PLoS One 9, e88738.

Lin, J., Kramna, L., Autio, R., Hyöty, H., Nykter, M., Cinek, O., 2017. Vipie : web pipeline for parallel characterization of viral populations from multiple NGS samples. BMC Genomics 18, 1–11.

Lipkin, W.I., 2013. The changing face of pathogen discovery and surveillance. Nat Rev Microbiol 11, 133–141.

Liu, H., Fu, Y., Xie, J., Cheng, J., Ghabrial, S. a, Li, G., Peng, Y., Yi, X., Jiang, D., 2011. Widespread endogenization of densoviruses and parvoviruses in animal and human genomes. J. Virol. 85, 9863–76.

Lukan, M., Bullova, E., Petko, B., 2010. Climate warming and tick-borne encephalitis, Slovakia. Emerg. Infect. Dis. 16, 524–526.

Lwoff, A., 1957. The Concept of Virus. J. Gen. Microbiol. 17, 239–253.

M, B., Delwart, E., Rosario, K., Segalés, J., Varsani, A., Consortium., I.R., 2017. ICTV Virus Taxonomy Profile : Circoviridae. J. Gen. Virol. 98, 1997–1998.

Ma, M., Huang, Y., Gong, Z., Zhuang, L., Li, C., Yang, H., Tong, Y., Liu, W., Cao, W., 2011. Discovery of DNA viruses in wild-caught mosquitoes using small RNA high throughput sequencing. PLoS One 6, e24758.

Martin, D.P., Biagini, P., Lefeuvre, P., Golden, M., Roumagnac, P., Varsani, A., 2011.

Recombination in eukaryotic single stranded DNA viruses. Viruses 3, 1699–1738.

McElroy, K., Thomas, T., Luciani, F., 2014. Deep sequencing of evolving pathogen populations: applications, errors, and bioinformatic solutions. Microb. Inform. Exp. 4.

McGraw, E. a, O'Neill, S.L., 2013. Beyond insecticides: new thinking on an ancient problem. Nat. Rev. Microbiol. 11, 181–193.

Metegnier, G., Becking, T., Chebbi, M.A., Giraud, I., Moumen, B., Schaack, S., Cordaux, R., Gilbert, C., 2015. Comparative paleovirological analysis of crustaceans identifies multiple widespread viral groups. Mob. DNA 6.

Meyer, M., Kircher, M., 2010. Illumina Sequencing Library Preparation for Highly Multiplexed Target Capture and Sequencing. Cold Spring Harb Protoc. 6.

Misof, B., Liu, S., Meusemann, K., Peters, R., Donath, A., Mayer, C., Frandsen, P., Ware, J., Flouri, T., Beutel, R.G., Niehuis, O., Petersen, M., Izquierdo-Carrasco, F., Wappler, T., Rust, J., Aberer, A., Aspöck, U., Aspöck, H., Bartel, D., Blanke, A., Berger, S., Böhm, A., Buckley, T., Calcott, B., Chen, J., Friedrich, F., Fukui, M., Fujita, M., Greve, C., Grobe, P., Gu, S., Huang, Y., Jermiin, L., Kawahara, A., Krogmann, L., Kubiak, M., Lanfear, R., Letsch, H., Li, Y., Li, Z., Li, J., Lu, H., Machida, R., Mashimo, Y., Kapli, P., McKenna, D., Meng, G., Nakagaki, Y., Navarrete-Heredia, J., Ott, M., Ou, Y., Pass, G., Podsiadlowski, L., Pohl, H., von Reumont, B., Schütte, K., Sekiya, K., Shimizu, S., Slipinski, A., Stamatakis, A., Song, W., Su, X., Szucsich, N., Tan, M., Tan, X., Tang, M., Tang, Timelthaler, G., Tomizuka, S., Trautwein, M., Tong, X., Uchifune, T., Walzl, M., Wiegmann, B., Wilbrandt, J., Wipfler, B., Wong, T., Wu, Q., Wu, G., Xie, Y., Yang, S., Yang, Q., Yeates, D., Yoshizawa, K., Zhang, Q., Zhang, R., Zhang, W., Zhang, Y., Zhao, J., Zhou, C., Zhou, L., Ziesmann, T., Zou, S., Li, Y., Xu, X., Zhang, Y., Yang, H., Wang, J., Wang, J., Kjer, K., Zhou, X., 2014. Phylogenomics resolves the timing and pattern of insect evolution. Science. 346, 763–767.

Mokili, J.L., Rohwer, F., Dutilh, B.E., 2012. Metagenomics and future perspectives in virus discovery. Curr. Opin. Virol. 2, 63–77.

Monjane, A.L., Martin, D.P., Lakay, F., Muhire, B.M., Pande, D., Varsani, A., Harkins, G., Shepherd, D., Rybicki, E., 2014. Extensive recombination-induced disruption of genetic interactions is highly deleterious but can be partially reversed by small numbers of secondary recombination events. J. Virol. 88, 7843–7851.

Mora, C., Tittensor, D.P., Adl, S., Simpson, A.G.B., Worm, B., 2011. How many species are there on Earth and in the ocean? PLoS Biol. 9, e1001127.

Moutailler, S., Popovici, I., Devillers, E., Eloit, M., 2016. Diversity of viruses in Ixodes ricinus, and characterization of a neurotropic strain of Eyach virus. New Microbes New Infect. 11, 71–81.

Mukha, D. V, Schal, K., 2003. A Densovirus of German Cockroach Blattella germanica : Detection , Nucleotide Sequence, and Genome Organization 37, 513–523.

Munang'andu, H.M., 2016. Environmental viral metagenomics analyses in aquaculture: Applications in epidemiology and disease control. Front. Microbiol. 7, 1–10.

Naccache, S.N., Greninger, A.L., Lee, D., Coffey, L.L., Phan, T., Rein-Weston, A., Aronsohn, A., Hackett, J., Delwart, E.L., Chiu, C.Y., 2013. The perils of pathogen discovery: origin of a novel parvovirus-like hybrid genome traced to nucleic acid extraction spin columns. J. Virol. 87, 11966–77.

Nasir, A., Kim, K.M., Caetan-Anolles, G., 2017. Insights & Perspectives Long-term evolution of viruses: A Janus-faced balance. BioEssays 1700026, 1–7.

Ng, T., 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. J. Virol. 86, 12161–12175.

Ng, T.F., Alavandi, S., Varsani, A., Burghart, S., Breitbart, M., 2013. Metagenomic identification of a nodavirus and a circular ssDNA virus in semi-purified viral nucleic acids from the hepatopancreas of healthy Farfantepenaeus duorarum shrimp. Dis. Aquat. Organ. 105, 237–242.

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., Rohwer, F., Breitbart, M., 2011b. Broad surveys of DNA viral diversity obtained through viral metagenomics of mosquitoes. PLoS One 6, e20579.

Nieuwenhuijse, D.F., Koopmans, M.P.G., 2017. Metagenomic Sequencing for Surveillance of Food- and Waterborne Viral Diseases. Front. Microbiol. 8, 1–11.

OILB-SROP, 1973. Statuts. Bull. SROP. 1.

Oliver, K., Degnan, P., Hunter, M., Moran, N., 2009. Bacteriophages encode factors required for protection in a symbiotic mutualism. Science. 325, 992–994.

Ostfeld, R.S., Brunner, J.L., 2015. Climate change and Ixodes tick-borne diseases of humans. Philos. Trans. R. Soc. B Biol. Sci. 370, 20140051–20140051.

Oulas, A., Pavloudi, C., Polymenakou, P., Pavlopoulos, G.A., Papanikolaou, N., Kotoulas, G., Arvanitidis, C., Iliopoulos, I., 2015. Metagenomics: tools and insights for analyzing next-generation sequencing data derived from biodiversity studies. Bioinform Biol Insights 9, 75–88.

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

Paez-espino, D., Pavlopoulos, G.A., Ivanova, N.N., Kyrpides, N.C., 2017. Nontargeted virus sequence discovery pipeline and virus clustering for metagenomic data. Nat. Protoc. 12, 1673–1682.

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

Radford, A.D., Chapman, D., Dixon, L., Chantrey, J., Darby, A.C., Hall, N., 2012. Application of next-generation sequencing technologies in virology. J. Gen. Virol. 93, 1853–1868.

Raoult, D., Forterre, P., 2008. Redefining viruses: lessons from Mimivirus. Nat. Rev. Microbiol. 6, 315–319.

Rascovan, N., Duraisamy, R., Desnues, C., 2016. Metagenomics and the Human Virome in Asymptomatic Individuals. Annu. Rev. Microbiol. 70, 125–141.

Remmert, M., Biegert, A., Hauser, A., Soding, J., 2012. HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment. Nat Meth 9, 173–175.

Ren, J., Ahlgren, N.A., Lu, Y.Y., Fuhrman, J.A., Sun, F., 2017. VirFinder : a novel k -mer

based tool for identifying viral sequences from assembled metagenomic data. Microbiome 5, 69.

Rinke, C., Schwientek, P., Sczyrba, A., Ivanova, N.N., Anderson, I.J., Cheng, J.-F., Darling, A., Malfatti, S., Swan, B.K., Gies, E. a, Dodsworth, J. a, Hedlund, B.P., Tsiamis, G., Sievert, S.M., Liu, W.-T., Eisen, J. a, Hallam, S.J., Kyrpides, N.C., Stepanauskas, R., Rubin, E.M., Hugenholtz, P., Woyke, T., 2013. Insights into the phylogeny and coding potential of microbial dark matter. Nature 499, 431–437.

Rodriguez-Brito, B., Li, L., Wegley, L., Furlan, M., Angly, F., Breitbart, M., Buchanan, J., Desnues, C., Dinsdale, E., Rayhawk, S., Rodriguez-mueller, J., Rodriguez-valera, F., Salamon, P., Srinagesh, S., Thingstad, T.F., Tran, T., Thurber, R.V., Willner, D., Youle, M., Rohwer, F., 2010. Viral and microbial community dynamics in four aquatic environments. ISME 4, 739–751.

Rohwer, F., Thurber, R.V., 2009. Viruses manipulate the marine environment. Nature 459, 207–212.

Roossinck, M.J., 2016. Deep sequencing for discovery and evolutionary analysis of plant viruses. Virus Res. 239, 82–86.

Roossinck, M.J., 2015. Plants, viruses and the environment: Ecology and mutualism. Virology 479–480, 271–277.

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

Roossinck, M.J., 2011b. The big unknown: plant virus biodiversity. Curr. Opin. Virol. 1, 63–7.

Roossinck, M.J., Bazán, E.R., 2017. Symbiosis: Viruses as Intimate Partners. Annu. Rev. Virol. 4, annurev-virology-110615-042323.

Roossinck, M.J., Martin, D.P., Roumagnac, P., 2015. Plant Virus Metagenomics : Advances in Virus Discovery. Phytopathology 105, 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. Mol. Ecol. 19 Suppl 1, 81–8.

Rosario, K., Breitbart, M., 2011. Exploring the viral world through metagenomics. Curr. Opin. Virol. 1, 289–297.

Rosario, K., Capobianco, H., Fei, T., Ng, 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, e886748.

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, E36.

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. Viruses 7, 5553–5570.

Rosen, G.L., Reichenberger, E.R., Rosenfeld, A.M., 2011. NBC : the Naïve Bayes Classification tool webserver for taxonomic classification of metagenomic reads. Bioinformatics 27, 127–129.

Roumagnac, P., Granier, M., Bernardo, P., Deshoux, M., Ferdinand, R., Galzi, S., Fernandez, E., Julian, C., Abt, I., Filloux, D., Mesléard, 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., Hurwitz, B.L., Sullivan, M.B., 2015a. VirSorter: mining viral signal from microbial genomic data. PeerJ 3, e985.

Roux, S., Enault, F., Ravet, V., Colombet, J., Bettarel, Y., Auguet, J., Bouvier, T., Lucas-staat, S., Vellet, A., Prangishvili, D., Forterre, P., Debroas, D., Sime-ngando, T., 2016. Analysis of metagenomic data reveals common features of halophilic viral communities across continents. Environ. Microbiol. 18, 889–903.

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, e33641.

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

Ryabov, E. V, Keane, G., Naish, N., Evered, C., Winstanley, D., 2009a. Densovirus induces winged morphs in asexual clones of the rosy apple aphid, Dysaphis plantaginea. Proc. Natl. Acad. Sci. U. S. A. 106, 8465–8470.

Ryabov, E. V, Keane, G., Naish, N., Evered, C., Winstanley, D., 2009b. Densovirus induces winged morphs in asexual clones of the rosy apple aphid, Dysaphis plantaginea. Proc. Natl. Acad. Sci. U. S. A. 106, 8465–70.

Sakamoto, J.M., Fei, T., Ng, F., Suzuki, Y., Tsujimoto, H., Deng, X., Delwart, E., Rasgon, J., 2016. Bunyaviruses are common in male and female Ixodes scapularis ticks in central Pennsylvania. PeerJ 4, e2324.

Sanjua, R., Nebot, M.R., Chirico, N., Mansky, L.M., Belshaw, R., 2010. Viral mutation rates. J. Virol. 84, 9733–9748.

Schlub, T.E., Smyth, R.P., Grimm, A.J., Mak, J., Davenport, M.P., 2010. Accurately measuring recombination between closely related HIV-1 genomes. PLoS Comput. Biol. 6, e1000766.

Schmieder, R., Edwards, R., 2011. Fast identification and removal of sequence contamination from genomic and metagenomic datasets. PLoS One 6, e17288.

Scholthof, K.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. Mol. Plant Pathol. 12, 938–954.

Schulz, F., Yutin, N., Ivanova, N.N., Ortega, D.R., Lee, T.K., Vierheilig, J., Daims, H., Horn, M., Wagner, M., Jensen, G.J., Kyrpides, N.C., Koonin, E. V, Woyke, T., 2017. Giant viruses with an expanded complement of translation system components. Science. 356, 82–85.

Shi, C., Liu, Y., Hu, X., Xiong, J., Zhang, B., Yuan, Z., 2015. A metagenomic survey of viral abundance and diversity in mosquitoes from Hubei province. PLoS One 10, e0129845.

Shi, M., Lin, X., Tian, J., Chen, L., Chen, X., Li, C., Qin, X., Li, J., Cao, J., Eden, J., Buchmann, J., Wang, W., Xu, J., Holmes, E., Zhang, Y., 2016. Redefining the invertebrate RNA virosphere. Nature 540, 539–543.

Simmonds, P., Adams, M., Benkő, M., Breitbart, M., Brister, J., Carstens, E., Davison, A., Delwart, E., Gorbalenya, A., Harrach, B., Hull, R., King, A., Koonin, E., Krupovic, M.,

Kuhn, J., Lefkowitz, E., Nibert, M., Orton, R., Roossinck, M., Sabanadzovic, S., Sullivan, M., Suttle, C., Tesh, R., van der Vlugt, R., Varsani, A., Zerbini, F., 2017. Virus taxonomy in the age of metagenomics. Nat. Rev. Microbiol. 15, 161–168.

Skewes-Cox, P., Sharpton, T.J., Pollard, K.S., DeRisi, J.L., 2014. Profile hidden Markov models for the detection of viruses within metagenomic sequence data. PLoS One 9, e105067.

Smuts, H., Kew, M., Khan, A., Korsman, S., 2014. Novel hybrid parvovirus-like virus, NIH-CQV/PHV, contaminants in silica column-based nucleic acid extraction kits. J. Virol. 88, 1398.

Soueidan, H., Schmitt, L.-A., Candresse, T., Nikolski, M., 2014. Finding and identifying the viral needle in the metagenomic haystack: trends and challenges. Front. Microbiol. 5.

Stern Adi, Rotem, S., 2011. The phage-host arms-race : Shaping the evolution of microbes. BioEssays 33, 43–51.

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

Sweet, M., Bythell, J., 2017. The role of viruses in coral health and disease. J. Invertebr. Pathol. 147, 136–144.

Syvanen, M., 2012. Evolutionary implications of horizontal gene transfer. Annu. Rev. Genet. 46, 341–358.

Szathmary, E., Maynard Smith, J., 1997. From Replicators to Reproducers: the First Major Transitions Leading to Life. J. Theor. Biol. 187, 555–571.

Tangherlini, M., Dell'Anno, A., Zeigler Allen, L., Riccioni, G., Corinaldesi, C., 2016. Assessing viral taxonomic composition in benthic marine ecosystems: reliability and efficiency of different bioinformatic tools for viral metagenomic analyses. Sci. Rep. 6, 28428.

Temmam, S., Davoust, B., Berenger, J.-M., Raoult, D., Desnues, C., 2014. Viral metagenomics on animals as a tool for the detection of zoonoses prior to human infection? Int. J. Mol. Sci. 15, 10377–10397.

Temmam, S., Monteil-Bouchard, S., Robert, C., Baudoin, J., Sambou, M., Aubadie-ladrix, M., Labas, N., Raoult, D., Mediannikov, O., Desnues, C., 2016. Characterization of Viral Communities of Biting Midges and Identification of Novel Thogotovirus Species and Rhabdovirus Genus. Viruses 8, 77.

Temmam, S., Monteil-bouchard, S., Robert, C., Pascalis, H., Michelle, C., Jardot, P., Charrel, R., Raoult, D., Desnues, C., 2015. Host-Associated Metagenomics : A Guide to Generating Infectious RNA Viromes. PLoS One 10, e0139810.

Thézé, J., Leclercq, S., Moumen, B., Cordaux, R., Gilbert, C., 2014. Remarkable diversity of endogenous viruses in a crustacean genome. Genome Biol. Evol. 6, 2129–2140.

Thompson, C.C., Chimetto, L., Edwards, R.A., Swings, J., Stackebrandt, E., Thompson, F.L., 2013. Microbial genomic taxonomy. BMC Genomics 14.

Thurber, R. V, Haynes, M., Breitbart, M., Wegley, L., Rohwer, F., 2009. Laboratory procedures to generate viral metagenomes. Nat. Protoc. 4, 470–83.

Tilman, D., Cassman, K.G., Matson, P.A., Naylor, R., Polasky, S., 2002. Agricultural sustainability and intensive production practices. Nature 418, 671–677.

Tokarz, R., Williams, S.H., Sameroff, S., Sanchez Leon, M., Jain, K., Lipkin, W.I., 2014.

Virome analysis of Amblyomma americanum, Dermacentor variabilis, and Ixodes scapularis ticks reveals novel highly divergent vertebrate and invertebrate viruses. J. Virol. 88, 11480–11492.

Valles, S.M., Shoemaker, D., Wurm, Y., Strong, C. a., Varone, L., Becnel, J.J., Shirk, P.D., 2013. Discovery and molecular characterization of an ambisense densovirus from South American populations of Solenopsis invicta. Biol. Control 67, 431–439.

Van Leeuwen, T., Vontas, J., Tsagkarakou, A., Dermauw, W., Tirry, L., 2010. Acaricide resistance mechanisms in the two-spotted spider mite Tetranychus urticae and other important Acari: a review. Insect Biochem. Mol. Biol. 40, 563–572.

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.

Vuillaume, F., Thébaud, G., Urbino, C., Forfert, N., Granier, M., Froissart, R., Blanc, S., Peterschmitt, M., 2011. Distribution of the phenotypic effects of random homologous recombination between two virus species. PLoS Pathog. 7, e1002028.

Webster, C., Longdon, B., Lewis, S., Obbard, D., 2016. Twenty-Five New Viruses Associated with the Drosophilidae (Diptera). Evol Bioinform Online 12, 13–25.

Webster, C.L., Waldron, F.M., Robertson, S., Crowson, D., Ferrari, G., Quintana, J.F., Brouqui, J.-M., Bayne, E.H., Longdon, B., Buck, A.H., Lazzaro, B.P., Akorli, J., Haddrill, P.R., Obbard, D.J., 2015. The Discovery, Distribution, and Evolution of Viruses Associated with Drosophila melanogaster. PLoS Biol. 13, e1002210.

Weinbauer, M.G., Rassoulzadegan, F., 2004. Are viruses driving microbial diversification and diversity ? Environ. Microbiol. 6, 1–11.

Whon, T.W., Kim, M.-S., Roh, S.W., Shin, N.-R., Lee, H.-W., Bae, J.-W., 2012. Metagenomic characterization of airborne viral DNA diversity in the near-surface atmosphere. J. Virol. 86, 8221–8231.

Wild, S., 2017. Invasive pest hits Africa. Nature 543.

Wilhelm, S.W., Suttle, C.A., 1999. Viruses and Nutrient Cycles in the Sea: Viruses play critical roles in the structure and function of aquatic food webs. Bioscience 49.

Willner, D., Furlan, M., Haynes, M., Schmieder, R., Angly, F.E., Silva, J., Tammadoni, S., Nosrat, B., Conrad, D., Rohwer, F., 2009. Metagenomic Analysis of Respiratory Tract DNA Viral Communities in Cystic Fibrosis and Non-Cystic Fibrosis Individuals. PLoS One 4, e7370.

Wolkowicz, R., Schaechter, M., 2008. What makes a virus a virus ? Nat. Rev. Microbiol. Corresp. 319, 2008.

Woolhouse, M., Gaunt, E., 2007. Ecological Origins of Novel Human Pathogens. Crit. Rev. Microbiol. 33, 231–242.

Woolhouse, M.E.J., Gowtage-Sequeria, S., 2005. Host range and emerging and reemerging pathogens. Emerg. Infect. Dis. 11, 1842–7.

Wu, Z., Ren, X., Yang, L., Hu, Y., Yang, J., He, G., Zhang, J., Dong, J., Sun, L., Du, J., Liu, L., Xue, Y., Wang, J., Yang, F., Zhang, S., Jin, Q., 2012. Virome analysis for identification of novel mammalian viruses in bat species from Chinese provinces. J. Virol. 86, 10999–11012.

Xia, H., Hu, C., Zhang, D., Tang, S., Zhang, Z., Kou, Z., Fan, Z., Bente, D., Zeng, C., Li, T., 2015. Metagenomic profile of the viral communities in Rhipicephalus spp. ticks from Yunnan, China. PLoS One 10, e0121609.

Zablocki, O., van Zyl, L., Adriaenssens, E.M., Rubagotti, E., Tuffin, M., Cary, S.C., Cowan, D., 2014. High-level diversity of tailed phages, eukaryote-associated viruses, and virophage-like elements in the metaviromes of antarctic soils. Appl. Environ. Microbiol. 80, 6888–6897.

