UNIVERSITÉ FRANÇOIS – RABELAIS DE TOURS

ÉCOLE DOCTORALE SSBCV

INSTITUT DE RECHERCHE SUR LA BIOLOGIE DE L'INSECTE

THÈSE présentée par :

Diane BIGOT

soutenue le : **18 décembre 2017**

pour obtenir le grade de : **Docteur de l'Université François – Rabelais de Tours**

Discipline/ Spécialité : Science de la Vie / Biologie Evolutive

BIODIVERSITE ET EVOLUTION DES VIRUS PRESENTS DANS LES METAGENOMES ANIMAUX

THÈSE dirigée par :

Mme HERNIOU Elisabeth	Chargée de Recherche CNRS, Université de Tours

Co-encadrée par :

Mr GAYRAL Philippe	Maitre de conférences, Université de Tours

RAPPORTEURS :

Mr GOURBAL Benjamin	Maitre de conférences, Université de Perpignan, Perpignan
Mme PIGANEAU Gwenaël	Directrice de Recherche, UPMC-CNRS, Banyuls-sur-Mer

---

JURY :

Mr BOYER Stéphane	Professeur, Université de Tours
Mme FAILLOUX Anna-Bella	Directrice de Recherche, Institut Pasteur, Paris
Mr GAYRAL Philippe	Maitre de conférences, Université de Tours
Mr GILBERT Clément	Chargé de Recherche, CNRS, Gif-sur-Yvette
Mr GOURBAL Benjamin	Maitre de conférences, Université de Perpignan, Perpignan
Mme HERNIOU Elisabeth	Chargée de Recherche CNRS, Université de Tours
Mme OGLIASTRO Mylène	Directrice de Recherche, Université de Montpellier, Montpellier
Mme PIGANEAU Gwenaël	Directrice de Recherche, UPMC-CNRS, Banyuls-sur-Mer

"Life is like riding a bicycle. To keep your balance you must keep moving."

Albert Einstein, 1930

Remerciements

MERCI.

Un petit mot qui l'air de rien veut dire beaucoup.

De nature très sensible, cette partie est pour moi difficile à écrire et je l'écris même les larmes aux yeux. Je pense que cela ne va étonner personne. Tout ce travail et ces années de recherche n'auraient certainement pas été les mêmes sans vous. Que vous soyez des connaissances, des collaborateurs, des collègues, des stagiaires, des amis ou de la famille, chacun, à votre façon, avez eu de l'importance pour moi. J'espère n'oublier personne…

Je souhaite ainsi tous vous remercier très chaleureusement.

En commençant par l'Ecole doctorale Santé, Sciences Biologiques et Chimie du Vivant ainsi que le Ministère de l'Enseignement Supérieur et de la Recherche qui, à travers l'attribution d'une bourse doctorale, m'ont permis de réaliser ce travail de thèse. Je tiens à remercier l'Académie d'Agriculture de France qui m'a attribué une bourse de recherche Jean et Marie-Louise Dufrenoy me permettant de réaliser du terrain ainsi que de participer à mon premier congrès international à Vancouver au Canada.

Je remercie tous les membres de mon jury qui ont accepté d'évaluer mon travail et plus particulièrement mes deux rapporteurs Gwenaël Piganeau et Benjamin Gourbal qui ont eu la lourde responsabilité de m'évaluer et de lire ce volumineux travail de thèse.

Je remercie Jean-Paul Monge et David Giron, directeurs de l'IRBI pour leur accueil au sein de cet Institut de Recherche sur Biologie de l'Insecte. Sans oublier le soutien du personnel administratif sans qui nous sommes parfois un peu perdus : Aicha, Nadine, Sylvie, Marjorie et Nathalie.

Je remercie tous les collaborateurs, français et étrangers, que j'ai pu rencontrer lors de conférences ou au sein du laboratoire ; James Cook, Anne Dalmon, Marc Sitbon, Chunsheng Hou, Mylène Weill, Magali Ribière-Chabert… ainsi que toutes les personnes rencontrées lors du congrès de la *Society for Invertebrate Pathology* à Vancouver et à Tours ; une réelle communauté à échelle humaine.

Je remercie Joël Meunier, Clément Gilbert et Louis Lambrechts qui ont participé à mon comité de thèse et qui m'ont aiguillée à mi-parcours.

Je remercie les personnes qui m'ont aidé à collecter des abeilles, des fourmis, des frelons pour mes analyses ; Eric Darouzet, Jérémy Gévar, Alain Lenoir, Raphaël Boulay, Carlos Lopez-Vaamonde, Antoine Guiguet, Sébatien Moreau, Jean-Christophe Lenoir.

Je remercie bien sûr mon équipe d'accueil dans son ensemble ; Jean-Michel Drezen, Thibaut Josse, Karine Musset, Carole Labrousse, Yannis Moreau, Géraldine Dubreuil, Elisabeth Huguet…

Avec un remerciement plus particulier pour toi **Elisabeth**, ma directrice de thèse. Merci de m'avoir encadrée, épaulée, conseillée et bien plus encore durant ces trois années. La thèse n'est jamais un long fleuve tranquille. Je me souviendrai toujours et je pense que toi aussi, de notre séjour au Canada et surtout de la yourte en cèdre rouge. Nous avons eu bien peur ces jours-là !

Philippe. Que dire. Tu m'as encadrée, mais beaucoup plus que cela, durant ces trois années et même bien avant et sans toi rien n'aurait été pareil. Tu as toujours su trouver les mots pour me réconforter dans les moments durs mais aussi partager les moments de joie. Ta bonne humeur, tes sourires et ta philosophie de vie m'ont appris, je l'espère, à devenir quelqu'un de meilleur, en faisant toujours de mon mieux et en évitant de me prendre trop la tête. Merci pour tout !

Annie. Tu es bien plus qu'une collègue de bureau pour moi. Nous avons partagé plein de choses au bureau mais pas seulement. Merci pour les petites promenades autour du lac, pour tes coups de main, pour tes conseils et tes petites attentions à mon égard qui m'ont souvent mis du baume au cœur. Je te souhaite tout le meilleur possible (sans oublier Éric bien sûr) et j'espère que l'on va rester en contact encore longtemps.

J'ai eu l'occasion de co-encadrer plusieurs stagiaires durant cette thèse et je les remercie grandement. Ils m'ont appris à enseigner des compétences avec je l'espère toute la bienveillance nécessaire. Merci à vous Michèle, Manon, Jean-Philippe et Sélim.

Je remercie l'ensemble des doctorants et post-doc de l'IRBI que j'ai pu croiser. Bon courage à tous, je vous souhaite la meilleure réussite possible ; Aurélien, Jérémy, Guillaume, Marlène,

Marta, Virginie… et toi Mourad pour ta bonne humeur sans faille et pour avoir partagé avec moi (et Cristela bien sûr) un petit séjour sympa à Copenhague.

Matthieu. Tu es un sacré personnage ! Malgré ton sourire et ton rire à toute épreuve, tu es quelqu'un de très sensible et j'espère que tu vas pouvoir réussir tout ce que tu entreprends et entreprendras dans ta vie. Garde toujours le sourire !

Cristela. Nous avons presque tout partagé durant ces trois années, depuis le début nous nous sommes liées d'amitié qui je l'espère résistera au temps. Nous en avons fait du chemin ensemble, côté professionnel comme personnel, jusqu'à finir par partager le même quotidien. Merci d'avoir été là pour moi comme je l'ai été pour toi en toutes circonstances. Je n'aurais pas assez de mots pour te dire à quel point tu as eu de l'importance pour moi…

A vous mes amis, vous qui m'avez fourni un soutien à toute épreuve et ceci depuis presque une décennie déjà ! Et oui que le temps passe vite. Vous êtes surement ceux qui me connaissent le mieux. Merci Julien, Maxime, Adrien, Amina, Manon, Paola. A quand notre prochaine virée tous ensemble ?

Et finalement, je remercie tous ceux sans qui je ne serais pas là. J'embrasse très chaleureusement toute ma famille qui a toujours cru en moi et me soutient dans tous mes choix ; mes oncles et tantes, cousins et cousines, mes grands-parents, papa, maman et ma p'tite sœur Lucie.

Résumé

Les virus font partie des entités les plus abondantes sur Terre, mais cette abondance n'est cependant pas la représentation de la diversité virale réelle puisque les taxonomistes estiment connaitre seulement 1 % des espèces. De plus il existe un réel biais dans les séquences connues puisque plus de 78 % des séquences virales présentes dans les bases de données génomiques ne concernent finalement que les agents de cinq grandes maladies virales humaines. Ces quelques chiffres illustrent à quel point la diversité virale est si peu connue à ce jour.

L'étude des virus est étroitement liée aux avancées technologiques et l'apport des nouvelles techniques de séquençage permet maintenant d'obtenir des informations qui étaient alors tout simplement inaccessibles. Le but de mon travail de thèse a alors été d'avoir accès à la diversité virale présente au sein d'un grand nombre d'animaux non-modèles et de tenter de répondre à la question majeure qui est de savoir comment les virus sont représentés au sein du monde animal.

Pour répondre à cette question il m'a fallu mettre en place une méthodologie innovante de méta-transcriptomique permettant de mettre en évidence la présence virale à partir de transcriptomes d'animaux. Le jeu de données que j'ai à ma disposition est important puisqu'il est composé de 523 transcriptomes individuels de 135 espèces, représentant une part importante de la diversité du règne animal.

Ce travail m'as permis de montrer que la méthodologie mise en place est pertinente et permet de découvrir de nouveaux virus appartenant à des espèces, genres et même familles inconnus jusqu'alors, et ayant des caractéristiques génomiques particulières. Mes recherches permettent également de montrer que la gamme d'hôte de virus connus peut être plus étendue qu'attendue et que de nouveaux hôtes peuvent être trouvés pour des familles virales pourtant très étudiées. Durant ce travail, et grâce à l'apport complémentaire de nouvelles séquences détectées par RT-PCR, j'ai aussi pu aborder les questions liées aux associations hôtes-virus qu'ils existent entre les insectes hyménoptères et les virus d'abeilles. Finalement d'une manière plus synthétique, mon travail permet de combler quelques lacunes existantes dans les connaissances liées à l'étude de la diversité des virus et de mettre en évidence l'importance de l'étude des animaux non-modèles.

Abstract

Viruses are among the most abundant entities on the Earth but this abundance does not represent the real viral diversity, of which only 1% is estimated to be known. Indeed, molecular sequences available in database are biased, as 78% of viral genomic sequences correspond to the five major human viral diseases (e.g. HIV, Hepatitis and Influenza). To date viral diversity therefore remains relatively unknown.

Studies on viral discoveries and viral diversity are strictly dependent of technologies, and among them the Next Generation Sequencing techniques allow now access to so far inaccessible information by their efficiency and deep detection power. The main goal of my PhD work was to study viral diversity in a large range of non-model animal species in order to evaluate how viruses are distributed within animals.

I set up an innovative bioinformatics methodology to extract viral sequences from individual animal meta-transcriptomes. The dataset was composed by 523 individual transcriptomes from 135 non-model animal species representing the great diversity of metazoan.

This work underlines the efficiency of this new method to discover new viruses with particular genome characteristics. Also, this study shows that viral host range is larger than previously known for a lot of viruses and that new host can be found even for known viral families. A part of my work was focused on honey bee viruses and their distribution within wild hymenoptera. New viruses have been detected in wild bees and some new hymenopteran host (ants, hornets and bumble bees) have been found for honey bee viruses suggesting their wide distribution within pollinators and related species.

My work helps to fill some lack of knowledge on viral diversity within animals and it clearly remains important to continue investigation on viral diversity, viral distributions and impacts on hosts.

Table des matières

Liste des figures

Liste des tableaux

Liste des annexes

Liste des publications

INTRODUCTION

1.1. Préambule à l'introduction générale

Le but de cette partie introductive est de présenter le contexte dans lequel s'inscrit mon étude.

Je vais notamment commencer par définir ce qu'est la « biodiversité » et pourquoi finalement on ne la connaît que très peu en ce qui concerne les virus. Je reviendrai par la suite sur les caractéristiques qui sont propres aux virus afin d'avoir un aperçu général en termes de structure, de réplication, de classification et d'origine des virus.

Mon propos se poursuivra en abordant la notion d'évolution et de son application aux virus. Je montrerai en quoi l'étude des mécanismes qui régissent l'évolution des virus est intéressante de par le rôle qu'ils jouent dans l'évolution des espèces animales.

Je clôturerai cette introduction générale en évoquant quels sont les outils qui permettent actuellement d'étudier l'évolution et la biodiversité virale. Je m'appuierai sur des outils de génétique, de phylogénie, et montrerai en quoi mon approche est originale et comble les lacunes méthodologiques qui existent actuellement pour étudier la biodiversité des virus chez les animaux.

Cette introduction s'ouvrira sur la définition des objectifs et de la structure de la thèse.

1.2. Qu'est-ce que la biodiversité ?

1.2.1. Définition de la biodiversité

La biodiversité est littéralement la contraction de diversité biologique. Elle fait référence à la grande variété du monde vivant. C'est l'ensemble de la diversité des milieux de vie, de la richesse des espèces, et des variations des individus au sein des espèces, que ce soit en termes de contenu génétique, de forme ou de fonctions. La biodiversité se définit généralement en trois niveaux, i) la diversité génétique, qui s'intéresse aux différences entre individus d'une même espèce ; ii) la diversité spécifique, qui étudie la richesse des espèces de manière qualitative et quantitative dans un écosystème ; et iii) la diversité des écosystèmes, qui se caractérise par la diversité des milieux de vie et des groupements d'espèces en interaction les unes avec les autres. Ces trois niveaux sont reliés les uns aux autres mais chacun peut être étudié indépendamment puisqu'ils sont suffisamment distincts.

La biodiversité concerne toutes les catégories du monde vivant microscopiques et macroscopiques, eucaryotes (animaux, végétaux, champignons, protistes) et procaryotes (bactéries, archées). Comme David Tilman le souligne, "*The most striking feature of Earth is the existence of life, and the most striking feature of life is its diversity*". La vie est ainsi très riche et essentielle mais finalement elle est peu connue et menacée par les activités humaines (Tilman, 2000).

1.2.2. Quantification de la biodiversité

Au niveau spécifique, la biodiversité peut se quantifier en nombre d'espèces. La biodiversité spécifique sur Terre est le résultat d'un processus très long d'évolution des espèces, qui a démarré avec l'arrivée de la vie sur Terre, il y a environ 3,7 à 4,2 milliards d'années (Dodd *et al.*, 2017). En termes d'échelle temporelle, ce que l'on aperçoit aujourd'hui est le résultat d'une explosion de la vie et de l'évolution des espèces, de spéciations et extinctions qui durent depuis des millions d'années. Toutes les espèces quelles qu'elles soient ont chacune suivi une évolution qui leur est propre, la sélection naturelle leur permettant de s'adapter aux différents environnements changeants auxquels elles ont dû être confrontées durant cette longue période de temps. Une grande question que chacun se pose alors aujourd'hui est combien y-

a-t-il donc d'espèces actuelles sur Terre ? Cette question, pourtant si simple, n'a pas de réponse simple puisque la quantification du nombre d'espèces sur Terre est très complexe, mais il existe cependant des estimations de ce nombre (Tableau 1).

Tableau 1: Diversité du monde vivant en termes de nombre d'espèces.

Les données ont été collectées à partir de (Chapman, 2009; Mora et al., 2011). Le nombre d'espèces de virus décrites a été mis à jour avec la dernière version de l'ICTV (International Committee on Taxonomy of Viruses, Avril 2017).

GROUPE TAXONOMIQUE	NOMBRE D'ESPECES DECRITES DANS LE MONDE	NOMBRE D'ESPECES ESTIMEES DANS LE MONDE	POURCENTAGE D'ESPECES INCONNUES
INVERTEBRES	1 359 365	~6 755 830	80 %
PLANTES	310 129	~390 800	21 %
CHAMPIGNONS	98 998	~1 500 000	93 %
VERTEBRES	64 788	~80 500	20 %
BACTERIES	7 643	~1 000 000	99 %
VIRUS	4 404	~400 000	99 %
TOTAL	**1 845 327**	**10 127 130**	**82 %**

Les différentes estimations sont principalement basées sur des extrapolations de ce qui est déjà connu. Actuellement, près d'1,36 millions d'invertébrés sont répertoriés (majoritairement des arthropodes) et seulement 4404 espèces virales (Tableau 1). Ces chiffres sont extrêmement variables car chaque année de nouvelles espèces sont toujours découvertes. Ce fut le cas récemment d'une étude publiée en 2017 relatant de la découverte en Amazonie de 216 espèces de plantes, 93 poissons, 32 amphibiens, 19 reptiles, 1 oiseau et 20 mammifères (Valsecchi *et al.*, 2017). Les nouvelles découvertes ne se limitent pas au monde animal ou végétal et concernent aussi le monde viral (Bexfield & Kellam, 2011; Ho & Tzanetakis, 2014; Kreuze *et al.*, 2009; Kristensen *et al.*, 2010; Liu *et al.*, 2011; Mokili *et al.*, 2012; Rosario & Breitbart, 2011; Smits *et al.*, 2015).

Si l'on compare le nombre d'espèces actuellement décrites aux estimations qui sont faites quant au nombre réel d'espèces sur Terre, on se rend compte que finalement cette diversité spécifique est très peu connue. Près de 80 % des espèces de plantes et de vertébrés sont connues alors que seulement 1 % des espèces bactériennes et virales sont décrites à ce jour (Tableau 1).

En se focalisant plus sur la diversité virale, actuellement l'ICTV (International Committee on Taxonomy of Viruses) reconnait l'existence de 4404 espèces virales, 735 genres, 122 familles et 8 ordres (ICTV, Avril 2017). On peut aisément penser que s'il existe finalement environ 400 000 espèces virales (estimation), de nombreux genres et familles virales, dont on ne connaît pas les caractéristiques, restent encore à découvrir.

1.2.3. L'abondance des virus est-elle la représentation de la biodiversité virale ?

Les virus sont capables d'infecter tous les types d'organismes existants sur Terre (Bamford, 2003), que ce soit des archées (Prangishvili, 2013), des bactéries (Ackermann, 2003; Hendrix *et al.*, 2003), des eucaryotes (protistes unicellulaires, animaux, champignons et végétaux) (Koonin *et al.*, 2015; Nagasaki, 2008) et même jusqu'à d'autres virus ; ils sont appelés alors virophages (La Scola *et al.*, 2008; Sun *et al.*, 2010).

En plus de pouvoir infecter n'importe quel organisme, ils sont présents aussi bien en milieu aquatique (Suttle, 2007) qu'en milieu terrestre et infectent des animaux terrestres comme c'est le cas du virus de l'immunodéficience humaine (Barré-Sinoussi *et al.*, 1983). Les virus sont également capables de vivre dans des milieux extrêmes, telles que des conditions extrêmes de pression (Anantharaman *et al.*, 2014), de salinité (Boujelben *et al.*, 2012; Santos *et al.*, 2010), d'équilibre acidobasique (Jiang *et al.*, 2004; Rice *et al.*, 2001) ou encore de température (Kepner *et al.*, 1998; Prangishvili, 2003; Rice *et al.*, 2001).
L'abondance des virus a tout d'abord été quantifiée à partir de simples échantillons d'eau de mer, lorsqu'en 1989 Bergh *et al*. ont pu visualiser par microscope électronique la présence d'environ 10 millions de particules virales par millilitre (Bergh *et al.*, 1989). Plus tard, en 2005, Breitbart et Rohwer, ont estimé qu'il y avait environ 10^{31} particules virales dans les océans à chaque instant (soit cent millions de fois plus que d'étoiles dans l'univers ~10^{23}) (Breitbart & Rohwer, 2005). Finalement les virus sont très abondants puisqu'ils constituent la plus grande partie de l'abondance des organismes sur Terre et notamment dans les océans (Suttle, 2007).

Malgré cette très grande abondance des virus et le fait qu'ils puissent infecter tous types d'organismes et de milieux, la biodiversité virale actuellement connue est très fortement biaisée. En effet, si l'on regarde dans la base de donnée mondiale des séquences nucléotidiques (GenBank, NCBI), sur les près de 2,1 millions de séquences virales, les 4/5èmes

ne représentent en réalité que quelques virus humains : le virus du SIDA (HIV), les virus des hépatites (*Hepatitis*) et de la grippe (*Influenza*) (Figure 1).

Le biais des connaissances actuelles de la biodiversité virale et des études de virologie est très lié à l'intérêt qui est porté à la santé humaine mais également aux animaux domestiques (exemple du virus de la rage) ou d'élevages (exemple du virus de la fièvre aphteuse) et des maladies virales des plantes utilisées pour la consommation humaine (exemple des virus de la mosaïque du tabac, du chou-fleur ou de la tomate).

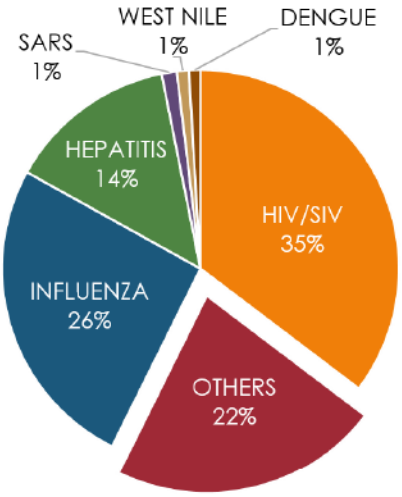


Figure 1 : Répartition des séquences nucléotidiques virales présentes dans GenBank (2,1 millions de séquences, Avril 2017).

1.3. Qu'est-ce qu'un virus ?

1.3.1. Définition des virus

Lwoff en 1957 a donné la première définition des virus, qui est toujours valable et peut se retenir en une liste de critères qui permettent de les définir et de les différencier de tout autre organisme vivant (King *et al.*, 2012; Lwoff, 1957) :

- les virus sont des **parasites moléculaires intracellulaires obligatoires**
- les virus sont de petits agents **infectieux, potentiellement pathogènes**
- les virus possèdent un **seul type d'acide nucléique (ADN ou ARN)** qui est la forme **transmissible** du virus (présente dans la particule virale)
- les virus sont **incapables de croitre** ou de subir de division binaire
- les virus ne disposent **pas de machinerie complète** capable de leur fournir de l'énergie via la respiration et doivent **obligatoirement utiliser les structures de la cellule hôte**.

Pour être plus précis, la définition peut s'étendre à ce qui est appelé le virion. Le virion est l'unité structurale d'un virus qui est composé (i) d'une « boite » protéique qui constitue la capside protéique, (ii) d'un génome constitué par l'acide nucléique (ADN ou ARN) et dans certains cas (iii) d'une enveloppe lipoprotéique (Figure 2). Le virion constitue le parasite obligatoire, qui en dehors d'une cellule est sous forme inerte et qui pour répliquer son génome et se multiplier doit nécessairement infecter une cellule hôte (Stanley, 1935).

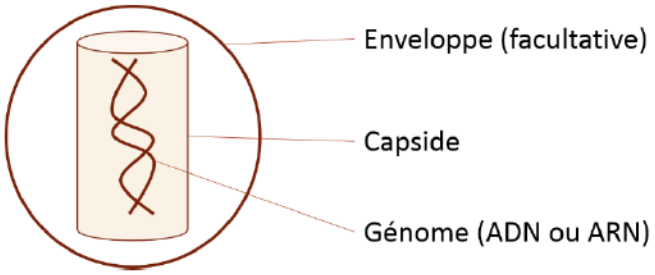


Figure 2: Schéma simplifié de la composition générale d'un virion.

Les virus ont une gamme de taille allant de 20 nm pour un picornavirus ou parvovirus, à 1000 nm pour les virus géants tels que le *Pandoravirus*. Ils sont principalement observés au microscope électronique ou optique (Figure 3). Comparé à la taille d'une bactérie ou d'une cellule animale ou végétale, le virus est la plus petite des entités organiques devant les protéines (Figure 3). Cette taille réduite des virus a fait que pendant de très nombreuses années ils n'étaient pas visibles pour l'homme jusqu'à leur première observation au microscope électronique (Ruska *et al.*, 1939), freinant ainsi leur étude.

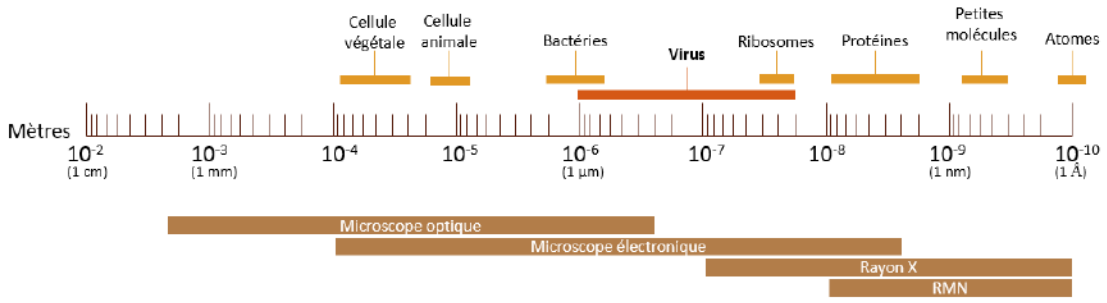


Figure 3: Echelle de taille des virus.
Comparaison de la taille des virus à celle des cellules, bactéries, protéines et atomes. Est représentée également la résolution des techniques qui permettent d'observer ces éléments. RMN : Résonance Magnétique Nucléaire.

1.3.2. Une stratégie commune à la propagation virale

Tous les virus possèdent la même stratégie pour leur propagation que l'on peut résumer en trois propriétés fondamentales (Flint *et al.*, 2015a) :

- Leur génome est empaqueté à l'intérieur d'une particule qui permet d'assurer leur transmission d'une cellule à une autre, et d'un hôte à un autre.
- Le génome viral contient les informations pour initier et compléter un cycle infectieux au sein d'une cellule sensible et permissive.
- Tous les virus nouvellement produits de manière correcte peuvent s'établir dans une population hôte afin que la propagation du virus soit assurée.

1.3.3. Classification des virus

Les virus sont classés, tout comme les animaux, en ordres, familles (et quelquefois sous-familles), genres et espèces. Enfin, les souches permettent de distinguer les lignées divergentes au sein d'une espèce. Contrairement à la classification binomiale proposée par Linée, les noms des espèces virales ne suivent cependant pas une règle commune. Afin de les classer correctement il est important de définir ce qu'est une espèce virale.

1.3.3.1. La classification de Baltimore

En 1971, David Baltimore (1938 -) a mis en place un système de classification des virus suivant leurs caractéristiques génomiques (Baltimore, 1971). Ce système prend en compte le type de génome (ADN ou ARN), le nombre de brin et le sens (simple (sens ou antisens) ou double brin), ainsi que le type de réplication et les intermédiaires réplicatifs. Le caractère linéaire ou circulaire ou encore la taille du génome viral ne sont pas pris en compte dans cette classification. Le seul point commun à toutes les classes est le but ultime de la réplication qui est de former un ARN messager (par convention en 5'-3', sens positif), support de la traduction nécessaire à la fabrication des protéines codées par l'ensemble des génomes viraux. Dans son article de 1971, Baltimore avait décrit arbitrairement 6 classes (I à VI), une septième classe (VII) est venue se rajouter à cette classification au fil du temps et des nouvelles avancées (Figure 4).

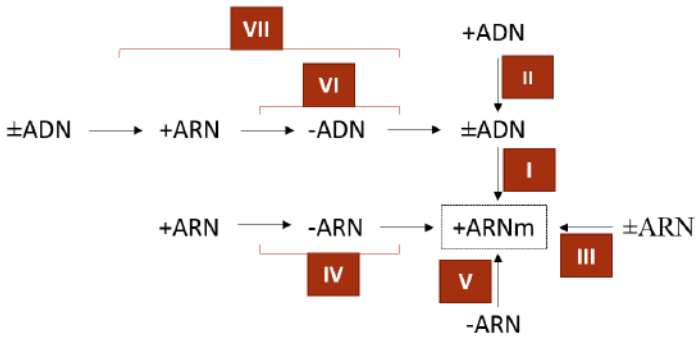


Figure 4: Représentation des sept groupes de virus selon la classification de Baltimore.

Les virus sont classés suivant leur type de réplication. + signifie le brin sens (5'-3'), - le brin anti-sens, ± signifie double brin. L'ARN messager (+ARNm) correspond à l'intermédiaire de traduction commun à tout type de virus.

1.3.3.2. La classification hiérarchique des virus

Bien avant la classification de Baltimore, en 1962, Lwoff, Horne et Tournier ont proposé une classification hiérarchisée permettant de classer les virus en phylum - classe – ordre (suffixe – *virales*) – famille (suffixe –*viridae*) – sous-famille (suffixe –*virinae*), genre (suffixe – *virus*) - espèces - souche/type (Lwoff et al., 1962).

Dans cette classification, les virus sont groupés suivant les propriétés qu'ils partagent et non en fonction des propriétés des cellules ou organismes qu'ils infectent. Les quatre critères qui forment cette hiérarchisation sont i) la nature de l'acide nucléique (ADN ou ARN, qui est maintenant plus détaillée suivant la classification de Baltimore), ii) la symétrie de la capside (voir section structure d'un virus), iii) la présence ou l'absence d'une enveloppe et iv) l'architecture du génome (segmenté/monopartite, circulaire/linéaire) (Figure 5).

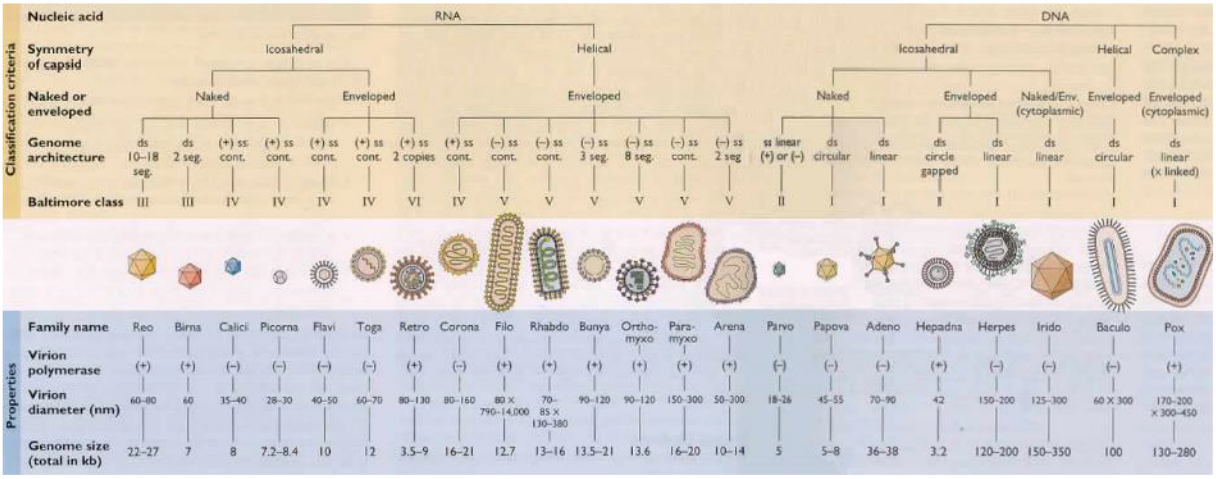


Figure 5: Représentation schématique de la classification hiérarchisée des virus.

La classe VII de Baltimore n'est pas représentée et correspond à la famille *Hepadnaviridae* (Flint et al., 2015b).

Cette classification hiérarchique des virus, combinée à la classification de Baltimore, est celle qui est utilisée actuellement dans la taxonomie des virus. Ell est reportée et mise à jour régulièrement dans l'ICTV, https://talk.ictvonline.org/ (King *et al.*, 2012). Les différents membres des familles virales sont, de plus, actuellement classés par les outils de génomique qui permettent de résoudre les relations évolutives par l'analyse des séquences nucléiques et protéiques.

1.3.3.3. Concept de l'espèce virale

L'actuelle version de l'ICTV définit une espèce virale comme : « *a polythetic class of viruses that constitutes a replicating lineage and occupies a particular ecological niche.* » d'après la définition donnée par Van Regenmortel en 1990 (Van Regenmortel, 1990). Cette définition, assez brève, inclut une notion importante qui est celle de la variabilité biologique. Cette notion est évoquée par le terme « polythétique », faisant référence à un ensemble de caractéristiques communes à toute une classe de virus, dont l'appartenance à ce groupe n'est pas dépendante du nombre de caractères qu'un virus possède, laissant place à une grande variabilité. La variabilité est aussi implicite par les termes « lignée réplicative », qui inclut la notion de variabilité génétique et d'évolution des virus au sein d'une même lignée. Et enfin, la notion de variabilité est incluse dans le terme de niche écologique qui est le critère essentiel des virus, ces derniers étant des parasites obligatoires dépendant de leur hôte.

Le concept d'espèce virale s'éloigne donc fortement du concept biologique de l'espèce, pour qui la notion de reproduction possible entre individus n'est pas applicable aux virus. Même si les lignées virales divergent au cours du temps, on ne pourra pas non plus parler de processus de spéciation comme le résultat des barrières aux flux de gènes entre populations. Par ailleurs, la très forte perméabilité des génomes viraux, favorisant les échanges génétiques avec leurs hôtes ou d'autres virus, diffère de la vision d'espèce biologique au sein de laquelle les individus sont plus cloisonnés génétiquement.

Le concept d'espèce virale reste cependant évolutif et doit s'adapter aux avancées technologiques qui permettent d'apporter des contre-exemples à la définition d'espèce par l'apport de la découverte de nouveaux virus remettant en cause les classifications actuelles. C'est dans ce cadre taxonomique que s'intègre mon travail de thèse sur la découverte de nouveaux virus. C'est en effet encore plus complexe lorsqu'il s'agit de définir des noms de virus, noms d'espèce(s), noms de souche(s) suivant une certaine taxonomie et classification

des virus qui n'est pas adaptée à l'arrivée massive de nouvelles espèces virales par les méta-analyses (Simmonds *et al.*, 2017). La classification est même controversée et remise en cause sous sa forme actuelle (Kuhn & Jahrling, 2010), avec une proposition de nomenclature binomiale comme celle des espèces animales ce qui permettrait de clarifier cette classification (Postler *et al.*, 2016).