Zhang, W., Yang, S., Shan, T., Hou, R., Liu, Z., Li, W., Guo, L., Wang, Y., Chen, P., Wang, X., Feng, F., Wang, H., Chen, C., Shen, Q., Zhou, C., Hua, X., Cui, L., Deng, X., Zhang, Z., Qi, D., Delwart, E., 2017. Virome comparisons in wild-diseased and healthy captive giant pandas. Microbiome 5, 1–19.

Zhou, N.N., Senne, D.A., Landgraf, J.S., Swenson, S.L., Erickson, G., Rossow, K., Liu, L.I.N., Yoon, K., Krauss, S., Webster, R.G., 1999. Genetic Reassortment of Avian , Swine , and Human Influenza A Viruses in American Pigs. J. Virol. 73, 8851–8856.

Webographie

CABI : http://www.cabi.org/

Ecophyto 2018 : http://agriculture.gouv.fr/ministere/le-plan-ecophyto-2018

European directives 2009/128/CE : https://www.legifrance.gouv.fr/affichTexte.do?cidTexte=JORFTEXT000022417742&categorieLien=id

FAO : http://www.fao.org/

Genohub : https://genohub.com/ngs-instrument-guide/

ICTV : https://talk.ictvonline.org/

ViralZone : http://viralzone.expasy.org/

Virus-Host DB : http://www.genome.jp/virushostdb/

Curriculum vitae scientifique

Publications en préparation

S. François, D. Mutuel, A. Duncan, L. Rodrigues, D. Filloux, E. Fernandez, P. Roumagnac, R. Froissart, M. Ogliastro. **Metagenomic analysis of the viral communities associated with the two-spotted mite *Tetranychus urticae*: identification of a novel mini densovirus and ten new viral species.** Soumis à Scientific Reports.

S. François, M. Kulikowski, M. Frayssinet, D. Filloux, E. Fernandez, P. Roumagnac, R. Froissart and M. Ogliastro. **Diversity and composition of pests' viral communities and their distribution in arthropod communities.** En préparation.

Publications parues ou sous presse dans des revues à comité de lecture

S. François, D. Filloux, E. Fernandez, M. Ogliastro and P. Roumagnac. **Viral metagenomics approaches for high resolution screening of multiplexed arthropod and plant viral communities.** In press, Methods in Molecular Biology.

S. François, M. Frayssinet, D. Filloux, M. Ogliastro and R. Froissart. **Increase in taxonomic assignment efficiency of viral reads in metagenomics.** (2017) Virus Research.

S. François, D. Filloux, P. Roumagnac, D. Bigot, P. Gayral, R. Froissart, M. Ogliastro. **Discovery of parvovirus sequences in an unexpected broad range of animals.** (2016) Scientific Reports.

S. François, P. Bernardo, D. Filloux, P. Roumagnac, N. Yaverkovski, R. Froissart, and M. Ogliastro. **A Novel Itera-Like Densovirus Isolated by Viral Metagenomics from the Sea Barley *Hordeum marinum.*** (2014) Genome Announcements.

A.S. Gosselin Grenet, F. Salasc, **S. Francois**, D. Mutuel, T. Dupressoir, C. Multeau , A. Perrin, M. Ogliastro. **Les densovirus: une « massive attaque » chez les arthropodes.** (2015) Virologie (Article en français).

P. Bernardo,[*] B. Muhire,[*] **S. François**, M. Deshoux, P. Hartnady, S. Kraberger, D. Filloux, M.J. Frilander, E. Fernandez, S. Galzi, R. Ferdinand, M. Granier, A. Marais, P. Monge, T. Candresse, F. Escriu, M. Peterschmitt, A.L. Laine, A. Varsani, G.W. Harkins, D.P. Martin, and P. Roumagnac. **Molecular characterization and prevalence of two capulaviruses: Alfalfa leaf curl virus from France and Euphorbia caput-medusae latent virus from South Africa.** (2016) Virology.

Communications à des congrès nationaux ou internationaux

Exploring agroecosystems arthropod pest's viral diversity. (2017) **Communication** – **Environmental Genomics National congress** (Marseille, France)

Discovery of Parvoviridae-related sequences in an unexpected broad range of animals. (2016) **Communication** – **XVIth Parvovirus International Workshop** (Ajaccio, France)

Exploring the viral world in arthropods using transcriptomics and metagenomics. (2015) **Communication** – **Microorganisms-hosts interactions National congress** (Montpellier, France)

Discovery of parvovirus sequences in an unexpected broad range of animals. (2015) **Poster** – **Pathobiome International congress** (Paris, France)

Exploring the viral world in arthropods using metagenomics. (2015) **Poster** – **Environmental Genomics National congress** (Montpellier, France)

Activités complémentaires

Enseignement: service de monitorat (192 h) effectué dans le cadre de modules d'initiation à la biologie, de biologie évolutive, de parasitologie et de microbiologie

Supervision d'un étudiant de Master 1

Annexes

Article de revue

Les densovirus : une « massive attaque » chez les arthropodes

Anne-Sophie Gosselin-Grenet[1,2], Fanny Salasc[2,3], **Sarah Francois**[1,2], Doriane Mutuel[2], Thierry Dupressoir[2,3], Cécilia Multeau[2,4], Aurélie Perrin[2,4], Mylène Ogliastro[2]

[1] Université de Montpellier, UMR 1333 DGIMI « Diversité, Génomes et Interactions Microorganismes-Insectes », place Eugène-Bataillon, 34095 Montpellier cedex 5, France

[2] Laboratoire « Diversité, Génomes et Interactions Microorganismes Insectes » (DGIMI) UMR 1333, INRA, Université Montpellier, 34095 Montpellier, France.

[3] EPHE, UMR 1333 DGIMI, 34095 Montpellier cedex 5, France

[4] InVivo AgroSolutions, route de Biot, 06560 Valbonne, France

Publié en janvier 2015 dans **Virologie** (19 (1) : 19-31)

Virologie 2015, 19 (1) : 19-31

Les densovirus : une « massive attaque » chez les arthropodes

Anne-Sophie Gosselin Grenet[1,2]
Fanny Salasc[2,3]
Sarah Francois[1,2]
Doriane Mutuel[2]
Thierry Dupressoir[2,3]
Cécilia Multeau[2,4]
Aurélie Perrin[2,4]
Mylène Ogliastro[2]

[1] Université de Montpellier, UMR 1333 DGIMI « Diversité, Génomes et Interactions Microorganismes-Insectes », place Eugène-Bataillon, 34095 Montpellier cedex 5, France
[2] INRA, UMR 1333 DGIMI, 34000 Montpellier, France
<asgossel@univ-montp2.fr>
[3] EPHE, UMR 1333 DGIMI, 34095 Montpellier cedex 5, France
[4] InVivo AgroSolutions, route de Biot, 06560 Valbonne, France

Résumé. Les densovirus (DV) sont des parvovirus d'arthropodes responsables d'épizooties chez les insectes et les crustacés. Structurellement simples, ces petits virus à ADN présentent une grande diversité de séquences et d'organisations génomiques, diversité probablement sous-estimée au regard des récentes découvertes de ces virus dans des hôtes inattendus. Les densovirus représentent un modèle de choix pour étudier à différentes échelles les interactions virus-hôtes et leurs évolutions. Nous proposons de revisiter les connaissances fondamentales sur les densovirus qui ont essentiellement été établies par des approches mécanistiques et envisageons les nouvelles perspectives d'études permises par des approches plus globales. Pour conclure, nous décrivons les applications possibles de ces virus comme outils biologiques, notamment pour le contrôle de populations d'insectes dits « nuisibles ».

Mots clés : densonucléose, épizooties, diversité virale, barrière intestinale, lutte biologique

Abstract. Densoviruses (DVs) are parvoviruses of arthropods and causative agents of natural epizootics in insects and crustaceans populations. Structurally simple, these small DNA viruses, display a large diversity of genomic sequences, structures and organizations. Such diversity, together with the diversity of their invertebrate hosts, from shrimps to mosquitoes and recently including sea stars, suggests that DVs are largely unknown and ubiquitous in the environment. Densoviruses are considered as a model of choice to study virus-host interactions and their evolution at different scales, from individuals to populations. This review summarizes the knowledge on densovirus biology obtained through mechanistic and global approaches. Finally, the potential use of these viruses as biological control agents against insect pests and disease-vectors are exposed.

Key words: densonucleosis, epizootics, viral diversity, gut barrier, biological pest control

De tout petits virus, un peu partout

Les densovirus ont initialement été découverts en France dans les années 1960, au cours d'épizooties affectant les élevages de chenilles de la fausse teigne de la ruche (*Galleria mellonella*) utilisées comme appât de pêche [30]. Ces virus doivent leur nom à la pathologie cellulaire observée lors de ces infections et décrite comme une « maladie à noyaux denses » ou « densonucléose », c'est-à-dire une hypertrophie nucléaire liée à l'accumulation de particules virales dans le noyau des cellules infectées [44]. Isolées et purifiées à partir de cadavres de ces insectes infectés, ces particules virales non enveloppées, parmi les plus petites connues chez les virus d'animaux, contenaient une unique molécule d'ADN linéaire et simple brin. Ces caractéristiques structurales et génomiques ont permis de classer ces virus dans la famille des *Parvoviridæ*, comprenant dès lors deux sous-familles, les *Densovirinæ* et les *Parvovirinæ*, définie selon leur spectre d'hôte puisqu'infectant, respectivement, les invertébrés et les vertébrés [10]. Le premier densovirus a été nommé *Galleria mellonella* densovirus (GmDV), nom d'espèce virale donné en référence au nom de son premier hôte. Depuis, les densovirus ont été découverts lors d'épizooties dans des populations d'autres lépidoptères, de moustiques, de blattes et de pucerons. Leur découverte a

Tirés à part : A.-S. Gosselin Grenet

doi:10.1684/vir.2015.0589

Pour citer cet article : Gosselin Grenet AS, Salasc F, Francois S, Mutuel D, Dupressoir T, Multeau C, Perrin A, Ogliastro M. Les densovirus : une « massive attaque » chez les arthropodes. *Virologie* 2015; 19(1) : 19-31 doi:10.1684/vir.2015.0589

ensuite dépassé la classe des insectes et ils ont été retrouvés chez les crustacés, notamment les crevettes, chez lesquels ils sont régulièrement responsables d'importantes mortalités dans les élevages. Très récemment, les densovirus ont été associés à des épizooties importantes d'échinodermes, élargissant ainsi la diversité et la distribution de cette famille de virus à plusieurs phyla animaux [21, 22]. À ce jour, une cinquantaine de densovirus ont été caractérisés, leur nom générique masque leur spécificité, stricte ou relative (une espèce virale peut infecter plusieurs espèces d'hôtes d'un même ordre). Leur distribution dans des hôtes phylogénétiquement et géographiquement distants, occupant des écosystèmes terrestres et aquatiques, suggèrent que ces virus sont ubiquitaires dans l'environnement et que l'essentiel de leur diversité reste probablement à découvrir. L'élargissement des échantillonnages environnementaux et le développement des analyses métagénomiques devraient enrichir nos connaissances sur la prévalence et la diversité de cette famille de virus et permettre d'améliorer notre compréhension de leur histoire évolutive.

Une classification en perpétuelle évolution

Les densovirus sont de petits virus non enveloppés, possédant une capside icosaédrique de 18 à 28 nm protégeant une molécule d'ADN monocaténaire linéaire non segmentée de 3,7 à 6,3 kilobases et dont les extrémités sont composées pour la plupart, de répétitions terminales inversées (ITRs) [15]. Leur génome code deux types de protéines : les protéines non structurales (NS), multifonctionnelles, qui permettent notamment la réplication du génome viral, et les protéines structurales (VP) qui s'assemblent pour former la capside virale (*figure 1A*).

Le concept d'espèce densovirale a été récemment redéfini, entraînant une modification de la classification des densovirus [16]. La définition d'une espèce est basée sur la séquence de la protéine de réplication NS1, protéine sous forte contrainte fonctionnelle et dont les motifs enzymatiques sont conservés chez tous les *Parvoviridæ*. À présent, une espèce est définie par rapport à une autre par une différence d'identité de séquence de plus de 15 % au niveau de NS1 et les espèces appartenant au même genre doivent posséder entre elles plus de 30 % d'identité sur NS1. Cette classification prend également en compte le spectre d'hôtes des virus et certaines de leurs caractéristiques moléculaires : le nombre et la taille de leurs cadres ouverts de lecture (ORF), l'organisation de leur génome (monosens ou ambisens) et les caractéristiques des extrémités du chromosome viral (des ITRs notamment). Ainsi, au sein des *Densovirinæ*, 15 espèces ont été définies et sont réparties en 5 genres reconnus par l'International Commitee on Taxonomy of Viruses (ICTV) : *Ambidensovirus*, *Brevidensovirus*, *Hepandensovirus*, *Iteradensovirus* et *Penstyldensovirus* [16] (*figure 1B*). Certains densovirus restent non classés car trop divergents pour être associés à l'un de ces 5 genres.

Cette classification évoluera encore. En effet, les densovirus sont essentiellement décrits dans des arthropodes d'intérêt économique, médical ou vétérinaire, comme les vers à soie, les crevettes, les moustiques, les tiques ou les phlébotomes, suggérant d'avantage un biais d'échantillonnage qu'une description exhaustive de la diversité de ces virus et de leurs hôtes.

De la pathogenèse aux pathologies

La pathogenèse virale décrit les processus par lesquels un virus induit une maladie, depuis l'entrée du virus dans un organisme hôte jusqu'à sa transmission à un autre organisme. À ce jour, la pathogenèse n'est connue que pour quelques densovirus, notamment le *Junonia coenia* densovirus JcDV, virus type du genre *Ambidensovirus*, infectant la chenille *Spodoptera frugiperda*, un lépidoptère ravageur de cultures (*figure 2*) [31].

La principale voie d'infection décrite pour les densovirus est alimentaire, les particules virales étant ingérées avec de la nourriture contaminée, et résulte d'une transmission entre individus, dite transmission horizontale. La transmission verticale, pour laquelle le virus est transmis à la descendance, a été décrite pour quelques densovirus [32, 37], mais nous nous limiterons ici à la description des mécanismes d'infection impliquant une transmission horizontale.

Une fois ingérées, les particules virales atteignent l'intestin de l'hôte grâce au péristaltisme et sont internalisées rapidement dans les cellules intestinales. Leur devenir dépend ensuite de l'espèce virale. Deux groupes de densovirus peuvent être distingués selon leur tropisme tissulaire, ceux dont le tropisme est large (polytropique) et ceux pour lesquels il est restreint à un seul tissu (monotropique). Les virus appartenant aux genres *Ambidensovirus*, *Brevidensovirus* et *Hepanvirus* sont polytropiques et pour la plupart traversent l'intestin sans se répliquer [7, 11, 29]. Les virus appartenant au genre *Iteradensovirus,* notamment les densovirus du ver à soie *Bombyx mori,* sont généralement monotropiques et restreints aux cellules columnaires de l'intestin moyen [24].

Le franchissement de la barrière intestinale

Comme pour tout virus à transmission orale, l'intestin est la première barrière que les densovirus doivent franchir pour initier l'infection. Chez les insectes, l'intestin est un épithélium monocouche composé de 3 parties d'origines

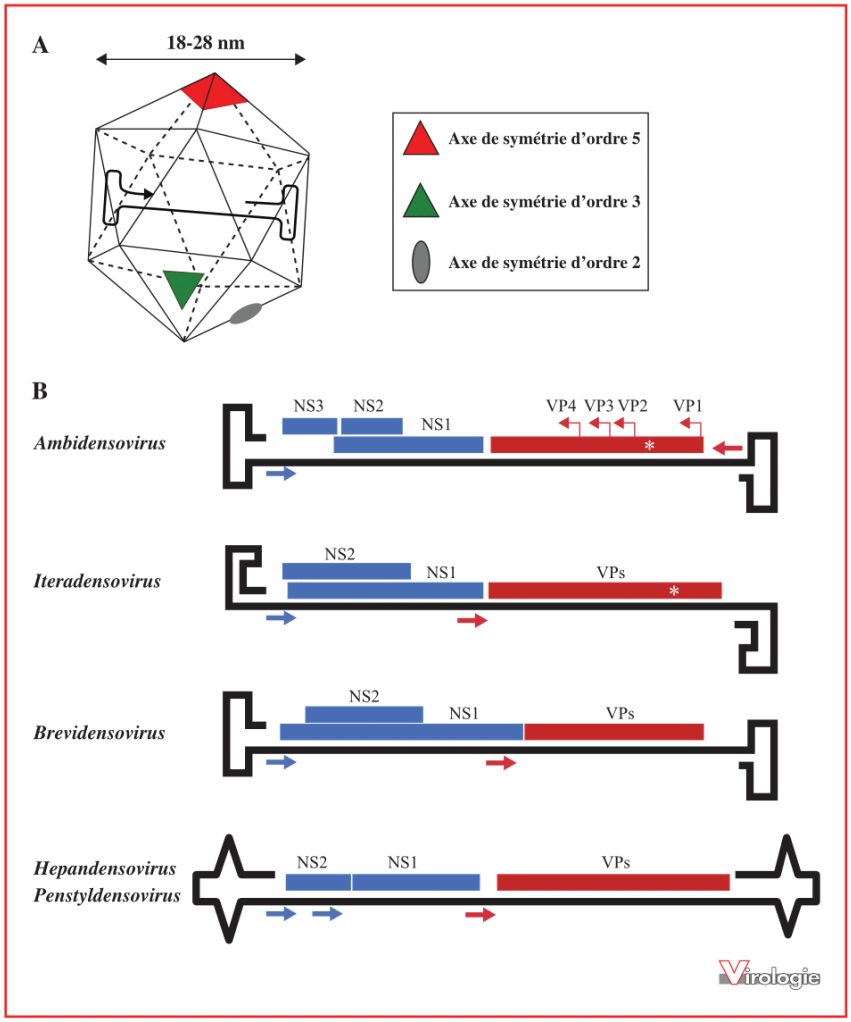


Figure 1. Structure et organisation génomique des *Densovirinæ*. A) Les densovirus sont des virus nus qui possèdent une capside icosaédrique de 18 à 28 nm de diamètre. Cette capside est constituée par l'assemblage de 60 protéines de structure (VP) et protège le génome viral constitué d'une molécule d'ADN monocaténaire et linéaire [15]. **B)** Le génome des densovirus est une molécule d'ADN simple brin de 3,7 à 6,3 kilobases, dont les extrémités contiennent des répétitions terminales inversées qui se replient en structures de type « épingles à cheveux » et servent à la réplication virale. Il possède des gènes codant deux types de protéines : les protéines non structurales (NS, en bleu) et les protéines de capside (VP, en rouge). Les flèches bleues et rouges indiquent les promoteurs initiant la transcription des gènes codant, respectivement, les protéines NS et les protéines VP. La localisation du motif phospholipase A2 (PLA2) est indiquée par une étoile. Les virus du genre *Ambidensovirus* possèdent un génome ambisens, celui des 4 autres genres (*Iteradensovirus*, *Brevidensovirus*, *Hepandensovirus* et *Penstyldensovirus*) est monosens.

embryonnaires et de fonctions physiologiques différentes. Les régions antérieures et postérieures, d'origine ectodermique, sont recouvertes d'une cuticule, une structure acellulaire imperméable qui empêche toute absorption et les protège du milieu extérieur, notamment des pathogènes. L'intestin moyen, d'origine endodermique, est dépourvu de cuticule mais protégé par une membrane péritrophique composée d'un réseau de fibrilles de chitine dans une matrice de carbohydrates et de protéines glycosylées de type mucines [42, 48]. Les interstices de cette membrane, de 20 à 40 nm selon les insectes, limitent l'entrée des pathogènes en fonction de leur taille. L'intestin moyen occupe la majeure partie de l'intestin, c'est le lieu d'absorption et d'échange avec le milieu extérieur et la porte d'entrée principale des pathogènes. Il est composé de 2 types cellulaires principaux associés par des jonctions serrées, dites

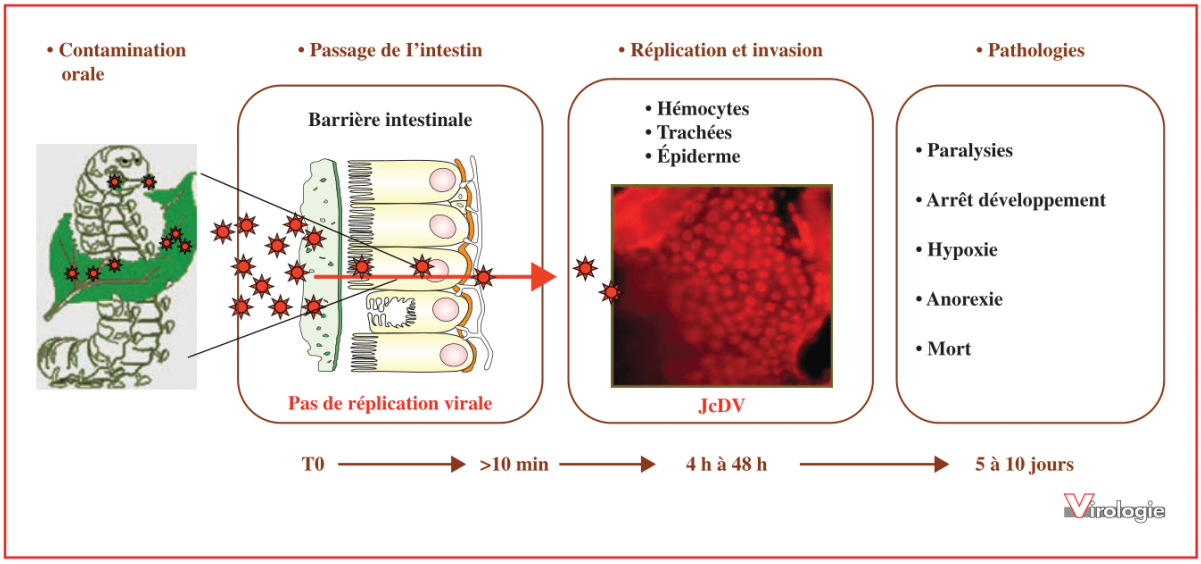


Figure 2. Pathogenèse du densovirus JcDV chez le lépidoptère *Spodoptera frugiperda*. Les particules virales, ingérées par la chenille *via* l'alimentation, rejoignent rapidement l'intestin moyen et traverse l'épithélium intestinal sans réplication virale pour atteindre les tissus cibles sous-jacents, principalement les hémocytes, les trachées et l'épiderme. La multiplication virale dans ces tissus bloque les mues et entraîne l'asphyxie des chenilles par obstruction des trachées, ce qui conduit à la mort de l'hôte, 5 à 10 jours après l'ingestion initiale. L'image d'immunofluorescence présentée au centre montre la multiplication virale dans l'épiderme d'une chenille à 5 jours p.i. (tirée de [31]). JcDV (en rouge) est mis en évidence à l'aide d'un anticorps primaire de souris dirigé contre les protéines de capside (VP) et d'un anticorps secondaire anti-souris Alexa Fluor® 594 (Invitrogen).

septées. Les cellules columnaires sont les plus abondantes et présentent des microvillosités apicales offrant une large surface d'absorption ; les cellules caliciformes (ou cellules « à mucus ») régulent notamment le pH de l'intestin (*figure 3A*). Un troisième type cellulaire, les cellules souches, assure le renouvellement des cellules intestinales. Ces cellules sont situées à la base de l'épithélium et n'établissent pas de jonctions septées avec les autres types cellulaires. La barrière intestinale est ainsi composée de deux types de structures, une membrane péritrophique acellulaire et un épithélium, qui impliquent des interactions moléculaires multiples et d'affinités variables avec les capsides virales.

Le franchissement de cette barrière par les densovirus n'a été étudié que pour JcDV [50] (*figure 3C*). Il semble impliquer dans ce cas une première étape de reconnaissance de la membrane péritrophique sur laquelle les particules virales se concentrent rapidement après leur ingestion (*figure 3B*). La taille des densovirus est, en théorie, suffisamment petite pour permettre leur passage passif à travers cette membrane péritrophique. La reconnaissance de cette structure suggère

Figure 3. Franchissement de la barrière intestinale du lépidoptère *Spodoptera frugiperda* par le densovirus JcDV (d'après [49]).
A) Image d'immunofluorescence sur coupe semi-fine d'intestin moyen de *S. frugiperda* montrant l'organisation de l'épithélium intestinal. L'actine (en vert) est visualisée à l'aide de phalloïdine-FITC et les noyaux (en bleu) à l'aide de Hoechst. **B)** Image d'immunofluorescence sur coupe semi-fine d'intestin moyen de *S. frugiperda* montrant la localisation des particules virales (en rouge) 15 min après leur ingestion par la chenille. La flèche indique la position de la membrane péritrophique (PM). **C)** Schéma résumant les mécanismes de franchissement de la barrière intestinale par JcDV. Dans l'intestin moyen, les particules virales se concentrent majoritairement sur la membrane péritrophique et les microvillosités apicales de l'épithélium intestinal. JcDV entre ensuite dans les cellules columnaires par un mécanisme d'endocytose dépendante de la dynamine et de la cavéoline et traverse l'épithélium intestinal par transcytose sans réplication virale. La transcytose du virus entraîne une augmentation de la perméabilité intestinale résultant de l'ouverture partielle des jonctions septées. Les mécanismes précis (directs ou indirects) permettant cette ouverture ne sont pas encore connus. **D)** Micrographie électronique sur coupe ultrafine d'intestin moyen de *S. frugiperda* montrant la structure en échelle des jonctions septées reliant les cellules intestinales. **E)** Micrographie électronique sur coupe ultra-fine d'intestin moyen de *S. frugiperda* montrant l'ouverture des jonctions septées consécutive à l'infection virale (10 min p.i.).
Barre d'échelle A et B : 20 µm. Barre d'échelle D et E : 100 nm.
Ni : non infecté ; cc : cellule columnaire ; gc : cellule caliciforme ; Lum : lumière de l'intestin ; HemoL : hémolymphe ; bl : lame basale ; Tr : trachée.

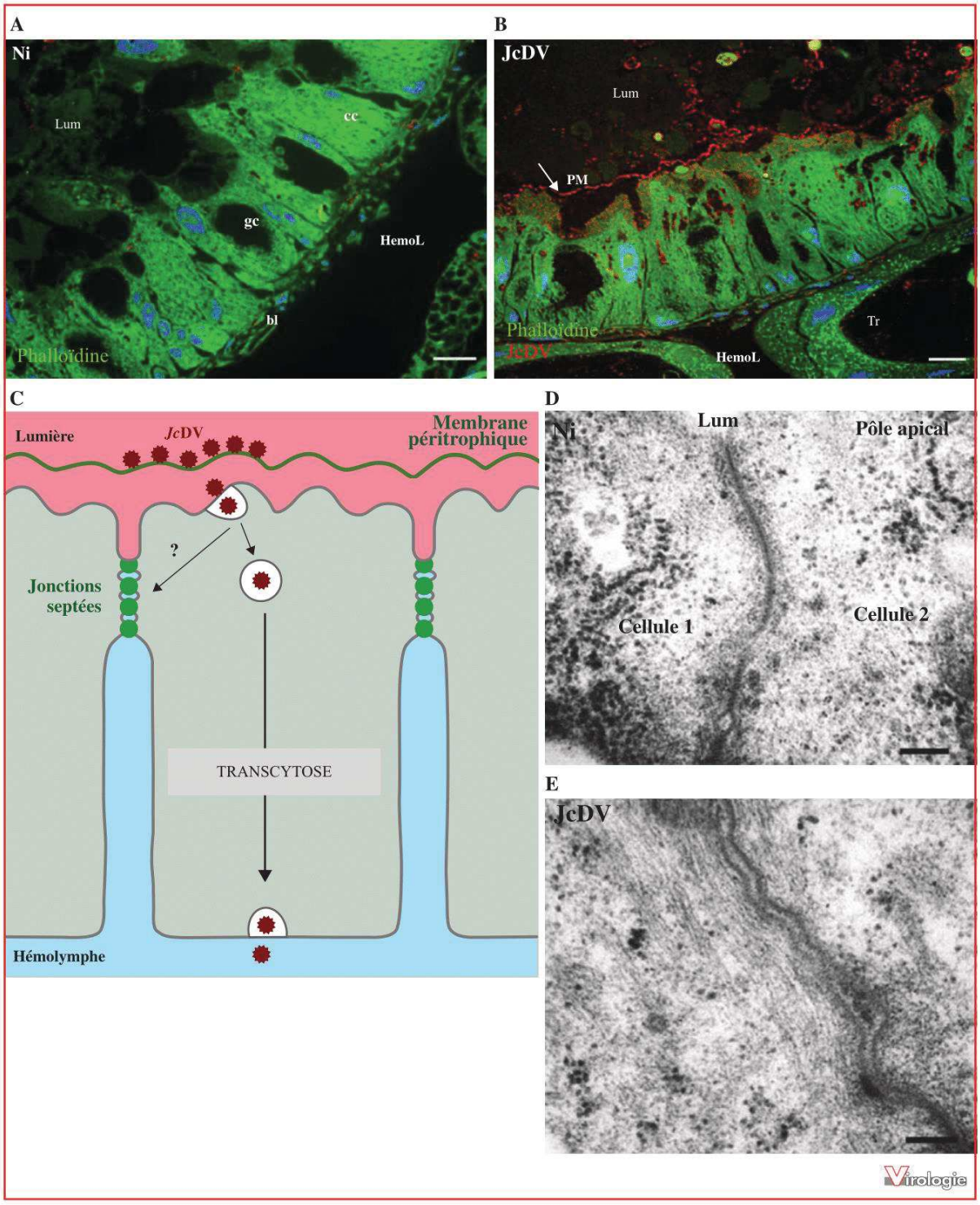


Figure 3. *Suite*

une forte affinité des capsides pour des molécules glycosylées, et pose donc la question du mécanisme par lequel les virions s'en détachent pour accéder à l'épithélium intestinal. On ne peut exclure un mécanisme actif de transport des virus au travers de la membrane qui impliquerait d'autres facteurs de l'environnement intestinal.

La deuxième étape concerne le franchissement de l'épithélium, impliquant la reconnaissance puis l'internalisation des virions par les cellules. Cette internalisation s'effectue par un mécanisme d'endocytose dépendant de la dynamine et de la cavéoline. Les particules virales sont ensuite transportées par transcytose à travers l'épithélium pour être délivrées dans le milieu intérieur de l'organisme (*figure 3C)*. Cette transcytose virale entraîne une augmentation de la perméabilité intestinale résultant d'une ouverture partielle des jonctions septées (*figure 3D et E*). De nombreuses questions restent encore en suspens, notamment la nature du/des récepteur(s) intestinaux des densovirus ainsi que le mécanisme de l'ouverture des jonctions septées et son rôle dans la pathogenèse virale.

La capside des densovirus : une simplicité relative

La capside est le premier composant viral interagissant avec l'hôte. Chez les densovirus, elle est constituée par l'assemblage de 60 protéines de structure VP (nombre de triangulation T=1) (*figure 1A*). Selon les densovirus, un à 2 gènes codent 2 à 4 VP (VP1 à VP4) qui partagent généralement une même région C-terminale mais possèdent des régions N-terminales différentes. Elles sont produites en proportions différentes probablement par un mécanisme de « *leaky scanning* ». La stabilité de la capside ainsi assemblée permet aux virions de résister aux variations de pH et de température en milieu naturel.

La topographie de surface de la particule virale des *Parvovirinæ* présente des reliefs et des motifs déterminants pour l'interaction du virion avec le(s) récepteur(s) cellulaire(s), et donc pour la spécificité d'hôtes, le tropisme tissulaire et la virulence (pour revue, voir [3]). La structure des capsides de *Densovirinæ* n'a été résolue que pour quelques espèces seulement, par cristallographie ou par cryo-microscopie électronique [10, 26, 39]. Contrairement aux *Parvovirinæ*, la surface de ces densovirus présente peu de reliefs, ce qui pourrait être une conséquence de pressions de sélection différentes chez les arthropodes, notamment celles du système immunitaire [43]. Les *Ambidensovirus,* GmDV et JcDV, présentent 94 % d'identité de séquence dans la protéine de capside majoritaire VP4 (*figure 4A*) mais un spectre d'hôte différent. Les analyses structurales comparées de ces 2 virus ont prédit notamment qu'un motif de 4 acides aminés (positions 121, 165, 172 et 175 de la séquence de VP4) formant un relief relatif dans l'axe de symétrie 5' du virus, était impliqué dans la spécificité d'hôtes (*figure 4B*) (ces 4 acides aminés correspondent à ceux annotés en position 123, 167, 174 et 177 dans [10, 30]). Cette prédiction structurale a été testée biologiquement en remplaçant les 4 résidus de JcDV par ceux de GmDV et en analysant les changements éventuels de spécificité des virus ainsi pseudotypés [30]. Les résultats montrent que ces substitutions n'inversent pas le spectre d'hôtes mais altèrent spécifiquement la capacité du virus JcDV pseudo-typé à franchir l'épithélium intestinal chez son hôte [30, 50]. La surface de ce motif semble donc être impliquée spécifiquement dans la reconnaissance d'un récepteur intestinal et en conséquence dans le trafic infectieux des virions à travers l'épithélium, l'étape initiale clé de la pathogenèse virale.

La multiplication virale et la prise de contrôle de la cellule

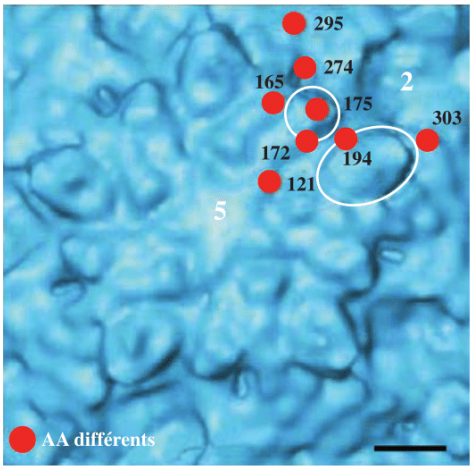
Les génomes de densovirus sont parmi les plus petits génomes connus chez les virus et ils expriment un nombre très limité de protéines virales (de 3 à 7). Les densovirus sont ainsi particulièrement dépendants de la cellule hôte et de leur capacité à détourner les grandes fonctions cellulaires pour assurer leur multiplication. Ces virus infectant des organismes non-modèles, notre compréhension des interactions moléculaires qui s'établissent entre les densovirus et la machinerie cellulaire est souvent limitée par la rareté des lignées cellulaires supportant la réplication virale et/ou d'outils moléculaires permettant de suivre l'odyssée des particules virales dans la cellule. Le processus infectieux de JcDV a été étudié dans une lignée cellulaire issue du lépidoptère *Lymantria dispar*, Ld652Y (*figure 5A*) [47]. L'entrée des virions dans ces cellules s'effectue par un mécanisme d'endocytose dépendante de puits recouverts de clathrine (*figure 5B*). Ce mécanisme, différent de l'entrée du virus dans les cellules intestinales, implique probablement d'autres récepteurs qui sont également à découvrir. Le trafic intracellulaire des virions s'effectue ensuite, *via* le cytosquelette, dans des endosomes précoces puis tardifs, dans lesquels l'acidification progressive joue probablement un rôle dans la libération du génome viral [47]. Chez les *Parvovirinæ*, l'exposition d'un motif de type phospholipase A2 (PLA2) au cours du trafic intracellulaire faciliterait l'introduction du génome viral dans le compartiment nucléaire [51]. La plupart des densovirus possèdent ce motif PLA2. Chez JcDV, il est situé dans l'extrémité C-terminale de la protéine VP1 , il est indispensable au cycle infectieux mais son rôle précis au cours du trafic intracellulaire reste à étudier [2].

La réplication virale se déroule dans le noyau et débute probablement pendant la phase S du cycle cellulaire. Elle est notamment orchestrée par l'ADN polymérase cellulaire et les activités enzymatiques des protéines virales non structurales. Les mécanismes de réplication des densovi-

A

```
VP4  JcDv  ----------MSLPGTGSGTSSGGGNTSGQEVYVIPRPFSNFGKKLSTYTKSHKFMI  47
VP4  GmDv  ----------MSLPGTGSGTSSGGGNTQGQDVYIIPRPFSNFGKKLSTYTKSHKFMI  47

VP4  JcDv  FGLANNVIGPTGTGTTAVNRLITTCLAEIPWQKLPLYMNQSEFDLLPPGSRVVECNVKVI  107
VP4  GmDv  FGLANNVIGPTGTGTTAVNRLLTTCLAEIPWQKLPLYMNQSEFDLLPPGSRVVECNVKVI  107

VP4  JcDv  FRTNRIAFETSSTATKQATLNQISNLQTAVGLNKLGWGIDRSFTAFQSDQPMIPTATSAP  167
VP4  GmDv  FRTNRIAFETSSTVTKQATLNQISNVQTAIGLNKLGWGINRAFTAFQSDQPMIPTATTAP  167

VP4  JcDv  KYEPITGTTTGYRGMIADYYGADSTNDAAFGNAGNYPHHQVGSFTFIQNYYCMYQQTNQGT  227
VP4  GmDv  KYEPVTGDTGYRGMIADYYGADSTNDTAFGNAGNYPHHQVSSFTFLQNYYCMYQQTNQGT  227

VP4  JcDv  GGWPCLAEHLQQFDSKTVNNQCLIDVTYKPKMGLIKPPLNYKIIGQPTAKGTISVGDNLV  287
VP4  GmDv  GGWPCLAEHLQQFDSKTVNNQCLIDVTYKPKMGLIKSPLNYKIIGQPTVKGTISVGDNLV  287

VP4  JcDv  NMRGAVVINPPEATQSVTESTHNLTRNFPANLFNIYSDIEKSQILHKGPWGHENPQIQPS  347
VP4  GmDv  NMRGAVVINPPEATQNVAESTHNLTRNFPADLFNIYSDIEKSQVLHKGPWGHENPQIQPS  347

VP4  JcDv  VHIGIQAVPALTTGALLVNSSPLNSWTDSMGYIDVMSSCTVMESQPTHFPFSTDANTNPG  407
VP4  GmDv  VHIGIQAVPALTTGALLINSSPLNSWTDSMGYIDVMSSCTVMEAQPTHFPFSTEANTNPG  407

VP4  JcDv  NTIYRINLTPNSLTSAFNGLYGNGATLGNV  437
VP4  GmDv  NTIYRINLTPNSLTSAFNGLYGNGATLGNV  437
```

B



JcDV

Virologie

Figure 4. Localisation des acides aminés de la protéine de capside VP4 impliqués dans le tropisme intestinal de JcDV. A) Alignement des séquences protéiques des VP4 de JcDV et de GmDV (437 acides aminés, 94 % d'identité). Les acides aminés qui diffèrent entre les 2 virus sont surlignés ; les 4 acides aminés (positions 121, 165, 172 et 175) impliqués dans le tropisme intestinal de JcDV sont surlignés en rouge [30]. B) Gros plan de la capside de JcDV montrant la topographie au niveau de l'axe de symétrie 5 et la localisation des 8 acides aminés de VP4, exposés à la surface, qui diffèrent de ceux de GmDV (en rouge) (modifié à partir d'une image fournie gracieusement par A. Bruemmer [10]). Les chiffres « 5 » et « 2 » en blanc sur la figure indiquent, respectivement, la position de l'axe de symétrie 5 (au centre du plateau en étoile) et celle de l'axe de symétrie 2. Barre d'échelle : 2 nm.
AA : acides aminés.

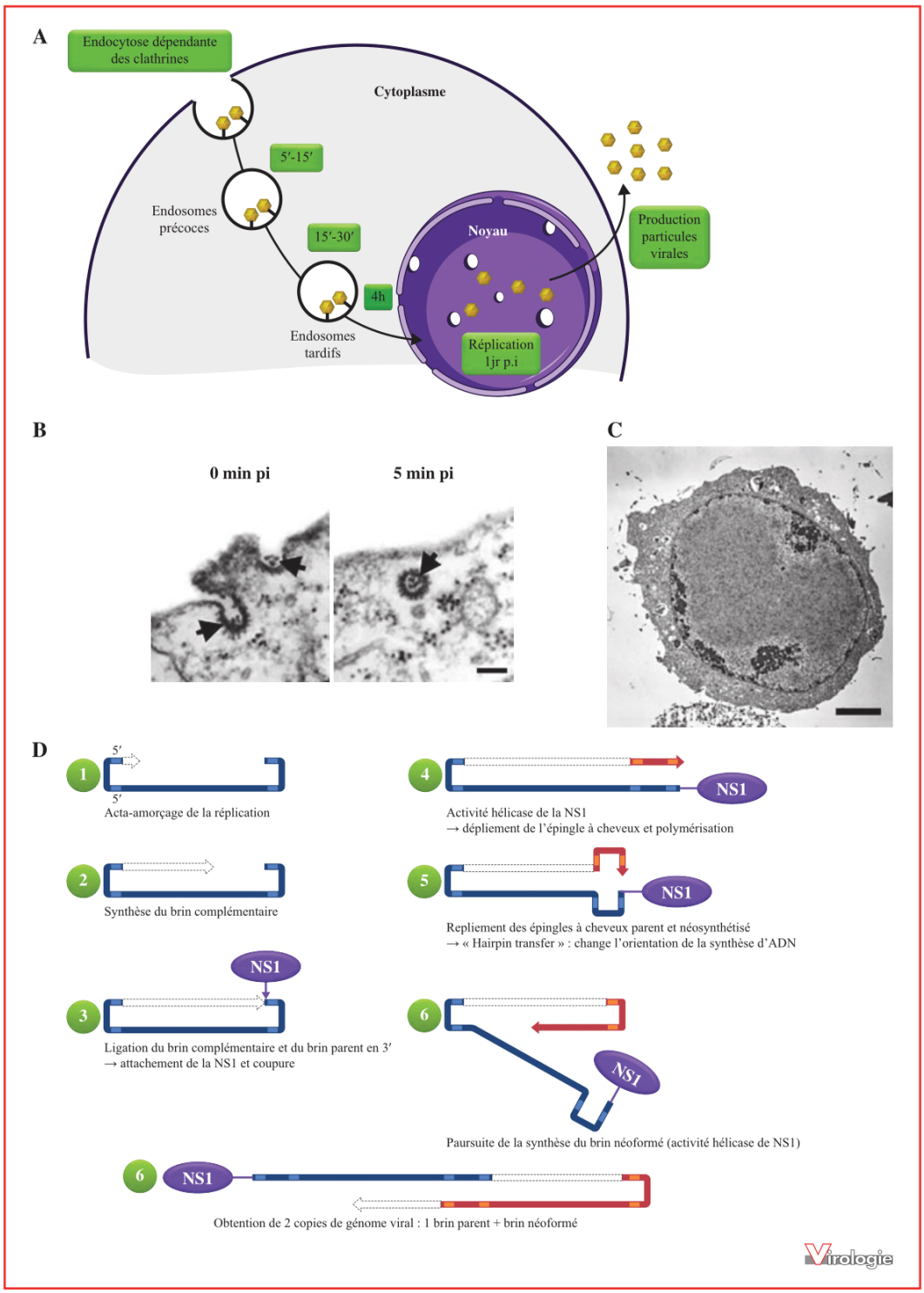


Figure 5. *Suite*

rus sont peu connus et les modèles proposés sont issus des connaissances acquises pour les *Parvovirinæ* (*figure 5D*). Ils font intervenir les régions terminales palindromiques en « épingle à cheveux » comme origines de réplication [15] et les activités endonucléase, ATPase et hélicase de la protéine NS1 [19]. Chez JcDV, la protéine NS3 est également indispensable à la réplication mais son rôle exact est inconnu [1]. Rien n'est connu concernant la protéine NS2 pourtant fortement exprimée lors du cycle infectieux de JcDV (données non publiées).

La multiplication virale est associée à des modifications morphologiques cellulaires importantes, notamment une augmentation drastique de la taille des cellules et une augmentation du rapport nucléo-cytoplasmique se traduisant par une densonucléose (*figure 5C*). Elle entraîne une demande accrue en énergie pour produire les constituants cellulaires nécessaires à la synthèse des protéines virales. Ainsi, les virus ciblent très souvent les voies de signalisation impliquées dans la croissance et la survie cellulaire dont les 3 acteurs principaux sont la phosphatidylinositol-3 kinase (PI3K), la sérine/thréonine kinase Akt et la protéine TOR [12, 14, 38]. Nos résultats récents montrent que l'infection des cellules Ld652Y par JcDV régule négativement la voie TOR et que l'inactivation de cette voie favoriserait la synthèse des protéines virales et la multiplication virale (F. Salasc, données non publiées).