1.3.4. Caractéristiques physiques et réplicatives des virus

1.3.4.1. Structure d'un virus

La capside virale, couplée avec une enveloppe dans certains cas, entoure et protège le génome viral. Cette capside est au cœur de l'interaction hôte-virus et possède un rôle dans l'attachement aux récepteurs cellulaires cibles qui sont à leur surface et qui permettent d'initier le cycle infectieux.

En 1956, Francis Crick et James Watson ont énoncé les critères permettant la formation des capsides virales pour les petits virus : i) la capside doit être assez large pour renfermer le génome viral et ii) un nombre limité de protéines de capside peut être codé par le génome (Crick & Watson, 1956). Ainsi, la meilleure façon de produire une protéine de capside codée par un petit génome est de l'utiliser de multiples fois à la manière de sous-unités, qui une fois assemblées forment une seule « coque » protéique : la capside. De manière mathématique, afin que chacune de ces sous-unités, une fois assemblées, garantissent un environnement identique à la protection de l'acide nucléique viral, seulement deux façons sont possibles : sous forme « cubique » et en « hélice ».

Les différentes observations qui ont ensuite été faites de la structure des virus ont corroboré cette théorie. Nous savons maintenant que la majeure partie des capsides virales existe sous la forme d'icosaèdre (=polygone de 20 facettes, Figure 6a) ou d'hélice (Figure 6b). Cependant, il existe des exceptions ; des virus plus complexes, tel l'exemple d'un bactériophage (Figure 6c) ou de très gros virus, ont plutôt des formes de capsides ovoïdes (exemple des *Poxviridae, Mimiviridae*).

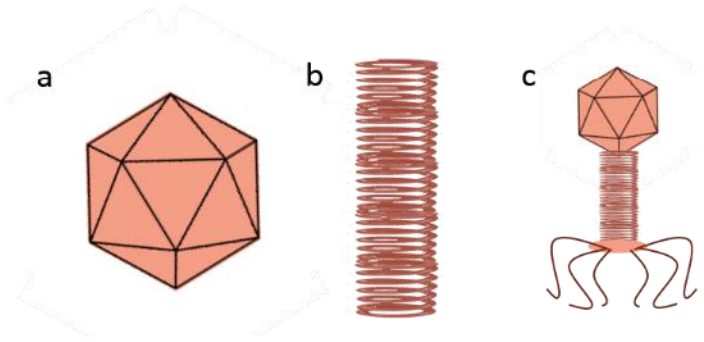


Figure 6: Représentation schématique des trois formes majeures de nucléocapside.
a) Icosaédrique, b) Hélicoïdale, c) Complexe (exemple d'un bactériophage).

1.3.4.2. Cycle réplicatif d'un virus

Quelle que soit la taille ou la nature du génome viral (simple ou double brin, linéaire ou circulaire, mono ou multipartite) ou la forme du virion (possédant ou pas une enveloppe), le génome viral doit être répliqué au sein d'une cellule hôte pour assurer un cycle infectieux. Le cycle typique d'un virus libre peut être résumé en 7 grandes étapes, schématisées Figure 7.

La première étape concerne l'attachement et la reconnaissance spécifique de récepteurs cellulaires, cibles des protéines virales de capside (ou d'enveloppe). Cette première étape est cruciale et permet de définir le tropisme (quel type cellulaire en particulier ils infectent) et souvent de définir la gamme d'hôtes des virus. Une fois la reconnaissance établie, l'entrée du virus dans la cellule peut avoir lieu (étape 2). Trois mécanismes, dépendant d'énergie, peuvent être utilisés : la translocation (entrée de l'entièreté du virion dans la cellule), l'endocytose (entrée via l'intermédiaire d'une vacuole intracellulaire) et la fusion (fusion de l'enveloppe virale et de l'enveloppe cellulaire, requièrant des protéines de fusion virales).

La décapsidation a généralement lieu juste après l'entrée du virus dans la cellule (étape 3). Le génome est libéré de la capside, ce dernier étant quelques fois associé à des protéines de protection de l'acide nucléique (VpG en 5' des picornavirus) ou des polymérases inverses (dans le cas des rétrovirus). La réplication a lieu suivant les différentes stratégies des génomes viraux (voir les différentes classes de Baltimore). Les nouveaux génomes sont ainsi produits par la machinerie cellulaire ainsi que l'expression des gènes permettant de produire l'ensemble des protéines nécessaires à la formation de nouvelles capsides, polymérases et autres protéines nécessaires au virion (étape 4). Une fois que toutes les molécules sont

disponibles, un grand nombre de virions sont assemblés (dans le cytoplasme, dans le noyau ou à la surface de la cellule) puis libérés (étape 5). Pour une majorité de virus (non-enveloppés), la sortie se produit par simple lyse de la cellule (étape 6). Pour les virus enveloppés, la membrane cellulaire est utilisée lors du bourgeonnement des virions afin de constituer leur propre enveloppe lipidique. Enfin, pour certains des virus, une étape supplémentaire de maturation a lieu au sein même du virion, comprenant des changements de conformation des protéines de capsides ou de condensation des nucléoprotéines avec le génome afin de devenir complètement infectieux pour commencer un nouveau cycle (étape 7).

Certains virus ont la nécessité de s'intégrer au génome hôte pour se répliquer comme les rétrovirus et d'autres sont des virus endogènes qui ont été intégrés au sein de génomes hôtes il y a plusieurs millions d'années pouvant toujours s'exprimer (Feschotte & Gilbert, 2012).

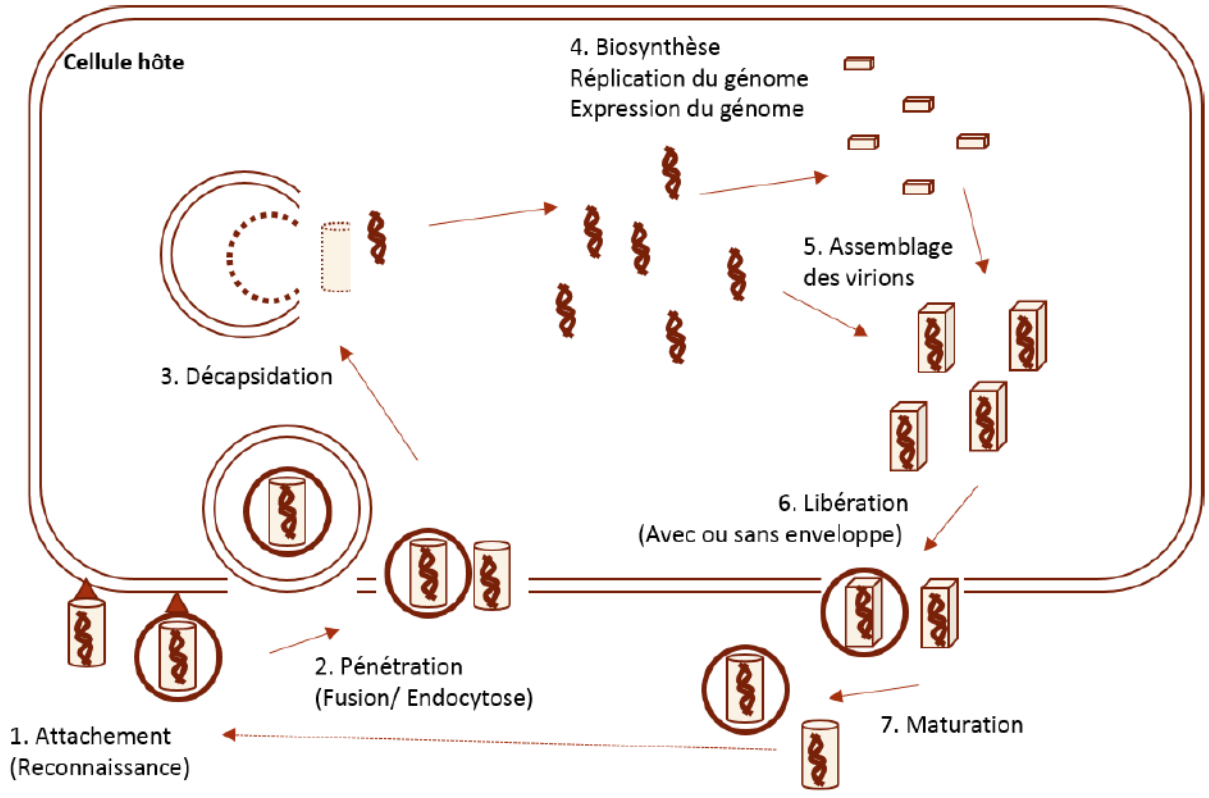


Figure 7: Schéma simplifié d'un cycle viral typique d'un virus libre.

1.3.5. L'origine des virus

Différentes hypothèses sont proposées pour définir l'origine des virus. Les deux hypothèses principales sont i) que l'origine des virus précède celle des cellules par l'intermédiaire de la

soupe primordiale ARN, *"virus-first hypothesis"* et ii) que les virus dérivent des cellules *"cell-first hypothesis"* par perte de fonction *"reduction hypothesis"* ou par échappement *"escape hypothesis"* (Forterre, 2006; Koonin *et al.*, 2006).

Quoiqu'il en soit l'évolution des virus et des organismes cellulaires sont étroitement liées puisque les virus sont capables d'infecter les trois domaines de la vie, les procaryotes, archées et eucaryotes (Bamford, 2003; Durzyńska & Goździcka-Józefiak, 2015; Koonin *et al.*, 2006). Les virus auraient alors une origine très ancienne, proche de l'origine du dernier ancêtre commun à toutes les cellules (LUCA, Last Universal Cellular Ancestor) (Claverie, 2006; Forterre & Prangishvili, 2009). Les virus auraient pu jouer un rôle important dans l'origine de l'ADN, jouer un rôle dans l'émergence des cellules eucaryotes et être à l'origine des trois domaines de la vie (Claverie, 2006; Forterre, 2006; Koonin *et al.*, 2015).

Deux visions s'opposent alors pour savoir si les virus font partie du monde vivant (Villarreal, 2004). D'un côté, il a été proposé de redéfinir le monde vivant en deux classes : les organismes codant pour les ribosomes (les eucaryotes, archées et bactéries) et les organismes codant pour les capsides ; les virus (Raoult & Forterre, 2008). D'autres s'opposent à l'inclusion des virus dans le domaine de la vie (Moreira & López-García, 2009). Toutes ces interrogations dépendent principalement de la définition même de la vie, de l'origine des virus et aussi de l'angle de vue sous lequel est étudiée l'évolution de la vie (Koonin & Dolja, 2013). Qu'ils soient vivants ou non, le plus important à savoir est que les virus sont capables d'évoluer et d'avoir un impact dans l'évolution de leurs hôtes.

1.4. Les découvertes des virus sont étroitement liée aux avancées technologiques

La diversité des virus est actuellement peu connue puisque l'étude des virus est étroitement liée aux avancées scientifiques et techniques. Il a fallu plusieurs siècles de mise au point des différentes techniques pour pouvoir maintenant identifier les virus, les caractériser, les visualiser et finalement étudier leur génétique et leur évolution.

1.4.1. Premières observations de maladies virales dès l'antiquité

Quelques écrits historiques relatent de la présence de maladies virales durant l'Egypte ancienne (environ 1200 avant J-C). C'est le cas de quelques hiéroglyphes montrant des adultes ayant des symptômes d'atrophies musculaires des membres inférieurs, probablement atteints

de poliomyélite (poliovirus) ou de momies présentant des traces de pustules caractéristiques de la variole (smallpox virus). Cependant, la présence de ce virus en Egypte durant cette période reste controversée puisque reposant sur très peu d'éléments (Galassi et al., 2017).

D'autres références attestent de la présence du virus de la rage (*Rabies lyssavirus*) dans des temps très anciens. C'est le cas par exemple du l'utilisation du terme « enragé » décrivant Hector dans l'Iliade d'Homer (probablement écrit entre 1200-1180 av J-C) ou d'une loi Mésopotamienne qui régit les responsabilités des propriétaires de chiens enragés (avant 1000 av J-C).

Une référence plus récente attestant d'un virus durant l'antiquité est celle d'un virus de plante. Au XVII^{ème} siècle en Hollande, eu lieu la dissémination involontaire d'un virus de tulipe (*tulip breaking virus* ou *tulip mosaic virus*) qui provoque des changements de couleurs des pétales des fleurs (Figure 8), provoquant des variations très recherchées par les amateurs et les collectionneurs de l'époque (Lesnaw & Ghabrial, 2000).

Figure 8: Représentation d'une infection virale.
Tulipe montrant un panachage de la couleur des pétales. Aujourd'hui cette caractéristique a été sélectionnée, est stable et n'est plus due au Tulip breaking virus, épidémie très répandue au XVIIème siècle.

1.4.2. Du XIX^{ème} au XXI^{ème} siècle : avancées technologiques et virologie

Au cours de l'Histoire, la notion et la définition de virus a évoluée au fil des découvertes et de l'évolution des technologies d'imagerie et de génétique. Plusieurs grandes étapes jalonnent ainsi ce parcours historique qui mène à l'heure actuelle aux connaissances que l'on a des virus et de leur diversité (Figure 9). Avant leur découverte chez les plantes en 1898 par Ivanovsky, les virus n'était alors défini que par des caractères négatifs ; par opposition aux bactéries ils sont non visibles, non cultivables, non arrêtés par les filtres. Il a fallu des décennies avant de montrer par cristallographie qu'ils sont constitués de protéine (Stanley, 1935) et de pouvoir les visualiser à l'aide du premier microscope électronique en 1939 (Ruska *et al.*, 1939). La caractérisation des génomes viraux, composés d'acide nucléiques (Fraenkel-Conrat *et al.*, 1957; Hershey & Chase, 1952) a permis d'amener à la première définition des virus en 1957. Enfin les différents progrès réalisés plus récemment en matière de biologie moléculaire et de séquençage haut débit permettent depuis quelques années d'avoir accès à de nombreuses informations relatives à la diversité virale et c'est dans ce cadre que s'intègre ainsi ma thèse (Figure 9).

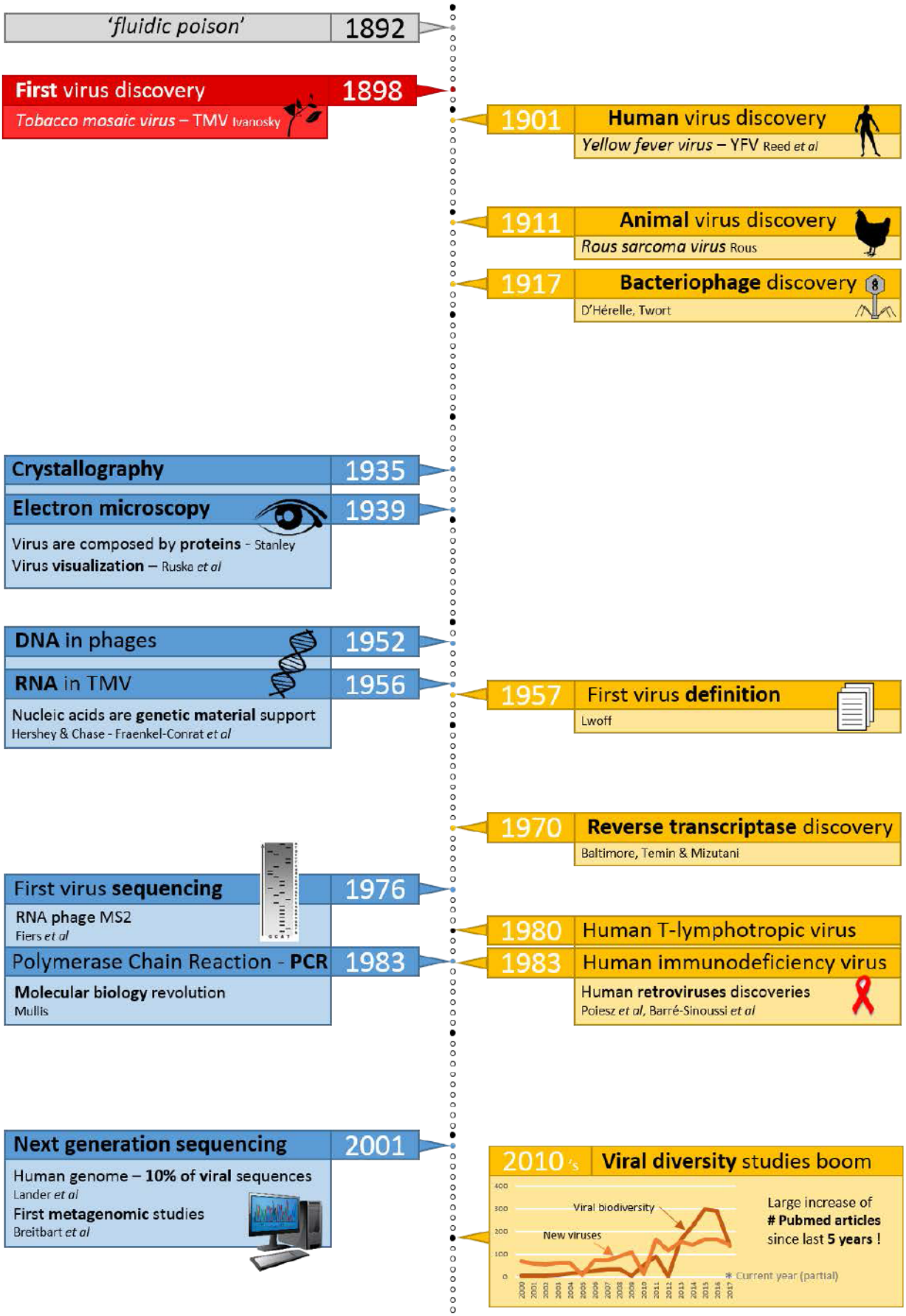


Figure 9 : Contexte historique des avancées technologiques majeures et leurs répercussions sur la virologie depuis la découverte des virus au XIX^{ème} siècle.

1.5. Le rôle des virus dans l'évolution animale

Notre propre génome est constitué de 8 % de séquences d'origine virale (Lander *et al.*, 2001) Ce simple exemple illustre le fait que les virus jouent un rôle majeur dans l'évolution de leurs hôtes.

1.5.1. La relation hôte-virus : une course aux armements constante

Cette notion de course aux armements a été appelée l'hypothèse de la reine rouge (*Red queen hypothesis*) par Leigh van Valen en 1973 (van Valen, 1973) en s'inspirant du livre de Lewis Caroll, *Through the Looking Glass and what Alice found there* écrit en 1871, suite du célèbre *Alice in Wonderland*. Ici cette course aux armements est illustrée par la course effrénée qu'Alice entreprend pour rester toujours au même endroit.

> *"Well, in OUR country," said Alice, still panting a little, "you'd generally get to somewhere else if you ran very fast for a long time, as we've been doing."; "A slow sort of country!" said the Queen. "Now, HERE, you see, it takes all the running YOU can do, to keep in the same place. If you want to get somewhere else, you must run at least twice as fast as that!"*
>
> *Lewis Carroll, Through the Looking Glass and what Alice found there, 1871*

Pour un virus et son hôte, cette course peut être schématisée avec l'exemple de l'interaction récepteur cellulaire-protéine d'enveloppe virale (Figure 10). Lorsque la protéine d'enveloppe virale est capable de se lier au récepteur cellulaire de sa cellule hôte, le virus est le gagnant de cette course puisqu'il est capable de poursuivre son cycle infectieux. Au contraire, si le récepteur cellulaire change (par une mutation d'acide aminé ou un changement de conformation), c'est l'hôte qui est le gagnant de l'interaction puisqu'il échappe à l'infection virale.

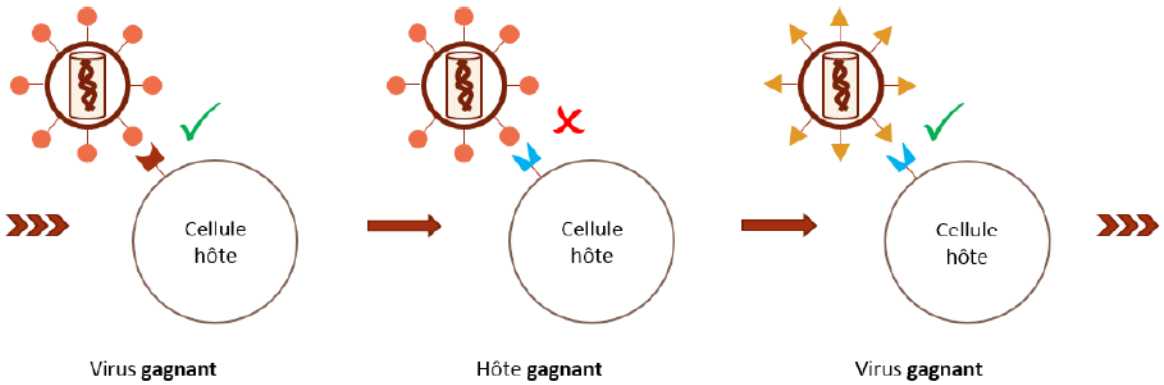


Figure 10 : Représentation schématique de la « course aux armements » de l'interaction hôte-virus.

1.5.2. Le cas des virus endogènes

Parmi les virus qui sont capables de s'intégrer au génome, les rétrovirus sont les plus connus. La phase d'intégration du génome viral au sein du génome hôte est une étape indispensable de leur cycle infectieux. Même si chaque espèce rétrovirale a une affinité pour un type cellulaire particulier (tropisme), l'intégration peut en pratique avoir lieu dans l'ADN de n'importe quel type cellulaire, que ce soit des cellules somatiques ou reproductrices (germinales) des animaux. Les cellules reproductrices sont celles qui transmettent le patrimoine génétique de l'hôte à la descendance et s'il y a eu intégration virale, le virus peut donc alors être transmis verticalement. C'est pourquoi il est commun de trouver dans les génomes animaux une certaine proportion de séquences rétrovirales et de nouveaux rétrovirus endogènes acquis au cours de l'évolution (Katzourakis *et al.*, 2007). L'un des meilleurs exemples de l'importance de l'intégration des rétrovirus dans les génomes animaux est l'émergence des mammifères placentaires. En effet, la formation du placenta est d'origine rétrovirale. Plus précisément, le détournement d'un gène d'enveloppe rétroviral, intégré au génome des animaux a permis d'utiliser la fonction de cette enveloppe, gène de la syncytine, dans la formation du placenta, barrière materno-fœtale devenue indispensable à la reproduction des mammifères placentaires (Dupressoir *et al.*, 2012).

Cependant, les rétrovirus n'ont pas le monopole de l'intégration aux génomes hôtes. Un autre exemple très particulier est celui de l'interaction guêpe parasitoïdes – polydnavirus. Les polydnavirus sont de grands virus à génome ADN. Il y a environ 100 millions d'années, l'intégration d'un nudivirus ancestral a été à l'origine de la domestication de ce génome viral par certaines guêpe parasitoïdes (*Braconidae* et *Ichneumonidae*), permettant à ces guêpe d'infecter des chenilles de lépidoptères hôtes (Drezen *et al.*, 2017b). En effet, ces petites guêpes, en plus de l'injection des œufs dans la chenille, injectent des particules virales contenant les cercles viraux qui portent des gènes de virulence, qui une fois exprimés, permettent de contrôler la réponse immunitaire de la chenille et d'assurer le bon développement des œufs de la guêpe dans la chenille parasitée (Herniou *et al.*, 2013).

Ces deux exemples ne représentent qu'une petite partie des éléments viraux endogènes (*Endogenous Viral Elements*, EVE) qui ont été « domestiqués » par les animaux. Finalement, ce sont de nombreux types de virus, qu'ils soient à ADN ou ARN ou qu'ils aient un grand ou petit génome, qui sont capables de s'intégrer au génome hôte et ainsi jouer un rôle important dans l'évolution des animaux (Katzourakis & Gifford, 2010).

1.5.3. Le changement d'hôte et les virus émergents

Le changement d'hôte ou saut d'hôte pour les virus revient à changer de niche écologique et à s'adapter à un nouvel hôte et ses caractéristiques. Les maladies virales émergeantes sont dans certains cas le résultat d'un changement d'hôte, permettant d'agrandir la gamme d'hôte du virus concerné (Geoghegan *et al.*, 2017). Il a été montré que les changements d'hôtes sont très fréquents dans une vingtaine de familles virales à ADN et ARN tels que les *Rhabdoviridae*, *Flaviviridae*, *Reoviridae* ou *Togaviridae* (Geoghegan *et al.*, 2017).

Le saut d'hôte peut être induit par des changements dans la population hôte initiale (forte proximité des hôtes), dans la population virale, ou encore par des changements liées à l'environnement ou au climat dans lesquels évoluent les virus et leurs hôtes (Jones *et al.*, 2008; Wasik & Turner, 2013). Le saut d'hôte à lieu le plus fréquemment parmi des hôtes phylogénétiquement proches, principalement car les virus peuvent infecter et se répliquer dans des hôtes qui possèdent des récepteurs cellulaires avec moins de divergence (Parrish *et al.*, 2008). Cependant, il existe des barrières à franchir pour qu'un virus puisse changer d'hôte : la barrière d'espèce hôte principalement et la barrière tissulaire via les récepteurs qui doivent être compatibles avec les capsides virales (Parrish *et al.*, 2008). Il a par exemple été montré que le *Canine parvovirus* (CPV) infectant les chiens aurait changé sa gamme d'hôte. Il est en effet capable de se lier spécifiquement au récepteur orthologue de la transferrine canine, alors que son ancêtre, le *Feline panleukopenia virus* (FPV) infecte les félins seulement (Hueffer *et al.*, 2003).

Enfin, deux notions sont importantes dans ces changements d'hôtes : la notion de vecteur, et la notion de réservoir (Mackenzie & Jeggo, 2013; Mandl *et al.*, 2015). Les réservoirs sont des hôtes qui possèdent de nombreux virus pouvant les infecter mais également infecter d'autres hôtes par transmission directe ou vectorisée. Les principaux réservoirs de virus humains sont les chauves-souris (*Paramyxoviridae*, *Filoviridae*, *Flaviviridae*), les rongeurs (*Arenaviridae*, *Hantavirus*) ou encore les oiseaux (*Influenza virus*) et d'autres primates (SIV) et mammifères (Calisher *et al.*, 2006; Quan *et al.*, 2013). Les vecteurs sont des organismes qui transmettent un pathogène d'un hôte (réservoir) à un nouvel hôte. Les vecteurs des virus d'animaux les plus communs sont les arthropodes, parmi lesquels les moustiques et les tiques qui sont vecteurs de nombreux arbovirus tels que les *Flaviviridae*, *Bunyavirales*, *Togaviridae*, *Reoviridae*, *Rhabdoviridae* (Conway *et al.*, 2014; Kuno & Chang, 2005; Shi *et al.*, 2015; Vasilakis & Tesh, 2015).

1.6. Les différents mécanismes qui permettent l'évolution virale

1.6.1. Les changements des séquences génomiques

Les changements d'acide nucléique dans une séquence génomique est appelée mutation. Les mutations sur les gènes peuvent ou non avoir un impact sur la fonction des protéines codées. Pour les virus, et plus particulièrement les virus à ARN, les mutations sont très fréquentes. En effet, les ARN polymérases sont peu fidèles et possèdent un taux de mutation très élevé : il y a en moyenne une erreur tous les 10^4-10^5 nucléotides polymérisés. Pour un génome ARN de 10 kb avec un taux de mutation 10^{-4} par nucléotide, il existe environ une erreur par génome néosynthétisé (Drake & Holland, 1999). Le taux de mutation pour les virus à ADN est 10 000 fois moins élevé à 10^{-8} par base et par génome copié. Toutes ces mutations, couplées à une large production de nouveaux virions permettent aux virus d'avoir un mélange de variant d'un même génome viral au sein d'un seul hôte, aussi appelé quasi-espèce. En effet, les virus sous forme de virions, sont très nombreux au sein d'un hôte et la population virale peut être très importante. A titre d'exemple, 10^9 à 10^{11} particules virales du VIH et du virus de l'hépatite B (HBV) sont respectivement présentes dans le sang d'un hôte et renouvelées à 50 % (HBV) et 90 % (VIH) toutes les 24 h (Flint *et al.*, 2015a).

La quasi-espèce en virologie est une notion controversée (Holmes, 2010; Holmes & Moya, 2002). La notion de quasi-espèce virale se caractérise comme étant un groupe de virus similaires, possédant des variations génomiques différentes, étant présents dans un même environnement (Lauring & Andino, 2010). Malgré les taux de mutations importants des génomes viraux, il s'agit d'un processus d'équilibre entre mutation et sélection naturelle qui génère une population de génomes variables, organisée autour d'un génotype consensus stable et conservé dans le temps (Holmes, 2010) (Figure 11)., Les mutations apparaissant au sein de la quasi-espèce peuvent alors être conservées par la sélection naturelle et pourront, en devenant majoritaires, être considérées comme des substitutions observées dans le génome consensus.

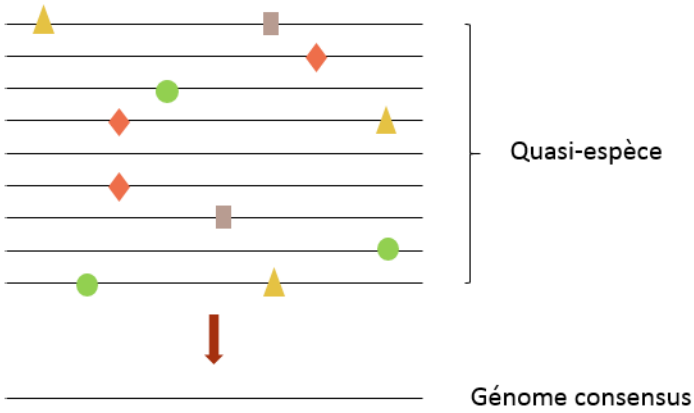


Figure 11 : L'effet de quasi-espèce.
Chaque génome viral présent au sein d'un hôte possède ses propres mutations qui n'affectent pas le génome viral consensus.

1.6.2. La dérive génétique

La dérive génétique est la force évolutive qui est due au hasard. La dérive génétique est le résultat de l'échantillonnage de la diversité génétique d'une population, cette dernière pouvant passer par un goulot d'étranglement génétique (*genetic bottleneck*). Ce goulot d'étranglement a lieu lorsqu'une population est drastiquement réduite et lorsqu'une petite partie de la population colonise un nouvel environnement ou un nouvel hôte/organe chez les virus (Zwart & Elena, 2015). Les variations au sein de la population virale peuvent alors être réduites lorsqu'il y a par exemple infection d'un nouvel hôte, seule une petite proportion de la population peut passer à travers ce goulot d'étranglement et l'effet de la dérive génétique est alors plus apparent dans ces petites populations.

Ce type de goulot d'étranglement à lieu très couramment en réalité pour les virus. Si on prend l'exemple du virus de la grippe, à chaque éternuement d'un individu infecté, une petite proportion de la population virale est libérée et une plus petite encore pourra infecter un nouvel individu. Mais les caractéristiques des virus font que les effets néfastes de la dérive (accumulation de mutations délétères) peuvent être amoindris lors des transmissions. En effet, le nombre de nouvelles particules virales peut être très important et ils sont capables de muter rapidement après un goulot d'étranglement, et donc de générer une nouvelle population constituée de nombreux virions fonctionnels.

1.6.3. Les transferts de gènes

Le transfert de gènes est un moyen pour les virus d'échanger leurs gènes, que ce soit à l'échelle intra ou inter-espèces virales (réassortiment ou recombinaison) ou entre un virus et son hôte (recombinaison, transfert horizontal).

Le réassortiment à lieu principalement lorsque deux virus segmentés similaires échangent une partie de leurs génomes durant la co-infection d'une même cellule (Vijaykrishna *et al.*, 2015). Le réassortiment à lieu systématiquement pour tous les virus segmentés et est particulièrement bien étudié pour les virus de la grippe (Marshall *et al.*, 2013; McDonald *et al.*, 2016).

La recombinaison, dont deux types existent, est le second moyen pour les virus d'échanger des gènes. La recombinaison entre une même espèce virale par recombinaison homologue à lieu chez de nombreux virus, que ce soit des virus à ADN et des rétrovirus (Martin *et al.*, 2011; Onafuwa-Nuga & Telesnitsky, 2009). Enfin, il existe la recombinaison entre un virus et le génome hôte eucaryote, appelé transfert horizontal de gènes. Ce type d'échange de gènes présent chez les virus à ARN (Liu *et al.*, 2010) est très fréquent chez les grands virus à ADN (Drezen *et al.*, 2017a; Gilbert *et al.*, 2014).

1.7. Quels sont les outils qui permettent la découverte et la description des virus ?

Différents outils permettent à l'heure actuelle d'étudier la biodiversité virale. La phylogénie permet de définir les relations de généalogie de gènes et les relations de parenté entre espèces virales. Elle permet également de résoudre la taxonomie et la systématique virale en proposant un outil de classification ayant un sens du point de vue de l'évolution. La génomique et surtout la métagénomique permettent actuellement d'avoir accès à des informations plus larges et plus importantes.

1.7.1. La phylogénie : relations et distance évolutives des gènes et des espèces

L'utilisation de la phylogénie moléculaire permet de résoudre les relations évolutives entre organismes ou gènes par une combinaison de techniques moléculaires et statistiques. Un

arbre phylogénétique représente alors les liens qui unissent plusieurs unités taxonomiques qu'il s'agissent d'espèces ou de gènes (feuilles de l'arbre). Le lien qui uni deux taxa correspond à leur ancêtre commun (nœuds internes de l'arbre), reliés par les branches de l'arbre et dont la longueur correspond à une distance représentant un nombre de changements entre les taxas (Figure 12).