Les symptômes et les pathologies

La virulence de l'infection et les pathologies associées dépendent du tropisme viral et donc, de l'espèce virale considérée. Concernant les *Ambidensovirus* pathogènes de lépidoptères, la mort de l'hôte survient une dizaine de jours après l'infection; ce temps varie selon la dose de virus inoculée et le stade de développement de l'hôte au moment de l'infection. Les symptômes génériques observés sont un arrêt du développement larvaire et le blocage des mues, suivis d'une anorexie, d'une léthargie et d'une lente mélanisation aboutissant à la mort de l'insecte [29, 31]. De façon caractéristique, les blattes infectées par le *Periplaneta fuliginosa* densovirus (PfDV) développent, dans la partie postérieure de l'intestin, un ulcère lié à l'accumulation d'hémocytes autour des cellules épithéliales infectées [40]. Chez les moustiques, l'infection induit des symptômes de paralysie, les larves perdent leur mobilité dans l'eau, le corps est déformé et perd sa pigmentation [7]. Enfin, les pathologies associées aux densovirus de crevettes sont particulièrement étudiées en raison des pertes économiques importantes causées par les épizooties (pour revue, voir [18]). Le *Penaeus stylirostris* densovirus (ou virus de la « nécrose hypodermique et hématopoïétique infectieuse ») provoque une mortalité massive (jusqu'à 90 %) chez la crevette bleue *P. stylirostris*. Ce virus est, en revanche, peu pathogène pour la crevette à pattes blanches *P. vannamei,* mais induit chez les larves infectées une réduction de la taille et une variété de déformations cuticulaires du rostre, des antennes, du thorax et de l'abdomen connues sous le nom de « syndrome de déformation de l'avorton ». Le virus hépatopancréatique (HPV) provoque, quant à lui, une atrophie de l'hépatopancréas, une anorexie et un important retard de croissance chez les crevettes infectées [18].

Analyse globale et découverte de nouveaux densovirus : vers un changement de paradigme ?

La diversité génomique et structurale des densovirus ainsi que la diversité de leurs hôtes contraste avec le faible nombre de génomes identifiés à ce jour. La phylogénie actuelle des densovirus présente une séparation attendue entre parvovirus et densovirus, mais sa robustesse s'avère toute relative, laissant présager que l'introduction de nouvelles séquences pourrait bousculer notre perception de leur histoire évolutive (*figure 6*).

L'essor du séquençage à haut débit, qui permet d'explorer sans *a priori* la diversité virale dans un environnement donné, a permis de montrer que les virus sont les entités les plus abondantes sur Terre [36]. Ces approches de métagénomiques virales ont permis de découvrir de nombreux virus dans des environnements et hôtes divers, qu'ils soient symptomatiques ou non (pour revue voir [35]). Ces découvertes ont montré la diversité des interactions virus-hôtes dans le règne animal et végétal, certains virus pouvant même avoir des effets positifs sur leur hôte [6, 34].

Le rythme des découvertes de nouveaux virus s'est ainsi considérablement accéléré ces dernières années. Les virus restant à découvrir seraient donc bien plus nombreux

Figure 5. Cycle de multiplication de JcDV et stratégie de réplication des *Parvoviridæ*. A) Cycle de multiplication de JcDV dans les cellules Ld652Y (d'après [47]). **B)** Micrographie électronique montrant l'entrée par endocytose des particules virales dans les cellules Ld652Y. Les flèches indiquent les puits (0 min p.i.) et les vésicules (5 min p.i.) recouvertes de clathrine. Barre d'échelle : 100 nm. [47]. **C)** Micrographie électronique montrant la densonucléose caractéristique de l'infection densovirale dans les cellules Ld652Y à 6 jours p.i. Barre d'échelle : 10 µm. (Marc Ravallec). **D)** Schéma présentant la stratégie de réplication des *Parvoviridæ*. Le brin parental est en bleu, le brin néosynthétisé en rouge. (Adapté de [15].).
NS1 : *Non-Structural protein* 1.

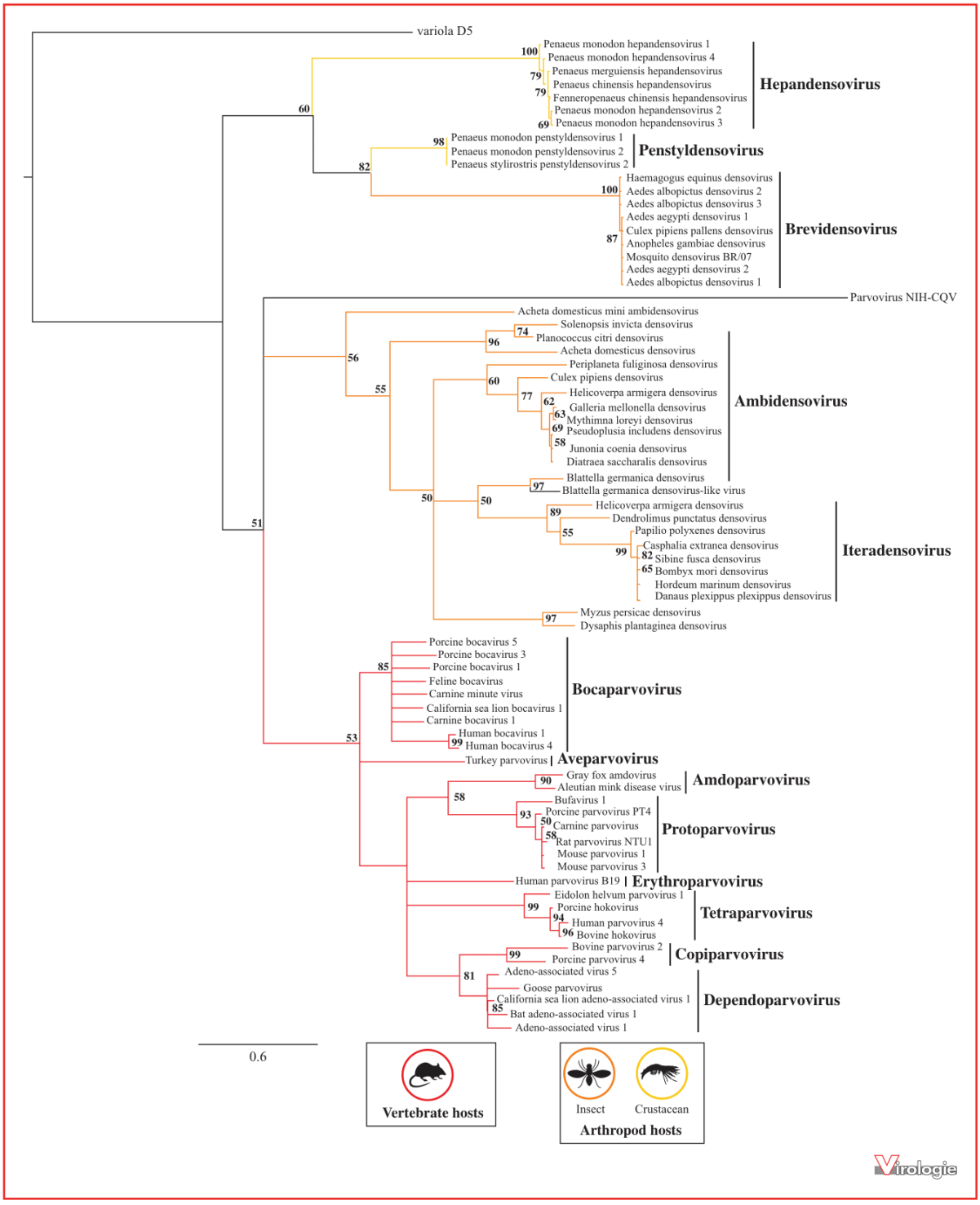


Figure 6. Arbre phylogénétique des *Parvoviridæ*.

que ceux déjà connus. Il a par exemple été estimé que 320 000 virus restaient à découvrir chez les quelques 5 500 espèces connues de mammifères [4]. Avec une estimation de 5 à 10 millions d'espèces d'arthropodes, on mesure l'immensité d'espèces virales à découvrir dans ce phylum parmi les plus diversifiés sur terre. Pour faire face à une future « explosion » de connaissance de génomes viraux, l'ICTV a récemment accepté d'inclure de nouvelles espèces virales uniquement sur la base d'une séquence génomique complète, en l'absence de toute propriété biologique, notamment la nature de l'hôte. Par exemple, une de nos études récentes menée sur des viromes de plantes de Camargue a permis d'identifier le génome complet d'un nouvel *Iteradensovirus* [20]. Mais nous ne savons pas si ce densovirus circulait dans la plante, comme le densovirus du puceron *Myzus persicæ* [45], ou s'il provient d'une contamination de la plante par un insecte infecté. Les données issues de la métagénomique posent donc de nouvelles questions sur l'écologie virale.

Par ailleurs, des séquences complètes ou partielles de densovirus ont été mis en évidence lors de recherches d'endogénisation virale [25, 28]. L'endogénisation consiste en l'intégration d'une partie (ou de l'ensemble) d'un génome viral au sein du génome de son hôte, suivie de la transmission du génome viral à la descendance de l'hôte. Ce phénomène est bien documenté pour les rétrovirus, car il constitue une composante essentielle de leur cycle de vie. Environ 8 % du génome humain serait ainsi issu de séquences de rétrovirus endogénéisées. Elle est également décrite chez d'autres types de virus, comme chez les *Circoviridæ*, les *Hepadnaviridæ* et les *Parvoviridæ* [23]. Des séquences virales apparentées aux densovirus ont ainsi été identifiées dans les génomes d'hôtes inattendus, comme un schistosome, une tique et la cione [28].

De manière générale, la mise en évidence de séquences densovirales au sein d'une diversité croissante de génomes d'animaux [8, 25, 43] suggère que le spectre d'hôte des densovirus est, ou a été, plus important que celui que l'on connaît actuellement. Ces nouvelles données remettent en question la validité taxonomique de la séparation entre *Parvovirinæ* et *Densovirinæ,* cette dichotomie étant basée sur une vision simplificatrice de leur spectre d'hôtes respectifs. Au regard de la diversité des hôtes, la diversité des séquences, structures et organisations génomiques des densovirus suggèrent, en effet, des histoires évolutives autrement plus complexes, et peut-être plus dynamiques, que celles illustrées par les phylogénies actuelles [28].

Ces observations soulignent la nécessité de caractériser de manière précise et approfondie les propriétés biologiques et l'écologie des virus nouvellement identifiés. Cela participe également du changement de paradigme en cours concernant les relations hôtes-pathogènes qui consiste à dépasser la notion de pathogénicité pour considérer tous les types de relations trophiques.

Les densovirus et la « chasse aux papillons »...

Depuis l'avènement de la chimie industrielle au milieu du 20e siècle, la lutte contre les insectes ravageurs de culture ou les vecteurs de maladies a recours aux insecticides de synthèse. L'utilisation intensive de ces produits phytosanitaires a entraîné l'accumulation dans l'environnement de molécules toxiques, notamment pour la santé humaine, ainsi que la sélection de phénotypes résistants dans les populations d'insectes ciblées.

Depuis les années 1970, la lutte biologique utilisant des virus entomopathogènes apparaît comme une stratégie alternative aux produits chimiques prometteuse en raison de son innocuité pour l'homme et de sa spécificité. Peu de ressources virales ont été jusque-là explorées et seuls les baculovirus sont actuellement commercialisés. Comme pour les insecticides, des résistances au baculovirus sont observées parmi les espèces cibles [5]. Dans ce contexte, les densovirus suscitent un regain de considération pour des applications en lutte biologique. Ils présentent l'intérêt d'être spécifiques et décrits dans les principaux groupes d'insectes d'importance agronomique, médicale ou vétérinaire chez lesquels ils sont transmissibles par voie orale, et pathogènes pour les stades larvaires.

Pour savoir si les densovirus peuvent être utilisés de manière durable et sans risque pour l'environnement, il est nécessaire de comprendre les mécanismes de spécificité et leur évolution à l'échelle 1) d'un insecte cible, 2) des

Figure 6. *Suite*
Cet arbre regroupe les densovirus répertoriés à ce jour par l'ICTV, ainsi que des parvovirus représentatifs de chacun des genres des *Parvovirinæ*. Il a été construit par la méthode du maximum de vraisemblance en utilisant PHYML 3.1 à partir d'un alignement de 127 acides aminés appartenant au domaine hélicase SF3 de la protéine NS1. La robustesse des nœuds de l'arbre a été testée par ré-échantillonnage selon la méthode du bootstrap (avec 500 itérations). La protéine D5 du virus de la variole humaine est utilisée comme groupe externe. Seules les valeurs de bootstrap supérieures à 50 % sont indiquées. L'échelle représente le taux de substitution par acide aminé et par position. Les genres auxquels appartiennent les virus sont indiqués à droite. Les hôtes associés sont représentés par des couleurs de branches différentes (vertébrés en rouge, insectes en orange et crustacés en jaune). Les branches terminales notées en noir correspondent à des *Parvoviridæ* dont les hôtes sont inconnus.

populations d'insectes cibles et 3) des populations naturelles d'insectes non-cibles. Cela suppose également de caractériser la capacité des insectes cibles à développer des résistances, ainsi que la capacité des virus à contourner ces résistances. La petitesse des génomes de densovirus constitue en ce sens un atout précieux pour explorer ces questions.

À l'instar des parvovirus associés aux adénovirus (AAV) développés comme vecteurs pour le transfert de gènes chez l'homme, les densovirus peuvent présenter un intérêt pour transférer des gènes chez l'insecte [13]. Ainsi, l'introduction de gènes hétérologues exprimant une toxine ou un ARN interférant permettant de contrôler le développement larvaire augmenterait la virulence et donc l'efficacité des densovirus en lutte biologique. La vectorisation de gènes *via* des capsides de densovirus a été validée chez les moustiques [32, 33]. L'idée d'améliorer l'efficacité de virus n'est pas nouvelle en matière de lutte biologique. Cela nécessite notamment des études de virologie fondamentale pour comprendre les mécanismes d'interactions afin de contrôler la rémanence des virus dans l'environnement et anticiper un risque de transmission à des espèces non cibles. Une alternative consistant à rendre ces virus déficients pour la réplication et/ou la transmission est actuellement à l'étude dans notre laboratoire.

Conclusion

L'étude des densovirus illustre la diversité des questions de virologie posées aujourd'hui et participe au changement de paradigme sur les virus amorcé par le développement du séquençage à haut débit. Initialement restreints aux arthropodes, les densovirus sont mis en évidence dans une diversité d'environnements et d'hôtes jusque-là insoupçonnés et pourraient illustrer la citation des microbiologistes Baas-Becking et Beijering écrivant dès 1934 « tout est partout mais l'environnement sélectionne » [17].

À la description de la diversité virale dans les écosystèmes, il faut désormais associer des concepts d'écologie des communautés et des approches mécanistiques d'étude des interactions à différentes échelles et notamment à l'échelle de l'hôte, qui est, en soi, un écosystème complexe. La réussite de l'infection d'un organisme dépend d'interactions cellulaires/moléculaires variées à chacune des étapes du processus infectieux. De nombreuses questions sont encore en suspens concernant les mécanismes de l'infection, par exemple celles concernant l'identification des récepteurs cellulaires et la réponse immunitaire des insectes face à l'infection virale. Décortiquer ces mécanismes permettra de reconstruire les réseaux d'interactions virus-cellule pour comprendre les contraintes structurant leur évolution.

Des travaux récents ont montré l'impact du microbiote sur le succès d'une infection virale [27]. Dans ce contexte, nous devons remplacer le concept de « pathogène » par celui de « pathobiome » reflétant d'avantage la complexité des interactions dans un environnement hôte [46].

Ces travaux sur les densovirus illustrent l'étroite relation qui doit exister entre virologie fondamentale et virologie appliquée pour permettre leur utilisation durable sur des cibles et sans risques pour des non cibles. En outre le développement d'élevages d'insectes pour l'alimentation animale et humaine demandera également de prévenir les risques d'épidémies causées par les densovirus [51]. Cela passe aussi par une connaissance de leur biologie et leur évolution.

Références

1. Abd-Alla A, Jousset FX, Li Y, *et al*. NS-3 protein of the Junonia coenia densovirus is essential for viral DNA replication in an Ld 652 cell line and Spodoptera littoralis larvae. *Journal of virology* 2004 ; 78 : 790-7.

2. Abd-Alla AM. Recherche sur un parvovirus d'insecte, le densovirus du lépidoptère Junonia coenia (JcDNV). Montpellier 2003.

3. Agbandje-McKenna MCM. Correlating structure with function in the viral capsid. In : Kerr JR, Cotmorre S, Bloom M, Linden M, Parrish C, eds. *Parvoviruses*. London : Hodder Arnold, 2006, p. 125-39.

4. Anthony SJ, Epstein JH, Murray KA, *et al*. 2013. A strategy to estimate unknown viral diversity in mammals. mBio 4:e00598-00513. doi:10.1128/mBio.00598-13.

5. Asser-Kaiser S, Fritsch E, Undorf-Spahn K, *et al*. Rapid emergence of baculovirus resistance in codling moth due to dominant, sex-linked inheritance. *Science* 2007 ; 317 : 1916-8.

6. Bao X, Roossinck MJ. A life history view of mutualistic viral symbioses: quantity or quality for cooperation? *Current opinion in microbiology* 2013 ; 16 : 514-8.

7. Barreau C, Jousset FX, Bergoin M. Pathogenicity of the Aedes albopictus parvovirus (AaPV), a denso-like virus, for Aedes aegypti mosquitoes. *Journal of invertebrate pathology* 1996 ; 68 : 299-309.

8. Belyi VA, Levine AJ, Skalka AM. Sequences from ancestral single-stranded DNA viruses in vertebrate genomes: the parvoviridae and circoviridae are more than 40 to 50 million years old. *Journal of virology* 2010 ; 84 : 12458-62.

9. Bergoin M, Tijssen P. Parvoviruses of Arthropods. In: Mahy BWJ & VanRegenmortel MHV (ed.). *Encyclopedia of Virology*, vol. 5. Oxford : Elsevier, 2008, p. 76-85.

10. Bruemmer A, Scholari F, Lopez-Ferber M, *et al*. Structure of an insect parvovirus (Junonia coenia Densovirus) determined by cryo-electron microscopy. *J Mol Biol* 2005 ; 347 : 791-801.

11. Buchatsky LP. Densonucleosis of bloodsucking mosquitoes. *Diseases of aquatic organisms* 1989 ; 6 : 145-50.

12. Buchkovich NJ, Yu Y, Zampieri CA, Alwine JC. The TORrid affairs of viruses: effects of mammalian DNA viruses on the PI3K-Akt-mTOR signalling pathway. *Nat Rev Microbiol* 2008 ; 6 : 266-75.

13. Carlson J, Suchman E, Buchatsky L. Densoviruses for control and genetic manipulation of mosquitoes. *Adv Virus Res* 2006 ; 68 : 361-92.

14. Cooray S. The pivotal role of phosphatidylinositol 3-kinase-Akt signal transduction in virus survival. *The Journal of general virology* 2004 ; 85 : 1065-76.

15. Cotmore S, Tattersall, P. A rolling-hairpin strategy: basic mechanisms of DNA replication in the parvoviruses. In : Kerr JR, Cotmorre S., Bloom M., *et al* (ed.). *Parvoviruses*. London (GB) : Hodder Arnold, 2006, p. 171-88.

16. Cotmore SF, Agbandje-McKenna M, Chiorini JA, *et al*. The family Parvoviridae. *Archives of virology* 2014 ; 159 : 1239-47.

17. de Wit R, Bouvier T. 'Everything is everywhere, but, the environment selects'; what did Baas Becking and Beijerinck really say? *Environmental microbiology* 2006 ; 8 : 755-8.

18. Dhar AK, Robles-Sikisaka R, Saksmerprome V, Lakshman DK. Biology, genome organization, and evolution of parvoviruses in marine shrimp. *Adv Virus Res* 2014 ; 89 : 85-139.

19. Ding C, Urabe M, Bergoin M, Kotin RM. Biochemical characterization of Junonia coenia densovirus nonstructural protein NS-1. *Journal of virology* 2002 ; 76 : 338-45.

20. Francois S, Bernardo P, Filloux D, *et al*. A Novel Itera-Like Densovirus Isolated by Viral Metagenomics from the Sea Barley Hordeum marinum. *Genome announc* 2014 ; 2 : e01196-1214.

21. Gudenkauf BM, Eaglesham JB, Aragundi WM, Hewson I. Discovery of urchin-associated densoviruses (family Parvoviridae) in coastal waters of the Big Island, Hawaii. *The Journal of general virology* 2014 ; 95 : 652-8.

22. Hewson I, Button JB, Gudenkauf BM, *et al*. Densovirus associated with sea-star wasting disease and mass mortality. *PNAS* 2014 ; 111 : 17278-83.

23. Horie M, Tomonaga K. Non-retroviral fossils in vertebrate genomes. *Viruses* 2011 ; 3 : 1836-48.

24. Ito K, Kidokoro K, Shimura S, Katsuma S, Kadono-Okuda K. Detailed investigation of the sequential pathological changes in silkworm larvae infected with Bombyx densovirus type 1. *Journal of invertebrate pathology* 2013 ; 112 : 213-8.

25. Kapoor A, Simmonds P, Lipkin WI. Discovery and characterization of mammalian endogenous parvoviruses. *Journal of virology* 2010 ; 84 : 12628-35.

26. Kaufmann B, El-Far M, Plevka P, *et al*. Structure of Bombyx mori densovirus 1, a silkworm pathogen. *Journal of virology* 2011 ; 85 : 4691-7.

27. Kuss SK, Best GT, Etheredge CA, *et al*. Intestinal microbiota promote enteric virus replication and systemic pathogenesis. *Science* 2011 ; 334 : 249-52.

28. Liu H, Fu Y, Xie J, *et al*. Widespread endogenization of densoviruses and parvoviruses in animal and human genomes. *Journal of virology* 2011 ; 85 : 9863-76.

29. Meynadier G, Vago C, Plantevin G, Atger P. Virose d'un type inhabituel chez le lépidoptère *Galleria mellonella*. *L Rev Zool Agri Appl* 1964 ; 63 : 207-8.

30. Multeau C, Froissart R, Perrin A, *et al*. Four amino acids of an insect densovirus capsid determine midgut tropism and virulence. *Journal of virology* 2012 ; 86 : 5937-41.

31. Mutuel D, Ravallec M, Chabi B, *et al*. Pathogenesis of Junonia coenia densovirus in Spodoptera frugiperda: a route of infection that leads to hypoxia. *Virology* 2010 ; 403 : 137-44.

32. Ren X, Rasgon JL. Potential for the Anopheles gambiae densonucleosis virus to act as an "evolution-proof" biopesticide. *Journal of virology* 2010 ; 84 : 7726-9.

33. Ren X, Hoiczyk E, Rasgon JL. Viral paratransgenesis in the malaria vector Anopheles gambiae. *PLoS Pathog* 2008 ; 4 : e1000135.

34. Roossinck MJ. The good viruses: viral mutualistic symbioses. *Nat Rev Microbiol* 2011 ; 9 : 99-108.

35. Roossinck MJ. Plant virus ecology. *PLoS Pathog* 2013 ; 9 : e1003304.

36. Rosario K, Breitbart M. Exploring the viral world through metagenomics. *Current opinion in virology* 2011 ; 1 : 289-97.

37. Ryabov EV, Keane G, Naish N, *et al*. Densovirus induces winged morphs in asexual clones of the rosy apple aphid, Dysaphis plantaginea. *PNAS* 2009 ; 106 : 8465-70.

38. Sanlioglu S, Benson PK, Yang J, *et al*. Endocytosis and nuclear trafficking of adeno-associated virus type 2 are controlled by rac1 and phosphatidylinositol-3 kinase activation. *Journal of virology* 2000 ; 74 : 9184-96.

39. Simpson AA, Chipman PR, Baker TS, *et al*. The structure of an insect parvovirus (Galleria mellonella densovirus) at 3.7 A resolution. *Structure* 1998 ; 6 : 1355-67.

40. Suto C. Characterization of a virus newly isolated from the smoky-brown cockroach, Periplaneta fuliginosa (Serville). *Nagoya journal of medical science* 1979 ; 42 : 13-25.

41. Terra WR. The origin and functions of the insect peritrophic membrane and peritrophic gel. *Arch Insect Biochem Physiol* 2001 ; 47 : 47-61.

42. Theze J, Leclercq S, Moumen B, Cordaux R, Gilbert C. 2014. Remarkable diversity of endogenous viruses in a crustacean genome. *Genome biology and evolution* 2014 ; 6 : 2129-40.

43. Tijssen P, Bando H, Li Y, *et al*. Evolution of densoviruses. In : Kerr JR, Cotmorre S, Bloom M, eds. *Parvoviruses*. London : Hodder Arnold, 2006, p. 55-68.

44. Vago C. The Utilization of Virus against Injurious Insects and the Possible Adaptation of This Method to Control Insect Disease Vectors. *Bull World Health Organ* 1964 ; 31 : 513-7.

45. van Munster M, Janssen A, Clerivet A, van den Heuvel J. Can plants use an entomopathogenic virus as a defense against herbivores? *Oecologia* 2005 ; 143 : 396-401.

46. Vayssier-Taussat M, Albina E, Citti C, *et al*. Shifting the paradigm from pathogens to pathobiome: new concepts in the light of meta-omics. *Frontiers in cellular and infection microbiology* 2014 ; 4 : 29.

47. Vendeville A, Ravallec M, Jousset FX, *et al*. Densovirus infectious pathway requires clathrin-mediated endocytosis followed by trafficking to the nucleus. *Journal of virology* 2009 ; 83 : 4678-89.

48. Wang P, Granados RR. Molecular structure of the peritrophic membrane (PM): identification of potential PM target sites for insect control. *Arch Insect Biochem Physiol* 2001 ; 47 : 110-8.

49. Wang Y, Gosselin Grenet AS, Castelli I, *et al*. Densovirus crosses the insect midgut by transcytosis and disturbs the barrier epithelial function. *J Virol* 2013 ; 87 : 12380-91.

50. Zadori Z, Szelei J, Lacoste MC, *et al*. A viral phospholipase A2 is required for parvovirus infectivity. *Developmental cell* 2001 ; 1 : 291-302.

51. Arnold van Huis, Joost Van Itterbeeck, Harmke Klunder, *et al*. Edible insecte. Future prospects for food and feed security. *FAO Forestry Paper (Rome)* 2013 : 171 (www.fao.org/docrep/018/i3253e/i3253e.pdf).

Article de recherche 5

A Novel Itera-Like Densovirus Isolated by Viral Metagenomics from the Sea Barley *Hordeum marinum*

Sarah François[1,2], Pauline. Bernardo[3], Denis Filloux[3], Philippe Roumagnac[3], Nicole Yaverkovski[4], Rémy Froissart[5], Mylène Ogliastro[2]

[1] Université de Montpellier, UMR 1333 DGIMI « Diversité, Génomes et Interactions Microorganismes-Insectes », place Eugène-Bataillon, 34095 Montpellier cedex 5, France

[2] Laboratoire « Diversité, Génomes et Interactions Microorganismes Insectes » (DGIMI) UMR 1333, INRA, Université Montpellier, 34095 Montpellier, France.

[3] Laboratoire « Biologie et Génétique des Interactions Plante-Parasite » UMR BGPI, CIRAD-INRA-SupAgro, Campus International de Montferrier-Baillarguet, Montpellier Cedex-5, France.

[4] Service Écologie Végétale, Fondation Tour du Valat, Le Sambuc, Montpellier, France

[5] Laboratoire « Maladies Infectieuses et Vecteurs: Écologie, Génétique, Évolution et Contrôle » (MIVEGEC) UMR 5290, CNRS, IRD, Université Montpellier, 911 avenue Agropolis, 34394 Montpellier, France.

Publié le 4 décembre 2014 dans **Genome Announc.** (2(6):e01196-14)

genomeA
Journals.ASM.org

A Novel Itera-Like Densovirus Isolated by Viral Metagenomics from the Sea Barley *Hordeum marinum*

S. François,[a] P. Bernardo,[b] D. Filloux,[b] P. Roumagnac,[b] N. Yaverkovski,[c] R. Froissart,[b,d] M. Ogliastro[a]

INRA, UMR 1333, DGIMI, Montpellier, France[a]; INRA-CIRAD-SupAgro, UMR 385, BGPI, Campus International de Baillarguet, Montpellier, France[b]; Service Écologie Végétale, Fondation Tour du Valat, Le Sambuc, Montpellier, France[c]; CNRS-IRD-UM1-UM2, UMR 5290, MIVEGEC, Montpellier, France[d]

S.F. and P.B. contributed equally to the work.

Densoviruses (DVs) infect arthropods and belong to the *Parvoviridae* family. Here, we report the complete coding sequence of a novel DV isolated from the plant *Hordeum marinum* (*Poaceae*) by viral metagenomics, and we confirmed reamplification by PCR. Phylogenetic analyses showed that this novel DV is related to the genus *Iteradensovirus*.

Received 7 October 2014 **Accepted** 24 October 2014 **Published** 4 December 2014

Citation François S, Bernardo P, Filloux D, Roumagnac P, Yaverkovski N, Froissart R, Ogliastro M. 2014. A novel itera-like densovirus isolated by viral metagenomics from the sea barley *Hordeum marinum*. Genome Announc. 2(6):e01196-14. doi:10.1128/genomeA.01196-14.

Copyright © 2014 François et al. This is an open-access article distributed under the terms of the Creative Commons Attribution 3.0 Unported license.

Address correspondence to M. Ogliastro, ogliastr@supagro.inra.fr.

Densoviruses (DVs) are small nonenveloped icosahedral viruses infecting arthropods, including pests and vectors for which they are considered biocontrol agents. They contain a single-strand linear DNA genome ranging from 4 to 6 kb, ended by inverted terminal repeats (ITRs) (1). Only 15 DV species are referenced in GenBank so far (2); they display a large diversity of sequences, structures, and organizations. Such diversity, together with the diversity of their invertebrate hosts, suggests that DVs are largely unknown and ubiquitous in the environment. It is crucial to understand the densovirus diversity and prevalence for both fundamental and applied virology issues.

A novel densovirus was detected from sea barley (*Hordeum marinum*) using a virion-associated nucleic acid (VANA) viral metagenomic approach (3). To complement this genome, we performed Rapid Amplification of cDNA Ends (RACE) (Roche), and the products were cloned in the pGEM-T Easy Vector (Promega) and sequenced. The sequences were assembled using Geneious 7.1.4 and compared to database sequences using BLASTn, BLASTp, and tBLASTx (4). The results were considered to be indicative of significant homology when BLAST *E* values were $<10^{-3}$. The genome of this novel DV consists of 4,734 nucleotides (nt), with short ITRs of 130 and 77 nucleotides (nt) at the 3′ and 5′ ends, respectively. The iteravirus genome size is about 5 kb, with ITRs of 250 nt, suggesting that the ITRs of this novel densovirus are not complete (5). The genomic organization of this densovirus is monosense, with three predicted intronless open reading frames (ORFs) encoding two nonstructural proteins (NS) and one structural protein (VP). ORF1 (nt 253 to 2505) has a coding capacity of 750 amino acids (aa) and contains the typical nonstructural 1 (NS1) helicase superfamily III. ORF2 (nt 2559 to 4568) encodes a 669-aa protein corresponding to VP, and it contains the characteristic phospholipase A2 motif (6). ORF3 (nt 380 to 1729) has a coding capacity for NS2 of 449 aa and typically overlapped NS1. The alignment of the VP and NS protein sequences using Clustal W 1.8.1 (7) revealed that this genome had the highest identity (84.9%) with *Danaus plexippus plexippus* densovirus (DpplDV) (GenBank accession no. KF963252) (8). This genome was independently purified from leaves of the original plant stored at −80°C (Qiagen plant DNeasy kit). The PCR products were obtained from different leaves using 15 pairs of primers covering the whole genome that were sequenced using Sanger's method (Cogenics). Recombination analyses using RDP4.18 (9) revealed that this novel DV might result from an intragenus recombination event between DpplDV and *Dendrolimus punctatus* densovirus (DpDV).

No insect has been found in any part of this plant, no reads obtained from this plant were assigned to arthropods, and no products were obtained using an insect DNA bar coding based on the PCR amplification of a fragment of the mitochondrial cytochrome *c* oxidase subunit I gene (10). This densovirus might come from contamination of the plant aerial part by infected arthropods or circulate systemically *in planta*, as already reported (11). This virus was tentatively named *H. marinum* densovirus (HormaDV).

Nucleotide sequence accession number. The GenBank accession no. of HormaDV is KM576800.

ACKNOWLEDGMENTS

S.F. was supported by a scholarship from the Institut National de la Recherche Agronomique (INRA). A fellowship to P.B. was funded by the Languedoc-Roussillon Region and the DGA (Département Général des Armées, France). R.F. acknowledges the support of the Center National de Recherche Scientifique (CNRS) and the Institut de Recherche pour le Développement (IRD).

REFERENCES

1. **Bergoin M, Tijssen P.** 2010. Densoviruses: a highly diverse group of arthropod parvoviruses, p 59–72. *In* Asgari S, Johnson KN (ed), Insect virology. Caster Academic Press, Norwich, United Kingdom.
2. **Cotmore SF, Agbandje-McKenna M, Chiorini JA, Mukha DV, Pintel DJ, Qiu J, Soderlund-Venermo M, Tattersall P, Tijssen P, Gatherer D, Davison AJ.** 2014. The family *Parvoviridae*. Arch. Virol. **159**:1239–1247. http://dx.doi.org/10.1007/s00705-013-1914-1.
3. **Candresse T, Filloux D, Muhire B, Julian C, Galzi S, Fort G, Bernardo P, Daugrois J-H, Fernandez E, Martin DP, Varsani A, Roumagnac P.** 2014. Appearances can be deceptive: revealing a hidden viral infection

with deep sequencing in a plant quarantine context. PLoS One **9:**e102945. http://dx.doi.org/10.1371/journal.pone.0102945.

4. **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 analysis of sequence data. Bioinformatics **28:**1647–1649. http://dx.doi.org/10.1093/bioinformatics/bts199.
5. **Yu Q, Tijssen P.** 2014. Gene expression of five different iteradensoviruses: BmDV, CeDV, PpDV, SfDV and DpIDV. J. Virol. **88:**12152–12157. http://dx.doi.org/10.1128/JVI.01719-14.
6. **Zádori Z, Szelei J, Lacoste MC, Li Y, Gariépy S, Raymond P, Allaire M, Nabi IR, Tijssen P.** 2001. A viral phospholipase A2 is required for parvovirus infectivity. Dev. Cell **1:**291–302. http://dx.doi.org/10.1016/S1534-5807(01)00031-4.
7. **Larkin MA, Blackshields G, Brown NP, Chenna R, McGettigan PA, McWilliam H, Valentin F, Wallace IM, Wilm A, Lopez R, Thompson JD, Gibson TJ, Higgins DG.** 2007. Clustal W and Clustal X version 2.0. Bioinformatics **23:**2947–2948. http://dx.doi.org/10.1093/bioinformatics/btm404.
8. **Yu Q, Tijssen P, Khlebnikova TA, Pikul' DA.** 2014. Iteradensovirus from the monarch butterfly, *Danaus plexippus plexippus*. Med Parazitol (Mosk) **2:**6–7.
9. **Martin DP, Lemey P, Lott M, Moulton V, Posada D, Lefeuvre P.** 2010. RDP3: a flexible and fast computer program for analyzing recombination. Bioinformatics **26:**2462–2463. http://dx.doi.org/10.1093/bioinformatics/btq467.
10. **Wilson JJ.** 2012. DNA bar codes for insects. Methods Mol. Biol. **858:**17–46. http://dx.doi.org/10.1007/978-1-61779-591-6_3.
11. **van Munster M, Janssen A, Clérivet A, van den Heuvel J.** 2005. Can plants use an entomopathogenic virus as a defense against herbivores? Oecologia **143:**396–401. http://dx.doi.org/10.1007/s00442-004-1818-6.

Article de recherche 6

Molecular characterization and prevalence of two capulaviruses: Alfalfa leaf curl virus from France and Euphorbia caput-medusae latent virus from South Africa

Pauline Bernardo[1*], Brejnev Muhire[2*], **Sarah François**[1,3,4], Maëlle Deshoux[1], Penelope Hartnady[2], Kata Farkas[5], Simona Kraberger[5], Denis Filloux[1], Emmanuel Fernandez[1], Serge Galzi[1], Romain Ferdinand[1], Martine Granier[1], Armelle Marais[6,7], Pablo Monge Blasco[8], Thierry Candresse[6,7], Fernando Escriu[8,9], Arvind Varsani[5,10,11], Gordon W. Harkins[12], Darren P. Martin[2], Philippe Roumagnac[1]

[1] CIRAD-INRA-SupAgro, UMR BGPI, Campus International de Montferrier Baillarguet, Montpellier Cedex-5, France

[2] Computational Biology Group, Institute of Infectious Disease and Molecular Medicine, Faculty of Health Sciences, University of Cape Town, Observatory, South Africa

[3] INRA, UMR 1333, DGIMI, Montpellier, France

[4] CNRS-IRD-UM1-UM2, UMR 5290, MIVEGEC, Avenue Agropolis, Montpellier, France

[5] School of Biological Sciences and Biomolecular Interaction Centre, University of Canterbury, Private Bag 4800, Christchurch, New Zealand

[6] INRA, UMR 1332 Biologie du Fruit et Pathologie, Villenave d'Ornon Cedex, France

[7] Université de Bordeaux,UMR 1332 Biologie du Fruit et Pathologie, Villenave d'Ornon Cedex, France

[8] Unidad de Sanidad Vegetal, Centro de Investigaciony Tecnologıa Agroalimentaria de Aragon (CITA), Av. Montañana 930, 50059 Zaragoza, Spain

[9] Unidad de Sanidad Vegetal, Instituto Agroalimentario de Aragón IA2 (CITA-Universidad de Zaragoza), Av. Montañana 930, 50059 Zaragoza, Spain

[10] Department of Plant Pathology and Emerging Pathogens Institute, University of Florida, Gainesville, USA

[11] Structural Biology Research Unit, Department of Clinical Laboratory Sciences, University of Cape Town, Observatory, South Africa

[12] South African National Bioinformatics Institute, MRC Unit for Bioinformatics Capacity Development, University of the Western Cape, Cape Town, South Africa

Publié le 31 mars 2016 dans **Virology** (493 142–153)

Virology 493 (2016) 142–153

Contents lists available at ScienceDirect

Virology

journal homepage: www.elsevier.com/locate/yviro

Molecular characterization and prevalence of two capulaviruses: Alfalfa leaf curl virus from France and Euphorbia caput-medusae latent virus from South Africa

Pauline Bernardo [a,1], Brejnev Muhire [b], Sarah François [a,c,d], Maëlle Deshoux [a], Penelope Hartnady [b], Kata Farkas [e], Simona Kraberger [e], Denis Filloux [a], Emmanuel Fernandez [a], Serge Galzi [a], Romain Ferdinand [a], Martine Granier [a], Armelle Marais [f,g], Pablo Monge Blasco [h], Thierry Candresse [f,g], Fernando Escriu [h,i], Arvind Varsani [e,j,k], Gordon W. Harkins [l], Darren P. Martin [b], Philippe Roumagnac [a,*]

[a] CIRAD-INRA-SupAgro, UMR BGPI, Campus International de Montferrier-Baillarguet, Montpellier Cedex-5, France
[b] Computational Biology Group, Institute of Infectious Disease and Molecular Medicine, Faculty of Health Sciences, University of Cape Town, Observatory, South Africa
[c] INRA, UMR 1333, DGIMI, Montpellier, France
[d] CNRS-IRD-UM1-UM2, UMR 5290, MIVEGEC, Avenue Agropolis, Montpellier, France
[e] School of Biological Sciences and Biomolecular Interaction Centre, University of Canterbury, Private Bag 4800, Christchurch, New Zealand
[f] INRA, UMR 1332 Biologie du Fruit et Pathologie, Villenave d'Ornon Cedex, France
[g] Université de Bordeaux, UMR 1332 Biologie du Fruit et Pathologie, Villenave d'Ornon Cedex, France
[h] Unidad de Sanidad Vegetal, Centro de Investigacion y Tecnologıa Agroalimentaria de Aragon (CITA), Av. Montañana 930, 50059 Zaragoza, Spain
[i] Unidad de Sanidad Vegetal, Instituto Agroalimentario de Aragón IA2 (CITA - Universidad de Zaragoza), Av. Montañana 930, 50059 Zaragoza, Spain
[j] Department of Plant Pathology and Emerging Pathogens Institute, University of Florida, Gainesville, USA
[k] Structural Biology Research Unit, Department of Clinical Laboratory Sciences, University of Cape Town, Observatory, South Africa
[l] South African National Bioinformatics Institute, MRC Unit for Bioinformatics Capacity Development, University of the Western Cape, Cape Town, South Africa

ARTICLE INFO

Article history:
Received 20 December 2015
Returned to author for revisions
22 March 2016
Accepted 23 March 2016

Keywords:
Geminiviridae
Alfalfa leaf curl virus
Euphorbia caput-medusae latent virus
Prevalence
Recombination
Secondary structure
Genome organization
France
Spain
South Africa

ABSTRACT

Little is known about the prevalence, diversity, evolutionary processes, genomic structures and population dynamics of viruses in the divergent geminivirus lineage known as the capulaviruses. We determined and analyzed full genome sequences of 13 Euphorbia caput-medusae latent virus (EcmLV) and 26 Alfalfa leaf curl virus (ALCV) isolates, and partial genome sequences of 23 EcmLV and 37 ALCV isolates. While EcmLV was asymptomatic in uncultivated southern African *Euphorbia caput-medusae*, severe alfalfa disease symptoms were associated with ALCV in southern France. The prevalence of both viruses exceeded 10% in their respective hosts. Besides using patterns of detectable negative selection to identify ORFs that are probably functionally expressed, we show that ALCV and EcmLV both display evidence of inter-species recombination and biologically functional genomic secondary structures. Finally, we show that whereas the EcmLV populations likely experience restricted geographical dispersion, ALCV is probably freely moving across the French Mediterranean region.

© 2016 Elsevier Inc. All rights reserved.

1. Introduction

Next Generation Sequencing and metagenomics-based study designs have impacted our appreciation of the prevalence, pervasiveness and diversity of environmental single stranded DNA (ssDNA) viruses (Candresse et al., 2014; Filloux et al., 2015b; Kraberger et al., 2015; Ng et al., 2014, 2009, 2011; Roossinck et al., 2015; Rosario and Breitbart, 2011). Among the best studied of these ssDNA viruses have been the plant-infecting viruses in the family *Geminiviridae*. The past four decades have seen the worldwide emergence of several major plant diseases caused by geminiviruses and also the discovery of hundreds of previously unknown, and sometimes highly divergent, geminiviral species (Bernardo et al., 2013; Briddon et al., 2010; Liang et al., 2015; Loconsole et al.,

* Corresponding author.
E-mail address: philippe.roumagnac@cirad.fr (P. Roumagnac).
[1] P.B. and B.M. contributed equally to this work.

http://dx.doi.org/10.1016/j.virol.2016.03.016
0042-6822/© 2016 Elsevier Inc. All rights reserved.

2012; Ma et al., 2015; Roumagnac et al., 2015; Varsani et al., 2009; Yazdi et al., 2008). The rate at which new geminiviruses are being discovered has recently been accelerated through the development and application of sequence-non-specific virus discovery approaches (Roossinck et al., 2015; Rosario et al., 2012) such as rolling circle amplification (RCA) based virus cloning and sequencing (Haible et al., 2006; Inoue-Nagata et al., 2004; Shepherd et al., 2008) and VANA (virion-associated nucleic acids) or siRNA based metagenomics (Candresse et al., 2014; Filloux et al., 2015a; Kreuze et al., 2009). In studies exploring the diversity of these viruses, such approaches have facilitated the expansion of sampling efforts to include both insects (both virus vectors and their predators) and uncultivated host species (Bernardo et al., 2013; Ng et al., 2011; Rosario et al., 2011, 2015). The application of these improved sampling and virus discovery strategies have revealed that geminivirus diversity exceeds that which is currently known (Haible et al., 2006; Ng et al., 2011; Schubert et al., 2007). Such studies have also led to a reevaluation of the known geographical ranges of the different geminivirus genera with, for example, the overturning of the long-held view that members of the genus *Mastrevirus* do not occur in the Americas (Agindotan et al., 2015; Candresse et al., 2014; Kreuze et al., 2009).