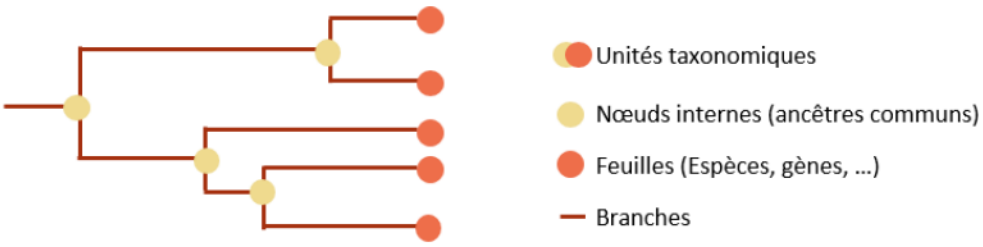


Figure 12 : Schéma d'un arbre phylogénétique et de ses caractéristiques.

L'utilisation de l'outil phylogénétique permet à l'heure actuelle de pouvoir définir la classification du vivant, comme ce qui a été le cas pour la classification des insectes à partir de 103 espèces, en utilisant 1478 gènes codants, permettant de dater la diversification de cette classe il y a environ 479 millions d'années (Misof *et al.*, 2014).

Les données moléculaires sont très abondantes et se révèlent particulièrement utiles lorsque l'on étudie les microorganismes pour lesquels des données morphologiques ou physiologiques ne sont pas ou peu existantes. Les approches moléculaires ont ainsi complètement révolutionné la taxonomie des microorganismes tels que les bactéries notamment (Woese, 1987).

Pour les virus, l'utilisation de l'outil phylogénétique est tout aussi important puisque la phylogénie est l'un des critères, et souvent le seul disponible, qui permet de replacer et définir la systématique des virus. Pour les virus à génome ARN par exemple, l'ARN polymérase est la protéine importante qui permet cette classification que ce soit avec la séquence nucléotidique ou protéique (Baker & Schroeder, 2008; Koonin, 1991; Koonin & Dolja, 1993). Puisqu'elle est partagée par un grand nombre de familles virales, elle est devenue le marqueur phylogénétique de choix pour les virus à ARN. Enfin, la phylogénie permet de résoudre les questions liées à l'origine et l'évolution des virus de manière plus générale (Nasir & Caetano-Anolles, 2015).

1.7.2. Distinguer la génomique de la génomique environnementale

Au XXI^{ème} siècle, les dernières avancées en termes de séquençage haut débit ont permis l'accès à de nouvelles informations qui étaient alors inaccessibles, de par notamment l'augmentation de la puissance des techniques (en nombre de séquence, profondeur de séquençage, taux d'erreur plus bas).

On peut alors distinguer la génomique de la génomique environnementale. La génomique est classiquement l'étude d'un organisme particulier, du séquençage du génome d'un ou de quelques individus de cette espèce. La génomique environnementale se place à une autre échelle spatiale, à l'échelle d'un environnement complexe, qui peut correspondre au sol d'une forêt, à un litre d'eau de mer ou à un organisme entier incluant ses symbiotes, ses virus et bactéries commensaux et ceux avec qui il est en interaction. Ce type d'analyse fournit un grand nombre de données d'origines multiples et permet d'apporter des connaissances approfondies qui nécessitent de la multidisciplinarité (Ungerer *et al.*, 2008).

1.7.3. Quelles sont les options disponibles pour l'étude des virus ?

1.7.3.1. Le code-barres ADN ?

Lorsque l'on parle de diversité microbienne, les technologies de séquençage, déjà utilisées à la fin des années 1980, ont été une avancée majeure dans la description de la flore bactérienne, pour laquelle une très grande partie était alors inconnue car tout simplement non cultivable en laboratoire. Ce fut notamment une révolution initiée par l'introduction du séquençage de l'ARN ribosomique 16S (Lane *et al.*, 1985), permettant l'analyse de tous types de procaryotes et montrant le grand intérêt de ce type d'études, permettant d'avoir accès à la diversité d'organismes.

Des études de diversités animales utilisent principalement le gène de la Cytochrome Oxydase I pour déterminer quelles espèces sont présentes dans un échantillon mixte (Hebert *et al.*, 2003a, b). Cette méthodologie dite de code-barres à ADN (*DNA barcoding*) est d'autant plus efficace qu'une base de donnée spécialisée, qui regroupe un très large ensemble de données, permet facilement à l'aide de la courte séquence du COI (700bp) d'identifier une espèce (Ratnasingham & Hebert, 2007).

Le problème est cependant plus compliqué lorsque l'on s'intéresse à la diversité des virus. Il n'existe pas d'équivalence au 16S ou au COI ni de gènes ou d'éléments génétiques commun à tous les virus (Rohwer & Edwards, 2002). La méthodologie de code-barres n'est donc pas appropriée.

1.7.3.2. La métagénomique virale et la métaviromique

La métaviromique est l'étude spécifique des virus présents dans un environnement complexe, comme évoqué précédemment il s'agit bien ici d'étude type de génomique environnementale.

La métaviromique, de manière concise, consiste à purifier et séquencer toutes les particules virales présentes dans un même environnement. Les premières études de ce type ont été réalisées à partir d'eau de mer (Breitbart *et al.*, 2002), de sédiments marins (Breitbart *et al.*, 2004) ou encore de fèces humaines (Breitbart *et al.*, 2003) et équines (Cann *et al.*, 2005). Parmi les études les plus récentes, une étude à l'échelle globale de la Terre a permi de montrer que la grande diversité virale découverte n'a en réalité aucunes similitudes avec ce qui est connue à ce jour à près de 99% (Paez-Espino *et al.*, 2016).

Ainsi de très nombreuses études ont montré la grande importance de l'utilisation de données de séquençage haut-débit et d'étude de métagénomique pour la découverte de nouveaux virus d'une manière générale (Bexfield & Kellam, 2011; Delwart, 2007; Edwards & Rohwer, 2005; Lipkin & Firth, 2013; Marz *et al.*, 2014; Mokili *et al.*, 2012; Radford *et al.*, 2012; Rosario & Breitbart, 2011; Rose *et al.*, 2016; Suttle, 2016).

De nombreux exemples montrent ainsi la grande diversité de virus présents par exemple dans le milieu marin (Brum *et al.*, 2015; Paul & Sullivan, 2005), chez les plantes (Adams *et al.*, 2009; Barba *et al.*, 2014) ou chez les insectes (Bichaud *et al.*, 2014; Junglen & Drosten, 2013; Li *et al.*, 2015; Liu *et al.*, 2011).
Des recherches plus ciblées sur les arthropodes et les invertébrés ont mis en évidence que l'étude d'un groupe d'espèce particulier permet d'observer la richesse spécifique des virus présents dans ces espèces et également d'observer des gammes d'hôtes plus importantes que précédemment connues pour certaines familles virales (Li et al., 2015; Shi et al., 2016).

Il a été proposé que les pratiques en termes de définition des séquences virales, de génome consensus et les notions de génomes partiels ou complets soient standardisées afin de répondre à l'arrivée des analyses à grande échelle (Ladner *et al.*, 2014). La classification et la définition de la taxonomie virale ont commencé a évoluer par l'arrivée d'un nombre important de données générées par la métagénomique virale et la métaviromique (Simmonds *et al.*, 2017).

Différents outils d'analyse spécialisés dans la métaviromique ont vu le jour avec l'avancée de ces techniques (Ho & Tzanetakis, 2014; Sharma *et al.*, 2015). Ce qui fait que les études de métaviromique ont maintenant de multiples applications : l'étude des virus en milieux aquacole (Alavandi & Poornima, 2012) ou en clinique (Quiñones-Mateu *et al.*, 2014), ou encore l'étude et la prévention des zoonoses (Temmam *et al.*, 2014).

1.8. Objectifs et structure de la thèse

1.8.1. Objectifs de la thèse

Le but principal de mon travail de thèse a été de répondre à la question :

Comment les virus sont-ils représentés au sein du monde animal ?

Pour cela je me suis intéressée à l'étude de transcriptomes d'animaux non-modèles me permettant d'étudier la diversité et l'évolution des virus dans ces hôtes.
Les hypothèses principales de cette étude d'animaux peu connus sont que cette approche permettrait d'avoir accès à de nouveaux virus pour lesquels la taxonomie, l'organisation génomique, les hôtes sont inconnus. Cette étude pourrait également permettre d'élargir nos connaissances sur des virus connus, comme la description de nouvelles associations hôtes-virus.

1.8.2. Mon approche – analyse de la diversité virale via la méta-transcriptomique

Comme évoqué précédemment, différentes méthodologies permettent actuellement d'étudier la biodiversité virale, c'est le cas notamment de la métaviromique. Cependant le principal inconvénient est qu'elles ne permettent pas d'avoir accès aux informations relatives aux hôtes, s'agissant généralement de mélanges d'organismes séquencés. Chaque échantillon du projet TARA Oceans par exemple est issu de la purification des particules virales provenant de 20L d'eau de mer (Brum *et al.*, 2015).

C'est pourquoi j'ai décidé d'utiliser une autre technique basée sur des transcriptomes. Les transcriptomes correspondent au séquençage des transcrits ARN messager d'un hôte et sont donc le reflet des gènes exprimés chez cet hôte. Ils sont ainsi très généralement utilisés pour des études fonctionnelles et d'expression des gènes.

Ce qui est intéressant avec les organismes quels qu'ils soient, c'est qu'ils sont complexes. Ce que j'entends par complexe est le fait qu'un organisme peut héberger toute une flore microbienne (bactéries intestinales par exemple) mais également transporter avec eux toute les bactéries et virus présents dans leur environnement. Un organisme est alors considéré comme un environnement complexe à part entière, aussi appelé holobionte (Theis *et al.*, 2016). Ainsi, le séquençage du transcriptome d'un hôte animal permet d'avoir accès au méta-transcriptome avec des transcrits bactériens ou fongiques, des transcrits de virus à ADN et surtout ce qui m'intéresse plus particulièrement ici des génomes de virus à ARN, par le biais des transcrits qui leurs sont propres.

1.8.3. Structure de la thèse

La thèse s'articule en cinq sections se terminant par une discussion générale sur l'ensemble du travail réalisé lors de cette thèse.

La seconde section de cette thèse est celle du **Matériels et Méthodes.** Elle consiste en premier lieu à la description du matériel biologique utilisé, quelles sont les espèces animales que j'ai étudiées et quelles sont les données que j'ai eues à ma disposition pour réaliser ces analyses.

Dans un second temps, sont abordées la description complète et détaillée de toutes les analyses qui permettent d'étudier la diversité virale.

La troisième section correspond au **Chapitre 1.** Cette section illustre la preuve de faisabilité de la découverte de nouveaux virus libres **(Article 1)** et endogènes **(Article 2 & Article 3)** par l'utilisation de cette méthodologie.

La quatrième section constitue le **Chapitre 2.** Il s'agit de faire un zoom spécifique de la démonstration de la nécessité de découvrir de nouveaux virus dans le cas de l'étude des abeilles **(Article 4)** et de l'épidémiologie des virus d'abeilles **(Article 5)**. Dans cette section, est abordé de manière plus ciblée la découverte et la description d'un nouveau virus d'abeille sauvage et de la présence d'un virus d'abeille chez des fourmis (**Article 4**) ainsi qu'une étude épidémiologique de virus d'abeilles dans un large échantillonnage d'hyménoptères sauvages et de la description de nouveaux hôtes de ces virus (**Article 5**).

Enfin, la cinquième et dernière section est une **Discussion générale** qui traite de manière plus synthétique de l'évaluation globale de la diversité virale au sein des transcriptomes d'animaux non-modèles étudiés ici. Cette discussion sur l'ensemble de ce travail mettra en perspective l'utilité de ce type d'étude en regard de l'immensité de ce qu'il reste à découvrir sur la diversité virale au sein des métagénomes animaux.

MATERIELS ET METHODES

523 raw transcriptomes

135 animal species

Methods

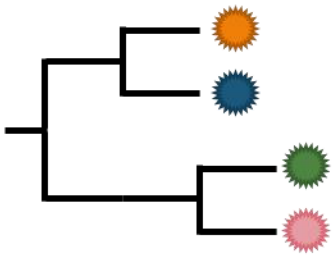
?

Viral genome coverage

New virus ?

Viral genomic organization

Viral phylogenetic analyses



2.1. Préambule au matériels et méthodes

Cette partie méthodologique a pour but de définir quelle est exactement la méthodologie que j'ai mise en place par ce travail.

Pour cela, dans un premier temps je vais définir quelles sont les données biologiques que j'ai utilisées, leur origine, la manière dont elles ont été acquises et ce qu'elles représentent (en terme de diversité animale et en terme de volume de données informatiques).

Puis, je décrirais en détail, étape par étape, le cheminement qui m'a permi à partir de données brutes, issues de 135 espèces d'animaux sauvages, de répondre aux questions : Comment trouve-t-on des virus dans les transcriptomes ? Quelles sont les informations qui me permettent de quantifier la diversité virale dans ces données ? Quelles sont analyses qui permettent de caractériser les nouveaux virus (organisation du génome, relations phylogénétiques…).

Je finirai par montrer en quoi il a été nécessaire de créer une nouvelle méthodologie pour l'étude et la découverte de nouveaux virus à partir des transcriptomes.

2.2. Matériel biologique

Les données dont je dispose sont issues d'un projet nommé PopPhyl (http://kimura.univ-montp2.fr/PopPhyl/) initié par Nicolas Galtier à l'Université de Montpellier, qui se proposait d'étudier la génomique des populations à partir de transcriptomes hauts débits. Ce projet avait pour but de caractériser l'évolution moléculaire chez un nombre important de métazoaires non modèles. Ces données sont utilisées ici dans un autre but : les espèces animales ne sont pas étudiées pour leur diversité génétique, mais en tant qu'hôte de virus. Une analyse de méta-transcriptomique virale sera mise en place afin de quantifier la biodiversité virale présente au sein de la partie exprimée des génomes animaux. Je dispose de 523 transcriptomes obtenus par la méthode Illumina, appartenant à 135 espèces d'animaux sauvages (répartis en 59 groupes taxonomiques) et incluant une fraction non négligeable de la diversité des métazoaires, dont des tortues des Galápagos, des huitres, des abeilles… (Cahais *et al.*, 2012; Gayral *et al.*, 2011, 2013, Romiguier *et al.*, 2014a, b) (Figure 13, Annexe 1).

Pour chacun des groupes taxonomiques, 10 individus vivants ont été échantillonnés à travers l'étendue géographique de l'espèce principale, ainsi que deux individus par espèces, chez deux à trois espèces phylogénétiquement proches servant de groupes externes (Figure 14). Les animaux représentent une grande diversité d'animaux non-modèles provenant des huit principaux phyla métazoaires (Annelida, Arthropoda, Chordata, Cnidaria, Echinodermata, Mollusca, Nematoda et Nemerta). Le phylum Haptophyta, correspondant à des algues unicellulaires, représente un clade externe (Figure 14, Tableau 2). Il s'agit d'animaux sauvages échantillonnés de manière à représenter la diversité géographique et génétique de chacune des espèces.

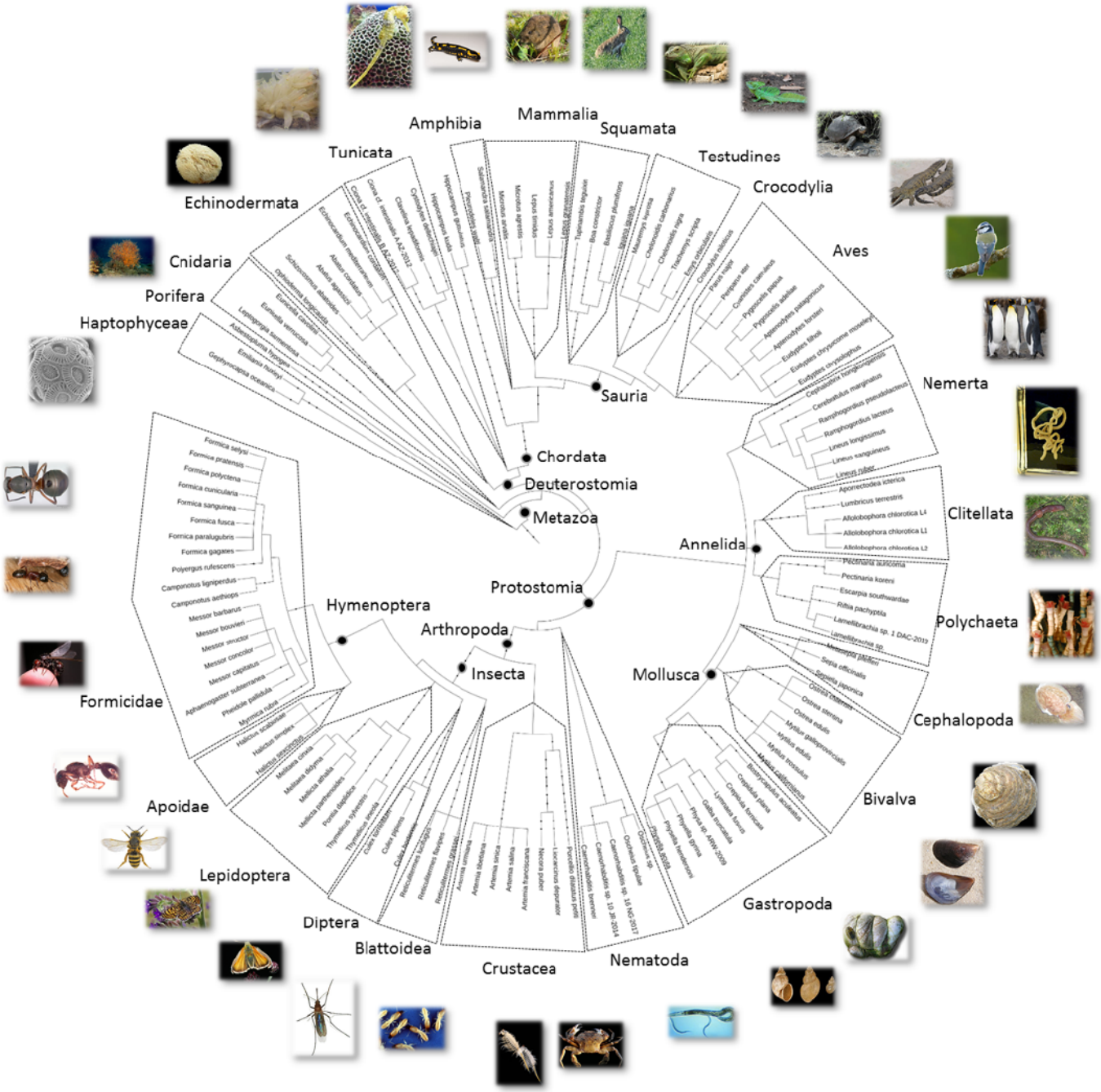


Figure 13: Arbre phylogénétique des 135 espèces étudiées.

Les feuilles de l'arbre représentent les espèces, les nœuds intermédiaires et les encadrés correspondent aux différents rangs taxonomiques étudiés. L'arbre a été réalisé à l'aide de l'outil PHY-FI (http://cgi-www.daimi.au.dk/cgi-chili/phyfi/go) (Fredslund, 2006) en utilisant les identifiants taxonomiques du NCBI. Visualisation de l'arbre par iTOL (http://itol.embl.de) (Letunic & Bork, 2007, 2011). Images : http://species.wikimedia.org/wiki/

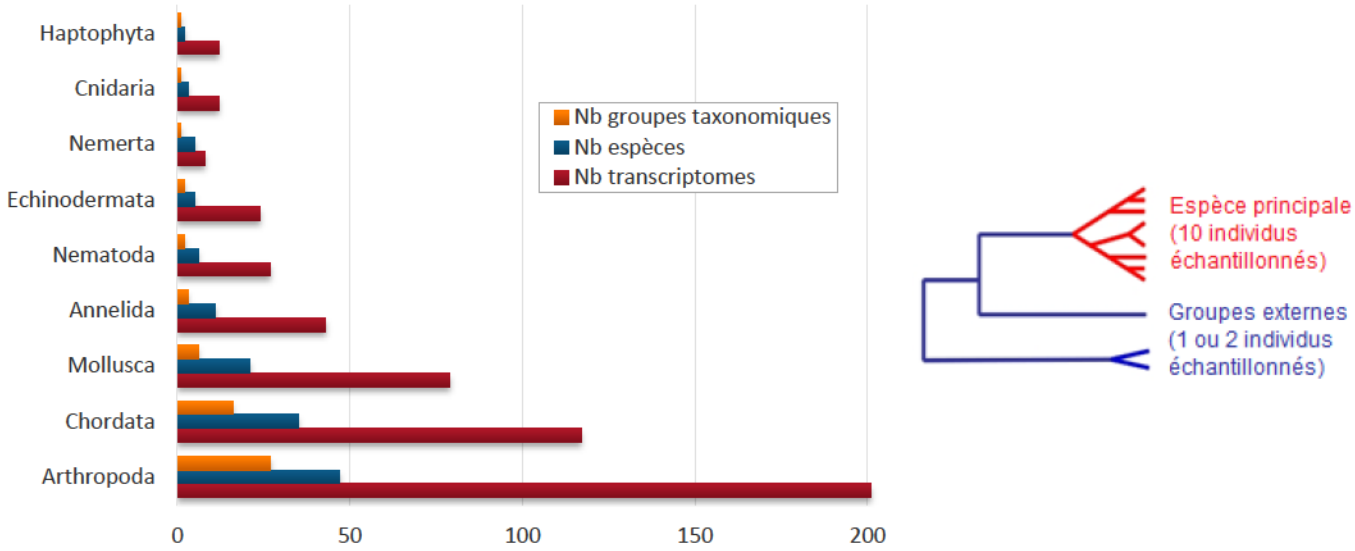


Figure 14 Répartition du nombre de groups taxonomiques et du nombre de transcriptomes en fonction des huit grands phyla représentatifs des métazoaires et du phylum externe Haptophyta (*Emiliania huxleyi*).

Tableau 2: Nombre d'espèces et de transcriptomes échantillonnés dans chacun des phyla animaux et répartition des espèces suivant leur habitat d'origine.

PHYLUM	NB ESPECES	HABITAT TERRESTRE	HABITAT AQUATIQUE	NB TRANSCRIPTOMES
ARTHROPODA	47	41	6	201
CHORDATA	35	28	7	117
MOLLUSCA	21	0	21	79
ANNELIDA	11	5	6	43
NEMATODA	6	6	0	27
ECHINODERMATA	5	0	5	24
NEMERTA	5	5	0	8
CNIDARIA	3	0	3	12
HAPTOPHYTA (GROUPE EXTERNE)	2	0	2	12
TOTAL	**135**	**85**	**50**	**523**

Les ARN totaux ont été extraits des 523 individus adultes échantillonnés dans le projet PopPhyl en suivant des protocoles standards et modifiés (Gayral *et al.*, 2011), soit à partir d'organes (ou de prélèvements sanguins) uniques pour les grands animaux, soit à partir d'individus entiers pour les plus petits. Puis les transcriptomes individuels ont été séquencés à l'aide de la technologie Illumina (~ 4.000.000 reads ~ 50-100 bp par individus), en single ou pair-end (séquençage d'un seul fragment par deux reads différents espacé d'une taille connue) pour la plupart des individus. Le volume du jeu de données initiales représente plus 11,3 milliard de reads ce qui correspond à un très large volume de données informatiques compressées (>2,5 To).

2.3. Méthodologies bio-informatiques

2.3.1. Phase Automatique

2.3.1.1. Assemblage des lectures de séquençage

L'assemblage *de novo* des lectures a été réalisé individuellement, c'est-à-dire un transcriptome par individu séquencé, ceci pour chacune des espèces des différents groupes taxonomiques. Deux logiciels: ABYSS (Assembly By Short Sequences) (Birol *et al.*, 2009; Simpson *et al.*, 2009) et CAP3 (Huang & Madan, 1999) ont été utilisés consécutivement (Figure 15).

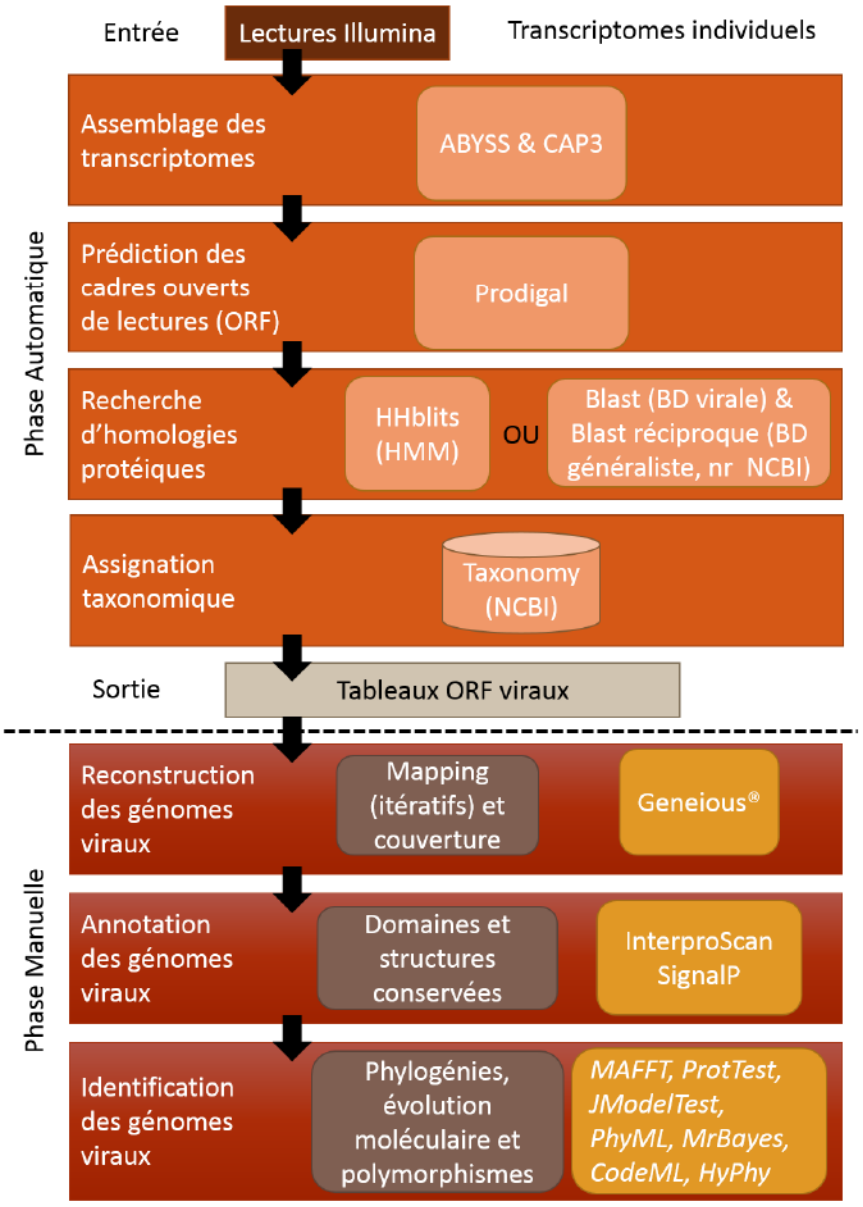


Figure 15: Pipeline bioinformatique développé lors de ce travail.
La phase automatique correspond à un seul script mise en place afin d'obtenir une liste d'ORF viraux à partir des transcriptomes bruts. La phase manuelle correspond aux différentes analyses réalisées *a posteriori* afin d'obtenir et de caractériser de nouveaux génomes viraux. Les différents logiciels et base des données utilisées sont indiqués par les bulles à l'intérieur de chaque étape représentée par les rectangles. Chacune des étapes illustrées ici sont développées dans le texte associé.

Le programme ABYSS possède un algorithme d'assemblage basé sur la théorie des graphes de De Bruijn permettant d'assembler les lectures en définissant un k-mer qui correspond à la taille minimale de bases qui doivent être identiques entre deux lectures pour pouvoir les assembler. Le k-mer utilisé est de k=60 ; c'est une valeur parmi celles possibles qui convient à nos données ayant des lectures de 100 bases (Cahais *et al.*, 2012), sauf pour les jeux de données avec des lectures de 50 bases, pour lesquelles cas un k-mer de k=40 a été utilisé (Robertson *et al.*, 2010). Les valeurs par défaut ont été choisies pour les autres paramètres du programme. Le programme CAP3 pour sa part utilise une méthode d'alignement multiple de séquences pour générer une séquence consensus, donne ensuite un poids aux bases suivant la qualité de l'alignement. Chaque jeu de données de lectures a été assemblé consécutivement par ABYSS puis CAP3. Il a en effet été montré sur ce même jeu de données que cette procédure améliorait la qualité des assemblages par rapport à des assemblage par ABYSS ou CAP3 seuls, ou par d'autres assembleurs tels que Mira, Trinity, Newbler ou Soap (Cahais *et al.*, 2012). La qualité de l'assemblage est déterminée par le nombre de contigs assemblés, la taille du plus long contig, la taille médiane des contigs (longueur des contigs pour laquelle la moitié des contigs ont une taille > médiane) et la statistique N50 (longueur du contig pour laquelle la moitié des nucléotides assemblés se trouvent dans des contigs de tailles > N50).

Cette méthodologie a été appliquée à tous les transcriptomes auxquels j'ai eu accès, y compris ceux de termite *Reticulitermes* qui ont fait l'objet d'une publication à lauqelle je suis associée, dans le cadre d'une autre étude (Dedeine *et al.*, 2015) (Annexe 5).

2.3.1.2. Prédictions des cadres ouverts de lecture

Une fois les assemblages réalisés, une recherche d'ORF (Open Reading Frame, cadre ouvert de lecture) est effectuée à l'aide du logiciel Prodigal (PROkaryotic DYnamic programming Gene-finding ALgorithm), conçut pour détecter des ORF bactériens (Hyatt *et al.*, 2010), et adapté à la métagénomique sur d'autres espèces (Hyatt *et al.*, 2012) (Figure 15). Prodigal recherche des codons d'initiation et stop mais peut également trouver des ORF tronquées en 5' et/ou 3'. Les séquences issues de cette analyse sont par la suite appelées ORF putatifs. L'étape suivante consiste, après la traduction de ces ORF putatifs, à procéder à une étape de recherche d'homologie protéique.

2.3.1.3. Homologies protéiques

Deux méthodologies ont été utilisées au cours de ce travail (Figure 15). La première en début de thèse utilisait le programme HHBlits (**Articles 1, 2 et 4**) et la seconde par la suite utilisait le programme BLAST (**Article 3**). La seconde à l'avantage de permettre la découverte de nouveaux virus en un temps plus court et d'obtenir des résultats plus fiables avec moins de faux-positifs que la première, qui fût néanmoins fiable (voir Chapitre 1 et 2).

HHBlits

L'annotation des ORF putatifs a été réalisée par recherche d'homologies protéiques à l'aide du programme HHblits (Figure 15) implémenté dans le logiciel HHsuite (Remmert *et al.*, 2011; Söding, 2005; Söding *et al.*, 2005). Cet algorithme utilise des profils HMM (Hidden Markov Model = modèle de Markov caché) pour détecter des similarités de séquences et de structure secondaire avec les protéines de la base de données nr20. Cette base de données est issue de la base non-redondante (nr) du NCBI (National Center for Biotechnology Information), pour laquelle les séquences sont regroupées en clusters de protéines s'alignant sur un minimum de 80 % de leur longueur et ayant des identités de séquence > 20 %. Ces clusters forment des alignements multiples représentés de façon concise en profils HMM ; ceux-ci contiennent pour chaque position de la séquence principale les probabilités d'observer chacun des 20 acides aminés dans les protéines homologues et d'observer (et de prolonger) une insertion ou une délétion après cette position. Les valeurs par défaut ont été choisies pour tous les paramètres du programme. Le programme calcule alors pour chaque ORF potentiel la e-value (nombre moyen de faux positifs avec un score meilleur que celui de la cible lors de la recherche dans la base de données), ainsi qu'une p-value (probabilité qu'une mauvaise comparaison par paire ait un meilleur score que la cible). J'ai écrit un script en langage python afin de récupérer le premier résultat d'homologies protéiques pour chacun des ORF potentiels. Afin de diminuer le nombre de faux positifs, les ORF ayant une E-value > 10^{-3}, et une probabilité < 95 % ont été éliminés.

BLAST & BLAST réciproque

Afin de cibler plus spécifiquement les ORF viraux d'intérêt, les ORFs putatifs ont été soumis à deux étapes de recherche par homologies de séquence successives ; contre une base de données spécifique virale dans un premier temps, puis lors d'un BLAST réciproque

contre la base de données complète non-redondante du NCBI (Figure 15). Ces deux étapes séquentielles de recherche de similarité ont eu pour but de diminuer les faux positifs, mais aussi de diminuer les temps de calcul en éliminant un grand nombre de séquences à l'issue de la première recherche BLAST.

En effet, la première étape, réalisée avec le programme BLAST+ version 2.2.29 (Basic Local Alignment Search Tool+) (Camacho *et al.*, 2009), a débuté par la construction d'une base de données protéique virale réalisée à l'aide d'une liste d'identifiants (numéro GI) spécifiques des virus (séquences complètes et incomplètes) récupérée sur le site du NCBI. Cette base de données virale comportant 3 308 095 séquences a été extraite au format FASTA de la base de données protéique nr du NCBI. L'algorithme BLASTp du logiciel BLAST a permis la recherche d'homologies entre les séquences d'ORF et les séquences constituant cette base de données virales. Seules les ORF ayant au moins un résultat d'homologie avec une e-value inférieure ou égale à 10^{-1} ont été retenues dans la suite de la méthodologie.

Une seconde recherche de similarité par BLAST réciproque utilisant BLASTp, a été effectuée par comparaison des séquences obtenues lors de la phase précédente et des séquences de la base de données nr. Le seuil d'e-value choisi pour cette étape (10^{-3}), plus bas que celui de l'étape précédente, permet de diminuer le taux de faux-négatifs. Dans les deux étapes, seul le meilleur résultat ('best hit') d'homologie est conservé. Les ORF sont alors dits viraux lorsque le premier hit du BLAST réciproque conserve son homologie avec une séquence virale de la base de données nr.

2.3.1.4. Assignation taxonomique virale

Pour avoir une idée de l'espèce virale des ORF potentiels, la taxonomie complète (Règne, Embranchement, Classe, Ordre, Famille, Genre et Espèce ; appelée « full lineage » par la base de donnée « NCBI taxonomy database ») de la meilleure protéine homologue (obtenue par HHBlits ou BLAST/BLAST réciproque) lui a été attribué (Figure 15). Pour cela, la commande *blastdbcmd* du logiciel BLAST+ (Camacho *et al.*, 2009) a été utilisée afin de rechercher les identifiants taxonomiques du NCBI (les TaxID) associés aux identifiants protéiques du NCBI (les GI) issus de la recherche d'homologie. Ensuite, la taxonomie complète des TaxID viraux associée aux ORF prédites a été récupérée à l'aide d'un script que j'ai développé en langage Python. Ce programme utilise les fichiers de taxonomie «nodes.dmp», «division.dmp» et «names.dmp» disponibles sur le site Taxonomy du NCBI

(ftp://ftp.ncbi.nih.gov/pub/taxonomy/). Un fichier final a été créé résumant les informations associées à chaque ORF prédit provenant des étapes précédentes, ayant une séquence virale pour premier résultat de recherche d'homologie. Ce fichier résume les caractéristiques des ORF prédits et de leurs protéines homologues (individu, nom d'ORF, numéro accession protéine…), les statistiques de la recherche d'homologie (% identité, e-value, p-value, probabilité…) ainsi que la taxonomie assignée à cet ORF (Ordre, Famille, Genre et Espèce). Ce tableau final, réalisé pour chaque groupe taxonomique, est le point de départ de la phase manuelle de l'analyse.

2.3.2. Phase manuelle

Dans les tableaux finaux issus de la phase automatique, les ORF viraux seront classés en deux types : les ORF disséminés et les génomes entiers. Lorsqu'une espèce virale est représentée par un seul ORF au sein d'un même transcriptome, cet ORF est dit disséminé ou unique. Lorsque la même espèce virale est représentée par au moins deux ORF dans le même transcriptome, ces ORF sont classés dans la catégorie génomes viraux putatifs et serviront de base aux analyses suivantes de reconstruction, annotation et identification des génomes viraux (Figure 15). Certains virus ne codent que pour un seul ORF représentant une polyprotéine qui sera par la suite clivée, comme les *Iflaviridae* (ARNsb+). Dans ces cas spécifiques, une attention particulière est faite pour vérifier la présence de virus complets en cas de la détection d'un seul hit viral. Dans la suite de l'analyse, seules les séquences correnspondant à des génomes entiers ou quasi-complets seront analysées. Même si l'étude des ORF viraux disséminés détectés dans les transcriptomes est digne d'intérêt, leur analyse (non détaillée dans cette thèse) sur quelques transcriptomes nous a montré qu'ils correspondaient pour partie à des faux-positifs. Pour certains, le résultat d'homologie n'était pas tellement pertinent du fait de leur faible taille. Pour d'autres, ces gènes disséminés correspondaient en fait à des familles de gènes présents à la fois dans le génome viral et dans celui de leur hôte, leur assignation virale n'a donc pas pu être confirmée.

2.3.2.1. Reconstruction des génomes viraux: *mapping* et analyse des couvertures

Afin de confirmer la présence de virus entiers (libres ou intégrés dans le génome hôte), les lectures Illumina ont été alignées par *mapping* à l'aide du logiciel Geneious® (Kearse *et al.*, 2012) version 8.1.7 & 9 (http://www.geneious.com/) en utilisant la fonction « Map To

Reference » sur un génome correspondant à l'espèce désignée lors de l'assignation taxonomique récupéré dans GenBank (Figure 15).

Le paramètre "allow gap" n'est pas coché, et celui "Maximum Mistakes Per Read" est fixé à 10 %. Ce premier paramètre n'autorise pas de gaps dans les séquences alignées et permet un assemblage de meilleure qualité des lectures Illumina ; le deuxième paramètre rejette les lectures s'il y a plus de 10 % de mésappariement entre les lectures Illumina et la séquence de référence. L'utilisation de ces deux paramètres assure une stringence suffisamment importante pour garder toute la pertinence du *mapping* et éviter l'alignement de lectures Illumina non-homologues.

Ce *mapping* permet i) de confirmer les assemblages initiaux ayant conduits aux contigs de départ, ii) de rallonger les extrémités du génome viral consensus, iii) de vérifier l'absence de contigs chimères en visualisant l'homogénéité et la qualité des alignements des lectures, et iv) de valider la réelle présence de cette séquence à l'aide du calcul de la couverture à chaque position nucléotidique le long du génome et de la couverture moyenne.

A partir d'un premier *mapping*, un nouveau consensus est alors créé, servant à son tour de séquence de référence pour d'autres *mapping* itératifs (dont le nombre est dépendant de la complétude des génomes) réalisés de la même manière. Ces étapes ont permis la reconstruction de génomes viraux putatifs.

2.3.2.2. Annotation des génomes viraux : détection des domaines et structures conservés

Des recherches de domaines protéiques conservés ont été réalisées en utilisant trois programmes (Figure 15). Cela a permis d'en apprendre davantage sur la fonction probable des ORF des génomes nouvellement reconstruits. Le programme InterProScan (Jones *et al.*, 2014) (version5) utilisant la base de données de domaines protéiques InterPro (composées de signatures de 19 788 familles de protéines et 8 439 domaines protéiques) a été choisi afin de rechercher des signatures de domaines conservés dans les séquences protéiques. Il utilise pour cela une combinaison de différentes méthodologies telles que l'utilisation d'expression régulière et de profils de Markov cachés (Hidden Markov Models). SMART (Letunic *et al.*, 2012; Schultz *et al.*, 1998) (version7) utilise quant à lui les bases de données Swiss-Prot, SP-TrEMBL et les protéomes présents dans Ensembl pour effectuer ces recherches à l'aide de profils

HMM. Enfin, la base de données de domaines conservés du NCBI, NCBI Conserved Domain and Conserved Domain Database v3.14 a été utilisée (Marchler-Bauer *et al.*, 2011, 2015).

A ces analyses s'est ajoutée une détection de motifs particuliers de la structure primaire des génomes viraux. La détection de peptide signal a été faite par SignalP (Petersen *et al.*, 2011) (version 4.0) par l'intermédiaire de l'interface SMART. La détection de la localisation affectée aux peptides signaux fut ensuite déterminée par TargetP (Emanuelsson et al., 2000) (version 1.1). La détection de potentielles queue poly(A) a été faite par PolyApred (Ahmed *et al.*, 2009). La recherche de motifs spécifiques aux ARN et d'éléments régulateurs tels que les structures tRNA-like ont été réalisée via RegRNA 2.0 (Chang *et al.*, 2013). Enfin les structures du type Internal ribosomal entry sites (IRES) ont été prédites en utilisant Viral IRES Prediction System (VIPS), spécifiques aux virus en utilisant les paramètres par défauts (Hong *et al.*, 2013).

2.3.2.3. Identification des génomes viraux: phylogénies et évolution moléculaire

Phylogénies

Afin de déterminer la place du génome viral parmi les virus qui lui sont proches, des phylogénies ont été réalisées à partir du génome viral entier (phylogénie en nucléotide) ou des gènes (particulièrement les domaines conservés) utilisées dans la littérature comme marqueurs phylogénétiques, comme les polymérases et les capsides (phylogénie en acide aminés) (Figure 15).

Les alignements multiples nucléotidiques ou protéiques ont été réalisés avec MAFFT version 7.31 (Katoh *et al.*, 2002) en utilisant les paramètres par défauts. Les alignements ont été corrigés manuellement si nécessaire, et les sites non-homologues localisés aux extrémités ont été enlevés. Dans le but de retirer de l'alignement les sites trop variables pouvant présenter un risque de saturation masquant le signal phylogénétique, le programme Gblocks v0.91b (Castresana, 2000) a été utilisé sur les alignements avec des paramètres peu stringents permettant de conserver un ensemble de blocs de séquences les plus pertinents et informatifs.

Le choix du meilleur modèle de substitution de chaque alignement a été fait après calcul de différents modèles par JModelTest v2 (Darriba *et al.*, 2012) pour les alignements

nucléotidiques et ProtTest v3.2 (Abascal *et al.*, 2005) pour les alignements protéiques. Le meilleur modèle choisi est celui qui avait le plus faible AIC (Critère d'information d'Akaike), qui permet de comparer des modèles statistiques satisfaisant au critère de parcimonie.

Les phylogénies ont été réalisées, en utilisant le meilleur modèle, par la méthode du Maximum de Vraisemblance (Maximum Likelihood, ML) à l'aide du programme PhyML (Guindon & Gascuel, 2003) implémenté dans le logiciel Seaview (Gouy *et al.*, 2010) et les paramètres par défauts du logiciel. La robustesse de chaque nœud est calculée par la méthode statistiques des aLRT (SH-like branch supports) (Anisimova & Gascuel, 2006).

Les phylogénies par Maximum de Vraisemblance ont été confirmées par méthode Bayésienne utilisant MrBayes version 3.2.6 (Huelsenbeck & Ronquist, 2001). Les paramètres sont de 4 chaines sur 10^6 générations. Les probabilités postérieures ont été calculées à partir du consensus de la règle de la majorité des arbres échantillonnés tous les 100 générations une fois que les chaînes de Markov étaient devenues stationnaires (déterminées par un contrôle empirique des valeurs de vraisemblance).

Les test de Shimodaira–Hasegawa (SH) (Shimodaira & Hasegawa, 1999) et de Kishino-Hasegawa (1sKH) (Kishino & Hasegawa, 1989) ont été réalisés avec TREE-PUZZLE (version 5.6.rc16) (Schmidt *et al.*, 2002) afin de détecter de possible incongruences entre différents domaines d'une même protéine ou différents gènes d'un même génome.

Evolution moléculaire et polymorphismes

La dernière étape a été d'étudier l'évolution moléculaire de ces virus (Figure 15). Les pressions de sélection qui agissent sur les génomes viraux sont calculées à partir des codons. Le programme PAL2NAL (Suyama *et al.*, 2006) a été utilisé afin de transformer un alignement protéique en alignement de codons correspondants à l'aide des séquences nucléotidiques initiales. Seuls les codons partagés dans l'alignement par tous les virus les plus proches du/des virus d'intérêt(s) ont été conservés à cette étape afin d'éviter les erreurs dans les estimations dues à des distances phylogénétiques trop importantes.

L'évaluation des pressions de sélection a été faite en utilisant des modèles qui calculent le ratio ω sur des branches particulières de l'arbre phylogénétique ou sur un sous arbre d'intérêt.

Ce ratio ω est le rapport du taux de substitution des sites non-synonymes (dN) sur le taux de substitution des sites synonymes (dS) (Kryazhimskiy & Plotkin, 2008; Nei & Gojobori, 1986). Lorsque ω est inférieur à 1 alors le gène est soumis à une sélection négative ou stabilisante qui permet à ce gène de conserver sa fonction, les mutations sont alors contre-sélectionnées. Lorsque ω est proche de 1 alors le gène est sous sélection neutre, aucune force sélective n'influence ce gène qui n'a pas de fonction importante pour l'organisme. Enfin, lorsque ω est supérieur à 1 alors le gène est soumis à sélection positive, ce qui indique que ce gène est très important et que les variants de ce gène sont conservés par la sélection naturelle, permettant une adaptation. Cette étape a été réalisée avec CodeML (Yang, 1998; Yang & Bielawski, 2000) implémenté dans PAML version 4.9c (Yang, 2007).

La comparaison des modèles a été réalisée par des tests de rapport de vraisemblance (Likelihood Ratio Test, LRT) en utilisant la statistique de test $\chi 2 = 2\Delta LnL$ (soit deux fois la différence du Log de la vraisemblance de chaque modèle) et un seuil de significativité $\alpha=0,05$, ddl=1 (c'est-à-dire le degré de liberté, la différence du nombre de paramètres entre deux modèles).

La diversité nucléotidique (paramètre π) a été estimé à chaque position du génome viral en utilisant PoPoolation (Kofler et al., 2011) (version 1.2.2) avec les paramètres par défaut. Cette diversité permet de quantifier le polymorphisme d'une souche virale à l'intérieur d'un hôte unique. Dans le cas où plusieurs génomes proches ont été détectés sur des hôtes distincts, une comparaison manuelle des allèles entre les différentes souches permet de visualiser les polymorphismes majoritaires ou fixés de type SNPs (*Single Nucleotide Polymorphisms*).

2.4. Pourquoi créer une nouvelle méthode ?

Le plein essor des NGS a permi le développement de différents programmes en ligne ou en utilisation locale pour l'étude des viromes et métagénomes, permettant la recherche de la présence virale et de sa caractérisation (Tableau 3). Au début de mon travail de thèse en 2014, seul Metavir (Roux *et al.*, 2011) et surtout Metavir2 (Roux *et al.*, 2014) auraient pu me permettre d'analyser une partie de mes données, sous l'angle de la description globale des virus présents, mais ce programme n'est cependant pas adapté pour la découverte de

nouveaux virus. Du fait de la taille trop importante du jeu de données initial (plus de 2,5 To de données compressés), l'utilisation de ce programme en ligne n'était pas possible. J'ai donc mis en place ma propre stratégie permettant de répondre aux questions d'ordre pratiques et scientifiques que soulevaient la détection et l'étude de la biodiversité virale à partir de transcriptomes.

Le principal inconvénient de la méthodologie développée ici est qu'elle n'est pas disponible en tant que pipeline complet (en ligne ou téléchargeable) comme le sont les autres outils (Tableau 3). En effet, la phase automatique est certes facilement reproductible mais la phase manuelle, la phase la plus importante en termes de temps d'analyse, n'est pas automatisable puisqu'elle requiert une expérience dans le tri et dans l'analyse des données (annotation des génomes) ainsi que dans le choix des séquences à inclure pour établir des phylogénies pertinentes. Ainsi, aucune phylogénie pertinente permettant l'étude de gènes et familles virales diverses ne peut être créée automatiquement. Les programmes Metavir et Metavir2 sont capable d'établir des phylogénies mais sur certains gènes marqueurs spécifiques uniquement, ce qui ne permet pas d'inclure toute la biodiversité virale théoriquement détectable par les NGS.

D'autres programmes ont par la suite été développés, mais n'ont pas pu être testés faute de temps. Les programmes VMGAP (Lorenzi *et al.*, 2011), VIROME (Wommack *et al.*, 2012) ou HoloVir (Laffy *et al.*, 2016) permettent par exemple de réaliser une annotation fonctionnelle automatique des hits détectés. Les programmes VIROME, ProVIDE (Ghosh *et al.*, 2011) ou HoloVir permettent en sortie d'obtenir une liste de tous les hits viraux incluant la taxonomie complète sous forme de tableau, similaire à la méthode développée ici. Certains, comme Metavir, ViromeScan (Rampelli *et al.*, 2016) ou MetaShot (Fosso *et al.*, 2017) sont additionnés de sorties graphiques et de calculs d'abondance des pathogènes détectés tels que produits par Metavir, ViromeScan (Rampelli *et al.*, 2016) ou MetaShot (Fosso *et al.*, 2017).

En revanche, la méthode développée lors de cette thèse permet d'utiliser les lectures brutes de séquençage puisque la première étape consiste à de l'assemblage de novo. Seuls certains programment ont été conçus pour en faire de même, comme drVM (Lin & Liao, 2017), Vipie (Lin *et al.*, 2017), VIP (Li *et al.*, 2016) et VirusTAP (Yamashita *et al.*, 2016). De plus, ma méthode peut être utilisée sur des données déjà assemblées en contigs.

Tableau 3 : Liste et caractéristiques des outils actuellement dédiés à la métaviromique et comparaison avec la méthode développée ici

Données en entrée					
Nom du programme	Données d'entrée	Type d'analyse	Génome viral détecté	Origine des séquences	
PHACCS	Reads	Viromes	ADN	Filtration virale	
Metavir	Reads	Viromes	ADN	Filtration virale	
VMGAP	Contigs	Viromes	ADN	Filtration virale	
ProViDE	Sortie BLASTX	Métagénomes	ADN	-	
VIROME	Reads/ ORFs	Viromes	ARN	Filtration virale	
Metavir 2	Reads/ Contigs	Viromes	ADN	Filtration virale	
RIEMS	Reads	Métagénomes	ADN	Généraliste	
VirSorter	Génomes	-	ADN	Bactéries et archées	
VIP	Reads	Métagénomes	ADN	Données cliniques	
VirusTAP	Reads/ Contigs	Métagénomes	ADN	Humain ou animal séquencé	
ViromeScan	Reads	Viromes	ADN/ARN	Filtration virale	
MG-Digger	Reads	Métagénomes	ADN	-	
HoloVir	Reads/ Contigs	Viromes	ADN	Holobiont marins	
Taxonomer	Reads/ Contigs	Métagénomes	ADN/ARN	Métagénome humain	
MetaShot	Reads	Métagénomes	ADN/ARN	Microbiote humain	
Vipie	Reads	Viromes	ADN/ARN	-	
drVM	Reads	Métagénomes	ADN/ARN	-	
***	**Reads/ Contigs**	**Méta-transcriptomes**	**ARN**	**Généraliste**	
Données en sortie					
	Assemblage	Taxonomie	Graphs / Tables	Phylogénies	Annotation fonctionnelle
PHACCS	×	×	×/✓	×	×
Metavir	×	✓	✓/✓	✓	×
VMGAP	×	×	×/×	×	✓
ProViDE	×	✓	×/✓	×	×
VIROME	×	✓	×/✓	×	✓
Metavir 2	×	✓	✓/✓	✓	×
RIEMS	×	✓	×/✓	×	×
VirSorter	×	×	×/×	×	×
VIP	✓	✓	×/×	✓	×
VirusTAP	✓	✓	×/×	×	×
ViromeScan	×	✓	✓/✓	×	×
MG-Digger	×	✓	×/✓	×	×
HoloVir	×	✓	×/✓	×	✓
Taxonomer	×	✓	✓/✓	×	×
MetaShot	×	✓	✓/✓	×	×
Vipie	✓	✓	✓/✓	×	×
drVM	✓	✓	×/×	×	×
***	✓	✓	**×/✓**	**×**	**×**

Détection virale				Source	
	Spécificité	Méthode détection virale	Cible virale	Disponibilité	Référence
PHACCS	Structure & diversité	Contig spectrum	Phages	Web	(Angly *et al.*, 2005)
Metavir	Composition viromes et phylogénies gènes spécifiques	BLAST	Tous virus	Web	(Roux *et al.*, 2011)
VMGAP	Annotation fonctionnelle	BLAST	Tous virus	na	(Lorenzi *et al.*, 2011)
ProViDE	Assignation taxonomique	Tri fichier	Tous virus	Download	(Ghosh *et al.*, 2011)
VIROME	Classification séquences virales	BLAST	Tous virus	Web	(Wommack *et al.*, 2012)
Metavir 2	Comparaison viromes et composition taxonomique	BLAST	Tous virus	Web	(Roux *et al.*, 2014)
RIEMS	Assignation taxonomique	BLAST	Généraliste	Download	(Scheuch *et al.*, 2015)
VirSorter	Détection signaux spécifiques intégration virale	HMMER et BLASTp	Prophages (intégrés)	Web	(Roux *et al.*, 2015)
VIP	Couverture génomes et phylogénies	BLAST et mapping	Tous virus	Download	(Li *et al.*, 2016)
VirusTAP	Assemblage viral *de novo*	BLAST	Virus sauf phages	Web	(Yamashita *et al.*, 2016)
ViromeScan	Abondance des virus connus	Mapping	Virus sauf phages	Download	(Rampelli *et al.*, 2016)
MG-Digger	Abondance des virus géants	BLAST	*Megavirales*	Download	(Verneau *et al.*, 2016)
HoloVir	Prédiction gène, taxonomie et annotation	BLAST	Tous virus	Download	(Laffy *et al.*, 2016)
Taxonomer	Abondance des pathogènes et expression ARNm hôtes	Mapping	Généraliste	Web	(Flygare *et al.*, 2016)
MetaShot	Abondance des pathogènes	Mapping	Généraliste	Download	(Fosso *et al.*, 2017)
Vipie	Parallel processing	BLAST et mapping	Tous virus	Web	(Lin *et al.*, 2017)
drVM	Assemblage des virus connus	BLAST	Virus sauf phages	Download	(Lin & Liao, 2017)
***	**Assemblage à taxonomie virale**	**BLAST & BLAST réciproque**	**Tous virus**	**na**	**na**

***correspond à la méthodologie mise en place durant ma thèse.

L'utilisation d'un BLAST sur une base de données virale suivie d'un BLAST réciproque sur une base de données généraliste confère à cette méthode une validation et surtout une suppression d'éventuels faux positifs qui pourraient être générés en utilisant uniquement un BLAST sur une base de données virale, comme c'est le cas pour Metavir ou VIROME.

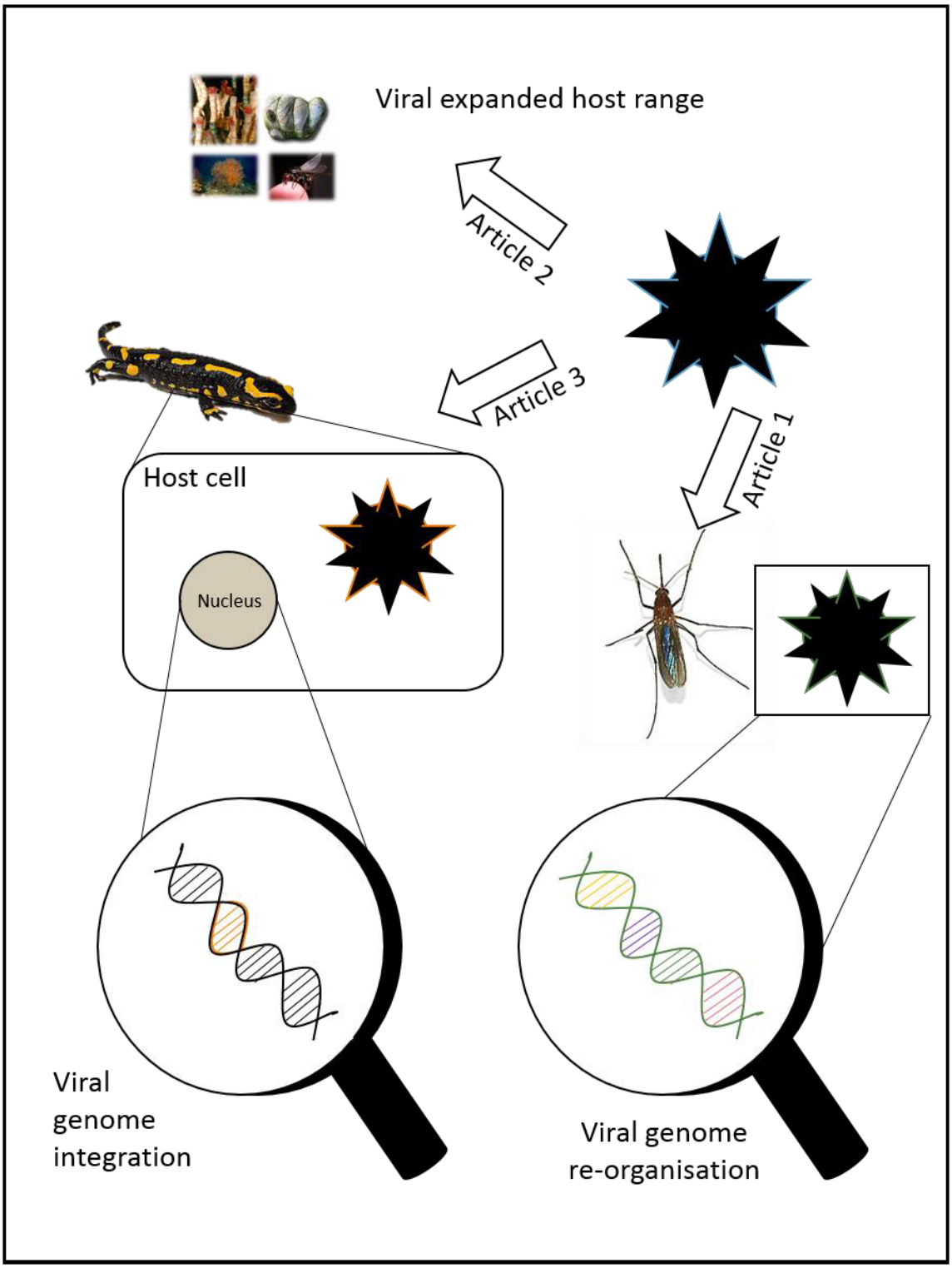
Ma méthode utilise spécifiquement des ARN par le biais du séquençage des transcriptomes, ce qui est différent des autres outils publiés, qui permettent principalement l'analyse des viromes ou métagénomes ADN. Certains programmes, notamment les plus récents, font exception à cette règle et permettent l'analyse des ADN aussi bien que des ARN (ViromeScan, Taxonomer (Flygare *et al.*, 2016), MetaShot, Vipie, drVM). Ainsi, le biais principal de l'utilisation de transcriptomes couplé à ma méthode d'analyse est la détection spécifique des hits viraux ARN.

Finalement, ma méthode reste généraliste, comme celle utilisée par REIMS (Scheuch *et al.*, 2015), dans le sens où il ne s'agit pas de l'analyse de viromes (comme pour VIROME, Metavir/Metavir2, ViromeScan, ou VMGAP), et donc il n'y a pas besoin de réaliser une filtration des virus avant séquençage. Le caractère généraliste de ma méthode s'applique aussi aux types de virus détectés et recherchés. En effet, certaines méthodes excluent la détection des phages (VirusTAP, ViromeScan, drVM) ou au contraire ne sont spécifiques que des phages ou prophages (PHACCS (Angly *et al.*, 2005), VirSorter (Roux *et al.*, 2015)), ou de familles virales particulières (*Megavirales* pour MG-Digger (Verneau *et al.*, 2016)).

Enfin, l'avantage majeur de cette méthodologie est qu'elle a fourni la preuve que ce concept fonctionne pour la découverte et l'étude de la diversité virale présente dans des transcriptomes. Les deux chapitres qui suivent montrent que cette méthode (en deux phases automatique et manuelle) permet cette description de nouveaux virus majoritairement à ARN.

CHAPITRE 1

Découverte et description de nouveaux virus libres et endogènes


Viral expanded host range
Article 2
Article 3
Article 1
Host cell
Nucleus
Viral genome integration
Viral genome re-organisation

3.1. Préambule au Chapitre 1

"Viral metagenomic approaches provide novel opportunities to generate an unbiased characterisation of the viral populations in various organisms and environments. […] With increasing use of sequence-independent amplification and efficient sequencing methodologies it seems likely that new viral species will be identified at a rate considerably greater than the knowledge of their biology."

E. Delwart, Viral metagenomics, Reviews in Medical Virology, 2007

Ce chapitre permet de montrer que la méthodologie employée dans ce travail permet la découverte de virus libres et endogènes, que ce soit des virus à génome ARN, ADN ou des rétrovirus.

Le premier travail concerne la découverte d'un nouveau virus libre ARN avec une organisation génomique particulière qui rejoint de récentes découvertes et qui permet de définir la présence d'une nouvelle famille virale spécifique des insectes (**ARTICLE 1**).

Le second travail décrit dans ce chapitre montre que l'étude d'animaux non-modèles permet d'élargir la gamme d'hôte de virus connus, tels que les *Parvoviridae,* virus à ADN (**ARTICLE 2**).

Enfin, cette technique permet également l'étude de rétrovirus endogènes, avec la découverte d'un nouveau Spumavirus intégré dans le génome de la salamandre tachetée *Salamandra salamandra* et d'autres urodèles, découverte confirmée par des méthodes de biologie moléculaire (**ARTICLE 3**).

3.2. Découverte d'un nouveau virus ARN chez un moustique (**Article 1**)

Ce travail pour lequel je suis première auteure est présenté sous forme d'un article en cours de révision dans *Virus Evolution*, et représente la preuve de concept de la méthodologie mise en place par ce travail de thèse sur l'exemple de la découverte d'un nouveau virus ARN, représentant un nouveau genre viral spécifique des insectes.

Discovery of Culex pipiens Associated Tunisia Virus, a new ssRNA(+) virus representing a new insect associated virus family.

Diane Bigot[1], Célestine M. Atyame[2], Mylène Weill[3], Elisabeth A. Herniou[1§] and Philippe Gayral[1§]

1 Institut de Recherche sur la Biologie de l'Insecte, UMR 7261, CNRS, Université François-Rabelais, 37200 Tours, France ; 2 Université de La Réunion, UMR PIMIT (Processus Infectieux en Milieu Insulaire Tropical) CNRS 9192, INSERM U1187, IRD 249; Sainte-Clotilde, Ile de La Réunion, France ; 3 Institut des Sciences de l'Evolution UMR5554, Université Montpellier–CNRS–IRD–EPHE, Montpellier, France

§ Equal contribution

Abstract

In the global context of arboviral emergence, deep sequencing unlocks the discovery of new mosquito-borne viruses. Mosquitoes of the species *Culex pipiens*, *C. torrentium* and *C. hortensis* were sampled from 22 locations worldwide for transcriptomic analyses. A virus discovery pipeline was used to analyze the dataset of 0.7 billion reads comprising 22 individual transcriptomes. Two closely related 6.8kb viral genomes were identified in *C. pipiens* and named as Culex pipiens Associated Tunisia Virus (CpATV) strains Moknine and ElHabibia. The CpATV genome contained four ORFs. ORF1 possessed helicase and RNA-dependent RNA polymerase (RdRp) domains related to new viral sequences recently found mainly in dipterans. ORF2 and 4 contained a capsid protein domain independently acquired from Virgaviridae plant viruses. ORF3 displayed similarities with eukaryotic Rhoptry domain and a merozoite surface protein (MSP7) domain only found in mosquito-transmitted *Plasmodium*, suggesting possible interactions between CpATV and vertebrate cells. Estimation of a strong purifying selection exerted on each ORFs and the presence of a polymorphism maintained in the coding region of ORF3 suggested that both CpATV sequences are genuine functional viruses. CpATV is part of an entirely new and highly diversified group of viruses recently found in insects, and that bears the genomic hallmarks of a new viral family.

Key-words: CpATV, *Culex pipiens* mosquitoes, *Plasmodium*, RNA virus, *Virgaviridae*, Virus discovery.

Introduction

Viral biodiversity remains largely unexplored. Viruses are found in all types of organisms (archaea, bacteria, eukaryota and some large dsDNA viruses) and are the most abundant micro-organisms on Earth (1). During the past decade, the development of high-throughput next-generation sequencing technologies (NGS) and the use of bioinformatics, metagenomics and phylogenetic analyses have allowed the discovery of many new viruses, particularly of phages in aquatic and mammal gut environments (2, 3). Although NGS have become a sensitive and reliable method for virus discovery, recent virus discoveries in arthropods in general and insects in particular are scarcer compared to other hosts or ecosystems (4–6) and often target viral families within ssRNA(+) viruses such as *Flaviviridae* (genus *Flavivirus*), *Togaviridae* (genus *Alphavirus*), *Nidovirales* (family *Mesoniviridae*) or ssRNA(-) viruses as *Bunyaviridae* (7). Indeed, arthropods were reported to be a great source of viral diversity (8), especially of RNA viruses (9, 10). Still, several recent studies discovered new insect-associated viruses in other recognized viral families (10), such as *Rhabdoviridae* (11–13), *Reoviridae* (14–16), *Tymoviridae* (17), *Nodaviridae* (18) or in non-recognized families as in the case of the negeviruses (Sandewavirus / Nelorpivirus) (19–24). The growth of insect-associated virus discoveries provides a fertile ground for the understanding of the complexity and dynamics of arboviral communities in an ecological and evolutionary perspective (25, 26).

Despite the health interest surrounding the study of mosquitoes, few studies have used NGS for virus-discovery in mosquitoes. In 2011, the first metagenomic approach by 454 pyrosequencing on wild mosquitoes revealed 6 new DNA viruses (27). Deep sequencing of small RNA (18-30bp) also allowed the discovery of new RNA and DNA viruses, in several mosquito species (6, 28, 29). Two novel rhabdoviruses and three novel bunyaviruses were also discovered in several Australian mosquitoes using deep sequencing of pooled insect viromes passaged in cell culture (30). Finally, the used of transcriptome sequencing of a large range of potential hosts has improved and increased the discovery of many new RNA viruses infecting invertebrates (10) including the discovery of the new *Gamboa mosquito virus* (GMV) from a mix of Culicidae mosquitoes (31). Yet, *Culex* mosquitoes may still host largely unexplored viral diversity and thus be potential reservoir of unknown viruses that might become relevant for human health (32, 33). These studies illustrate how underestimated is viral diversity associated with mosquitoes and outline a way to bridge the gap between our current state of knowledge and the vast viral biodiversity occurring in nature.

Here we report the screening of 22 Illumina transcriptomes of wild individuals of *Culex pipiens*, *C. hortensis*, and *C. torrentium* species and the subsequent discovery and genetic characterization of a new mosquito virus species associated with *C. pipiens*. Results of genome annotation and phylogenies of Culex pipiens Associated Tunisia Virus were used to decipher its evolutionary relationships to a new group of viruses recently found in insects.

Materials and Methods

Mosquito sampling

Twenty-two transcriptomes were obtained from single adult female mosquitoes, belonging to the species *Culex pipiens*, *C. hortensis*, and *C. torrentium* (Supplementary Table 1). All mosquitoes were sampled as larvae in fresh water puddles and grown in laboratory conditions. After emergence, females were conserved in liquid nitrogen, or kept under laboratory conditions before they had the opportunity to take any blood meal.