Since some of the geminiviruses discovered over the past few years are highly divergent, and, in some cases, have unique genome architectures (Briddon et al., 2010; Loconsole et al., 2012; Varsani et al., 2009; Yazdi et al., 2008), three new geminivirus genera were created and approved in 2014 (*Becurtovirus*, *Turncurtovirus*, and *Eragrovirus*; Varsani et al., 2014b) and at least three others are likely to follow (Bernardo et al., 2013; Krenz et al., 2012; Loconsole et al., 2012); among which are the capulaviruses (Bernardo et al., 2013). Three distinct species of this new lineage were discovered between 2010 and 2011 infecting, respectively, a wild spurge, *Euphorbia caput-medusae* in South Africa (Euphorbia caput-medusae latent virus, EcmLV; Bernardo et al., 2013), alfalfa (*Medicago sativa*) in France (Alfalfa leaf curl virus, ALCV; Roumagnac et al., 2015), and French bean (*Phaseolus vulgaris*) in India (French bean severe leaf curl virus, FbSLCV; Accession number NC_018453). In addition to their high degree of sequence divergence, these viruses also exhibit a genome organization that is unique among the geminiviruses (Bernardo et al., 2013; Roumagnac et al., 2015). ALCV and FbSLCV cause severe symptoms in alfalfa and French bean, respectively, and it has recently been shown that ALCV is transmitted by *Aphis craccivora* (Roumagnac et al., 2015); an invasive aphid species with an almost global distribution (CIE, 1983).

We here collected hundreds of alfalfa and *E. caput-medusae* plants from France and South Africa, respectively, and comparatively analyze the genome sequences of 13 isolates of EcmLV and 26 ALCV along with that of FbSLCV. We show that both EcmLV and ALCV have high prevalence (12–13%) in their respective host species in the Western Cape region of South Africa and in three Southern regions of France. We also present a new codon-model based natural selection detection approach to reveal open reading frames that are probably functionally expressed in capulaviruses genomes. We further demonstrate that, as with other geminiviruses, capulavirus genomes display evidence of both inter-species recombination and biologically functional secondary structures.

2. Materials and methods

2.1. Plant sampling

In 2014, 238 *M. sativa* plants were randomly collected (i.e. irrespective of the presence of potential symptoms) from three regions of Southern France, including the Rhône delta (seven sampling locations), the Montpellier region (four sampling locations) and the Toulouse region (two sampling locations; Supplementary Fig. 1). The symptom status of the 238 plants was assessed and plants were then stored at −80 °C prior to virus detection and characterization. A further 43 symptomatic and 15 asymptomatic *M. sativa* plants were collected later in 2014 from the same three areas (Supplementary Fig. 1). An additional four symptomatic alfalfa plants collected in 2012 and 2013 from the Ebro valley region of the Zaragoza province of Spain were also included for further analysis. Collectively, 300 alfalfa plants were obtained from France and Spain between 2012 and 2014.

In 2012, 302 asymptomatic *Euphorbia caput-medusae* were randomly collected from seven separate locations within the Western Cape region of South Africa (Supplementary Fig. 1). These samples were stored at −80 °C. In 2015, 14 additional samples were collected from an eighth location at the University of the Western Cape Nature Reserve (Supplementary Fig. 1).

2.2. DNA extraction, amplification, cloning and sequencing

Total DNA from alfalfa and *E. caput-medusae* plant samples was extracted as previously described by Bernardo et al. (2013).

PCR-mediated detection of ALCV from the 296 alfalfa plants collected in France, and four plants from Spain was performed using two pairs of PCR primers (ALCV-187F forward primer 5'-TGG AAT ATT GTG CTG CTT GG-3' and ALCV-971R reverse primer 5'-ATT TTG GGA CTT GTG CTC CA-3'; and ALCV-986F forward primer 5'-ATG ATG GAT AAT TCA AAC CC-3' and ALCV-1202R reverse primer 5'-TTC TTC TGG GTA TTT GCA TA-3'). Amplification conditions consisted of 94 °C for 2 min; 30 cycles at 94 °C for 1 min, 58 °C for 1 min (primer pair 1)/55 °C for 30 s (primer pair 2), 72 °C for 50 s; and 72 °C for 5 min. Amplicons were directly sequenced using automated Sanger sequencing (Beckman Coulter Genomics). Circular DNA molecules from samples that tested positive by at least one of the two PCR assays were enriched using RCA (using TempliPhi™, GE Healthcare, USA) as previously described (Shepherd et al., 2008). The RCA products were used as a template for PCR using an abutting pair of primers designed from the 44-1E ALCV complete genome (Accession number KP732474; Roumagnac et al., 2015); Cap-ncoIF: 5'-CCA TGG CCT TCA AAG GTA GCC CAA TTC AAY ATG G-3' and Cap-ncoIR: 5'-CCA TGG GGC CTT ATY CCT CKG YGA TCG-3' using KAPA HiFi Hotstart DNA polymerase (Kapa Biosystems, USA). Amplification conditions consisted of: 96 °C for 3 min, 25 cycles at 98 °C for 20 s, 60 °C for 30 s, 72 °C for 165 s, and 72 °C for 3 min. The amplicons were gel purified, cloned into pJET2.1 (Thermo Fisher, USA) and Sanger sequenced by primer walking at Macrogen Inc. (Korea). In addition, RCA products were digested with *Eco*RI, *Bam*HI, *Dra*I, *Nco*I or *Nde*I for 3 h at 37 °C in order to screen for the presence of a potential DNA-B geminiviral component or satellite sequences.

PCR-mediated detection of EcmLV from the 316 *E. caput-medusae* plants was performed using two pairs of PCR primers: (i) Dar-136F forward primer 5'-CGA AGA GGT CAT TGG GAC AT-3' and Dar-730R reverse primer 5'-CGG GTC TGG CTA AGA GAG TG-3' as previously described by Bernardo et al. (2013) and (ii) Dar-1775F forward primer 5'-TTG AAT TGC ATG GGC ACT TA-3' and Dar-2433R reverse primer 5'-GCC CTT TTG GTC ATT TTG AA-3'. Amplification conditions consisted of: 95 °C for 5 min; 30 cycles at 94 °C for 1 min, 56 °C for 1 min, 72 °C for 50 s; and 72 °C for 5 min. Circular DNA molecules from samples that tested positive by at least one of the two PCR assays was enriched using RCA as described above for alfalfa. RCA products were all digested with *Eco*RI for 3 h at 37 °C. Subsequently, samples that could not be cleaved using *Eco*RI were digested with *Bam*HI for 3 h at 37 °C. Geminivirus-like genomes from 13 *E. caput-medusae* samples were cloned in pGEM-T Easy (Promega Biotech) using methods described by Bernardo et al. (2013).

Prevalence was defined as the proportion of alfalfa or *E. caput-medusae* plants being infected by ALCV or EcmLV from the alfalfa or *E. caput-medusae* populations that were randomly collected in France or South Africa, respectively.

2.3. Sequence analysis

Sequence contigs were assembled using BioNumerics Applied Maths V6.5 (Applied. Maths, Ghent, Belgium) and were compared to sequences in the GenBank database using BLASTn and BLASTx (Altschul et al., 1990). All pairwise identity analyses of the full genome nucleotide sequences, capsid protein (CP) amino acid sequences, and replication associated protein (Rep) amino acid sequences were carried out using the MUSCLE-based pairwise alignment (Edgar, 2004) option implemented in SDT v1.2 (Muhire et al., 2014b).

2.4. Detection of conserved secondary-structural elements within capulavirus genomes

The computer program Nucleic Acid Secondary Structure Predictor (NASP) (Semegni et al., 2011) was used as previously described (Muhire et al., 2014a) to identify the conserved secondary-structural elements present within 45 capulavirus genomes (EcmLV, $n=16$; ALCV, $n=27$; FbSLCV, $n=2$). In each of the data sets, secondary-structural elements were first inferred using a minimum free-energy (MFE) approach implemented in hybrid-ssmin (a component of the UNAFold package; (Markham and Zuker, 2008). From amongst sets of plausible whole genome secondary structures (approximately ~10 alternative folds per genome) NASP identified subsets of conserved high-confidence structural elements - referred to as high-confidence structure sets (HCSSs). For the NASP analysis, sequences were folded as circular ssDNA at 25 °C under 1 M sodium. In subsequent analyses, only nucleotides identified as being base-paired within the HCSSs were treated as paired sites, whereas all other nucleotides were treated as unpaired sites. The ALCV and EcmLV datasets contained enough sequences that were sufficiently divergent to test for evidence of evolutionary pressures favoring the maintenance of base–pairing interactions within the HCSSs (Muhire et al., 2014a). Three different tests, all designed to test whether sequences were evolving in a way consistent with the evolutionary preservation of biologically functional structural elements, were applied exactly as recently described (Muhire et al., 2014a). First, at the whole-genome-scale an allele frequency spectrum permutation test (Fu and Li, 1993; Tajima, 1989), which compares frequencies of alternative alleles at paired vs unpaired sites, was used to compare degrees of negative selection (selection disfavoring change) at paired vs unpaired sites. Second, within the CP and Rep gene coding regions, a Maximum Likelihood codon-model based test (based on the FUBAR method (Murrell et al., 2013) implemented in HyPhy (Pond et al., 2005) was used to compare synonymous substitution rates in codons containing paired and unpaired third codon position nucleotides. Third, the complementary coevolution detection method (Muhire et al., 2014a) (based on the SPIDERMONKEY method (Poon et al., 2008), also implemented in HyPhy) was used to test whether pairs of nucleotides that were base-paired within the HCSSs also displayed any evidence of complementary coevolution (i.e. coevolution specifically favoring the maintenance of base-pairing). Structures were visualized and, based on the various analyses performed, ranked in order of their likely biological functionality using the computer program, DOOSS (http://dooss.computingforbiology.org; Golden and Martin, 2013).

2.5. Identification of open reading frames (ORFs) and capulavirus genome organizations

Identification of open reading frames (ORFs) was initially performed using the ORF Finder ncbi graphical analysis tool (http://www.ncbi.nlm.nih.gov/gorf/gorf.html). In addition, we devised a computational tool, named "ORFunc" that detects all ORFs above a user defined length that are both conserved within a multiple sequence alignment (in this case an alignment of all available capulavirus sequences) and are evolving under a detectable degree of purifying selection (i.e. selection that disfavoring non-synonymous substitutions). ORFunc ranks ORF in order of their likely functional expression based on the degree of detected purifying selection against changes in the potentially encoded amino acid sequence. Specifically, given a multiple sequence alignment in FASTA format, ORFunc performs a preliminary scan of the alignment to generate a list of unique ORFs sharing greater than a predetermined pairwise sequence identity (in the present analyses 0.80 to ensure the sequences within each of the resulting sub-alignments were credibly aligned) and above a given sequence length (in the present analyses 100nt; a size slightly below that of any currently known geminivirus gene). A local BLAST search (Altschul et al., 1990) is performed using the command-line version of NCBI-BLAST (available from ftp://ftp.ncbi.nlm.nih.gov/blast/executables/blast+/) whereby each ORF is used as a query to search a local BLAST database produced from the input alignment so as to obtain matches for other sequences: matches that are then used to produce a codon alignment. For each codon alignment thus produced, synonymous substitution rates (dS), non-synonymous substitution rates (dN) and the probability of negative selection ($dS > dN$) at each codon site are estimated using the FUBAR method (Murrell et al., 2013). This approach enabled the determination of whether inferred ORFs within the various EcmLV and ALCV genomes were evolving under negative selection (identified by their constituent codons on average displaying significantly higher synonymous substitution rates than non-synonymous substitution rates). ORF-wide probabilities of negative selection are then estimated by combining the p-values of non-negative selection at each individual codon-site (Theiler and Bloch, 1996). To test the reliability of this approach, two geminivirus datasets for which genes have already been well-characterized (*Tomato yellow leaf curl virus* [TYLCV] and *Maize streak virus* [MSV]; Dry et al., 1993; Lazarowitz et al., 1989) were used as controls. ORFunc is written in python and is available from: web.cbio.uct.ac.za/~brejnev/downloads/ORFunc.tar.gz

2.6. Analysis of potential recombination events

Evidence of potential recombination events was detected within the 45 capulavirus full-genome alignment using the RDP, GENECONV, BOOTSCAN, MAXIMUM CHI SQUARE, CHIMAERA, SISCAN and 3SEQ recombination detection methods that are implemented in RDP4.50 with default settings (Martin et al., 2015). Only recombination events detected with three or more distinct groups of methods (where RDP and GENECONV were each considered as distinct methods and BOOTSCAN/SISCAN and MAXIMUM CHI SQUARE/CHIMAERA/3SEQ were considered as single distinct method groups) and that also had significant phylogenetic support, were considered credible evidence of recombination.

2.7. Phylogenetic analysis

The evolutionary relationships of capulaviruses and other geminiviruses were reconstructed using Rep and CP amino acid sequences (i.e. the only proteins that are clearly homologous across all geminiviruses). Datasets consisting of 45 predicted capulavirus Rep and CP amino acid sequences together with the corresponding homologous sequences from Grapevine cabernet franc-associated virus (GCFaV, JQ901105; Krenz et al., 2012) chosen as a divergent non-capulavirus outlier used to root the Rep and CP phylogenies. Predicted Rep and CP amino acid sequences were aligned using MUSCLE. Maximum likelihood phylogenetic trees of the Rep and CP were inferred using PhyML3 (Guindon et al., 2009) implemented in MEGA with the rtREV+G+F amino acid substitution model chosen as the best-fit using ProtTest (Abascal et al., 2005). Five hundred bootstrap replicates were used to test the support of branches. In addition, 64 ALCV gene fragments (ALCV 44-1 genomic positions 260–806) 547 nt in length encompassing the V3 ORF and part of the

cp ORF were aligned using the MUSCLE method (Edgar, 2004) implemented in MEGA (with default settings). A maximum likelihood phylogenetic tree was constructed using PhyML3 with a K2+G+I nucleotide substitution model (selected as best fit by MEGA) and 500 bootstrap replicates were used to test the support of branches. Finally, all 45 currently available whole capulaviruses genome sequences were aligned using MUSCLE. A maximum likelihood phylogenetic tree was constructed using PhyML3 with a TN93+G nucleotide substitution model (selected as best fit by MEGA) with 500 bootstrap replicates used to test the support of branches.

2.8. Statistical analyses

Mantel (1967) tests conducted using XLSTAT (10.000 permutations) were used to test for evidence of correlation between genetic and geographic distances for 39 × 508 nt long EcmLV gene fragments (EcmLV-Dar10 genomic positions 1877–2384) encompassing the C3 ORF and part of the C1 ORF and for 41 × 547 nt long ALCV gene fragments (ALCV 44-1 genomic positions 260–806) encompassing the V3 ORF and part of the *cp* ORF. Samples collected at the regional scale (Southern regions of France for ALCV and Western Cape region for EcmLV) were used (Supplementary Fig. 1). The genetic distance matrix was obtained using MEGA 5.2.1 (CLUSTALW alignment followed by uncorrected pairwise distance estimation using the pairwise deletion option) and the geographic distance matrix was obtained using the program Geographic Distance Matrix Simulator 1.2.3 (http://biodiversityinformatics.amnh.org/open_source/gdmg).

3. Results and discussion

3.1. Characterization of a collection of ALCV isolates from France

Based on the discovery of ALCV in 2010 (Roumagnac et al., 2015), broad sampling surveys in three regions of Southern France and in one region of Spain were conducted during 2012–2014 (Supplementary Fig. 1). PCR analysis of the collected samples revealed that ALCV was present in all four regions. Because the virus was detected in 32/238 plants that were randomly collected in France the percentage of infected plants (13.4%) is likely a valid estimate of the prevalence of ALCV in alfalfa across the three French regions (Supplementary Fig. 1). The prevalence of ALCV at each of the thirteen French sampling sites ranged from 1.2% (La Tour du Valat, Rhône delta region) to 45.8% (Petit Bastières, Rhône delta region).

Visual comparisons of the 32 ALCV-positive plants with plants that tested negative revealed that, relative to non-infected plants, all ALCV-infected plants were stunted and consistently displayed varying degrees of leaf curling, crumpling and shriveling (Supplementary Fig. 2). These potential symptoms unequivocally resemble those observed in alfalfa plants infected by aphid-inoculation with the 44–1E infectious clone of ALCV (Roumagnac et al., 2015). It is noteworthy that enations such as those observed in 44–1E agroinoculated faba beans (Roumagnac et al., 2015) were neither observed in the sampled plants, nor in 44–1E aphid-inoculated alfalfa plants (Roumagnac et al., 2015). Based on the presence or absence of these conspicuous symptoms, an additional 43 symptomatic and 15 asymptomatic plants were collected later in 2014 from the three previously sampled regions of Southern France (Supplementary Fig. 1). A diagnostic PCR detected ALCV in 39/43 of the symptomatic plants but in none of the asymptomatic plants, suggesting that the symptoms observed in the field are likely caused by ALCV infection.

Twenty-six ALCV complete genome sequences were obtained that ranged in size from 2737 nt to 2769 nt in length and shared > 82.5% genome-wide pairwise identity (Supplementary Fig. 3). This degree of similarity is above the species demarcation thresholds recommended for all of the geminivirus genera (Muhire et al., 2013; Varsani et al., 2014a, 2014b) except for the begomoviruses (which have a species demarcation threshold of 91% (Brown et al., 2015) showing that the 26 isolates from which these genomic sequences were obtained could be reasonably, albeit tentatively, classified as ALCV variants. The circular DNA molecules obtained using RCA were tentatively considered to be the complete genomes of geminiviruses infecting the alfalfa plants, because only one band was resolved by electrophoresis of the digested RCA products.

3.2. Characterization of a collection of EcmLV isolates from South Africa

In order to both assess the diversity of EcmLV genome sequences, and determine the prevalence of EcmLV within the Western Cape province of South Africa, 316 asymptomatic *E. caput medusae* plants from eight dispersed locations were sampled in this region (Supplementary Fig. 1). Using a diagnostic PCR, EcmLV was detected in 38/316 plants sampled in 5/8 of the sampling locations. This indicates an overall EcmLV prevalence of 12% at the Western Cape regional scale: a prevalence very similar to that of 13.4% found for ACLV in alfalfa in Southern France. The local prevalence of EcmLV ranged from 5% (at the Silver Stream sampling site) to 30% (at the Paternoster site).

The 13 EcmLV complete genome sequences that were obtained together with three full genomes previously described (Bernardo et al., 2013) had pairwise identities ranging between 92.8 and 99.7% (Supplementary Fig. 3), confirming that all isolates belong to the same species (Bernardo et al., 2013). As for ALCV, the circular DNA molecules obtained using RCA were tentatively considered to be the complete genomes of geminiviruses infecting the *E. caput-medusae* plants, because only one band was resolved by electrophoresis of the digested RCA products.

3.3. Characterization of capulavirus genome organizations

ORFunc was used to identify the potential functional genes of EcmlV and ALCV, with the well-known geminiviruses *Maize streak virus* (MSV) and *Tomato yellow leaf curl virus* (TYLCV) being used as controls for the validation of this new tool. ORFunc is designed to detect and rank ORFs in order of the likelihood of their functional expression based on overall evidence of negative selection detected across all of their constituent codon sites. It is expected that the majority of functional ORFs should be evolving under some degree of negative selection favoring the maintenance of functionally important amino acid sequences. By implementing such a selection-based approach, ORFunc goes beyond simply annotating genomic fragments delimited by start and stop codons. As a proof-of-concept, ORFunc consistently identified significant degrees of negative selection across all four known functional genes of MSV (Fig. 1 and Supplementary Dataset 1) and the *rep* (C1), transcription activator protein (C2), replication enhancer protein (C3) and coat protein (V1) genes of TYLCV (Fig. 1 and Supplementary Dataset 1). ORFunc did not identify any potentially functional genes other than those illustrated in Fig. 1. In addition, the failure to obtain statistically significant evidence of negative selection in the V2 and C4 ORFs in TYLCV should not be equated with the absence of negative selection in these ORFs. The C4 gene is completely contained within the *rep* gene, which means that very few possible substitutions in this ORF would be synonymous (i.e. almost all possible substitutions that would not change the encoded amino acid sequence encoded by C4 would change the amino acid sequence of Rep and would, therefore, not be synonymous). In the case of V2 whereas 10/134 individual codon sites are apparently evolving under a significant degree of negative selection (with a p-value cutoff of 0.1), 1/134 codon sites are apparently evolving under a significant degree

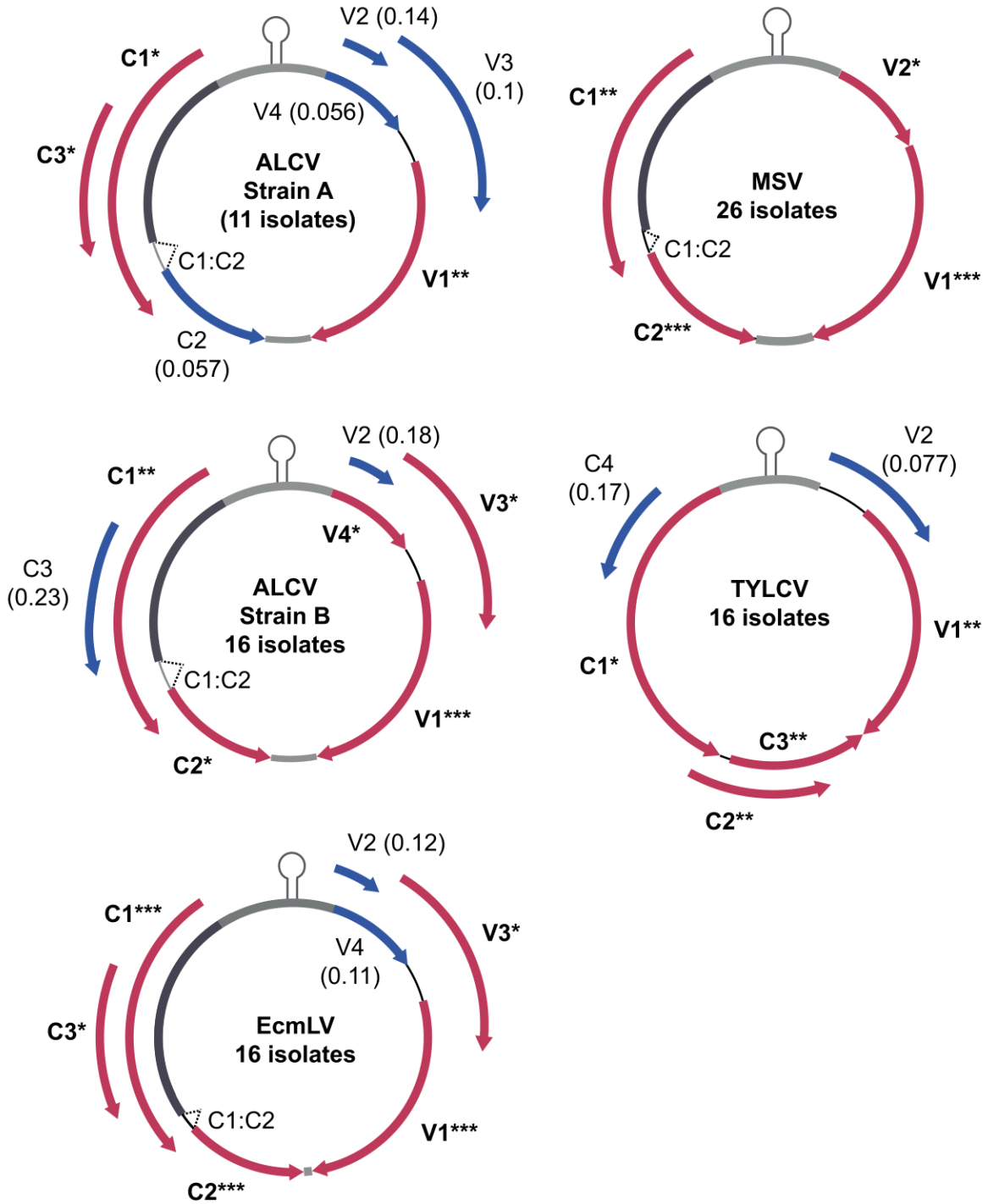


Fig. 1. Genomic organization of ALCV strain A, ALCV strain B, EcmLV, MSV and TYLCV showing the arrangements of potential genes. Open reading frames (ORFs) larger than 100 nucleotides that contain predicted codon sites that are collectively evolving under significant degrees of negative selection (i.e. for which ORF-wide estimates of synonymous substitution rates are significantly higher than non-synonymous substitution rate estimates) are indicated in pink (*=p-value <0.05, **=p-value <0.01, and ***=p-value <0.001; see Supplementary Dataset 1). ORFs containing predicted codons that are not evolving under significant degrees of negative selection ($p>0.05$) are indicated in blue (with the associated p-value for the synonymous substitution rate being lower than or equal to the non-synonymous rate indicated in brackets; see Supplementary Dataset 1).

of positive selection. Most importantly, however, ORFunc yielded no false-positive evidence of functionally expressed ORFs in MSV and TYLCV (Fig. 1 and Supplementary Dataset 1).