RNA extraction, transcriptome sequencing and assembly

Total RNA isolation was performed using RNeasy Mini Spin Columns (Qiagen, Chatsworth, CA, USA) on individual mosquitoes as previously described (34). RNA quality was assessed on Agilent Bioanalyzer 2100 system and a RNA 6000 Nano Lab-Chip (Agilent). Then, 5 µg total RNA was reverse-transcribed using the SMART cDNA library Construction kit (Clontech, Mountain View, USA). An oligo(dT)-primed first-strand synthesis followed by a cap-primed second-strand synthesis was performed. Eight libraries were sequenced per lane using an Illumina HiSeq 2000 sequencer to produce 50 bp single-end reads. All 22 transcriptomes were de novo assembled using a previously developed bioinformatics pipeline (35): a first assembly with ABYSS V 1.2.0 (36, 37) with Kmer set at 40 (38) was followed by contig re-assembly with CAP3 (39). Complete and 5'- and/or 3'-truncated ORFs were detected and translated with standard genetic code using Prodigal V2_60 software for metagenomic data (40, 41). ORFs displaying undetermined nucleotides were not discarded in subsequent analyses.

Virus Discovery

Protein homology searches were performed on all translated ORFs of the 22 transcriptomes using the accurate and sensitive HHblits program implemented in the HHSuite package (42, 43). The Nr20 NCBI protein database, a clustered version of the protein non-redundant database from NCBI down to a maximum pairwise sequence identity of 20% protein, was used

as a search database as recommended (42, 43). To minimize false-positive results, only ORFs displaying homology e-value <10-5 and probability >95% were kept. When a positive hit was detected, NCBI taxonomic identifier (TaxID; ftp://ftp.ncbi.nih.gov/pub/taxonomy) of the corresponding Nr protein was retrieved using the BLAST+ program (44) and assigned to the predicted ORFs. Only ORFs identified as 'viruses' in the superkingdom taxonomic rank were kept for further analysis. Then, to reduce false-discovery rate and ensure the detection of functional infectious viruses, isolated virus-like ORFs or to single ORFs of dubious viral homology were discarded and only putative full-length viral genomes were further analyzed. To verify the accuracy of viral contigs assembly, Illumina reads were mapped on the assembled viral genome using BWA (45) with default parameters. Mapping results (SAM files) were used to calculate the coverage at each nucleotide position along the viral genomes using Geneious 8.1.7 (http://www.geneious.com, (46)). The distribution of the quality score of mapped reads was plotted using the fastx_quality_stats and fastq_quality_boxplot programs implemented in FASTX-Toolkit (http://hannonlab.cshl.edu/fastx_toolkit/).

Lastly, nucleotide polymorphism at each genomic position (pi) was estimated on the mapping results using PoPoolation (version 1.2.2) with default parameters (47). Three nucleotides were trimmed from both 5' and 3' ends of all viral reads to ensure that sequencing errors would not bias the estimation of intra-host virus diversity as such errors are more frequent at the extremities of Illumina reads.

Prevalence in other hosts

We screened several public databases to evaluate if the virus described in this study might be widespread or infecting other hosts. First BLASTN searches were performed using the entire CpATV genome (e-value threshold: 10-5) on the *Culex* genome (*C. quinquefasciatus*) and on the 15 available mosquito transcriptomes (Supplementary Table 2). In addition, *Culex* reads produced in this study were mapped on the CpATV genome using BWA software with defaults parameters (45).

Further BLASTN and BLASTX searches (e-value threshold: 1) were performed against the Transcriptome Shotgun Assembly (TSA), Whole-genome Shotgun contigs (WGS) and Metagenomic proteins (env-nr) databases (October 2016), as well as the 60 Arachnida, Gastropoda and Insecta genomes available in VectorBase (September 2016, https://www.vectorbase.org/).

Genome annotation

Viral genomes were annotated based on conserved protein domains searched using InterProScan (version 5) (48, 49), NCBI Conserved Domain and Conserved Domain Database (v3.14) (50), as well as SMART (version 7) (51, 52). Signal peptides were detected using SignalP (version 4.0) (53) and TargetP (version 1.1) (54). The detection of potential poly(A) tail was performed using PolyApred (55). RNA motifs and cis-regulatory elements such as tRNA-like structures were searched using RegRNA 2.0 (56). Internal ribosomal entry sites (IRES) were predicted using Viral IRES Prediction System (VIPS) with default parameters (57).

BLAST searches (e-value threshold: 1) were performed against *Plasmodium* genomes available in PlasmoDB (58) (release 2015/07/23) to detect homologies with 13 available *Plasmodium* genomes (Supplementary Table 3) as well as the unpublished *P. relictum* genome (Ana Rivero, personal communication).

Phylogenetic Analyses

Amino acid sequences alignment of each conserved protein domain were performed with MAFFT (59) using default parameters and curated manually. Non-homologous sites located in ORFs extremities were discarded from alignments. The best substitution model was selected using ProtTest v3.2 (60). Maximum Likelihood (ML) phylogenetic trees were inferred for each alignment using PhyML (61) and robustness of nodes was assessed with aLRT statistics (SH-like branch supports) (62). Shimodaira–Hasegawa (SH) (63) and a one-sided Kishino-Hasegawa (1sKH) (64) tests were performed using TREE-PUZZLE (version 5.6.rc16) (65) to detect possible phylogenetic incongruence between different domains of the same ORF.

Molecular evolutionary analyses

PAL2NAL program (66) was used to obtain relevant codon alignments guided by protein alignments. Only the most closely related sequences to the taxon of interest were kept to ensure genuine homology of aligned sites. Selective pressures acting on viral coding sequences were assessed using branch-models which estimated the ratio of non-synonymous (dN) / synonymous (dS) substitution rates (67, 68) in a chosen branch or subtree, in codeml (69, 70) implemented in PAML version 4.9c (71). Likelihood ratio tests (LRTs) were employed using the $\chi 2$ tests statistics = $2\Delta LnL$, (i.e. twice the difference of the Log-Likelihood of each

model) and a type I error = 0.05, df= 1 (i.e. the difference of number of parameters between two models).

Results

Virus detection in *C. pipiens* GA35C from Moknine, Tunisia

Twenty-two individual mosquito transcriptomes, including eight new, were assembled from a total of 687 million Illumina reads. Average contig size (N50) was 230 bp (range 164-318 bp; Supplementary Table 4). Out of 69,756 ORFs predicted in this dataset, the virus detection pipeline identified only one 6,818 bp contig presenting high homology with viral sequences in the *Culex pipiens* individual GA35C, sampled in 2005 in Moknine, Tunisia.

This contig contained 4 ORFs of sizes 3,240 bp, 414 bp, 2,118 bp and 483 bp, consistent with a putative full-length viral genome (Figure 1A). Initial protein homology search results showed ORF1 and ORF4 were strongly homologous to structural proteins of *Tobamovirus*, a ssRNA(+) plant virus belonging to the *Virgaviridae* family (e-value=8.40e-210 and 4.70e-65, p=8.00e-210 and 4.20e-70 for ORF1 and ORF4, respectively). The coding region was surrounded by two non-coding sequences of 333 bp and 99 bp at the 5' and 3' extremities respectively, possibly representing two UTRs. Advanced analyses of genomic motifs did not reveal any regulatory sequences, such as poly(A) tail, IRES signals or tRNA-like structures.

To ensure this viral contig did not result from chimeric assembly, several verification steps were undertaken. First, read mapping quality was assessed. A total of 176,684 very good quality reads (median quality scores > 36 at each position of the reads; Supplementary Figure 1A) from *C. pipiens* GA35C mapped on the viral genome (Figure 1B). This indicated that at least 99.9% of bases were accurately called. Second, the viral contig was evenly covered along its entire length, without any coverage drop that could have indicated a chimera (Figure 1B). Third, the viral contig was very abundant in the mosquito transcriptome as its mean sequence coverage was 1,322X compared to the 227X mean coverage obtained for *C. pipiens* GA35C. Such difference in transcript abundance may indicate the presence of a replicating infectious virus. As our contig bared all the hallmarks of a viral genome, we named this new virus Culex pipiens Associated Tunisia Virus strain Moknine (CpATV_Moknine).

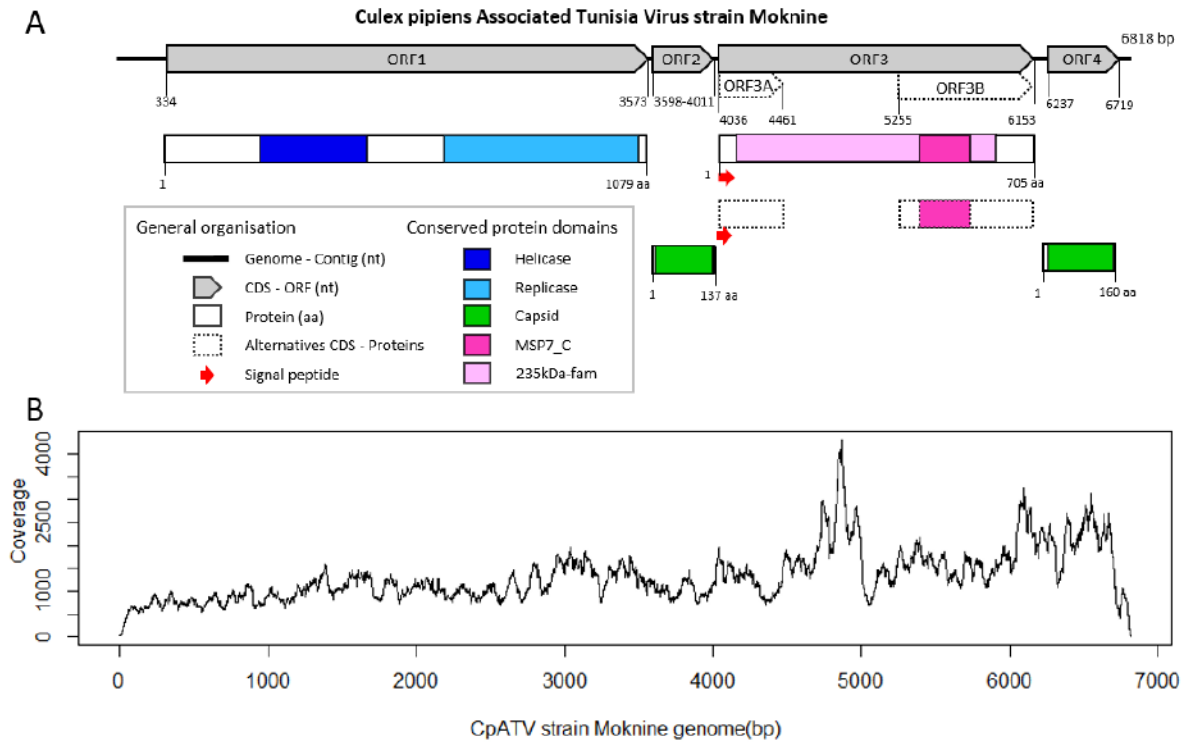


Figure 1: Culex pipiens Associated Tunisia Virus (CpATV) characteristics. A) Genome organization of CpATV with conserved domains and predicted ORFs. B) Read coverage of CpATV_Moknine genome.

CpATV_Moknine displays typical viral ORFs

The ORF1 of CpATV_Moknine contained a viral helicase domain (pfam01443, e-value 8.82e-35) and a RNA dependent RNA polymerase domain (RdRp_2 pfam00978, e-value 4.74e-110) (Figure 1A). Interestingly, both domains possessed all amino acids conserved within ssRNA(+) viruses of closely related genera in the 7 conserved domains of helicase (Supplementary Figure 2) as well as in the 8 RdRp domain (Supplementary Figure 3) as defined by Koonin and Dolja (72). RdRp catalyzes RNA replication while helicase facilitates RdRp initiation and elongation by unfolding secondary structures of ssRNA and unwinding dsRNA templates. Thus, CpATV ORF1 is putatively a functional viral replicase.

ORF2 and ORF4 both contained a complete capsid domain similar to the TMV capsid-like domain (pfam00721) found in *Tobacco Mosaic Virus* (*Tobamovirus*, *Virgaviridae*). This result was supported by strong homologies of protein domains: e-value= 5.45e-3 and 1.43e-11 for ORF2 and ORF4, respectively. Protein alignment also showed the 2 amino acids identified as functionally important to form a R(+) - E(-) salt bridge (72) were present in both CpATV capsid sequences (Supplementary Figure 4).

CpATV_Moknine ORF3 has similarities with Plasmodium domains

Two conserved protein domains were discovered in ORF3. First, homology was detected for protein 235kDa-fam (TIGR01612; 634 aa, e-value = 2.25e-4), which encodes a reticulocyte binding protein or Rhoptry protein. This *Plasmodium* protein plays a role in the red blood cell invasion process (73). Second, ORF3 displayed homology with a complete MSP7_C domain (Pfam id = pfam12948; 114 aa, e-value= 6.76e-3); which is the C-terminal part of the Merozoite Surface Protein 7 of *Plasmodium falciparum* (74). However, phylogenetic analyses did not reveal any direct relationships between the CpATV sequence and the *Plasmodium* genes (data not shown). Further BLASTN and BLASTX searches of the entire ORF3 against the *Plasmodium* genomes did not produce any significant hits towards more closely related sequences. The N-terminal region of ORF3 harbored a 36 bp signal peptide (secretory pathway score 0.764, reliability class 3/5 where 5 is the least reliable class, see (54)), indicating that the translated protein is likely secreted.

Comparative viral genomic analyses

The CpATV genome content was compared to genomes of closely related to Virgaviridae and 'Negeviruses'. Additionally, 14 new viral sequences associated with diverse insects (including many fruit flies, Diptera) (75, 76) were annotated for the first time in the present study (Figure 2). In all virus genomes, the large ORF1 contained a viral helicase domain adjacent to an RdRp domain. The 5' end of ORF1 CpATV did not harbor any methyltransferase domain (Figure 2), which plays a role in the biosynthesis of the 5' cap of RNA viral genomes (77, 78).

The remaining ORFs were variable both in number (1 to 5, depending on the virus genome) and size (67 to 681 aa) and contained structural and virulence domains (Figure 2).

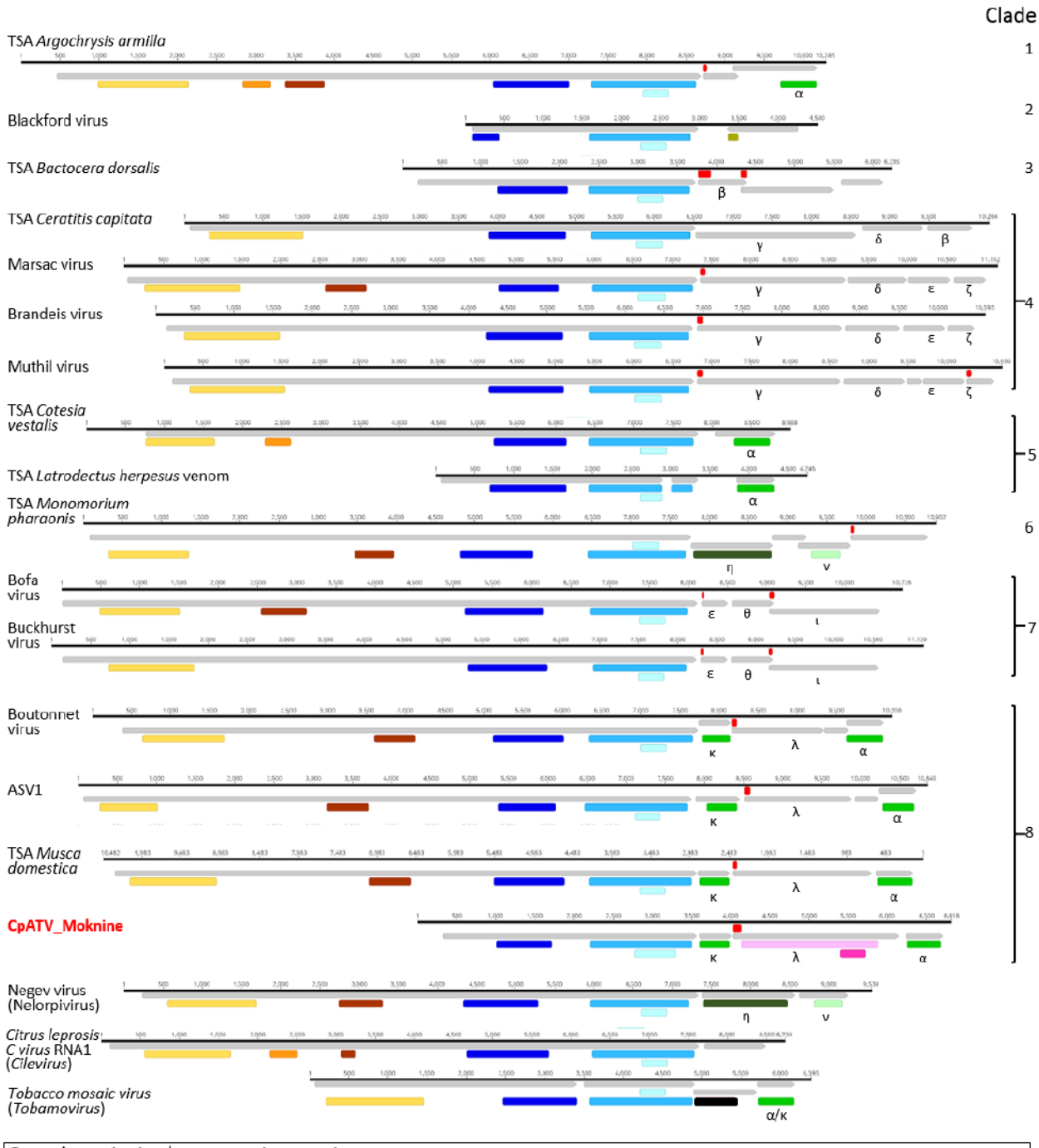


Figure 2: Comparative genomics of Culex pipiens Associated Tunisia Virus (CpATV) genome and related virus genomes. The 14 new insect viruses are from (75, 76) and the outgroups are *Citrus leprosis virus* C RNA1 (*Cilevirus*), Negev virus (Nelorpivirus), *Tobacco mosaic virus* (*Tobamovirus*, *Virgaviridae*). Sequences details and GenBank accessions numbers are available in Supplementary Table 6.

Capsid domains (TMV-like coat protein - IPR001337) similar to those of CpATV were found in the Boutonnet virus and the TSA *Musca domestica* viral sequence (both displayed two capsid ORFs), as well as in the TSA *Argochrysis armilla*, TSA *Cotesia vestalis* and TSA *Latrodectus hesperus* venom viral sequences (all 3 displayed one ORF) (Figure 2). The Blackford virus contained a putative F-Box protein (domain PF00646), which, in Poxviruses, is involved in the ubiquitination of proteins targeted for degradation in the proteasome and deregulation of host cells (79, 80). The TSA *Monomorium pharaonis* sequence contained a putative envelope glycoprotein and a putative virion membrane protein of plant and insect viruses (η and ν in Figure 2). Both genes were recently described in the RNA2 of Cilevirus and in Negev virus (Nelorpivirus) as well as in the unrelated *Chronic Bee paralysis virus* (CBPV) (81). The other ORFs could not be associated to specific functions in public databases. However, reciprocal BLASTP analyses (evalue threshold=10-3) within the dataset revealed 39 ORFs could be clustered into 12 homologous groups (detailed by Greek alphabet in Figure 2), the 10 remaining ORFs were ORFans. Signal peptides towards the secretory pathway were observed at the 5' end of accessory ORFs of 11 viruses including CpATV as well as in ORF2 of the TSA *Bactocera dorsalis* viral sequence targeting the mitochondrion compartment.

The Boutonnet virus (GenBank KU754539) and a virus-like sequence from TSA *Musca domestica* comp20588_c0 transcribed RNA sequence (GenBank GARN01041480) have similar genomic organization to CpATV (Figure 2). However, their ORF3 did not possess the MSP7_C domain, or the 235kDa-fam domain similarities found in CpATV. The genome of CpATV, lacking the methyltransferase domain, is shorter than those of these two viruses. This likely represents the true biological size of CpATV, as the absence of the methyltransferase domain in the GA35C and GA35E mosquito transcriptomes was confirmed by BLASTN using the 5' end of the Boutonnet virus ORF1 as query against mosquito transcriptomes.

Replicase phylogeny

Phylogenetic analyses were performed to determine the relationships between CpATV and other viruses close to the Virgaviridae (82), including 15 new insect viruses recently discovered (75, 76, 83). *Tymovirales* (*Gammaflexiviridae*, *Betaflexiviridae*, *Alphaflexiviridae* and *Tymoviridae*) and *Benyviridae* were used as outgroup in order to resolve the relationships between the new viruses and unassigned genera such as *Cilevirus*, *Higrevirus*, Nelorpivirus and Sandewavirus as well as the *Virgaviridae* and the *Bromoviridae*. As the helicase and RdRp

domains had congruent phylogenetic signals (Helicase: p-SH=0.115; p-1sKH=0.097; RdRp: p-SH=0.499; p-1sKH=0.181), their alignments were concatenated into a 979 aa alignment for increased resolving power.

Maximum likelihood phylogenetic analyses, using Beet necrotic yellow vein virus (*Benyviridae*) as outgroup, showed the *Tymovirales* formed an early-diverging clade at the root of the tree (Figure 3A). All remaining sequences formed a highly supported monophyletic group (aLRT support =1, here called superclade), which based on branch length is as genetically diversified as the order *Tymovirales*. At the base of the superclade, the 16 new virus sequences, including CpATV formed 8 early-diverging clades. All these viruses are associated with insects and none are closely related to characterized plant viruses (*Virgaviridae*, *Bromoviridae*, *Closteroviridae*, *Idaeovirus*, as well as *Cilevirus* and *Higrevirus*) or described mosquito-infecting viruses (Sandewavirus and Nelorpivirus, grouped under the term negeviruses), which are not monophyletic unlike previously observed (21). CpATV clustered with the Boutonnet virus, the TSA *Musca domestica* sequence and ASV1 (*Adelphocoris suturalis-associated virus 1*) into clade 8 with high support 0.94. The use of appropriate outgroups in the phylogenetic analyses clearly showed the 16 new insect viruses did not evolved from a recent common ancestor with Nelorpivirus and Sandewavirus clades, and thus belong to different viral genera and families, even though they also infect mainly dipteran hosts.

Phylogeny of CpATV capsids

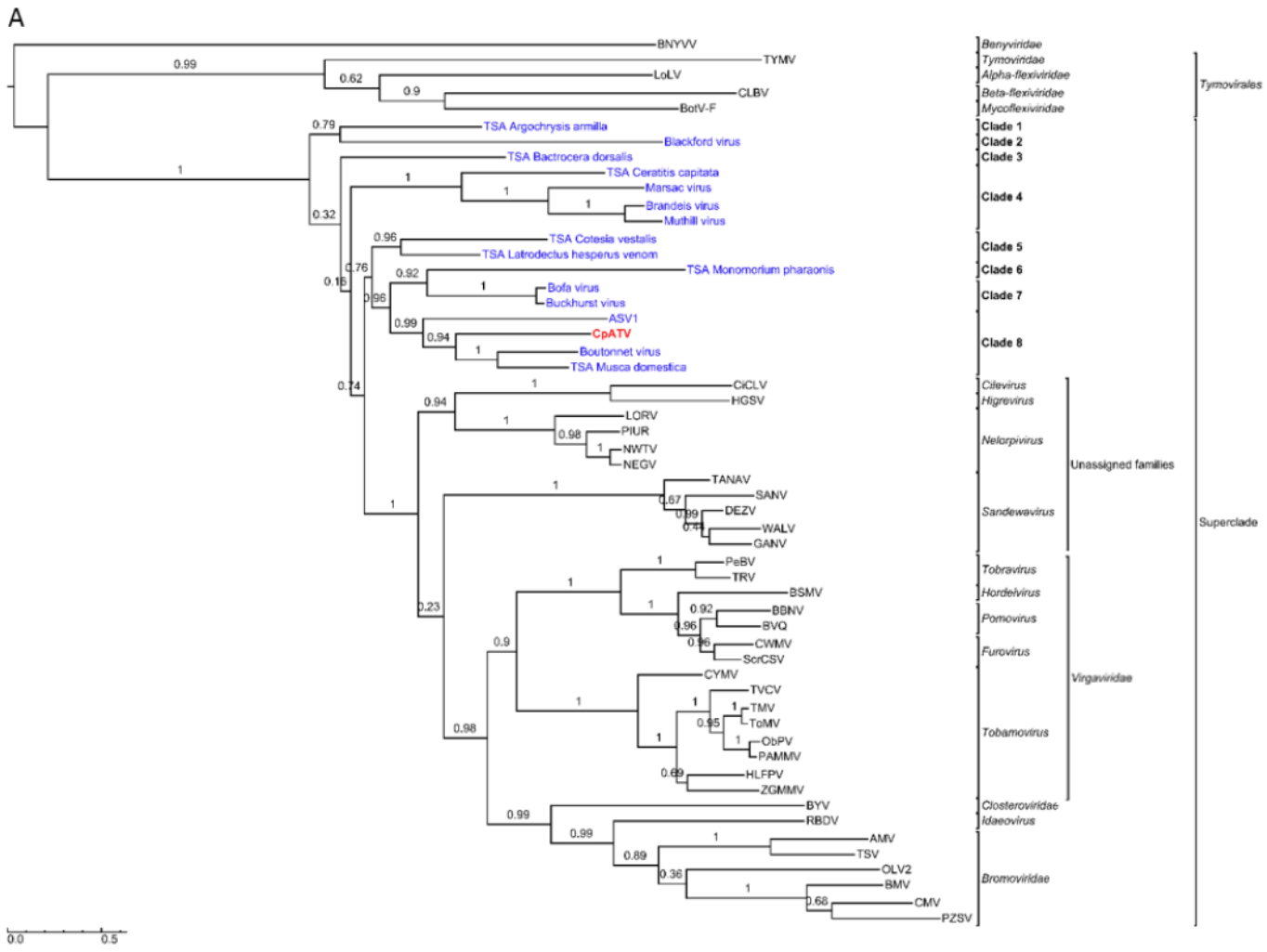
Both CpATV TMV coat capsids, as well as the 7 other homologous capsids identified in the new insect virus dataset were aligned to TMV coat capsid proteins from the *Virgaviridae* and *Closterovirus*. The capsid phylogeny (Figure 3B) was incongruent with the RdRp tree (Figure 3A) as all 9 insect-virus capsids fell inside the *Virgaviridae* clade instead of forming the ancestral lineages of the superclade (Figure 3A). The insect virus capsids formed two distinct evolutionary lineages. The first group of sequences corresponded to homologs of CpATV ORF4, (α in Figure 2), and encompassed ORF4 of the three closely related viruses Boutonnet virus, TSA *Musca domestica* and ASV1, as well as ORF2 of TSA *Cotesia vestalis*, and ORF3 of TSA *Latrodectus hesperus* venom and TSA *Argochrysis armilla*. The topology of this group reflected that of RdRp tree, suggesting a single capsid acquisition event from a *Virgaviridae* ancestor. The second group of sequence comprised the homologs of CpATV ORF2 (κ in Figure 2), and only included CpATV, Boutonnet virus, TSA Musca domestica and ASV1. This second

lineage also derived from a Virgaviridae ancestor, and the topology suggested that this four κ sequences derived from a single acquisition event. Altogether, our results suggest at least that two independent capsid acquisition events (α and κ) occurred in some of the new insect viruses. As the *Virgaviridae*, like most ancestral virus of the superclade (TSA *Argochrysis armilla*, TSA *Cotesia vestalis* and TSA *Lactrodectum herpesum* venom) possess a single TMV coat capsid (84), this might represent the ancestral genome content and indicate that a second capsid acquisition (ORF κ) occurred later in the lineage of CpATV.

Molecular evolution of the CpATV_Moknine genome

Ratios of non-synonymous over synonymous mutations (dN/dS) were used to estimate the selective pressures acting on the helicase, RdRp, and capsid conserved domains. All conserved domains identified in CpATV evolved under strong purifying selection with dN/dS values <0.1 (Table 1). Similar values were obtained for the entire clade of new insect viruses (clades 5, 6, 7, 8) as well as for the closely related infectious viruses' clades. LRT test showed that all the conserved viral domains of CpATV have the same molecular evolutionary rates as closely related infectious viruses (Table 1) p-value >0.40). Altogether, these results showed that CpATV and the other new viruses discovered in silico are probably functional since they harbor the molecular evolution hallmarks of known infectious viruses.

Figure 3: Maximum Likelihood phylogenies of conserved protein domains in the Culex pipiens Associated Tunisia Virus. (A) Concatenated helicase and RNA-dependent RNA polymerase domains from ORF1 (979 aa, substitution model LG+G). (B) Capsid domains from ORF2 and ORF4 (203 aa, substitution model LG+I+G). Taxon names represent virus acronyms. In red: CpATV; in blue: new viruses from Webster et al, 2016 (76). The scale bar represents the substitutions rate per site. aLRT statistics are indicated above nodes. Sequences details and GenBank accessions numbers are available in Supplementary Table 6.


A
BNYVV
TYMV
LoLV
CLBV
BotV-F
TSA Argochrysis armilla
Blackford virus
TSA Bactrocera dorsalis
TSA Ceratitis capitata
Marsac virus
Brandeis virus
Muthill virus
TSA Cotesia vestalis
TSA Latrodectus hesperus venom
TSA Monomorium pharaonis
Bofa virus
Buckhurst virus
ASV1
CpATV
Boutonnet virus
TSA Musca domestica
CiCLV
HGSV
LORV
PIUR
NWTV
NEGV
TANAV
SANV
DEZV
WALV
GANV
PeBV
TRV
BSMV
BBNV
BVQ
CWMV
ScrCSV
CYMV
TVCV
TMV
ToMV
ObPV
PAMMV
HLFPV
ZGMMV
BYV
RBDV
AMV
TSV
OLV2
BMV
CMV
PZSV
Benyviridae
Tymoviridae
Alpha-flexiviridae
Beta-flexiviridae
Mycoflexiviridae
Tymovirales
Clade 1
Clade 2
Clade 3
Clade 4
Clade 5
Clade 6
Clade 7
Clade 8
Cilevirus
Higrevirus
Nelorpivirus
Sandewavirus
Unassigned families
Tobravirus
Hordeivirus
Pomovirus
Furovirus
Tobamovirus
Virgaviridae
Closteroviridae
Idaeovirus
Bromoviridae
Superclade
0.0
0.5

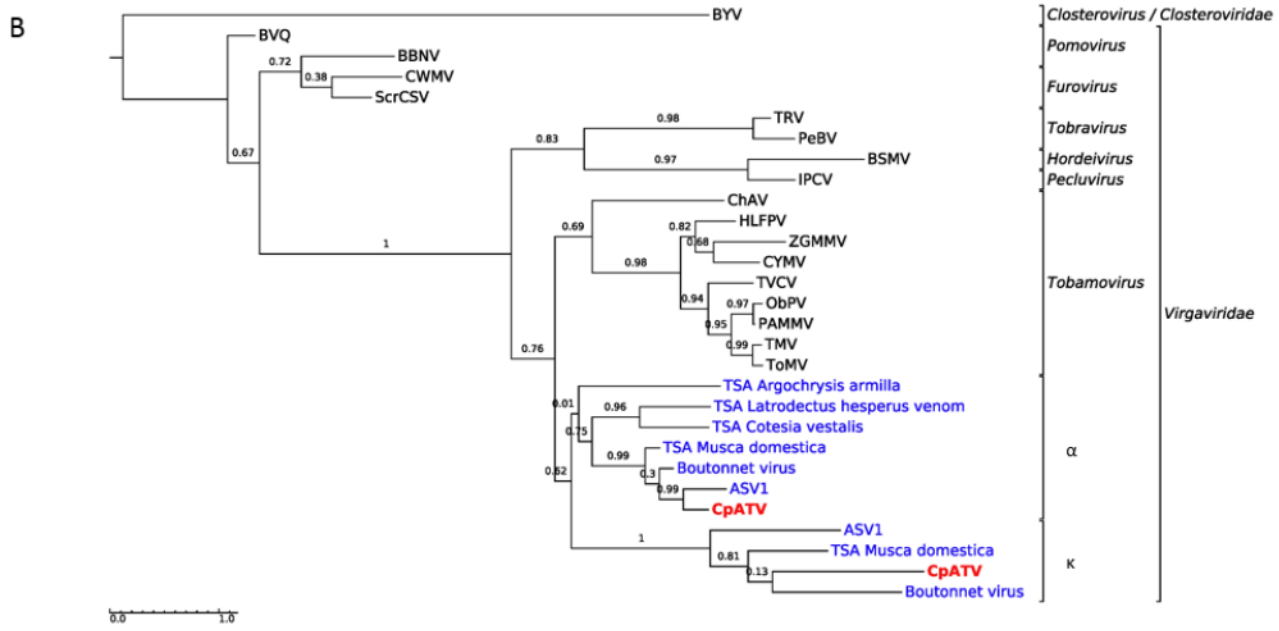

B
BYV
BVQ
BBNV
CWMV
ScrCSV
TRV
PeBV
BSMV
IPCV
ChAV
HLFPV
ZGMMV
CYMV
TVCV
ObPV
PAMMV
TMV
ToMV
TSA Argochrysis armilla
TSA Latrodectus hesperus venom
TSA Cotesia vestalis
TSA Musca domestica
Boutonnet virus
ASV1
CpATV
ASV1
TSA Musca domestica
CpATV
Boutonnet virus
Closterovirus / Closteroviridae
Pomovirus
Furovirus
Tobravirus
Hordeivirus
Pecluvirus
Tobamovirus
Virgaviridae
α
K
0.0
1.0

Table 1: dN/dS ratio estimations and LRT comparisons between CpATV and closely related viruses.

ORF / Branch or subtree of interest	Helicase ORF1	RdRp ORF1	Capsids α and κ (shared estimation)	Capsids α and κ (separate estimation)
CpATV	0.004	0.003	0.10	ORF2 (κ): 0.10 ORF4 (α): 0.08
New insect viruses	*Clades 5, 6, 7, 8*		*All sequences with α and κ domains*	
	0.01	0.008	0.01	0.01
Closely-related infectious viruses	Nelorpivirus, Sandewavirus, *Higrevirus, Cilevirus*		*Virgaviridae*	
	0.004	0.006	0.02	0.02
p*-value	0.93	0.56	0.40	0.95

* Likelihood ratio test comparison with a null model assuming a common rate between the CpATV branch and branches of closely related infectious viruses and an alternative model assuming independent rates for CpATV and closely related infectious viruses. New insect viruses were excluded of the comparison.