Of the seven large ORFs identified by the ncbi ORF Finder graphical analysis tool that are apparently conserved between the various capulavirus genomes (V1–V4 and C1–C3; Fig. 1), three (V1, C1 and C3), five (V1, V3, V4, C1 and C2) and five (V1, V3, C1, C2 and C3) were detectably evolving under significant ORF-wide negative selection (associated p-value <0.05) in ALCV strain A (see below regarding ALCV strain A and B demarcation), ALCV strain B, and EcmLV, respectively (Fig. 1 and Supplementary Dataset 1). The V2 ORF did not display significant ORF-wide evidence of negative selection in any of the three groups, suggesting that this ORF might not be functionally expressed. It should be noted, however, that absence of a significant negative selection signal is not proof that this ORF is not functionally expressed. Bigger and more

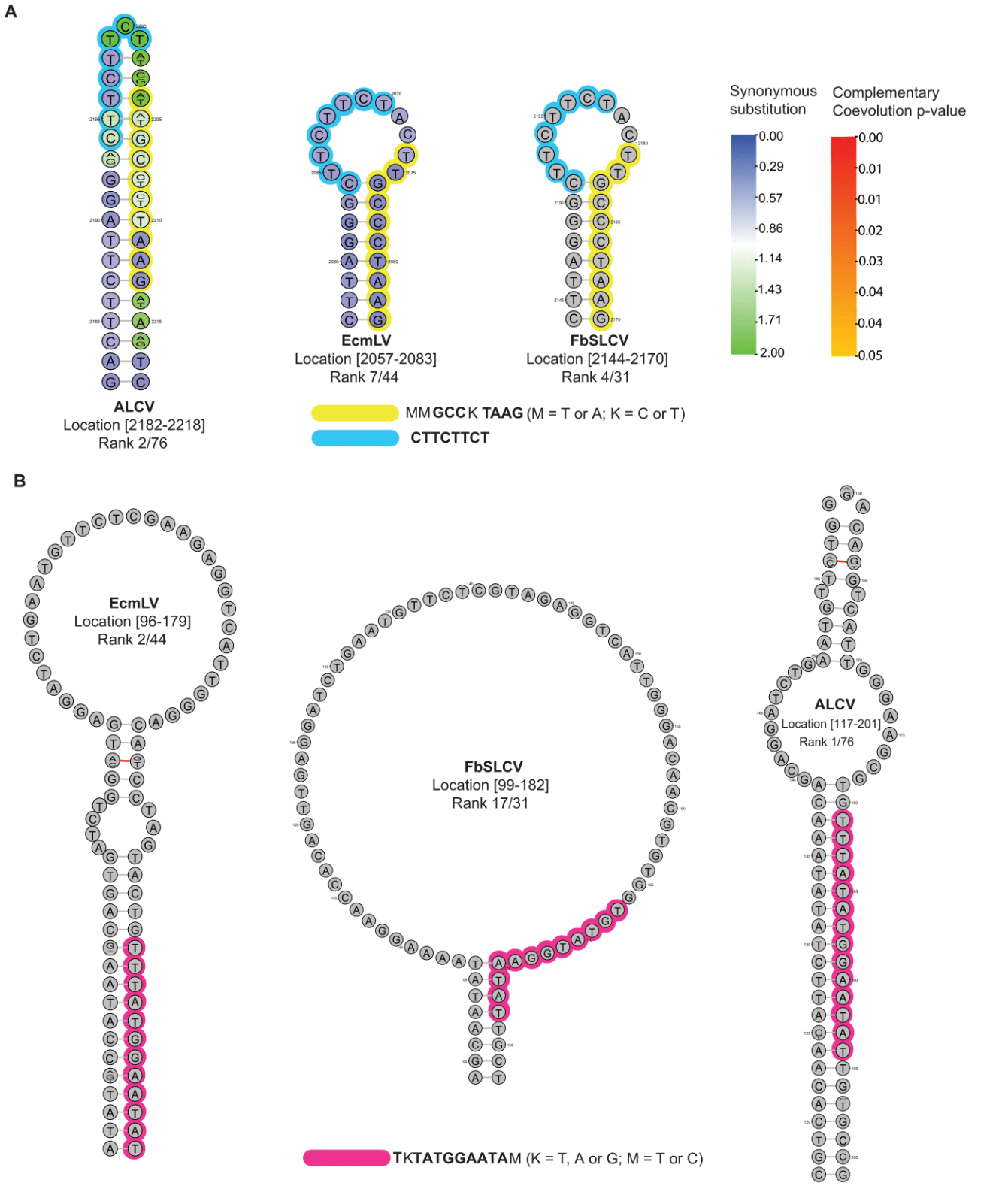


Fig. 2. (A) Secondary structures associated with the 5' end of capulavirus C2 ORFs. The similarities between these structures include homologous stem-loop sequences (highlighted in yellow), and a highly conserved sequence motif found in the ALCV, EcmLV and FbSLCV data sets (highlighted in blue). This structure is ranked highly within the high confidence structure sets of all data sets (seventh out of 44 for EcmLV, second out of 76 for ALCV and fourth out of 31 for FbSLCV). Nucleotide sequence variability at each nucleotide site is reflected by a sequence logo at each position. For the ALCV and EcmLV structures substitution rate estimates are represented by the color of nucleotide triplets falling within individual predicted codon sites (ranging from blue to green). Low synonymous substitution rates are observable in the stem region, which is consistent with a high degree of evolutionary conservation of the nucleotides comprising these codons. Nucleotides falling outside of the gene or for which synonymous substitution rates could not be estimated due to too few sequences being available for analysis (as in the FbSLCV), are shaded gray. (B) Capulavirus secondary structure predicted in the long intergenic region (LIR). The AT-rich stem-loop structure, with a 12 nt conserved sequence (highlighted in pink), is highly ranked in all of the analyzed capulavirus datasets (first out of 76 HCSS structures for ALCV, second out of 44 in EcmLV and 17th out of 31 for FbSLCV). In the case of ALCV and EcmLV, base-paired sites displaying significant degrees of complementary coevolution (P value < 0.05) are indicated by the red lines between the nucleotides in the stem region of each structure. It is plausible that this highly conserved structural element either contains the complementary strand origin of replication or is involved in the regulation of virion and/or complementary gene expression.

diverse capulavirus genomic sequence datasets will substantially increase the power of ORFunc to detect low degrees of negative selection within this ORF: possibly to the point where it is able to confirm the functional expression of V2.

3.4. Detection of conserved secondary-structural elements within capulavirus genomes

Recent computational analyses of geminiviruses in the genera *Mastrevirus* and *Begomovirus* have revealed evidence of evolutionarily conserved (and hence likely biologically functional) genomic secondary structures (Muhire et al., 2014a). Using NASP, we identified 76, 44, and 31 high-confidence structural elements within the ALCV, EcmLV, and FbSLCV genome datasets, respectively.

Besides a conserved stem-loop structure at the presumed virion-strand origin of replication that resembles those found in other geminiviruses (Lazarowitz, 1992), a further 12 uncharacterized but potentially biologically functional genomic and/or mRNA structural elements that are clearly conserved across all three of the analyzed capulavirus species were identified. Amongst these are two particularly interesting uncharacterized structural elements that display high degrees of evolutionary conservation across both the capulaviruses and geminiviruses in other genera (Fig. 2). The first of these structural elements, which is highly ranked within the HCSSs of all three capulavirus species, is a conserved hairpin-loop structure with a 10–20 nt long stem within the C2 ORF (Fig. 2). This structure is possibly homologous to a similarly situated secondary structure previously identified using these same computational methods at the 5' end of MSV C2 ORF.

The second particularly conserved new structural element identified is another hairpin located within the long intergenic region (LIR) of all the analyzed capulavirus species. It contains an A-T rich sequence (Fig. 2) and resembles a structure previously identified in two different mastrevirus species (MSV and *Panicum streak virus*). In diverse ssDNA viruses, virion strand origins of replication (*v-oris*) consist of hairpin structures with highly conserved AT-rich loop sequences that generally occur within intergenic regions (IRs). It is plausible that this second conserved structural element may be either associated with the as yet undiscovered capulavirus complementary strand origin of replication (which in begomoviruses is close to the *v-ori*), or to the regulation of transcription and replication (both of which are known to be regulated by sequence elements within the IRs of various other geminiviruses; Arguello-Astorga et al., 1994; Gutierrez et al., 2004).

It is noteworthy that these analyses also identified what appears to be a repeated 8 bp sequence (with the consensus 5'-AGGCCCAA-3') within the stem regions of multiple structural elements within the HCSSs identified in the various capulavirus datasets. The consensus of the repeated sequence is almost identical to functional sequence motifs previously detected in three other settings. Specifically, it bears a striking resemblance to (i) the 3' ends (3'-AGGCCCA-5') of predicted viral miRNA hairpins (Li et al., 2008); (ii) a regulatory promoter heptamer element involved in the development of plant tissues (5'-AGGCCCA-3') (Obayashi et al., 2007), and (iii) a 7 bp sequence motif (5'-AGGCCCA-3') located upstream of ribosomal protein transcription initiation sites in *Arabidopsis thaliana* (Thompson et al., 1992). It is plausible therefore, that many of the secondary structures identified in these capulavirus genomes (Supplementary Dataset 2) may play a role in modulating the sensitivity of capulavirus genomes to RNA interference (Schubert et al., 2005).

Although their high degree of interspecific evolutionary conservation suggests that the highest ranked of the structural elements identified within the various capulavirus HCCSs are indeed biologically functional, we also tested whether evidence of this biological functionality was apparent within the patterns of nucleotide substitution that the EcmLV and ALCV genomes have undergone (the two capulavirus species with sufficient available data to perform these analyses). Towards this end, the ALCV and EcmLV genomes were partitioned into “paired” and “unpaired” site sets and, focusing only on variable sites (invariant sites were removed), the frequency spectra of the “minor alleles” (i.e. those present at the lowest frequencies within the sampled viruses) were compared at these sites within the paired and unpaired genome partitions. This revealed that while minor allele frequencies were lower at paired sites in both the ALCV and EcmLV genomes (a finding consistent with stronger negative selection acting on paired sites than on unpaired sites), the difference was statistically significant only for EcmLV (permutation test p-value$=0.01$, Fu and Li's F test; Supplementary Table 1). Also consistent with the hypothesis that paired sites are evolving under stronger negative selection than unpaired sites was the detection within both the ALCV and EcmLV Rep and CP coding regions of significantly reduced synonymous substitution rates in codons, which have base-paired third-position nucleotides (respective multiple comparison-corrected Mann–Whitney U test p-values$=0.013$ and 0.032 for ALCV and 0.0005 and 0.0235 for EcmLV; Supplementary Table 2). Furthermore, within the EcmLV dataset a very strong association between nucleotide sites that are complementary coevolving and nucleotide sites that are base-paired (Chi squared p-value$=0.0001096$) was detected (Supplementary Table 3).

Collectively, these results provide evidence that many of the detected structural elements within the EcmLV and ALCV genomes are likely being actively preserved by natural selection, and, in the case of EcmLV at least, that there is substantial evolutionary pressure for the maintenance of specific biologically important base–pairing interactions. Further experimental assays should of course be carried out both to test whether these predicted functional secondary structural elements are indeed functional, and, if they are, to determine the precise aspects of capulavirus biology that they impact.

3.5. Phylogenetic and recombination analyses

Both CP and Rep phylogenetic trees indicate that the new ALCV and EcmLV isolates cluster with the other described capulaviruses isolates (Fig. 3).

Whereas the EcmLV Rep sequences all cluster within a single group of variants sharing $>92.8\%$ genome-wide nucleotide sequence identity (Fig. 3), ALCV Rep sequences cluster within two distinct groups (tentatively referred to here as strain A and strain B), with the isolates in each group differing at $\sim 17.5\%$ of sites relative to those in the other group (Supplementary Fig. 3). Strain A isolates are slightly over-represented among the 64 ALCV isolates for which a 547 nt long fragment encompassing the V3 ORF and part of the *cp* ORF was sequenced (37/64; 58%). Whereas ALCV strain A isolates were found at 11/14 of the sampling sites (including the Spanish one; Supplementary Fig. 4), strain B isolates were only found at 8/14 of the sampling sites (including the Spanish one; Supplementary Fig. 4).

Efforts were made to detect and characterize recombination events that could have occurred among the current dataset of 45 capulavirus full genomes. Fourteen apparently unique recombination events were detected, including one event in the EcmLV genomes and 13 in the ALCV genomes (Table 1, Supplementary Table 4 and Fig. 4). Interestingly, all of the examined ALCV and EcmLV isolates display traces of recombination events, with ALCV isolates displaying, on average, evidence of 2.7 events (Supplementary Table 4).

Three out of the 13 ALCV recombination events apparently involved exchanges of sequences between ALCV variants (events 2, 7 and 14) whereas the other detected events (10 in ALCV and one in EcmLV), apparently involved inter-species sequence transfers (events 1, 3 to 6 and 8–13; Table 1 and Fig. 4). Interestingly, these inter-species sequence transfers all appear to have involved at least one parent that is related to a currently known capulavirus

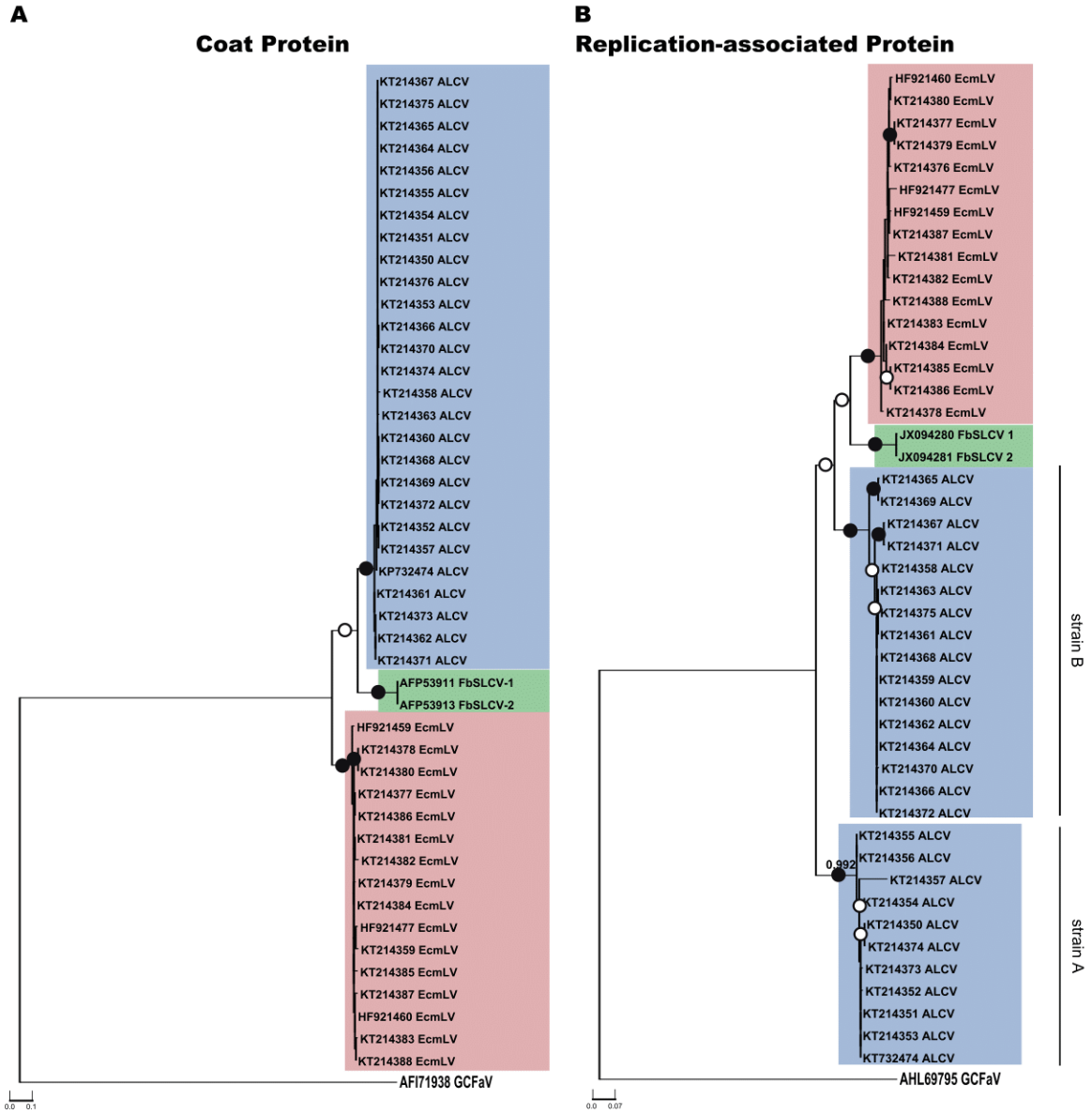


Fig. 3. (A) and (B) Maximum-likelihood phylogenetic trees of, respectively predicted CP and Rep amino acid sequences of 45 isolates of the three capulavirus species. Branches associated with a filled dot have bootstrap supports above 85% whereas those with an unfilled dot have bootstrap supports above 50%. The maximum-likelihood phylogenetic trees (A) and (B) are rooted with the Grapevine cabernet franc-associated virus (GCFaV, AHL69795) Rep and (GCFaV, AFI71938) CP, respectively.

Table 1
Recombination events detected in capulaviruses.

Event	Recombinant(s)	Major parent	Minor parent	Methods[a]	Breakpoints positions[b]
1	ALCV_VAU14_LUZ142	ALCV_ASS34_Assas1	Unknown	RGBMC	1433-2708
2	ALCV_PB14_LUZ178	ALCV_ASS34_Assas1	ALCV_PB14_LUZ171	RGBMCST	866-1046
3	ALCV_PB14_GS6	Unknown	ALCV_ASS14_Assas1	RGBMCST	1357–2112
4	ALCV_PB14_LUZ165	Unknown	ALCV_ASS14_Assas1	RGBMCST	2757–908
5	ALCV_PB14_LUZ171	Unknown	ALCV_PB14_LUZ184	RGBMCST	1033 (nad)[c] – 1716
6	ALCV_PB14_LUZ184	ALCV_GAG13_LUZ193	EcmLV_CM251	RGBM	884–969
7	ALCV_ASS14_Assas2	ALCV_GAG13_LUZ193	ALCV_TDV12_48-2 A	RGMCT	3011 (nad) – 346 (nad)
8	ALCV_PB14_GS4	ALCV_GAG13_LUZ193	Unknown	RMST	2375 (nad) – 642 (nad)
9	ALCV_PB14_LUZ178	EcmLV_MP4C	Unknown	RBMCS	531–1009 (nad)
10	ALCV_PB14_LUZ171	ALCV_PB13_LUZ165	Unknown	RGMC	2382–2726 (nad)
11	EcmLV_CM243	Unknown	FbSLCV-1	GBC	2366–2652 (nad)
12	ALCV_VAU14_LUZ136	ALCV_PB13_LUZ188	Unknown	RGBMCST	451 (nad) – 1599 (nad)
13	ALCV_ASS14_Assas1	ALCV_PB13_LUZ188	Unknown	GBMS	1196 (nad) – 1599 (nad)
14	ALCV_PB14_LUZ188	ALCV_TDV13_48-2 A	ALCV_ALB14_LUZ147	RGM	451 (nad) – 1599 (nad)

[a] RDP (R), GENECONV (G), BOOTSCAN (B), MAXIMUM CHI SQUARE (M), CHIMAERA (C), SISCAN (S) and 3SEQ (T) recombination detection methods.
[b] Begin and end breakpoints positions in the recombinant sequence.
[c] nad: not accurately determined.

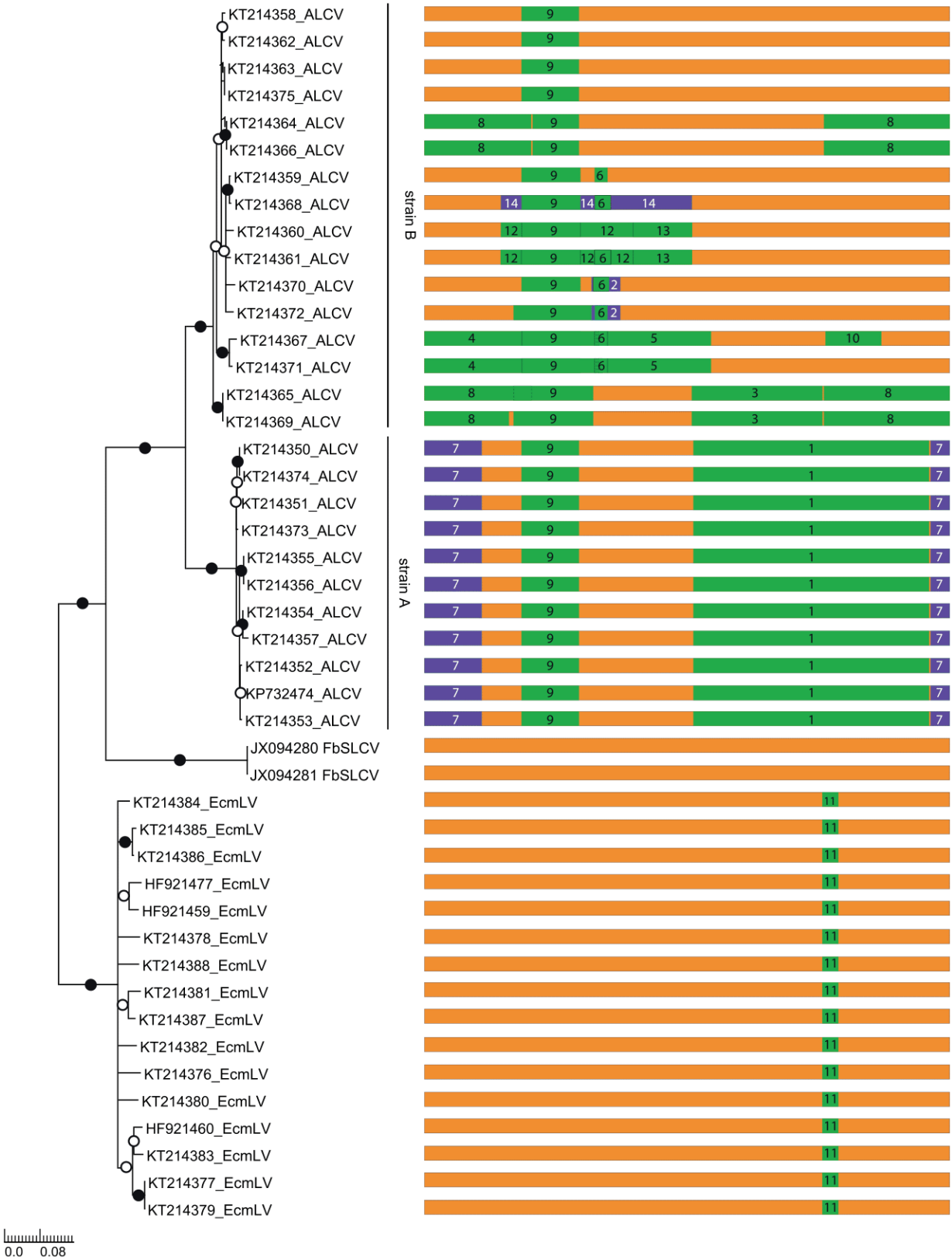


Fig. 4. Maximum-likelihood phylogenetic tree depicting the relatedness of the non-recombinant genomic regions of 45 isolates of the three capulavirus species. The fourteen unique recombination events detected within these sequences are presented to the right of the tree. Green and purple colors indicate genome regions that have likely been acquired by intra- and inter-species sequence transfers. Recombination events are numbered according to Table 1. Branches associated with a filled dot have bootstrap supports above 90% whereas those with an unfilled dot have bootstrap supports above 70%. All branches with less than 50% bootstrap support have been collapsed.

(i.e. EcmLV, FbSLCV or ALCV; Table 1) and another involving either an undescribed capulavirus species, or even perhaps a virus from an undescribed geminivirus genus.

Events 9 and 11 are respectively shared by all ALCV and EcmLV isolates (Supplementary Table 4), which indicates that these two events predate the most recent common ancestors of the analyzed ALCVs and EcmLVs and are, therefore, more ancient than the rest of the identified events. In addition, inter-species recombination event 1 and inter-ALCV strain recombination event 7 may account for the clear divergence of the ALCV-A and -B strains from one another.

3.6. Comparisons between the spatial distribution and genetic diversity of ALCV and EcmLV

Mantel tests were used to determine whether there was any correlation between the genetic and geographical distances of the sampled isolates. For ALCV, no significant correlation was observed between geographic distance and genetic distance for the fragment encompassing the V3 ORF and part of the *cp* ORF (Mantel R correlation scores of −0.034 with an associated *p*-value=0.230; Supplementary Fig. 5). This result suggests free movement of ALCV across the French Mediterranean region.

It is interesting that symptoms in alfalfa plants that resemble in some respect those described here (including plant stunting and leaf curling, crumpling and shriveling) have been described throughout the Mediterranean basin (France, Bulgaria, Romania, Spain and Saudi Arabia) from the late 1950s to the 1980s (Alliot et al., 1972; Blattný, 1959; Cook and Wilton, 1984; Leclant et al., 1973; Rodriguez Sardiña and Novales Lafarga, 1973) and, more recently, in Argentina (Bejerman et al., 2011). However, enations, which were associated with alfalfa leaf curling in these reports, have not been observed in ALCV infected plants either in the field or the laboratory. Attempts to identify viral particles by electron microscopy in some of the earliest reports revealed that the tissues within enations contained bullet-shaped, rhabdovirus-like particles with the viral species producing these particles being tentatively named Lucerne enation virus (LEV) (Alliot et al., 1972; Rodriguez Sardiña and Novales Lafarga, 1973). Additional experiments indicated that both grafting and *A. craccivora* (but not *Acyrthosiphon pisum*) transmission could successfully spread LEV symptoms between alfalfa plants. On the other hand, mechanical and leafhopper (*Calligypona pellucida* F.) transmission failed (Alliot et al., 1972; Blattný, 1959; Leclant et al., 1973; Rodriguez Sardiña and Novales Lafarga, 1973). It was therefore concluded that LEV was a circulatively transmitted virus within the family *Rhabdoviridae* (Alliot et al., 1972). However, because differences in the types of symptoms observed depended on the mode of transmission, Rodriguez Sardiña and Novales Lafarga (1973) hypothesized that another virus might frequently be present in co-infection with LEV (Rodriguez Sardiña and Novales Lafarga, 1973). To test if, apart from *Alfalfa mosaic virus*, which is ubiquitous in alfalfa, another virus could be present in co-infections, we examined four rhabdovirus-infected alfalfa plants collected from Spain and discovered that all of them were indeed co-infected with ALCV. Besides indicating that the geographical range of ALCV extends beyond the borders of France, this result is consistent with the hypothesis of Sardiña and Lafarga: i.e. that the disease attributed to LEV in the 1970s could potentially be caused by a complex of two or more viruses, one of which we now know is likely to be ALCV. This hypothesis will need to be further tested by examining further symptomatic alfalfa plants from around the Mediterranean basin.

For EcmLV, geographic distances between sampling locations were significantly correlated with genetic distances between the *rep* gene fragments of the viral isolates (Mantel *r* correlation score=0.142, *p*-value=0.002; Supplementary Fig. 5). This indicates that a degree of differentiation exists between EcmLV populations at a sub-regional scale within the Western Cape, which in turn suggests that there are likely restrictions on the free movement of EcmLV across the region.

The observed differences between the spatial distribution and genetic diversity of ALCV and EcmLV might reflect general differences between viruses that infect cultivated and non-cultivated hosts. Relative to viruses such as ALCV that infect cultivated hosts, the population genetic structures of viruses such as EcmLV that preferentially infect uncultivated hosts are likely to be impacted by a more complex combination of biotic parameters. Variable distributions and population densities of suitable host plants within natural environments can influence the probability that viruliferous insect vectors will successfully transmit viruses to an appropriate host (Keesing et al., 2006). Also, the variable life-spans of host plants, the possibility of long-term vertical transmission chains when hosts are vegetatively propagated, and the potential for sporadic vector transmission will all contribute to the selection processes that ultimately shape the population genetic structure of viruses that are adapted to infecting uncultivated species such as *E. caput-medusae*. It is entirely plausible, however, that, as with ALCV, EcmLV is also transmitted by *A. craccivora* (which is polyphagous, very widely distributed and has even been repeatedly observed on *E. caput-medusae*; Supplementary Fig. 6), further studies are needed to test this hypothesis.

4. GenBank accession numbers

Full genomes of: Alfalfa leaf curl virus (KT214350-KT214375); Euphorbia caput-medusae latent virus (KT214376-KT214388). V3 ORF and part of the *cp* ORF of Alfalfa leaf curl virus (KT214391-KT214427). C3 ORF and part of the C1 ORF of Euphorbia caput-medusae latent virus (KT964062-KT964084).

Acknowledgments

We wish to express our sincere thanks and appreciation to Mr Paul Loubser and colleagues from Buffelsfontein Game & Nature Reserve and to Mrs. Hestelle Melville and Miss Laurenda Van Breda from the University of the Western Cape Nature Reserve Unit. We also thank Michel Peterschmitt for helpful discussions and Stéphane Blanc for effective manuscript review. DPM, AV and GWH are supported by the National Research Foundation of South Africa (Grant N° TTK1207122745). PH is supported by the Polyomyelitis Research Foundation (Grant N° 15/102). This work was supported by Direction Générale de l'Armement (Grant N° 201160060) (Ministère de la Défense, France), The Métaprogramme INRA «Meta-omics of microbial ecosystems» (Grant N° 24000466) and CIRAD.

Appendix A. Supplementary material

Supplementary data associated with this article can be found in the online version at http://dx.doi.org/10.1016/j.virol.2016.03.016.

References

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

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.

Alliot, B., Signoret, P.A., Giannott., J., 1972. Presentation of bacilliform virus-particles associated with enation disease of alfalfa (Medicago-Sativa L). Cr Acad. Sci. D Nat. 274, 1974–1976.

Altschul, S.F., Gish, W., Miller, W., Myers, E.W., Lipman, D.J., 1990. Basic local alignment search tool. J. Mol. Biol. 215, 403–410.
Arguello-Astorga, G.R., Guevara-Gonzalez, R.G., Herrera-Estrella, L.R., Rivera-Bustamante, R.F., 1994. Geminivirus replication origins have a group-specific organization of iterative elements: a model for replication. Virology 203, 90–100.
Bejerman, N., Nome, C., Giolitti, F., Kitajima, E., de Breuil, S., Pérez Fernández, J., Basigalup, D., Cornacchione, M., Lenardon, S., 2011. First report of a rhabdovirus infecting alfalfa in Argentina. Plant Dis. 95 771-771.
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.
Blattný, C., 1959. Virus papillosity of the leaves of lucerne. Folia Microbiol. 4, 212–215.
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.
Brown, J.K., Zerbini, F.M., Navas-Castillo, J., Moriones, E., Ramos-Sobrinho, R., Silva, J. C., Fiallo-Olive, E., Briddon, R.W., Hernandez-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. Arch. Virol. 160, 1593–1619.
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, in: Proceedings of the CAB International, N.W., Wallingford, Oxfordshire, OX10 8DE, UK. (Ed.). CAB International Wallingford UK.
Cook, A.A., Wilton, A.C., 1984. Alfalfa enation virus in the Kingdom of Saudi Arabia. FAO Plant Prot. Bull. 32, 139–140.
Dry, I.B., Rigden, J.E., Krake, L.R., Mullineaux, P.M., Rezaian, M.A., 1993. Nucleotide sequence and genome organization of tomato leaf curl geminivirus. J. Gen. Virol. 74, 147–151.
Edgar, R.C., 2004. MUSCLE: a multiple sequence alignment method with reduced time and space complexity. BMC Bioinf. 5, 113.
Filloux, D., Dallot, S., Delaunay, A., Galzi, S., Jacquot, E., Roumagnac, P., 2015a. Metagenomics approaches based on virion-associated nucleic acids (VANA): an innovative tool for assessing without a priori viral viversity of plants. Methods Mol. Biol. 1302, 249–257.
Filloux, D., Murrell, S., Koohapitagtam, M., Golden, M., Julian, C., Galzi, S., Uzest, M., Rodier-Goud, M., D'Hont, A., Vernerey, M.S., Wilkin, P., Peterschmitt, M., Winter, S., Murrell, B., Martin, D.P., Roumagnac, P., 2015b. The genomes of many yam species contain transcriptionally active endogenous geminiviral sequences that may be functionally expressed. Virus Evol. 1, 1–17.
Fu, Y.X., Li, W.H., 1993. Statistical tests of neutrality of mutations. Genetics 133, 693–709.
Golden, M., Martin, D., 2013. DOOSS: a tool for visual analysis of data overlaid on secondary structures. Bioinformatics 29, 271–272.
Guindon, S., Delsuc, F., Dufayard, J.F., Gascuel, O., 2009. Estimating maximum likelihood phylogenies with PhyML. Methods Mol. Biol. 537, 113–137.
Gutierrez, C., Ramirez-Parra, E., Mar Castellano, M., Sanz-Burgos, A.P., Luque, A., Missich, R., 2004. Geminivirus DNA replication and cell cycle interactions. Veter. Microbiol. 98, 111–119.
Haible, D., Kober, S., Jeske, H., 2006. Rolling circle amplification revolutionizes diagnosis and genomics of geminiviruses. J. Virol. Methods 135, 9–16.
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 phi 29 DNA polymerase. J. Virol. Methods 116, 209–211.
Keesing, F., Holt, R.D., Ostfeld, R.S., 2006. Effects of species diversity on disease risk. Ecol. Lett. 9, 485–498.
Kraberger, S., Farkas, K., Bernardo, P., Booker, C., Arguello-Astorga, G.R., Mesleard, F., Martin, D.P., Roumagnac, P., Varsani, A., 2015. Identification of novel Bromus- and Trifolium-associated circular DNA viruses. Arch. Virol. 160, 1303–1311.
Krenz, B., Thompson, J.R., Fuchs, M., Perry, K.L., 2012. Complete genome sequence of a new circular DNA virus from grapevine. J. Virol. 86, 7715.
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.
Lazarowitz, S.G., 1992. Geminiviruses - genome structure and gene-function. Crit. Rev. Plant Sci. 11, 327–349.
Lazarowitz, S.G., Pinder, A.J., Damsteegt, V.D., Rogers, S.G., 1989. Maize streak virus genes essential for systemic spread and symptom development. EMBO J. 8, 1023–1032.
Leclant, F., Alliot, B., Signoret, P.A., 1973. Transmission et épidémiologie de la maladie à énations de la luzerne (LEV). Premiers résultats. Ann. Phytopathol. 5, 441–445.
Li, S.C., Shiau, C.K., Lin, W.C., 2008. Vir-Mir db: prediction of viral microRNA candidate hairpins. Nucleic Acids Res. 36, D184–D189.
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.
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, 162–172.
Ma, Y., Navarro, B., Zhang, Z., Lu, M., Zhou, X., Chi, S., Di Serio, F., Li, S., 2015. Identification and molecular characterization of a novel monopartite geminivirus associated with mulberry mosaic dwarf disease. J. Gen. Virol. 96, 2421–2434.
Mantel, N., 1967. The detection of disease clustering and a generalized regression approach. Cancer Res. 27, 209–220.
Markham, N.R., Zuker, M., 2008. UNAFold: software for nucleic acid folding and hybridization. Methods Mol. Biol. 453, 3–31.
Martin, D.P., Murell, B., Golden, M., Khoosal, A., Muhire, B., 2015. RDP4: detection and analysis of recombination patterns in virus genomes. Virus Evol. 1, vev003.
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). Arch. Virol. 158, 1411–1424.
Muhire, B.M., Golden, M., Murrell, B., Lefeuvre, P., Lett, J.M., Gray, A., Poon, A.Y., Ngandu, N.K., Semegni, Y., Tanov, E.P., Monjane, A.L., Harkins, G.W., Varsani, A., Shepherd, D.N., Martin, D.P., 2014a. Evidence of pervasive biologically functional secondary structures within the genomes of eukaryotic single-stranded DNA viruses. J. Virol. 88, 1972–1989.
Muhire, B.M., Varsani, A., Martin, D.P., 2014b. SDT: a virus classification tool based on pairwise sequence alignment and identity calculation. Plos One 9, e108277.
Murrell, B., Moola, S., Mabona, A., Weighill, T., Sheward, D., Pond, S.L.K., Scheffler, K., 2013. FUBAR: a fast, unconstrained bayesian approximation for inferring selection. Mol. Biol. Evol. 30, 1196–1205.
Ng, T.F., Chen, L.F., Zhou, Y., Shapiro, B., Stiller, M., Heintzman, P.D., Varsani, A., Kondov, N.O., Wong, W., Deng, X., Andrews, T.D., Moorman, B.J., Meulendyk, T., MacKay, G., Gilbertson, R.L., Delwart, E., 2014. Preservation of viral genomes in 700-y-old caribou feces from a subarctic ice patch. Proc. Natl. Acad. Sci. USA 111, 16842–16847.
Ng, T.F., Manire, C., Borrowman, K., Langer, T., Ehrhart, L., Breitbart, M., 2009. Discovery of a novel single-stranded DNA virus from a sea turtle fibropapilloma by using viral metagenomics. J. Virol. 83, 2500–2509.
Ng, T.F.F., Duffy, S., Polston, J.E., Bixby, E., Vallad, G.E., Breitbart, M., 2011. Exploring the diversity of plant DNA viruses and their satellites using vector-enabled metagenomics on whiteflies. Plos One 6, e19050.
Obayashi, T., Kinoshita, K., Nakai, K., Shibaoka, M., Hayashi, S., Saeki, M., Shibata, D., Saito, K., Ohta, H., 2007. ATTED-II: a database of co-expressed genes and cis elements for identifying co-regulated gene groups in Arabidopsis. Nucleic Acids Res. 35, D863–D869.
Pond, S.L.K., Frost, S.D.W., Muse, S.V., 2005. HyPhy: hypothesis testing using phylogenies. Bioinformatics 21, 676–679.
Poon, A.F., Lewis, F.I., Frost, S.D., Kosakovsky Pond, S.L., 2008. Spidermonkey: rapid detection of co-evolving sites using Bayesian graphical models. Bioinformatics 24, 1949–1950.
Rodriguez Sardiña, J., Novales Lafarga, J., 1973. Una virosis de la alfalfa con producción de "enations". An. INIA/Ser. Prot. veg. 3, 132–146.
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., Duffy, S., Breitbart, M., 2012. A field guide to eukaryotic circular single-stranded DNA viruses: insights gained from metagenomics. Arch. Virol. 157, 1851–1871.
Rosario, K., Marinov, M., Stainton, D., Kraberger, S., Wiltshire, E.J., Collings, D.A., Walters, M., Martin, D.P., Breitbart, M., Varsani, A., 2011. Dragonfly cyclovirus, a novel single-stranded DNA virus discovered in dragonflies (Odonata: Anisoptera). J. Gen. Virol. 92, 1302–1308.
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. Viruses 7, 5553–5570.
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.
Schubert, J., Habekuss, A., Kazmaier, K., Jeske, H., 2007. Surveying cereal-infecting geminiviruses in Germany–diagnostics and direct sequencing using rolling circle amplification. Virus Res. 127, 61–70.
Schubert, S., Grunweller, A., Erdmann, V.A., Kurreck, J., 2005. Local RNA target structure influences siRNA efficacy: systematic analysis of intentionally designed binding regions. J. Mol. Biol. 348, 883–893.
Semegni, J.Y., Wamalwa, M., Gaujoux, R., Harkins, G.W., Gray, A., Martin, D.P., 2011. NASP: a parallel program for identifying evolutionarily conserved nucleic acid secondary structures from nucleotide sequence alignments. Bioinformatics 27, 2443–2445.
Shepherd, D.N., Martin, D.P., Lefeuvre, P., Monjane, A.L., Owor, B.E., Rybicki, E.P., Varsani, A., 2008. A protocol for the rapid isolation of full geminivirus genomes from dried plant tissue. J. Virol. Methods 149, 97–102.
Tajima, F., 1989. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics 123, 585–595.
Theiler, J., Bloch, J., 1996. Nested test for point sources. In: Babu, G.J., Feigelson, E.D. (Eds.), Statistical Challenges in Modern Astronomy II. Springer-Verlag, New York, pp. 407–408.

Thompson, M.D., Jacks, C.M., Lenvik, T.R., Gantt, J.S., 1992. Characterization of rps17, rp19 and rpl15: three nucleus-encoded plastid ribosomal protein genes. Plant Mol. Biol. 18, 931–944.
Varsani, A., Martin, D.P., Navas-Castillo, J., Moriones, E., Hernandez-Zepeda, C., Idris, A., Murilo Zerbini, F., Brown, J.K., 2014a. Revisiting the classification of curtoviruses based on genome-wide pairwise identity. Arch. Virol. 159, 1873–1882.
Varsani, A., Navas-Castillo, J., Moriones, E., Hernandez-Zepeda, C., Idris, A., Brown, J. K., Murilo Zerbini, F., Martin, D.P., 2014b. Establishment of three new genera in the family Geminiviridae: Becurtovirus, Eragrovirus and Turncurtovirus. Arch. Virol. 159, 2193–2203.
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.
Yazdi, H.R.B., Heydarnejad, J., Massumi, H., 2008. Genome characterization and genetic diversity of beet curly top Iran virus: a geminivirus with a novel non-anucleotide. Virus Genes 36, 539–545.