Detection of CpATV in other NGS datasets

The CpATV sequence was screened by BLAST search in our remaining 21 Culex transcriptomes, as well as in 60 arthropod and gastropod NGS datasets publicly available. Only one dataset returned a significant hit. A complete viral sequence was found in the transcriptome of the *Culex pipiens* individual GA35E. This mosquito had been sampled in El Habibia, Tunisia (160 km from Moknine) and six years later than the one in which CpATV_Moknine was detected (Supplementary Table 1). A mapping step on the CpATV_Moknine sequence as reference genome allowed the reconstruction of a 6,816 bp full-length genome. This second CpATV genome of the strain named ElHabibia (CpATV_ElHabibia) was covered by 42,924 good quality reads for a mean coverage of 320X (Supplementary Figure 1B), whereas the mean coverage of GA35E individual transcriptome was 205X.

Comparison of the CpATV Moknine and ElHabibia strains

Comparison of the CpATV_Moknine and CpATV_ElHabibia consensus sequences showed 27 mutations; 4 indels occurring in intergenic regions and 23 Single Nucleotide Polymorphism (SNPs) (Supplementary Table 5). One SNP occurred in intergenic regions and the remaining 22 were dispersed in ORFs 1, 3 and 4. Only 6 SNPs were non-synonymous substitutions, but they did not occur in any of the identified functional domains (helicase, RdRp, capsids). This suggests that both strains may have the same activity. The genetic divergence between the

Moknine and ElHabibia strains excludes the possibility that cross-contamination from a single biological sample happened during RNA extraction, library construction or sequencing, and confirmed two independent CpATV discoveries in distinct mosquitoes sampled at different times and places.

Intra-host diversity

Deep sequence coverage of 1,322X for CpATV_Moknine and 320X for CpATV_ElHabibia allowed the analysis of viral genetic variations within each mosquito host (ID GA35C and GA35E). Both CpATV strains displayed similar intra host genetic diversity with mean pi values of 1.0x10-3 and 0.96x10-3. Interestingly, two high frequency variants were observed.

A first SNP (83% G - 17% A) at genome position 2021 was found only in CpATV_Moknine. In the ElHabibia strain the Adenine was fixed at this position (100% A). This SNP, which is located within ORF1 outside the helicase and RdRp domains, caused an Asn-Ser amino acid replacement. Both amino acids share similar biochemical properties (polar, hydrophilic, neutral and relatively small), suggesting no major functional change associated with this SNP.

The second high frequency polymorphism was detected in both CpATV_Moknine and CpATV_ElHabibia. The variants presented the insertion of a 7th Adenine in ORF3 between the positions 4449 and 4454 (CpATV_Moknine). This indel caused a frameshift resulting in a premature stop codon at position 4461 for CpATV_Moknine (Figure 1A). As a result, two additional ORFs (ORF3A and ORF3B) could be predicted in the variants. ORF3A contained the full signal peptide, and ORF3B the full-length MSP7_C domain (Figure 1A). Remarkably, this indel was observed at very similar frequencies of 19.2% and 21.5% mapped reads in CpATV_Moknine and CpATV_ElHabibia respectively.

Discussion

High throughput sequencing accelerates virus discovery (85). Recent metaviromic approaches on invertebrates, based on pools of several species, have shown the virosphere extends far beyond the current taxonomy of viruses (10, 75). Here, we focused our analyses on the transcriptomes of individual mosquitoes to warrant virus-host interactions.

We discovered a new mosquito virus, named Culex pipiens Associated Tunisia Virus (CpATV), through systematic bioinformatic search of individual transcriptomes. The 6.8 kb genome of CpATV contains four ORFs, which encode a RNA replicase with helicase and RdRp domains

(ORF1), two capsids (ORF2 and 4) and an accessory gene (ORF3, λ) with similarities to *Plasmodium* domains. This genomic organization shows some similarities but also differences with that of *Virgaviridae*, ssRNA(+) viruses, to which CpATV is related. Within ORF1, the synteny between the helicase and replicase domains is highly conserved in all viruses investigated. However in contrast to all other genera, CpATV lacked a 5' methyltransferase, protease or movement protein domains (19, 82). This could indicate that the CpATV genome does not possess a covering 5' cap, which is mainly synthetized by this protein (86). Other ssRNA(+) viruses such as *Picornaviridae* or *Calicivirus* use other 5'-end protection such as internal ribosome entry site (IRES) and/or a viral genome-linked protein (VPg) recruitment (87), however no such elements could be detected in CpATV. Regulatory elements normally found in *Virgaviridae*, such as IRES, poly(A) tail, and tRNA-like cis-regulatory elements (i.e. hairpin-type pseudoknots) could not be identified in CpATV. These features appear facultative in several unassigned genera. For example, the mosquito infecting Negev virus (Nelorpivirus) has a 30-34 nt Poly (A) tail (19), a putative IRES structure in the 5'-UTR (88) but not hairpin-type pseudoknot.

The accessory genome of CpATV is distinguishable from those of other described viruses. It contains 2 capsid genes (α and κ) as well as ORF λ, with similarities to rhoptry proteins and *Plasmodium* MSP7 domain. We found this accessory genome composition in only three others recently described insect virus sequences: Boutonnet virus, TSA *Musca domestica* virus and ASV1, which together with CpATV form clade 8 (75, 76, 83). Both capsid genes derive from *Virgaviridae*, which only contains one capsid gene.

The capsid tree topology shows distinct phylogenetic origins for α and κ that apparently derive from two independent acquisitions rather than a unique ancestral acquisition followed by a duplication event producing two paralogs. Both horizontal gene transfer events occurred well after the initial diversification of the Virgaviridae, but not recently since ORF α was acquired before the divergence of clade 1, 7 and 8. Of note, not all viruses within this large clade possess a TMV-like coat protein. This suggests frequent capsid losses occurred in this group, potentially linked to adaption to new hosts, as we observed several host switches between plants and insects in the superclade. Viral capsid proteins are known source of evolutionary innovations in ssRNA(+) virus as they are implicated in multiple functions such as virus infectivity, pathogenicity, virus movement and transmission (89).

ORF λ was found in all viruses from clade 8. It shows high similarities with Rhoptry proteins (reticulocyte-binding proteins) involved in the process of invasion of the red blood cells by *Plasmodium* (73). Interestingly, only ORF 3 of CpATV, the mosquito virus in the group, displayed similarities to the Merozoite Surface Protein 7 domain, which is involved in vertebrate red blood cells invasion by the malaria agent (merozoite) during the *Plasmodium* life cycle (74). The MSP7 domain is not normally associated with Rhoptry proteins. It could derive from horizontal gene transfer of *Plasmodium* MSP7 sequence within a mosquito host. *Culex pipiens*, in which CpATV was found, is a natural vector of the avian malaria *Plasmodium relictum* (90–92). However BLAST search in the *P. relictum* genome and in other *Plasmodium* species did not reveal any MSP7 homologs. Alternatively the CpATV MSP7 domain could result from convergent evolution. At the molecular level, this evolutionary process occurs when two proteins, coded by unrelated genes in unrelated organisms, are facing similar specific environmental conditions. Evolution can retain a similar tertiary structure in unrelated proteins, which fulfil the same function (e.g. catalytic activity of an enzyme). For example, the convergent orientation of the active site of serine and cysteine proteases independently retained the same catalytic triad in 20 enzyme superfamilies (93). The hypothesis of interaction of the MSP7 domain in CpATV ORF λ with vertebrate host cells should now be addressed through functional studies.

Together with the other sequences from clade 8, CpATV does not bare the hallmarks of any viral group currently recognized by the ICTV. Comparative genomics and phylogenetic analysis indicate these sequences should belong to a new viral family. In support of this proposal, there are several lines of evidence indicating CpATV is a genuine infectious virus. First CpATV was sampled twice in 2005 and 2011 in *C. pipiens* collected as larvae in different regions of Tunisia. In both cases, the larvae were reared in the lab to adulthood, so the infection was not lethal to the larvae. Both mosquitoes were part of a larger effort to sample different *Culex* populations, which were not infected. This indicates that the CpATV sequence is not fixed in *Culex* genomes and that CpATV infections were initiated in the field rather than in the lab. Furthermore, as the sequence coverage of the virus is six times higher than the average transcriptome of the host in the Moknine sample, this points towards ongoing CpATV replication in the mosquito.

As the CpATV genome includes complete ORFs with “missing ends” and is covered by over 100 reads, it matches the definition of a “coding complete” (CC) viral genome, which fits the

recommendations for the description of novel viruses as it allows proper identifications and characterizations of the viral proteome and accurate phylogenetic analyzes (94). This virus seemed to have a restricted local distribution so far, since no traces of CpATV were found in other sequence database or in broader environmental samples databases. CpATV prevalence thus seems relatively low, but larger targeted field sampling of mosquitoes could reveal its full extent. Both CpATV_Moknine and CpATV_ElHabibia share 99.6% of nucleotide identity and display low intra-host nucleotide polymorphism (pi =1.0x10-3 and 0.96x10-3 respectively). Low polymorphism appears as a biological characteristic of CpATV since both estimations were independent and yet very similar. Taking the high mutation rate expected by error-prone RNA-dependent RNA polymerase of RNA viruses into account (95), it is therefore likely that CpATV may have small population size or suffered from large, recent or recurrent bottlenecks during vertical transmission (96, 97). It is thus possible that the two strain-specific SNPs observed at high frequency in the genomes of CpATV_Moknine and CpATV_ElHabibia originated by genetic drift only. However, the biological meaning of maintaining the indel polymorphism found in ORF3 seems incompatible with genetic drift, as the same pattern was observed twice. The high frequency (circa 20%) of Adenine insertion observed in ORF3 is in addition not compatible with known Illumina error rates in 7 bases homopolymers, were 0.02 to 0.002% error rate is expected (98). This indel polymorphism rather seems the consequence of natural selection. The finer evaluation of the strength of selection pressure using either dN/dS estimations as a proxy of selective pressures or the positions of non-synonymous polymorphisms on functional domains on the CpATV genome shows that all conserved domains are under strong purifying selection. The conservation of important protein sites suggests that both CpATV strains are fully functional and adapted to their hosts. These results strongly suggest that both Moknine and ElHabibia CpATV strains are infectious viruses.

Phylogenetic analyses retraced the evolution of CpATV among other recently discovered ssRNA(+) insect viruses (76). CpATV does not cluster with the clade comprising the mosquito specific Negev virus (Nelorpivirus). It is most closely related to the Boutonnet virus within a larger group of new insect viruses, several of which are associated with Diptera. These viruses do not form a monophyletic group and are early-diverging in the phylogeny. They do not belong to the same clade as *Virgaviridae* / *Bromoviridae* / *Closteroviridae* / *Idaeovirus* and Nelorpivirus / Sandewavirus / *Higrevirus* / *Cilevirus*, but clearly form distinct lineages. Overall, phylogenetic relationships show frequent host shifts between insect and plant shaped the

evolution of these viruses, as recently pointed out (10, 99). A more exhaustive screening using degenerated PCR primers on a broad insect sample is now required to gain knowledge of host range, prevalence and biodiversity of these viruses. It would also be interesting to look at the real replication competence of these new viruses in plants and insects, to determine if the hosts in which they were first found are vectors or final hosts. Both Sandewavirus and Nelorpivirus have been shown to replicate efficiently in mosquitoes, and are true dipteran viruses (19). Similar analyzes should be done for the newly discovered viruses, but this implies lifting technical issues on virus isolation.

It was recently proposed to integrate in the taxonomic classification viruses discovered solely through metagenomics (100). The current classification of RNA viruses is mostly based on the replicase phylogeny (101). The analyses showed that CpATV represents a new viral species that does not belong to any currently described virus families or unassigned genera. The other 3 sequences from clade 8 (Boutonnet virus, TSA *Musca domestica*, and ASV1) found in different insect species, also represent different viral species based on phylogenetic distances (from 46.1% to 30.2% nucleotide identity between each pair of viruses). Together these 4 viruses form a monophyletic group that corresponds to higher classification rank at least at the level of genus within a new family. Furthermore, our study enabled the distinction of seven other clades of insect viruses, also representing new families or genera. To sum up this study described a powerful methodology for high-throughput virus discovery using transcriptomic analysis, which completed by comparative analyses accelerates viral taxonomy inference.

Availability of supporting data

Reads used for this analysis originated from a previous study (35) on *Culex hortensis* (SRA accession no. SRX565078 and SRX565079), *Culex torrentium* (SRA accession no. SRX565090 and SRX565091), *Culex pipiens* (SRA accession no. SRX565080 to SRX565089) and from this study for *Culex pipiens* (SRA accession SRX1453901, SRX1453908, SRX1457280 to SRX1457282, SRX1457496, SRX1457498 and SRX1457500). CpATV strains Moknine and ElHabibia genome sequences are in the Supplementary File 1 and were not deposited on GenBank according to submission policies.

Authors' contributions

PG and EAH performed the design and the coordination of the study. CA and MW carried out the sample collection and performed RNA isolation. DB carried out the bioinformatics analysis. DB, EAH and PG analyzed the result and wrote the manuscript. All authors read and approved the final manuscript.

Funding information

This work was supported by European Research Council (ERC) grants EAH (ERC GENOVIR 205206). The funders had no role in study design, data collection and interpretation, or the decision to submit the work for publication.

Acknowledgements

Data used in this work were partly produced through molecular genetic analysis technical facilities of the SFR "Montpellier Environnement Biodiversite", thanks to Dr. P. Clair (UM2-Montpellier GenomiX). Analyses largely benefited from the ISEM computing cluster platform. We are also grateful to the Genotoul Bioinformatics Platform Toulouse Midi-Pyrenees for providing computing and storage resources. We would like to thank J. Thézé and A. Bézier for fruitful discussions, A. Rivero for sharing unpublished data of P. relictum and N. Galtier, M. Ballenghien, J. Romiguier, K. Belkhir, R. Dernat, J. Veyssier for help in data accessibility and M. Rivera for English improvement. This work has been supported by a European Research Council (ERC) grant to Nicolas Galtier (ERC PopPhyl 232971).

The authors declare that they have no competing interests.

References

1. Paul JH, Sullivan MB. 2005. Marine phage genomics: what have we learned? Curr. Opin. Biotechnol. 16:299–307.
2. Ng TFF, Marine R, Wang C, Simmonds P, Kapusinszky B, Bodhidatta L, Oderinde BS, Wommack KE, Delwart E. 2012. High variety of known and new RNA and DNA viruses of diverse origins in untreated sewage. J. Virol. 86:12161–12175.
3. Phan TG, Kapusinszky B, Wang C, Rose RK, Lipton HL, Delwart EL. 2011. The fecal viral flora of wild rodents. PLoS Pathog. 7:e1002218.
4. Chandler JA, Liu RM, Bennett SN. 2015. RNA shotgun metagenomic sequencing of northern California (USA) mosquitoes uncovers viruses, bacteria, and fungi. Front. Microbiol. 6:185.
5. Bishop-Lilly K, Frey K, Biser T, Hamilton T, Santos C, Pimentel G, Mokashi V. 2016. Bioinformatic characterization of mosquito viromes within the eastern United States and Puerto Rico: discovery of novel viruses. Evol. Bioinforma. 12:1.
6. Cook S, Chung BY-W, Bass D, Moureau G, Tang S, McAlister E, Culverwell CL, Glucksman E, Wang H, Brown TDK, Gould EA, Harbach RE, de Lamballerie X, Firth AE. 2013. Novel virus discovery and genome reconstruction from field RNA samples reveals highly divergent viruses in dipteran hosts. PLoS One 8.
7. Junglen S, Drosten C. 2013. Virus discovery and recent insights into virus diversity in arthropods. Curr. Opin. Microbiol. 16:507–513.
8. Liu S, Chen Y, Bonning BC. 2015. RNA virus discovery in insects. Curr. Opin. Insect Sci. 8:54–61.

9. Bichaud L, de Lamballerie X, Alkan C, Izri A, Gould EA, Charrel RN. 2014. Arthropods as a source of new RNA viruses. Microb. Pathog. 77:136–41.
10. Shi M, Lin X-D, Tian J-H, Chen L-J, Chen X, Li C-X, Qin X-C, Li J, Cao J-P, Eden J-S, Buchmann J, Wang W, Xu J, Holmes EC, Zhang Y-Z. 2016. Redefining the invertebrate RNA virosphere. Nature 1–12.
11. Quan P-L, Junglen S, Tashmukhamedova A, Conlan S, Hutchison SK, Kurth A, Ellerbrok H, Egholm M, Briese T, Leendertz FH, Lipkin WI. 2010. Moussa virus: A new member of the Rhabdoviridae family isolated from Culex decens mosquitoes in Côte d'Ivoire. Virus Res. 147:17–24.
12. Kuwata R, Isawa H, Hoshino K, Tsuda Y, Yanase T, Sasaki T, Kobayashi M, Sawabe K. 2011. RNA splicing in a new Rhabdovirus from Culex mosquitoes. J. Virol. 85:6185–6196.
13. Vasilakis N, Castro-Llanos F, Widen SG, Aguilar P V., Guzman H, Guevara C, Fernandez R, Auguste AJ, Wood TG, Popov V, Mundal K, Ghedin E, Kochel TJ, Holmes EC, Walker PJ, Tesh RB. 2014. Arboretum and Puerto Almendras viruses: two novel rhabdoviruses isolated from mosquitoes in Peru. J. Gen. Virol. 95:787–792.
14. Attoui H, Mohd Jaafar F, Belhouchet M, Biagini P, Cantaloube JF, De Micco P, De Lamballerie X. 2005. Expansion of family Reoviridae to include nine-segmented dsRNA viruses: Isolation and characterization of a new virus designated aedes pseudoscutellaris reovirus assigned to a proposed genus (Dinovernavirus). Virology 343:212–223.
15. Hermanns K, Zirkel F, Kurth A, Drosten C, Junglen S. 2014. Cimodo virus belongs to a novel lineage of reoviruses isolated from African mosquitoes. J. Gen. Virol. 95:905–909.
16. Auguste AJ, Kaelber JT, Fokam EB, Guzman H, Carrington CVF, Erasmus JH, Kamgang B, Popov VL, Jakana J, Liu X, Wood TG, Widen SG, Vasilakis N, Tesh RB, Chiu W, Weaver SC. 2015. A newly isolated reovirus has the simplest genomic and structural organization of any reovirus. J. Virol. 89:676–87.
17. Wang L, Lv X, Zhai Y, Fu S, Wang D, Rayner S, Tang Q, Liang G. 2012. Genomic characterization of a novel virus of the family Tymoviridae isolated from mosquitoes. PLoS One 7:e39845.
18. Schuster S, Zirkel F, Kurth A, van Cleef KWR, Drosten C, van Rij RP, Junglen S. 2014. A unique nodavirus with novel features: mosinovirus expresses two subgenomic RNAs, a capsid gene of unknown origin, and a suppressor of the antiviral RNA interference pathway. J. Virol. 88:13447–59.
19. Vasilakis N, Forrester NL, Palacios G, Nasar F, Savji N, Rossi SL, Guzman H, Wood TG, Popov V, Gorchakov R, González AV, Haddow AD, Watts DM, Rosa APAT da, Weaver SC, Lipkin WI, Tesh RB. 2013. Negevirus: a proposed new taxon of insect-specific viruses with wide geographic distribution. J. Virol. 87:2475–2488.
20. Fujita R, Kuwata R, Kobayashi D, Bertuso AG, Isawa H, Sawabe K. 2016. Bustos virus, a new member of the negevirus group isolated from a Mansonia mosquito in the Philippines. Arch. Virol. 1–10.
21. Kallies R, Kopp A, Zirkel F, Estrada A, Gillespie TR, Drosten C, Junglen S. 2014. Genetic characterization of goutanap virus, a novel virus related to negeviruses, cileviruses and higreviruses. Viruses 6:4346–57.
22. Auguste AJ, Carrington CVF, Forrester NL, Popov VL, Guzman H, Widen SG, Wood TG, Weaver SC, Tesh RB. 2014. Characterization of a novel Negevirus and a novel Bunyavirus isolated from Culex (Culex) declarator mosquitoes in Trinidad. J. Gen. Virol. 95:481–5.
23. Kawakami K, Kurnia YW, Fujita R, Ito T, Isawa H, Asano S, Binh ND, Bando H. 2016. Characterization of a novel negevirus isolated from Aedes larvae collected in a subarctic region of Japan. Arch. Virol. 161:801–809.
24. Nabeshima T, Inoue S, Okamoto K, Posadas-Herrera G, Yu F, Uchida L, Ichinose A, Sakaguchi M, Sunahara T, Buerano CC, Tadena FP, Orbita IB, Natividad FF, Morita K. 2014. Tanay virus, a new species of virus isolated from mosquitoes in the Philippines. J. Gen. Virol. 95:1390–1395.
25. Bolling BG, Weaver SC, Tesh RB, Vasilakis N. 2015. Insect-specific virus discovery: significance for the arbovirus community. Viruses 7:4911–28.
26. Hall RA, Bielefeldt-ohmann H, Mclean BJ, O'Brien CA, Colmant AMG, Piyasena TBH, Harrison JJ, Newton ND, Barnard RT, Prow NA, Deerain JM, Mah MGKY, Hobson-Peters J. 2016. Commensal viruses of mosquitoes : host restriction , transmission , and interaction with arboviral pathogens. Evol. Bioinforma. 12:35–44.
27. Ng TFF, Willner DL, Lim YW, 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.
28. 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.
29. Aguiar ERGR, Olmo RP, Paro S, Ferreira FV, de Faria IJ da S, Todjro YMH, Lobo FP, Kroon EG, Meignin C, Gatherer D, Imler J-L, Marques JT. 2015. Sequence-independent characterization of viruses based on the pattern of viral small RNAs produced by the host. Nucleic Acids Res. 43:6191–6206.
30. Coffey LL, Page BL, Greninger AL, Herring BL, Russell RC, Doggett SL, Haniotis J, Wang C, Deng X, Delwart EL. 2014. Enhanced arbovirus surveillance with deep sequencing: Identification of novel rhabdoviruses and bunyaviruses in Australian mosquitoes. Virology 448:146–58.
31. Shi M, Lin X-D, Vasilakis N, Tian J-H, Li C-X, Chen L-J, Eastwood G, Diao X-N, Chen M-H, Chen X, Qin X-C, Widen SG, Wood TG, Tesh RB, Xu J, Holmes EC, Zhang Y-Z. 2015. Divergent viruses discovered in arthropods and vertebrates revise the evolutionary history of the Flaviviridae and related viruses. J. Virol. 90:659–69.
32. Fonseca DM, Keyghobadi N, Malcolm CA, Mehmet C, Schaffner F, Mogi M, Fleischer RC, Wilkerson RC. 2004. Emerging vectors in the Culex pipiens complex. Science 303:1535–8.
33. Mackenzie JS, Jeggo M. 2013. Reservoirs and vectors of emerging viruses. Curr. Opin. Virol. 3:170–9.
34. Gayral P, Weinert L, Chiari Y, Tsagkogeorga G, Ballenghien M, Galtier N. 2011. Next-generation sequencing of transcriptomes: a guide to RNA isolation in nonmodel animals. Mol. Ecol. Resour. 11:650–661.

35. Romiguier J, Lourenco J, Gayral P, Faivre N, Weinert LA, Ravel S, Ballenghien M, Cahais V, Bernard A, Loire E, Keller L, Galtier N. 2014. Population genomics of eusocial insects: the costs of a vertebrate-like effective population size. J. Evol. Biol. 27:593–603.
36. Simpson JT, Wong K, Jackman SD, Schein JE, Jones SJM, Birol I. 2009. ABySS: a parallel assembler for short read sequence data. Genome Res. 19:1117–1123.
37. Birol I, Jackman SD, Nielsen CB, Qian JQ, Varhol R, Stazyk G, Morin RD, Zhao Y, Hirst M, Schein JE, Horsman DE, Connors JM, Gascoyne RD, Marra M a., Jones SJM. 2009. De novo transcriptome assembly with ABySS. Bioinformatics 25:2872–2877.
38. Cahais V, Gayral P, Tsagkogeorga G, Melo-Ferreira J, Ballenghien M, Weinert L, Chiari Y, Belkhir K, Ranwez V, Galtier N. 2012. Reference-free transcriptome assembly in non-model animals from next-generation sequencing data. Mol. Ecol. Resour. 12:834–45.
39. Huang X, Madan A. 1999. CAP3: a DNA sequence assembly program. Genome Res. 9:868–877.
40. Hyatt D, Chen G-L, Locascio PF, Land ML, Larimer FW, Hauser LJ. 2010. Prodigal: prokaryotic gene recognition and translation initiation site identification. BMC Bioinformatics 11:119.
41. Hyatt D, Locascio PF, Hauser LJ, Uberbacher EC. 2012. Gene and translation initiation site prediction in metagenomic sequences. Bioinformatics 28:2223–2230.
42. Remmert M, Biegert A, Hauser A, Söding J. 2011. HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment. Nat. Methods 9:173–175.
43. Söding J. 2005. Protein homology detection by HMM-HMM comparison. Bioinformatics 21:951–960.
44. Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL. 2009. BLAST+: architecture and applications. BMC Bioinformatics 10:421.
45. Li H, Durbin R. 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 25:1754–60.
46. 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, Mentjies 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.
47. Kofler R, Orozco-terWengel P, De Maio N, Pandey RV, Nolte V, Futschik A, Kosiol C, Schlötterer C. 2011. PoPoolation: a toolbox for population genetic analysis of next generation sequencing data from pooled individuals. PLoS One 6:e15925.
48. Mitchell A, Chang H-Y, Daugherty L, Fraser M, Hunter S, Lopez R, McAnulla C, McMenamin C, Nuka G, Pesseat S, Sangrador-Vegas A, Scheremetjew M, Rato C, Yong S-Y, Bateman A, Punta M, Attwood TK, Sigrist CJA, Redaschi N, Rivoire C, Xenarios I, Kahn D, Guyot D, Bork P, Letunic I, Gough J, Oates M, Haft D, Huang H, Natale DA, Wu CH, Orengo C, Sillitoe I, Mi H, Thomas PD, Finn RD. 2015. The InterPro protein families database: the classification resource after 15 years. Nucleic Acids Res. 43:D213–D221.
49. Jones P, Binns D, Chang H-Y, Fraser M, Li W, McAnulla C, McWilliam H, Maslen J, Mitchell A, Nuka G, Pesseat S, Quinn AF, Sangrador-Vegas A, Scheremetjew M, Yong S-Y, Lopez R, Hunter S. 2014. InterProScan 5: genome-scale protein function classification. Bioinformatics 30:1236–1240.
50. Marchler-Bauer A, Derbyshire MK, Gonzales NR, Lu S, Chitsaz F, Geer LY, Geer RC, He J, Gwadz M, Hurwitz DI, Lanczycki CJ, Lu F, Marchler GH, Song JS, Thanki N, Wang Z, Yamashita RA, Zhang D, Zheng C, Bryant SH. 2015. CDD: NCBI's conserved domain database. Nucleic Acids Res. 43:D222-6.
51. Schultz J, Milpetz F, Bork P, Ponting CP. 1998. SMART, a simple modular architecture research tool: Identification of signaling domains. Proc. Natl. Acad. Sci. 95:5857–5864.
52. Letunic I, Doerks T, Bork P. 2012. SMART 7: recent updates to the protein domain annotation resource. Nucleic Acids Res. 40:D302–D305.
53. Petersen TN, Brunak S, von Heijne G, Nielsen H. 2011. SignalP 4.0: discriminating signal peptides from transmembrane regions. Nat. Methods 8:785–786.
54. Emanuelsson O, Nielsen H, Brunak S, von Heijne G. 2000. Predicting subcellular localization of proteins based on their N-terminal amino acid sequence. J. Mol. Biol. 300:1005–1016.
55. Ahmed F, Kumar M, Raghava GPS. 2009. Prediction of polyadenylation signals in human DNA sequences using nucleotide frequencies. In Silico Biol.
56. Chang T-H, Huang H-Y, Hsu JB-K, Weng S-L, Horng J-T, Huang H-D. 2013. An enhanced computational platform for investigating the roles of regulatory RNA and for identifying functional RNA motifs. BMC Bioinformatics 14 Suppl 2:S4.
57. Hong J-J, Wu T-Y, Chang T-Y, Chen C-Y. 2013. Viral IRES prediction system - a web server for prediction of the IRES secondary structure in silico. PLoS One 8:e79288.
58. Aurrecoechea C, Brestelli J, Brunk BP, Dommer J, Fischer S, Gajria B, Gao X, Gingle A, Grant G, Harb OS, Heiges M, Innamorato F, Iodice J, Kissinger JC, Kraemer E, Li W, Miller JA, Nayak V, Pennington C, Pinney DF, Roos DS, Ross C, Stoeckert CJ, Treatman C, Wang H. 2009. PlasmoDB: a functional genomic database for malaria parasites. Nucleic Acids Res. 37:D539–D543.
59. Katoh K, Misawa K, Kuma K, Miyata T. 2002. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. Nucleic Acids Res. 30:3059–3066.
60. Abascal F, Zardoya R, Posada D. 2005. ProtTest: Selection of best-fit models of protein evolution. Bioinformatics 21:2104–2105.
61. Guindon S, Gascuel O. 2003. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. Syst. Biol. 52:696–704.

62. Anisimova M, Gascuel O. 2006. Approximate likelihood-ratio test for branches: a fast, accurate, and powerful alternative. Syst. Biol. 55:539–52.
63. Shimodaira H, Hasegawa M. 1999. Multiple comparisons of log-likelihoods with applications to phylogenetic inference. Mol. Biol. Evol. 16:1114–1116.
64. Kishino H, Hasegawa M. 1989. Evaluation of the maximum likelihood estimate of the evolutionary tree topologies from DNA sequence data, and the branching order in hominoidea. J. Mol. Evol. 29:170–179.
65. Schmidt HA, Strimmer K, Vingron M, von Haeseler A. 2002. TREE-PUZZLE: maximum likelihood phylogenetic analysis using quartets and parallel computing. Bioinformatics 18:502–504.
66. Suyama M, Torrents D, Bork P. 2006. PAL2NAL: robust conversion of protein sequence alignments into the corresponding codon alignments. Nucleic Acids Res. 34:W609-12.
67. Nei M, Gojobori T. 1986. Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. Mol. Biol. Evol. 3:418–426.
68. Kryazhimskiy S, Plotkin JB. 2008. The population genetics of dN/dS. PLoS Genet. 4:e1000304.
69. Yang Z. 1998. Likelihood ratio tests for detecting positive selection and application to primate lysozyme evolution. Mol. Biol. Evol. 15:568–573.
70. Yang Z, Bielawski JP. 2000. Statistical methods for detecting molecular adaptation. Trends Ecol. Evol. 15:496–503.
71. Yang Z. 2007. PAML 4: Phylogenetic Analysis by Maximum Likelihood. Mol. Biol. Evol. 24:1586–1591.
72. Koonin E V, Dolja V V. 1993. Evolution and taxonomy of positive-strand RNA viruses: implications of comparative analysis of amino acid sequences. Crit. Rev. Biochem. Mol. Biol. 28:375–430.
73. Counihan NA, Kalanon M, Coppel RL, de Koning-Ward TF. 2013. Plasmodium rhoptry proteins: why order is important. Trends Parasitol. 29:228–236.
74. Pachebat JA, Ling IT, Grainger M, Trucco C, Howell S, Fernandez-Reyes D, Gunaratne R, Holder AA. 2001. The 22 kDa component of the protein complex on the surface of Plasmodium falciparum merozoites is derived from a larger precursor, merozoite surface protein 7. Mol. Biochem. Parasitol. 117:83–89.
75. Webster CL, Waldron FM, Robertson S, Crowson D, Ferrari G, Quintana JF, Brouqui J-M, Bayne EH, Longdon B, Buck AH, Lazzaro BP, Akorli J, Haddrill PR, Obbard DJ. 2015. The discovery, distribution, and evolution of viruses associated with Drosophila melanogaster. PLoS Biol. 13:e1002210.
76. Webster C, Longdon B, Lewis S, Obbard D. 2016. Twenty-five new viruses associated with the Drosophilidae (Diptera). Evol. Bioinforma. 13.
77. Rozanov MN, Koonin E V, Gorbalenya AE. 1992. Conservation of the putative methyltransferase domain : a hallmark of the "Sindbis-like" supergroup of positive-strand RNA viruses. J. Gen. Virol. 73:2129–2134.
78. Byszewska M, Śmietański M, Purta E, Bujnicki JM. 2014. RNA methyltransferases involved in 5′ cap biosynthesis. RNA Biol. 11:1597–1607.
79. Mercer AA, Fleming SB, Ueda N. 2005. F-box-like domains are present in most poxvirus ankyrin repeat proteins. Virus Genes 31:127–133.
80. Barry M, van Buuren N, Burles K, Mottet K, Wang Q, Teale A. 2010. Poxvirus exploitation of the ubiquitin-proteasome system. Viruses 2:2356–2380.
81. Kuchibhatla DB, Sherman WA, Chung BYW, Cook S, Schneider G, Eisenhaber B, Karlin DG. 2014. Powerful sequence similarity search methods and in-depth manual analyses can identify remote homologs in many apparently "orphan" viral proteins. J. Virol. 88:10–20.
82. King AMQ, Adams MJ, Carstens EB, Lefkowitz EJ. 2012. Virus taxonomy. Ninth report of the International Committee on Taxonomy of Viruses.Virus Taxonomy. Academic Press.
83. Li X, Xu P, Yang X, Yuan H, Chen L, Lu Y. 2017. The genome sequence of a novel RNA virus in Adelphocoris suturalis. Arch. Virol. 1.
84. King, A.M.Q., Adams, M.J., Carstens, E.B. and Lefkowitz EJ. 2012. Virgaviridae family, p. 1139–1162. In King, A.M.Q., Adams, M.J., Carstens, E.B. and Lefkowitz, EJ (ed.), Virus taxonomy: classification and nomenclature of viruses: Ninth report of the International Committee on Taxonomy of Viruses. Elsevier Academic Press.
85. Paez-Espino D, Eloe-Fadrosh EA, Pavlopoulos GA, Thomas AD, Huntemann M, Mikhailova N, Rubin E, Ivanova NN, Kyrpides NC. 2016. Uncovering Earth's virome. Nature 536:425–430.
86. Decroly E, Ferron F, Lescar J, Canard B. 2011. Conventional and unconventional mechanisms for capping viral mRNA. Nat. Rev. Microbiol. 10:51.
87. Goodfellow I, Chaudhry Y, Gioldasi I, Gerondopoulos A, Natoni A, Labrie L, Laliberté J-F, Roberts L. 2005. Calicivirus translation initiation requires an interaction between VPg and eIF4E. EMBO Rep. 6:968–972.
88. Gorchakov R V, Tesh RB, Weaver SC, Nasar F. 2014. Generation of an infectious Negev virus cDNA clone. J. Gen. Virol. 95:2071–4.
89. Weber PH, Bujarski JJ. 2015. Multiple functions of capsid proteins in (+) stranded RNA viruses during plant-virus interactions. Virus Res. 196:140–149.
90. Melrose WD. 2002. Lymphatic filariasis: new insights into an old disease. Int. J. Parasitol. 32:947–960.
91. Turell MJ. 2012. Members of the Culex pipiens complex as vectors of viruses. J. Am. Mosq. Control Assoc. 28:123–126.
92. Tate P, Vincent M. 2009. The susceptibility of autogenous and anautogenous races of Culex pipiens to infection with avian malaria (Plasmodium relictum). Parasitology 26:512.

93. Buller AR, Townsend C a. 2013. Intrinsic evolutionary constraints on protease structure, enzyme acylation, and the identity of the catalytic triad. Proc. Natl. Acad. Sci. 110:E653–E661.
94. Ladner JT, Beitzel B, Chain PSG, Davenport MG, Donaldson E, Frieman M, Kugelman J, Kuhn JH, O'Rear J, Sabeti PC, Wentworth DE, Wiley MR, Yu G-Y, Sozhamannan S, Bradburne C, Palacios G. 2014. Standards for sequencing viral genomes in the era of high-throughput sequencing. MBio 5:e01360-14-e01360-14.
95. Drake JW, Holland JJ. 1999. Mutation rates among RNA viruses. Proc. Natl. Acad. Sci. 96:13910–13913.
96. Zwart MP, Elena SF. 2015. Matters of Size: Genetic Bottlenecks in Virus Infection and Their Potential Impact on Evolution. Annu. Rev. Virol. 2:161–179.
97. Elena SF, Sanjuán R, Bordería A V, Turner PE. 2001. Transmission bottlenecks and the evolution of fitness in rapidly evolving RNA viruses. Infect. Genet. Evol. 1:41–48.
98. Minoche AE, Dohm JC, Himmelbauer H. 2011. Evaluation of genomic high-throughput sequencing data generated on Illumina HiSeq and Genome Analyzer systems. Genome Biol. 12:R112.
99. Nunes MRT, Contreras-Gutierrez MA, Guzman H, Martins LC, Barbirato MF, Savit C, Balta V, Uribe S, Vivero R, Suaza JD, Oliveira H, Nunes Neto JP, Carvalho VL, da Silva SP, Cardoso JF, de Oliveira RS, da Silva Lemos P, Wood TG, Widen SG, Vasconcelos PFC, Fish D, Vasilakis N, Tesh RB. 2017. Genetic characterization, molecular epidemiology, and phylogenetic relationships of insect-specific viruses in the taxon Negevirus. Virology 504:152–167.
100. Simmonds P, Adams MJ, Benkő M, Breitbart M, Brister JR, Carstens EB, Davison AJ, Delwart E, Gorbalenya AE, Harrach B, Hull R, King AMQ, Koonin E V., Krupovic M, Kuhn JH, Lefkowitz EJ, Nibert ML, Orton R, Roossinck MJ, Sabanadzovic S, Sullivan MB, Suttle CA, Tesh RB, van der Vlugt RA, Varsani A, Zerbini FM. 2017. Consensus statement: Virus taxonomy in the age of metagenomics. Nat. Rev. Microbiol. 15:161–168.
101. Koonin E V. 1991. The phylogeny of RNA-dependent RNA polymerases of positive-strand RNA viruses. J. Gen. Virol. 72:2197–2206.

Appendix Table A1: Origins and characteristics of mosquito samples.

Culex species	Individual name	Country	Location	Year	Latitude	Longitude	SRA Accession number	References
Culex pipiens	GA34I	Algeria	Lac des Oiseaux	2008	36.7715922	8.1092816	SRX565088	Previous work (Romiguier, *et al*, 2014)
	GA34C	Burkina Faso	Ouagadougou	1997	12.364637	-1.5338639	SRX565082	(Romiguier, *et al*, 2014)
	GA34J	China	Zhu-Shang	2003	24.6446709	102.686988	SRX565089	(Romiguier, *et al*, 2014)
	GA34H	Costa Rica	Pueto Viero	2006	9.646168	-82.7490229	SRX565087	(Romiguier, *et al*, 2014)
	GA34F	France	Ganges	2011	43.934614	3.708787	SRX565085	(Romiguier, *et al*, 2014)
	GA34G	France	Triolet	2011	43.629366	3.860393	SRX565086	(Romiguier, *et al*, 2014)
	GA35A	USA	California (lab strain)	1950	na	na	SRX1453901	This study
	GA35B	USA	California (lab strain)	1950	na	na	SRX1453908	This study
	GA34E	Israel	Tel Aviv Yatouch	2011	32.070065	34.777508	SRX565084	(Romiguier, *et al*, 2014)
	GA34B	Réunion island	Saint Benoît	2011	-21.043737	55.717857	SRX565081	(Romiguier, *et al*, 2014)
	GA34D	Philippine	Manille	2003	14.5995124	120.9842195	SRX565083	(Romiguier, *et al*, 2014)
	GA34A	Tunisia	Grombalia	2005	36.601541	10.50034	SRX565080	(Romiguier, *et al*, 2014)
	GA35C *	Tunisia	Moknine	2005	35.626433	10.912933	SRX1457280	This study
	GA35D	Tunisia	Grombalia	2005	36.590900	10.490600	SRX1457281	This study
	GA35E *	Tunisia	El Habibia	2011	36.803817	9.942400	SRX1457282	This study
	GA35F	Tunisia	Kef2	2008	36.159550	8.718633	SRX1457496	This study
	GA35G	Tunisia	Azib	2011	37.222633	9.934417	SRX1457498	This study
	GA35H	Tunisia	Utique	2011	37.069600	10.007067	SRX1457500	This study
Culex hortensis	GA34K	France	Laroque	2011	43.922663	3.723849	SRX565078	(Romiguier, *et al*, 2014)
	GA34L	France	Mas de Foiton	2004	43.796794	4.593698	SRX565079	(Romiguier, *et al*, 2014)
Culex torrentium	GA34M	France	Polignac	1996	45.071348	3.859441	SRX565090	(Romiguier, *et al*, 2014)
	GA34N	Sweden	Uppsala	1994	59.8585638	17.6389267	SRX565091	(Romiguier, *et al*, 2014)

*Individuals with a full-length viral genome detected

Romiguier J, **Lourenco J**, **Gayral P**, **Faivre N**, **Weinert LA**, **Ravel S**, **Ballenghien M**, **Cahais V**, **Bernard A**, **Loire E**, **Keller L**, **Galtier N**. 2014. Population genomics of eusocial insects: the costs of a vertebrate-like effective population size. J. Evol. Biol. **27**:593–603.

Appendix Table A2: Mosquitoes genomic and transcriptomic sequence data used in this study.

BioProject Accession	ID	Project Name	Type of data	Taxonomic ID	Scientific Name	Availability
PRJNA29017	29017	Reference sequences (RefSeq) of the *Culex quinquefasciatus* genome	Genome and transcriptome	7176	*Culex quinquefasciatus*	Release April 2014*
PRJNA18751	18751	The vector for *West Nile Virus*	Genome	7176	*Culex quinquefasciatus*	
PRJNA246260	246260	*Culex pipiens pallens* and *C. p. quinquefasciatus* with different resistant level	Transcriptome	7176	*Culex quinquefasciatus*	SRR1271734
				7176	*Culex quinquefasciatus*	SRR1271735
				7176	*Culex quinquefasciatus*	SRR1271736
				7176	*Culex quinquefasciatus*	SRR1271737
				42434	*Culex pipiens pallens*	SRR1271738
				42434	*Culex pipiens pallens*	SRR1271739
				42434	*Culex pipiens pallens*	SRR1271740
				42434	*Culex pipiens pallens*	SRR1271741
				42434	*Culex pipiens pallens*	SRR1271742
				42434	*Culex pipiens pallens*	SRR1271743
				42434	*Culex pipiens pallens*	SRR1271744
				7176	*Culex quinquefasciatus*	SRR1271745
PRJNA253629	253629	RNA-sequencing of *Culex pipiens molestus* and *C. pipiens quinquefasciatus*	Transcriptome	7176	*Culex quinquefasciatus*	SRR1462324
				233155	*Culex pipiens molestus*	SRR1462325

*http://www.broadinstitute.org/annotation/genome/culex_pipiens

Appendix Table A3: *Plasmodium* genomes available in PlasmoDB (release 2015/07/23)

***Plasmodium* species**	**Genome version**
P. berghei ANKA	2013/03/01
P. chabaudi chabaudi	2013/03/01
P. coatneyi hackeri	2014/07/21
P. cynomolgi strain B	2012/09/19
P. falciparum 3D7	2013/03/01
P. falciparum IT	2013/03/01
P. gallinaceum 8A	2006/12/31
P. knowlesi strain H	2013/03/01
P. reichenowi CDC	2014/09/03
P. vivax Sal-1	2015/05/01
P. yoelii yoelli 17X	2014/02/01
P. yoelii yoelli 17XNL	2005/09/01
P. yoelii yoelli YM	2013/03/01

Appendix Table A4: Summary statistics (means for each species) of transcriptome assembly quality, ORF prediction and viral homology search.

Genus	*Culex*		
Species	*pipiens*	*hortensis*	*torrentium*
No. of transcriptomes analyzed	18	2	2
Initial assembly (ABYSS)			
No. million reads	25.6	54.9	58.3
No. contigs (x 1000)	27.2	73.8	90.2
Median length	52	52	51
N50	69	54	52
Final assembly (ABYSS-CAP3)			
No. contigs (x 1000)	3.4	8.3	7.9
Median length	122	118	122
N50	233	196	238
Virus detection			
No. of ORF predicted per species (x 1000)	2.6	6.0	5.8
No. of transcriptomes with full-length viral genome	**2**	0	0

Appendix Table A5: Fixed mutations distinguishing CpATV_Moknine and CpATV_ElHabibia genomes.

ORF	Genomic Region	Mutation number	Position on CpATV_Moknine	Nucleotide in CpATV_Moknine	Position on CpATV_ElHabibia	Nucleotide in CpATV_ElHabibia	Mutation Type	Consequence of mutation	Amino acid changes
5' non-coding region	-	1	1	C	-	-	Insertion / Deletion	-	-
		2	2	C	-	-	Insertion / Deletion	-	-
		3	3	T	-	-	Insertion / Deletion	-	-
		4	110	C	107	T	Transition	-	-
ORF 1	No domain	5	498	C	495	T	Transition	Synonymous	AGC (S) <-> AGT (S)
	Viral Helicase	6	765	G	762	T	Transversion	Synonymous	TCG (S) <-> TCG (S)
		7	1,014	C	1,011	T	Transition	Synonymous	GTC (V) <-> GTT (V)
	No domain	8	1,497	A	1,494	G	Transition	Synonymous	CCA (P) <-> CCG (P)
		9	1,608	C	1,605	T	Transition	Synonymous	TTC (F) <-> TTT (F)
		10	1,699	A	1,696	C	Transversion	Synonymous	AGA (R) <-> CGA (R)
	RdRp	11	2,079	T	2,076	A	Transversion	Synonymous	ATT (I) <-> ATA (I)
		12	2,580	T	2,577	A	Transversion	Synonymous	CTT (L) <-> CTA (L)
		13	2,739	T	2,736	C	Transition	Synonymous	GCT (A) <-> GCC (A)
Intergenic region	-	14	4,025	-	4,022	T	Insertion / Deletion	-	-
ORF3	Signal peptide	15	4,111	T	4,109	C	Transition	Non-synonymous	TTT (F) <-> CTT (L)
	No domain	16	4,377	G	4,375	A	Transition	Synonymous	TTG (L) <-> TTG (L)
		17	4,410	T	4,408	C	Transition	Synonymous	TAT (Y) <-> TAC (Y)
		18	4,467	G	4,465	A	Transition	Synonymous	ACG (T) <-> ACA (T)
		19	4,607	G	4,605	A	Transition	Non-synonymous	AGA (R) <-> AAA (K)
		20	5,083	A	5,081	G	Transition	Non-synonymous	AAA (K) <-> GAA (E)
		21	5,117	C	5,115	T	Transition	Non-synonymous	CCT (P) <-> CTT (L)
	MSP7_C	22	5,598	G	5,596	A	Transition	Synonymous	AAG (K) <-> AAA (K)
	No domain	23	5,803	G	5,801	A	Transition	Non-synonymous	GTT (V) <-> ATT (I)
		24	5,964	T	5,962	C	Transition	Synonymous	ATT (I) <-> ATC (I)
		25	6,072	C	6,070	T	Transition	Synonymous	AAC (N) <-> AAT (N)
ORF 4	Capsid	26	6,416	C	6,414	T	Transition	Synonymous	GCC (A) <-> GCT (A)

Appendix Table A6: Details of viruses and viral sequences used in ORF1 (helicase and RNA-dependant RNA polymerase) and ORF2-ORF4 (capsids) phylogenies.

Acronyme	Complete virus name	Viral Family	Viral Genus	Genbank Accession	Host
AMV	*Alfalfa mosaic virus*	Bromoviridae	*Alfamovirus*	NC_002024, NC_001495	Plant
ASV1	Adelphocoris suturalis-associated virus 1	Not assigned	Not assigned	KX966285	Hemiptera
BBNV	*Broad bean necrosis virus*	Virgaviridae	*Pomovirus*	D86636, D86637	Plant
	Blackford Virus	Not assigned	Not assigned	KU754514	Drosophilidae (Diptera)
BMV	*Brome mosaic virus*	Bromoviridae	*Bromovirus*	NC_002027, KU726253	Plant
BNRBV	*Blueberry necrotic ring blotch virus*	Not assigned	*Blunervirus*	NC_016084, NC_016085	Plant
BNYVV	*Beet necrotic yellow vein virus*	Benyviridae	*Benyvirus*	NC_003514	Plant
	Bofa Virus	Not assigned	Not assigned	KU754515	Drosophilidae (Diptera)
BotV-F	*Botrytis virus F*	Gammaflexiviridae (Tymovirales)	*Mycoflexivirus*	NC_004063	Fungi
	Boutonnet Virus	Not assigned	Not assigned	KU754539	Drosophilidae (Diptera)
	Brandeis Virus	Not assigned	Not assigned	(Webster *et al.*, 2015)	Drosophilidae (Diptera)
BSMV	*Barley stripe mosaic virus*	Virgaviridae	*Hordeivirus*	MBSRNAGT, MBSARNA, X03854	Plant
	Buckhurst Virus	Not assigned	Not assigned	KU754516	Drosophilidae (Diptera)
BVQ	*Beet virus Q*	Virgaviridae	*Pomovirus*	AJ223596, AJ223597	Plant
BYV	*Beet yellows virus*	Closteroviridae	*Closterovirus*	NC_001598	Plant
ChAV	*Chara australis virus*	Not assigned	Not assigned	JF824737	Plant
CiCLV	*Citrus leprosis virus C*	Not assigned	*Cilevirus*	DQ157466	Plant (mites transmitted)
CLBV	*Citrus leaf blotch virus*	Betaflexiviridae (Tymovirales)	*Citrivirus*	NC_003877	Plant
CMV	*Cucumber mosaic virus*	Bromoviridae	*Cucumovirus*	NC_002035, KC019299	Plant
CWMV	*Chinese wheat mosaic virus*	Virgaviridae	*Furovirus*	AJ012005, AJ012006	Plant
CYMV	*Clitoria yellow mottle virus*	Virgaviridae	*Tobamovirus*	NC_016519	Plant
DEZV	*Dezidougou virus*	Not assigned	Sandewavirus	JQ675604	Insect (Diptera-Culicidae)
GANV	*Goutanap virus*	Not assigned	Sandewavirus	KF588035	Insect (Diptera-Culicidae)

HGSV	*Hibiscus green spot virus*	Not assigned	*Higrevirus*	HQ852052	Plant
HLFPV	*Hibiscus latent Fort Pierce virus*	Virgaviridae	*Tobamovirus*	FJ196834, AY250831	Plant
IPCV	*Indian peanut clump virus*	Virgaviridae	*Pecluvirus*	AF447397	Plant
LoLV	*Lolium latent virus*	Alphaflexiviridae (Tymovirales)	*Lolavirus*	NC_010434	Plant
LORV	*Loreto virus*	Not assigned	Nelorpivirus	JQ675611	Insect (Diptera-Culicidae)
	Marsac Virus	Not assigned	Not assigned	KU754518	Drosophilidae (Diptera)
	Muthill Virus	Not assigned	Not assigned	KU754517	Drosophilidae (Diptera)
NWTV	*Ngewotan virus*	Not assigned	Nelorpivirus	JQ686833	Insect (Diptera-Culicidae)
NEGV	*Negev virus*	Not assigned	Nelorpivirus	JQ675608	Insect (Diptera-Culicidae)
ObPV	*Obuda pepper virus*	Virgaviridae	*Tobamovirus*	MTVGRNA	Plant
OLV-2	*Olive latent virus 2*	Bromoviridae	*Oleavirus*	NC_003674, NC_003673	Plant
PAMMV	*Paprika mild mottle virus*	Virgaviridae	*Tobamovirus*	AB089381	Plant
PCV	*Peanup clump virus*	Virgaviridae	*Pecluvirus*	X78602	Plant
PeBV	*Pea early browning virus*	Virgaviridae	*Tobravirus*	X14006, X15883	Plant
PIUR	*Piura virus*	Not assigned	Nelorpivirus	JQ675607	Insect (Diptera-Culicidae)
PZSV	*Pelargonium zonate spot virus*	Bromoviridae	*Anulavirus*	NC_003650, AJ272327	Plant
RBDV	*Raspberry bushy dwarf virus*	Not assigned	*Idaeovirus*	NC_003739	Plant
SANV	*Santana virus*	Not assigned	Sandewavirus	JQ675606	Insect (Diptera-Culicidae)
ScrCSV	*Sorghum chlorotic spot virus*	Virgaviridae	*Furovirus*	AB033691, AB033692	Plant
TANAV	*Tanay virus*	Not assigned	Sandewavirus	KF425261	Insect (Diptera-Culicidae)
	TSA *Argochrysis armilla*	Not assigned	Not assigned	GAXO01029871	Unannotated virus-like (Webster *et al.*, 2016)
	TSA *Bactrocera dorsalis*	Not assigned	Not assigned	GAKP01015888	Unannotated virus-like (Webster *et al.*, 2016)
	TSA *Ceratitis capitata*	Not assigned	Not assigned	GAMC01007262	Unannotated virus-like (Webster *et al.*, 2016)
	TSA *Cotesia vestalis*	Not assigned	Not assigned	GAKG01005025	Unannotated virus-like (Webster *et al.*, 2016)

	TSA *Latrodectus hesperus* venom	Not assigned	Not assigned	GBCS01005187	Unannotated virus-like (Webster *et al.*, 2016)
	TSA *Monomorium pharaonis*	Not assigned	Not assigned	LA822010	Unannotated virus-like (Webster *et al.*, 2016)
	TSA *Musca domestica*	Not assigned	Not assigned	GARN01041480	Unannotated virus-like (Webster *et al.*, 2016)
TMV	*Tobacco mosaic virus*	Virgaviridae	*Tobamovirus*	V01408	Plant
ToMV	*Tomato mosaic virus*	Virgaviridae	*Tobamovirus*	X02144	Plant
TRV	*Tobacco rattle virus*	Virgaviridae	*Tobravirus*	AF166084, AF034621	Plant
TSV	*Tobacco streak virus*	Bromoviridae	*Ilarvirus*	NC_003842, FJ561302	Plant
TVCV	*Turnip vein-clearinf virus*	Virgaviridae	*Tobamovirus*	BRU03387	Plant
TYMV	*Turnip yellow mosaic virus*	Tymoviridae (Tymovirales)	*Tymovirus*	NC_004063	Plant
WALV	*Wallerfield virus*	Not assigned	Sandewavirus	KF042857	Insect (Diptera-Culicidae)
ZGMMV	*Zucchini green mottle mosaic virus*	Virgaviridae	*Tobamovirus*	AJ295949	Plant

1. Webster CL, Waldron FM, Robertson S, Crowson D, Ferrari G, Quintana JF, Brouqui J-M, Bayne EH, Longdon B, Buck AH, Lazzaro BP, Akorli J, Haddrill PR, Obbard DJ. 2015. The discovery, distribution, and evolution of viruses associated with Drosophila melanogaster. PLoS Biol. 13:e1002210.

2. Webster C, Longdon B, Lewis S, Obbard D. 2016. Twenty-five new viruses associated with the Drosophilidae (Diptera). Evol. Bioinforma. 13.

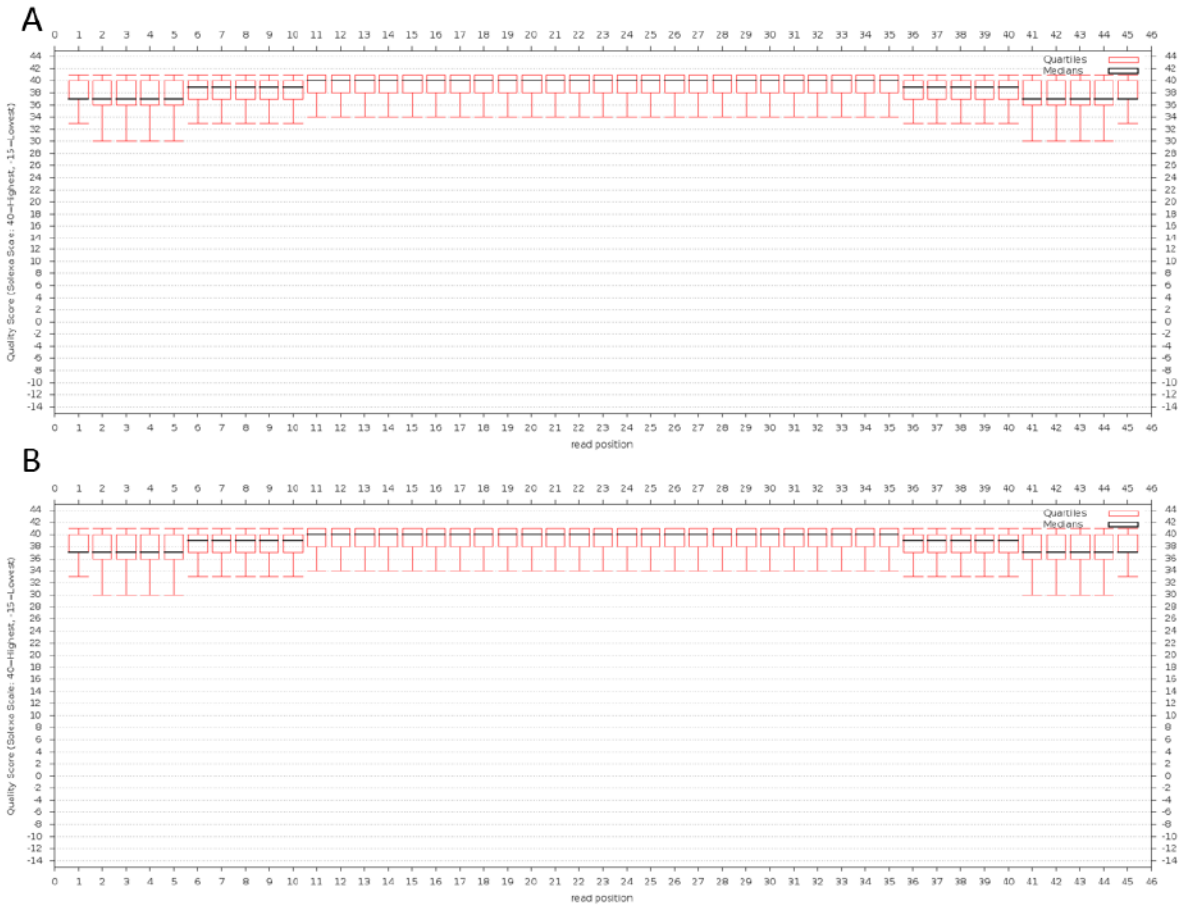


Figure A1: Distribution of quality scores using Solexa/Illumina scale along read position, for the reads mapped to the CpATV_Moknine (panel A) and CpATV_ElHabibia (panel B) genomes after nucleotide trimming of read ends.

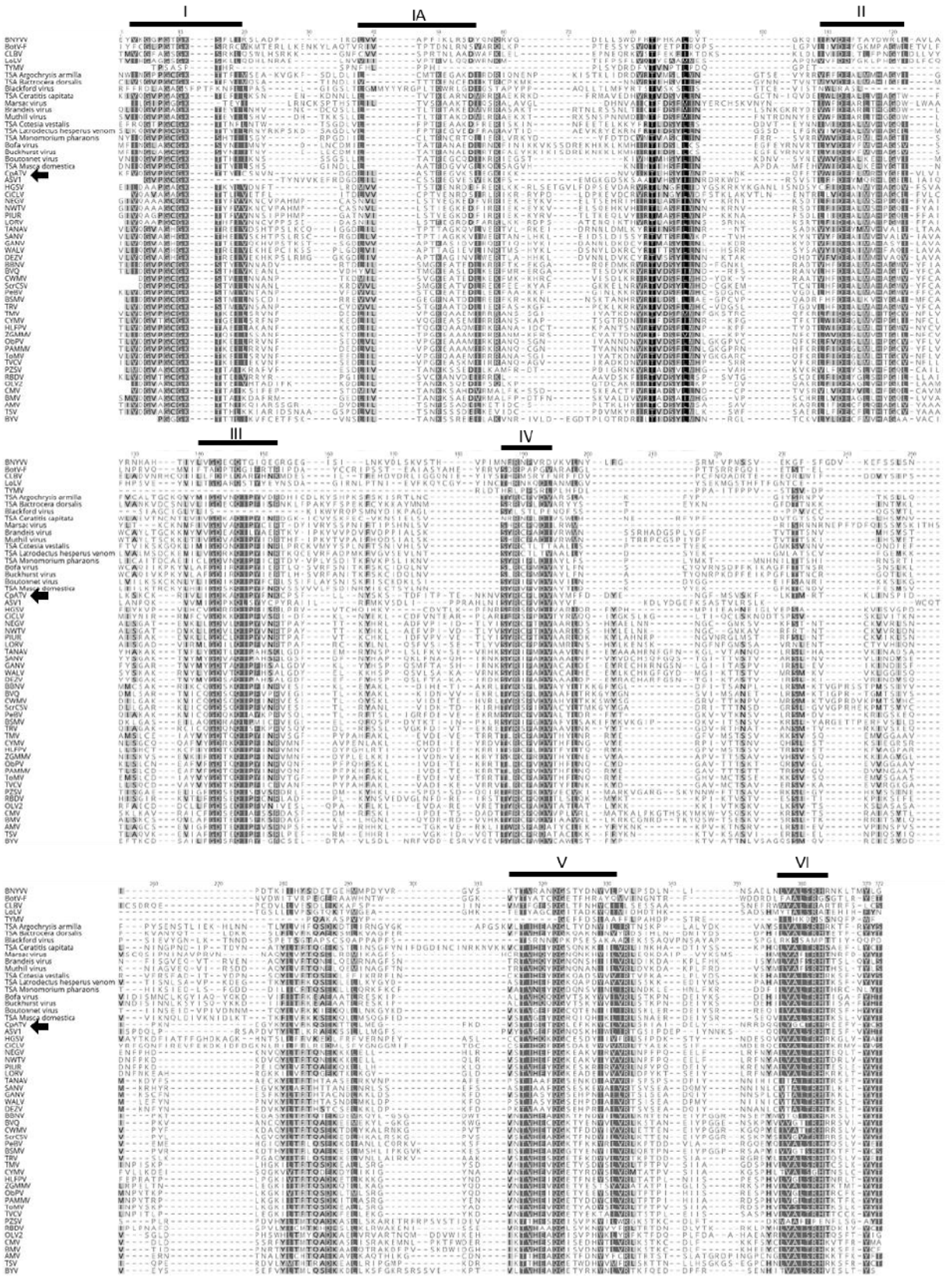


Figure A2: Alignment of conserved helicase domain of the genome of Culex pipiens Associated Tunisia Virus (344 aa). Taxon names are virus acronyms. Similarity in amino acids alignments is represented by shades of black (100% in black, 80% in dark gray and 60% in light gray). Arrows indicate sequences of CpATV. Conserved motifs I to VI as defined by Koonin and Dolja in 1993 are indicated above sequences (72). Sequences details and GenBank accessions numbers are available in Appendix Table A2.

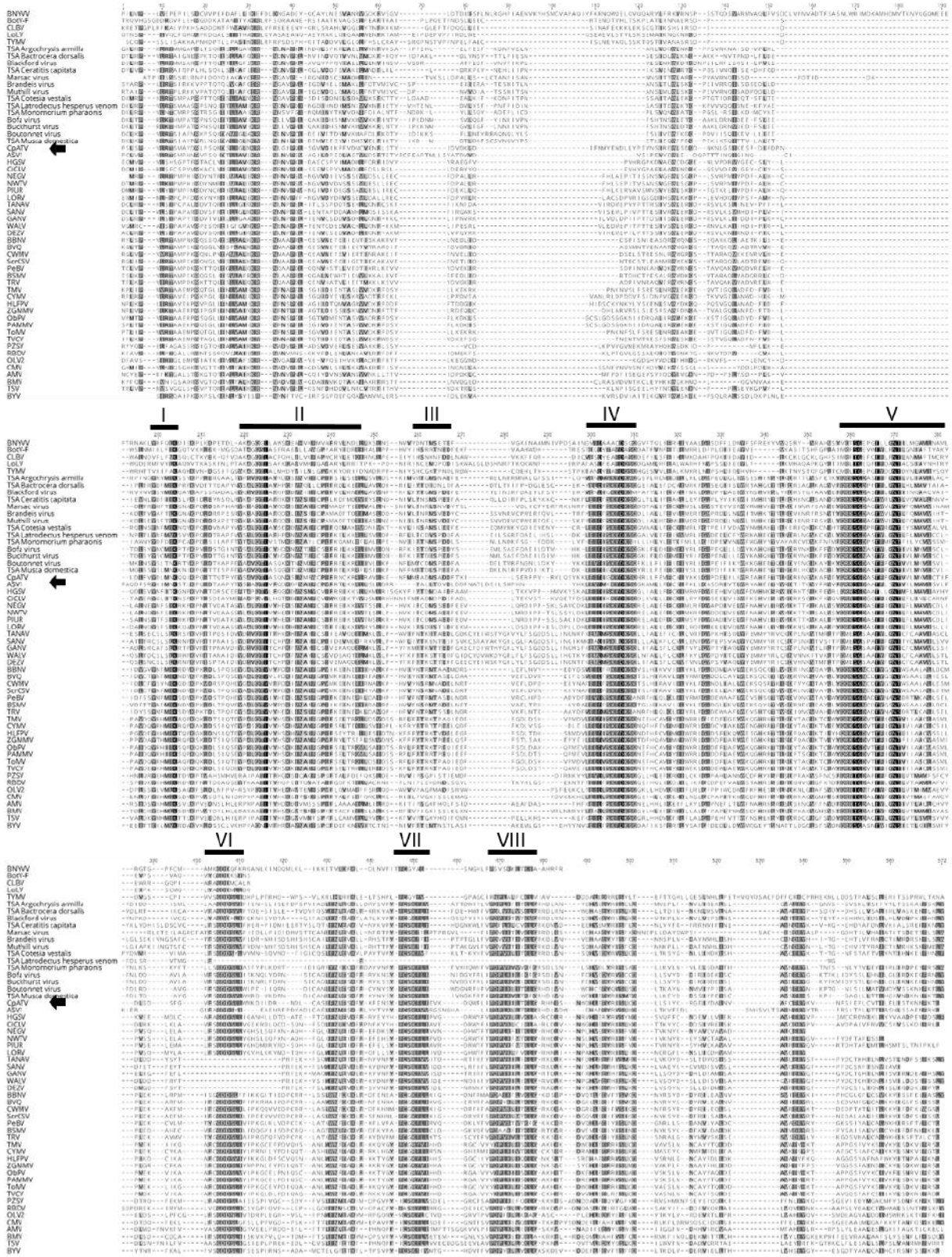


Figure A3: Alignment of conserved RNA-dependent RNA-polymerase domain of the genome of Culex pipiens Associated Tunisia Virus (597 aa). Taxon names are in acronyms. Similarity in amino acids alignments is represented by shades of black (100% in black, 80% in dark gray and 60% in light gray). Arrows indicate sequences of CpATV. Conserved motifs I to VI as defined by Koonin and Dolja in 1993 are indicated above sequences (72). Sequences details and GenBank accessions numbers are available in Appendix Table A2.

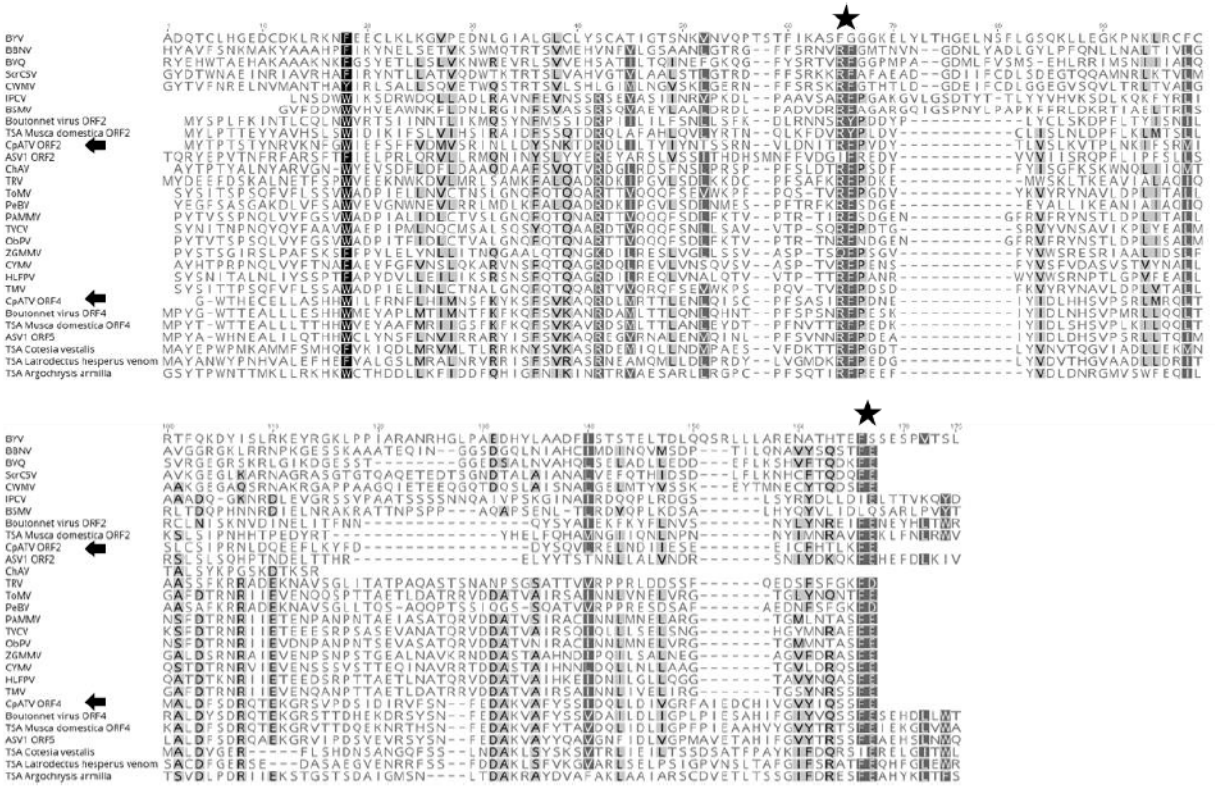


Figure A4: Alignment of conserved capsids domains of the genome of Culex pipiens Associated Tunisia Virus (218 aa). Taxon names are in acronyms. Similarity in amino acids alignments is represented by shades of black (100% in black, 80% in dark gray and 60% in light gray). Arrows indicate sequences of CpATV. Stars indicates the conserved residues as defined by Koonin and Dolja in 1993 which form a functionally important salt bridge (R(+) - E(-)) (72). Sequences details and GenBank accessions numbers are available in Appendix Table A2.

3.3. Etude d'une famille virale particulière, les *Parvoviridae* (**Article 2**)

Cette étude à laquelle j'ai participé dans le cadre d'une collaboration a été publiée en 2016 dans le journal *Scientific Report* (doi : 10.1038/srep30880). Cette étude est une analyse de la biodiversité des *Parvoviridae* (ssDNA). Ma participation à ce travail s'intègre pleinement dans le cadre de cette thèse. Après l'analyse d'une partie des transcriptomes d'animaux, toutes les séquences assignées par le programme HHBlits (Figure 15) comme appartenant à la famille des *Parvoviridae* (sous-familles Parvovirinae et Densovirinae) ont été récupérées et analysées lors dans cet article. Mon approche a permi d'identifier un total de 27 séquences provenant de 7 espèces (correspondant à 11 transcriptomes individuels), qui ont ainsi pu être ajoutées à cette étude plus vaste.

Les espèces ainsi incluses dans cette étude sont : un ver marin des profondeurs (Annelida, *Lamellibrachia* sp.), une crevette d'eau douce (Arthropoda, Brachiopoda, *Artemia franciscana)*, une crepidule (Mollusca, Gastropoda, *Crepidula fornicata*,), une gorgone (Cnidaria, Anthozoa, *Eunicella cavolinii*), un moustique commun (Arthropoda, Diptera, *Culex pipiens*) et deux espèces de fourmis moissonneuse (Arthropoda, Insecta, Hymenoptera, *Messor barbarus* et *Messor concolor*). A l'exception des fourmis, les espèces énoncées ici sont les seules représentantes de la famille animale échantillonnée (voire de l'ordre, de la classe et même du phylum) au sein des bases de données génomique, ce qui apporte des informations nouvelles sur la présence des séquences de parvovirus dans des animaux non-modèles. Les 7 espèces, ainsi que les 27 séquences incluses dans cette étude par mon analyse représentent près de 6 % des données de *Parvoviridae* ici décrite dans de nouveaux hôtes animaux.

Cette étude a permi de découvrir que cette famille virale est finalement très répandue dans le règne animal (François *et al.*, 2016).

www.nature.com/scientificreports

SCIENTIFIC REPORTS

OPEN

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

r eceived: 17 May 2016
a ccepted: 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 f ndings 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 dif erent species. T e 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. T e 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	8	18
Arthropoda	Arachnida	Araneae	Theridiidae	**Latrodectus hesperus**	8		8
		Trombidiformes	Tetranychidae	**Tetranychus urticae**	1		1
		Ixodida	Ixodidae	**Amblyomma americanum**	4		4
				Ixodes scapularis [23]	2		2
				Rhipicephalus appendiculatus [11]	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* [40]	1	2	3
			Diopsidae	**Teleopsis dalmanni**	1		1
			Drosophilidae	*Drosophila persimilis* [11*]	2		2
				Drosophila sechellia [11*]	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* [11]	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* [24]	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	**Micropterix calthella**	2	2	4
			Noctuidae	**Agrotis segetum**		1	1
				Helicoverpa armigera	1		1
			Nymphalidae	*Bicyclus anynana* [11]	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* [11]	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	Astacidae	**Pontastacus leptodactylus**	1		1
			Palaemonidae	**Macrobrachium nipponense**		1	1
			Penaeidae	*Penaeus monodon* [34*]	2		2
			Portunidae	**Scylla olivacea**	2		2
		Isopoda	Armadillidiidae	*Armadillidium nasatum* [34*]	3		3
				Armadillidium vulgare [34*]	2		2
	Maxillopoda	Arguloida	Argulidae	**Argulus siamensis**	4	4	8
		Siphonostomatoida	Caligidae	**Caligus rogercresseyi**	2		2
				Lepeophtheirus salmonis [11]	3		3
Chordata	Actinopterygii	Cyprinodontiformes	Fundulidae	**Fundulus grandis**	2		2
	Aves	Galliformes	Phasianidae	*Gallus gallus* [11]	1		1
	Mammalia	Artiodactyla	Bovidae	*Bos taurus* [11]		2	2
				Capra hircus		2	2
			Suidae	*Sus scrofa* [11]	1		1
		Diprotodontia	Macropodidae	*Macropus eugenii* [33]		1	1
			Phalangeridae	*Trichosurus vulpecula* [11]	1	1	2
		Lagomorpha	Leporidae	*Oryctolagus cuniculus* [34*]	4	3	7
		Primates	Hominidae	*Homo sapiens* [11]		1	1
		Rodentia	Chinchillidae	*Chinchilla lanigera* [33*]	4		4
			Muridae	*Rattus norvegicus* [33*]	5	4	9
			Octodontidae	*Octodon degus* [33*]	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 diminuta**	14		14
				Hymenolepis microstoma [34]	30		30
				Hymenolepis nana	2		2
			Taeniidae	*Echinococcus granulosus* [34]	3		3
				Echinococcus multilocularis	35		35
				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* [34]	3		3
				Schistosoma margrebowiei	1		1
				Schistosoma mattheei	1		1
	Turbellaria	Tricladida	Dugesiidae	**Schmidtea mediterranea**	1		1
Urochordata	Ascidiacea	Enterogona	Cionidae	*Ciona intestinalis* [11]		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 copticus*	1	NA
	Polychaeta	Canalipalpata	Sthoglinidae	*Lanellibrachia sp.*	16	NA
Arthropoda	Insecta	Araneae	Theridiidae	*Latrodectus hesperus*	A	1
		Trombidiformes	Tetranychidae	*Tetranychus urticae*	A	1
	Arachnida	Ixodida	Ixodidae	*Amblyomma americanum*	4	NA
				Ixodes scapularis [30]	1	1
				Rhipicephalus appendiculatus [31]	2	NA
				Haemaphysalis longicornis [32]	5	1
				Rhipicephalus pulchellus	2	NA
		Mesostigmata	Phytoseiidae	*Metaseiulus occidentalis*	2	1
	Branchiopoda	Anostraca	Artemiidae	*Artemia franciscana*	1	NA
		Diplostraca	Daphniidae	*Daphnia pulex*	A	1
		Collembola	Entomobryidae	*Pogonognathellus flavescens*	2	NA
		Coleoptera	Cerambycidae	*Prionus laticollis*	1	A
			Cupedidae	*Ipe hypogaeus*	1	NA
			Curculionidae	*Pissodes strobi*	2	NA
				Meligethes aeneus	1	NA
			Nitidulidae	*Onthophagus nigrivantris*	1	NA
			Scarabaeidae	*Onthophagus sinicus*	2	NA
		Diptera	Culicidae	*Culex pipiens* [26]	3	NA
			Ekpsidae	*Tokeyusu shirmanni*	1	A
			Drosophilidae	*Drosophila persimilis* [27]*	NA	2
				Drosophila sechellia [28]*	4	1
			Syrphidae	*Eristalis tenax*	1	NA
			Tephritidae	*Bactrocera dorsalis*	2	1
				Rhagoletis pomonella	1	NA
		Hemiptera	Aleyrodidae	*Bemisia tabaci*	7	NA
			Aphididae	*Acyrthosiphon pisum* [25]	2	29
				Sitobion avenae	1	NA
			Cicadellidae	*Graminella nigrifrons*	1	NA
			Coreidae	*Anasa tristis*	1	NA
			Coreidae	*Clavigralla tomentosicollis*	1	NA
			Delphacidae	*Nilaparvata lugens*	8	1
			Kerriidae	*Kerria lacca*	5	NA
			Miridae	*Lygus hesperus*	13	NA
			Pentatomidae	*Chinavia ubica*	1	NA
				Halyomorpha halys	2	3
			Psyllidae	*Diaphorina citri*	A	6
				Pachypsylla venusta	3	2
			Reduviidae	*Rhodnius prolixus* [29]	A	5
				Triatoma infestans	2	NA
		Hymenoptera	Apidae	*Bombus impatiens*	1	3
			Formicidae	*Acromyrmex echinatior*	NA	5
				Atta cephalotes	NA	6
				Lasius niger	NA	5
				Monomorium pharaonis	4	NA
				Pogonomyrmex barbatus	1	NA
				Messor barbarus	A	6
				Messor concolor	NA	4
				Tetramorium bicarinatum	1	NA
		Lepidoptera	Crambidae	*Chilo suppressalis*	5	8
			Micropterigidae	*Micropteryx calthella*	4	NA
			Noctuidae	*Agrotis segetum*	1	NA
				Helicoverpa armigera	1	NA
			Nymphalidae	*Bicyclus anynana* [33]	1	NA
			Thaumetopoeidae	*Thaumetopoea pityocampa*	4	NA
		Mecoptera	Nannochoristidae	*Nannochorista philpotti*	2	NA
		Megaloptera	Corydalidae	*Corydalinae sp.*	2	NA
		Neuroptera	Chrysopidae	*Chrysopa pallens*	1	NA
		Orthoptera	Acrididae	*Schistocerca gregaria*	3	NA
			Gryllidae	*Gryllus bimaculatus* [34]	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 ornata*	1	NA
	Malacostraca	Amphipoda	Ampeliscidae	*Ampelisca abdita*	5	NA
		Decapoda	Astacidae	*Pontastacus leptodactylus*	1	NA
			Palaemonidae	*Macrobrachium nipponense*	1	NA
			Penaeidae	*Penaeus monodon* [35]*	A	2
			Portunidae	*Scylla olivacea*	2	NA
		Isopoda	Armadillidiidae	*Armadillidium nasatum* [36]*	NA	2
				Armadillidium vulgare [36]*	A	2
	Maxillopoda	Arguloida	Argulidae	*Argulus siamensis*	8	NA
		Siphonostomatoida	Caligidae	*Caligus rogercresseyi*	2	1
				Lepeophtheirus salmonis [37]	A	1
Chordata	Actinopterygii	Cyprinodontiformes	Fundulidae	*Fundulus grandis*	2	NA
	Aves	Galliformes	Phasianidae	*Gallus gallus* [17]	1	A
	Mammalia	Artiodactyla	Bovidae	*Bos taurus* [38]	2	A
				Capra hircus	2	A
			Suidae	*Sus scrofa* [39]	1	A
		Diprotodontia	Macropodidae	*Macropus eugenii* [16]	1	12
			Phalangeridae	*Trichosurus vulpecula* [40]	2	NA
		Lagomorpha	Leporidae	*Oryctolagus cuniculus* [41]*	5	1
		Primates	Hominidae	*Homo sapiens* [24]	1	A
		Rodentia	Chinchillidae	*Chinchilla lanigera* [42]*	4	1
			Muridae	*Rattus norvegicus* [43]*	9	42
			Octodontidae	*Octodon degus* [44]*	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	1
				Octopus vulgaris	1	NA
		Sepiida	Sepiidae	*Sepia officinalis*	1	NA
		Sepiolida	Sepiolidae	*Euprymna scolopes*	4	NA
	Gastropoda	s	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 [45]	NA	11
				Hymenolepis nana	NA	4
			Taeniidae	*Echinococcus granulosus* [46]	A	1
				Echinococcus multilocularis	NA	69
				Hydatigera taeniaeformis	NA	2
				Taenia asiatica	NA	2
				Taenia multiceps	2	NA
	Monogenea	Monopisthocotylea	Capsalidae	*Neobenedenia melleni*	1	NA
	Trematoda	Echinostomida	Echinostomatidae	*Echinostoma caproni*	NA	10
		Opisthorchiida	Opisthorchiidae	*Clonorchis sinensis* [47]	A	16
				Opisthorchis felineus	9	NA
				Opisthorchis viverrini	A	18
		Plagiorchiida	Dicrocoeliidae	*Dicrocoelium dendriticum*	NA	6
		Strigeatida	Schistosomatidae	*Schistosoma mansoni* [48]	A	8
				Schistosoma margrebowiei	NA	2
				Schistosoma mattheei	NA	2
	Turbellaria	Tricladida	Dugesiidae	*Schmidtea mediterranea*	1	3
Urochordata	Ascidiacea	Enterogona	Cionidae	*Ciona intestinalis* [49]	1	23
				Number of PRSs	247	375
				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.

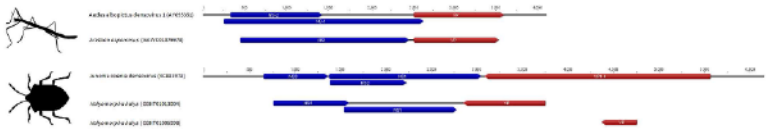


Figure 4. Organization of open reading frames of three PRSs (GAWC01079978 from *Aretaon asperrimus*, GBHT01013004 and GBHT01005998 from *Halyomorpha halys*). Arrowhead boxes indicate viral and predicted viral genes (NS are in blue and VP in red).

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.

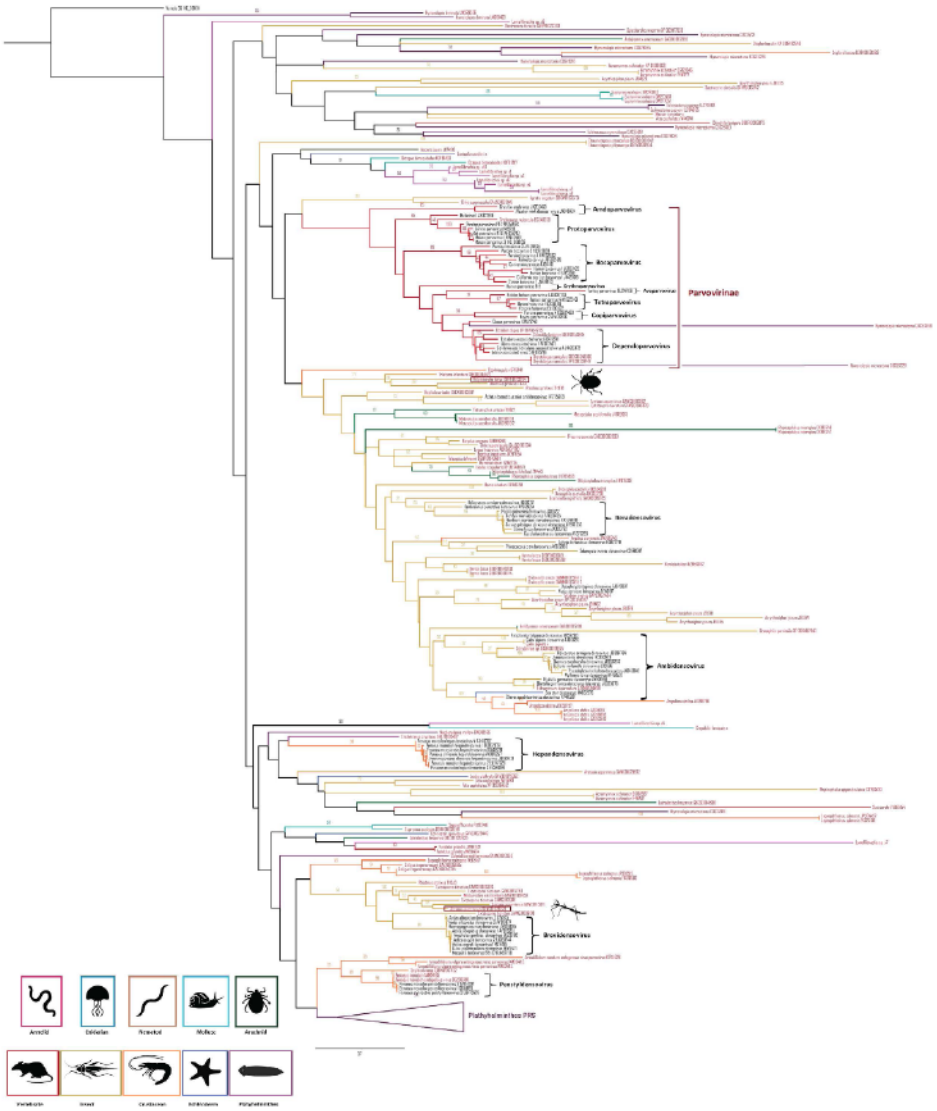


Figure 5. Maximum likelihood phylogenetic tree based on partial SF3 domains of the NS1 protein, including 74 parvovirus species (in black) and 264 PRSs (in red). The alignment was produced using MUSCLE 3.7 with default settings. The tree was rooted with the SF3 domain of the variola virus D5 protein. 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.E., 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).

3.4. Découverte d'un rétrovirus chez la Salamandre tachetée (**Article 3**)

L'objet de ce travail, ici écrit sous format article en cours de préparation, montre la présence de séquences de rétrovirus *Spumavirus* endogènes chez les urodèles et plus particulièrement la salamandre tachetée *Salamandra salamandra* et le triton de Waltl *Pleurodeles waltl*.

Une partie de ce travail, la partie bioinformatique sur la détection des rétrovirus, a été réalisée lors du stage de Master 2 de Jean-Philippe Vernadet en 2016 que j'ai co-encadré.

First foamy-like endogenous retroviruses discovery in early-diverging vertebrates: evidence from Caudata genome and transcriptomes.

Diane Bigot[1], Jean-Philippe Vernadet[1], Marc Sitbon[2], Jean-Luc Battini[2], Valérie Courgnaud[2], Gilles Labesse[3], Elisabeth A. Herniou[1] and Philippe Gayral[1]

[1] Institut de Recherche sur la Biologie de l'Insecte, UMR 7261, CNRS, Université François-Rabelais, 37200 Tours, France ; [2] Institut de Génétique Moléculaire de Montpellier, IGMM-CNRS, Université de Montpellier, 34293 Montpellier Cedex 5, France ; [3] Centre de Biochimie Structurale, INSERM U1054 - CNRS UMR5048 - UM1, 34090 MONTPELLIER Cedex, France

Abstract

Retroviruses are among the most studied virus. However, our knowledge of their diversity, their origin and evolution are still incomplete. The objective of this study is the research of the presence of retroviral sequences from transcriptomes of non-models animals. First, open readings frames were predicted in contigs assembled from 2 Caudata available transcriptomes, the fire salamander *Salamandra salamandra* and the Waltl newt *Pleurodeles waltl*. Second, homology search against a viral protein database was conducted using BLASTP tool, and reciprocal BLAST search against a general protein database was performed. Finally, viral genomes were reconstructed and annotated. Phylogenetic analyses were used to place the new sequences in a retrovirus phylogeny. This work allowed the reconstruction of a new complete spumavirus genome in the fire salamander, and a second partial viral genome in the Waltl newt. Broader homology searches in other organisms allowed the discovery of spumavirus sequences in the transcriptome of the Hokkaido salamander (*Hynobius retardatus*) and in the genome of the axolotl (*Ambystoma mexicanum*). Our phylogenetic analysis showed that these four retroviral sequences clearly belonged to Spumavirus and formed a monophyletic group. These sequences are distinct from other known amphibian endogenous virus. This work led to the discovery of a new endogenised retroviral lineage, confirmed by molecular detection in fire salamander, which seemed specific to Caudata.

Key-words: Retroviruses, Non-model animals, Transcriptomes, Phylogeny, Spumavirus, Caudata

Introduction

Paleovirology is an emerging field, supported by deep sequencing technologies of complete genomes (Patel et al., 2011), allowing discoveries of ancient interactions of vertebrates and viruses (Katzourakis & Gifford, 2010).

Here we show that transcriptomes can also be a good genetic resource for paleovirology; although it cannot lead to an exhaustive view of the genomic content of endogenous viruses, its affordability has led to numerous transcriptomes of non-model animals available, and furthermore any endogenous virus discovered in transcriptome is potentially expressed, hence possibly domesticated by the host (Feschotte & Gilbert, 2012).

The *Retroviridae* family comprises seven extant genera belonging to two subfamilies *Orthoretrovirinae* (*Alpharetrovirus*, *Betaretrovirus*, *Gammaretrovirus*, *Deltaretrovirus*, *Epsilonretrovirus* and *Lentivirus*) and *Spumaretrovirinae* (*Spumavirus*) (King et al., 2012). Their ability to integrate their genome as an obligatory part of their replication cycle in the germline has conducted to a spread of endogenous retroviruses (ERV) separated into three classes: ERV Class I related to Gamma- and Epsilonretroviruses, Class II to Aplha- and Betaretroviruses and Class III to Spumaretroviruses (Gifford & Tristem, 2003).

Known infectious foamy viruses (Spumavirus, FVs) are characterized by a replication strategy which differs from other retroviruses (Delelis et al., 2004), a lesser long-term pathogenicity than others retroviruses (Linial, 2000) and until recentlty were thought to be restricted in placental mammal hosts only: in primates (Falcone et al., 1999), Cetartiodactyla: horse and sheep (Materniak et al., 2013; Tobaly-Tapiero et al., 2000), and in Pilosa: sloth as an endogenous Foamy virus (EFV) (Katzourakis et al., 2009).

However, recent work on vertebrate genome content has shown the presence of a foamy-like endogenous retrovirus (FV-like ERV) outside mammals in two distant organisms: in the genome of the lobe-finned fish coelacanth (Han & Worobey, 2012), in ray-finned fishes and in sharks where 23 novel FV-like ERV were discovered (Ruboyianes & Worobey, 2016) and in amphibians (Aiewsakun & Katzourakis, 2017). Foamy viruses possessed the older origin among retroviruses which could have a marine origin >450 million years ago (Aiewsakun & Katzourakis, 2017).

Here we report the detection of FV-like elements in salamandra by database searches, by the sequencing and screen of caudate transcriptomes and PCR verification. Genome annotation and phylogenetic analyses allowed to complete our knowledges on amphibian spumaviruses.

Materials and Methods

Transcriptome data of Salamandridae

59.95 million single-end 100pb Illumina reads of the fire salamander (*Salamandra salamandra*) were retrieved from the SRA database (SRX793983) (Figuet et al., 2015). 389 million single-end 100pb Illumina reads of the shatp-ribbed newt (*Pleurodeles waltl*) (Jean-Yves Sire, pers. comm. ANR Jaws) were also used in this study. De novo transcriptomes were assembled using a strategy validated in previous works (Cahais et al., 2012; Figuet et al., 2015). A first assembly with ABYSS V 1.2.0 (Birol et al., 2009; Simpson et al., 2009) with Kmer set at 60 was followed by contig re-assembly with CAP3 (Huang & Madan, 1999).

Virus detection in Salamandridae transcriptomes

Open reading Frames (ORFs) were first predicted in the contigs assembled from *Salamandra salamandra* and *Pleurodeles waltl* using Prodigal V2_60 software for metagenomic data (Hyatt et al., 2010, 2012). Complete or fragmented ORFs (5'-end and/or 3'-end missing) were predicted this way. ORFs displaying internal undetermined nucleotides (several Ns) were kept for further analyses.

Viral homology was detected using a reciprocal Best BLAST hit approach, chosen to decrease the amount false positive hits detected and to optimize computation time. Homology searches on putative ORFs were first performed using BLASTp algorithm of BLAST+ program version 2.2.29 (Altschul et al., 1990; Camacho et al., 2009) using an e-value threshold of 0.1 against a custom database of viral proteins. The viral protein database was built from the 3,308,095 proteins of the NCBI Genbank database (nr version 20/01/2016) assigned to virus organism. The reciprocal BLAST was performed using BLASTp, an e-value threshold of 0.001 against the entire NCBI protein database (nr version 20/01/2016). For each ORF, hits with the best BLAST score were kept and were used to transfer their taxonomic assignation using NCBI TaxID and NCBI Taxonomy database (http://www.ncbi.nlm.nih.gov/taxonomy/). ORF assigned in the *Retroviridae* family were kept and further analyzed.

Construction and annotation of retroviral genomes

In order to reconstruct the most complete possible viral genomes from fragmented sequences, the contigs were aligned by the MAFFT v1.2 (Katoh et al., 2002) software with the default parameters, visualized by the Geneious® software (www.geneious.com, v8), and manually corrected. In the best case, several sequences of the same viral genus were found in the same individual, and it was possible to reconstruct viral genomes by aligning the predicted ORFs to the nearest known exogenous viral genome and replaced by “N” when the nucleotides of the areas missing. Consensus sequences were also performed when the sequences carrying the predicted ORFs were overlapping but too short to be analysed on their own. Phylogenetic reconstruction requires long enough sequences to reconstruct the evolutionary history of the retroviral sequence.

Retroviral sequences were then lengthened by mapping using the "Geneious Mapper" program and the "allow gap" parameters, "Maximum Mistakes per Read" at 10%. This first parameter does not allow gaps in the aligned sequences and allows a good quality assembly of the Illumina readings; the second parameter rejects reads if there is more than 10% mismatch between Illumina reads and the reference sequence. The use of these two parameters ensures a sufficiently high stringency to keep all the relevance of the mapping and to avoid the alignment of non-homologous Illumina reads. For each individual transcriptome, sequencing reads were mapped to the concatenated consensus reference obtained in the previous step. This mapping allows 1) to confirm the initial assemblies leading to the starting contigs, 2) to lengthen the extremities of the consensus, 3) to fill the missing regions (nucleotides N added in the previous step), 4) to confirm the absence of chimeric contigs by visualizing the homogeneity, the quality of the alignments of the reads, to validating the real coverage parameter along the reference and the mean coverage.

Protein annotation of the constructed retroviral genomes was carried out by searching for protein domains preserved using the InterproScan v 5.19 software with the default parameters (Jones et al., 2014) which uses the InterPro v58 database composed of signatures of 19788 families of proteins and 8439 protein domains. Prediction of transmembrane domain in envelope was performed using PhDHTM Transmembrane helices prediction (Rost, 1996).

Detection of FV-like sequence in other Amphibian genomes and transcriptomes

A deeper screen of FV-like sequence was performed in other amphibians with the *Salamandra* sequence previously obtained as query using BLAST homology searches with e-value thresholds set at 10^{-5}. TBLASTN searches were performed on transcriptomes of the three amphibian species available; the Hokkaido salamander (*Hynobius retardatus*, Japan, accession number: LE271194), the Chinese salamander (*Hynobius chinensis*) and the Chinese giant salamander (*Andrias davidianus*), through Transcriptome Shotgun Assembly (TSA) database.

Similar TBLASTN searches were performed on genome of three other amphibians, the Axolotl (*Ambystoma mexicanum*) and the Tibetan frog *Nanorana parkeri* through the Whole-Genome Shotgun contigs database (WGS) and the western clawed frog *Xenopus tropicalis* through the frog genomic database (http://www.xenbase.org/entry/).

Phylogenetic analyzes

Polymerase sequences of the new FV-like discovered in Amphibians were included in phylogenies with representative species of extant and endogenous retroviruses and related endogenous retroelements. ClustalW v1.81 (Thompson et al., 1994), MUSCLE v3.8.31 (Edgar, 2004) and MAFFT v7.273 (Katoh et al., 2002) programs were compared and the latter used with -- **auto** et -- **leavegappyregion** options showed the best sequence identity scores, as previously assessed (Pais et al., 2014). Alignments were cleaned by removing highly diverging sites which could contain substitution saturation masking the phylogenetic signal, using Gblocks v0.91b and low stringency parameters (Castresana, 2000).

Best amino acid substitution models were assessed by a Maximum Likelihood approach with ProtTest v3.4 (Abascal et al., 2005) using the following parameter sets: -S 2, -s BEST, -AIC, -AICC and -all-distributions. Maximum Likelihood phylogenies were constructed using the PhyML v3.1 program (Guindon & Gascuel, 2003) implemented in the Seaview v4.6 software (Gouy et al., 2010), and aLRT statistics were used to assess node supports (Anisimova & Gascuel, 2006).

Biological confirmation of FV-like ERV in *Salamandra*

Two specimens of palmate newt *Lissotriton helveticus* (TP1 and TP2; GPS N 47.3407; E 0.68919) and two fire salamander *Salamandra salamandra* (ST1 and ST2; GPS N 47.357256; E 0.701575) were sampled in Tours, France in 2016. This sampling of these protected species was authorized by the 'Direction Départementale des Territoires d'Indre-et-Loire'.

Non-invasive genomic DNA isolation was performed on ventral epithelial cells using HydraFlock® Flocked Swab (Puritan®) as described in (Pichlmüller et al., 2013). The Qiamp DNA isolation kit (Qiagen) and the DNA purification from buccal swabs (spin protocol) following the manufacturer's instructions were used. DNA was eluted in 150 µL AE Buffer. DNA integrity was assessed by a 0.8 % agarose gel electrophoresis and UV visualization using GelRed® staining (Interchim) and GelDoc® XR+ System (BioRad). DNA quantity was assessed using Qubit® Fluorimeter (Invitrogen) and the High Sensitivity (HS) dsDNA kit. *Salamandra* quantity ST1 was 0.22 µg and ST2 1.67 µg and *Lissotriton* TP1 was 0.066 µg and TP2 0.052 µg.

PCR amplifications of Spumavirus were only done on genomic DNA of ST2 sample, the DNA extraction resulting in the higher yield and better DNA quality. Amplifications of the polymerase gene and putative junction between the two concatenated genomes were chosen to detect and confirm *Spumavirus* presence in fire salamander.

The PCR mix of the polymerase amplification contained 0.125 µL of ExTaq Polymerase (5U/µL) (Takara®), 2.5 µL buffer 10x, 2 µL dNTP (2,5 mM each), 0.5 µL of each forward and reverse primers (PolSF and PolSR, 20mM) (Table 1), and 5 µL of cDNA (55ng), completed by water for a total volume 25 µL. The amplification cycle was: 3 min at 94°C, 35 cycles [30 sec at 94°C, 30 sec at 55°C, 120 sec at 72°C], 3 min at 72°C. PCR products were visualized in 1.5 % agarose gels stained with GelRed under UV light after 40 min migration at 100 V.

The PCR mix of the junction amplifications contained identical components as above but used only 2 µL of cDNA (22 ng) (Table 1 for primers). The amplification cycle for junctions 1-4 was 3 min at 94°C, 35 cycles [30 sec at 94°C, 30 sec at 55°C, 60 sec at 72°C], 3 min at 72°C. The amplification cycle for junctions 5-8 was 3 min at 94°C, 40 cycles [20 sec at 94°C, 20 sec at 55°C, 130 sec at 72°C], 3 min at 72°C. PCR products were visualized in 1.5 % agarose gels stained with GelRed under UV light after 40 min migration at 100 V.
Sanger sequencing of PCR product were done using forward and reverse primers by GATC Biotech (Germany) for PCRs PolS and J4F-J4R.

Table 1: PCR primers used for *Spumavirus* detection in genomic DNA of *Salamandra salamandra*.

PCR	Target	Forward primer	Reverse primer	Expected product size (nt)
PolS	Polymerase	PolS_F_2759: GTGTGCTCTTCTCCATCCTCC	PolS_R_4654 : AGGTGATGAATCCATGAGCGT	1,893
1	Junction	J1F_7762: TCAGCACGTGAAACCTACGA	J2R_8888: AAGGCCCAAAGAGTTCAGCT	1,127
2	Junction	J3F_8594: TCTTTTGATTCGGCCTAGGGA	J3R_9587: TCAGAAGAGACAAGCTAGGCC	994
3	Junction	J4F_9250: CTCTCGTTAGACACCGAGCG	J4R_10214: GCTTGGACAGCTGCCAGATA	965
4	Junction	J5F_9919: TGGTCTAAGCTGAGCACGTG	J1R_10870: GGACATACTGTGCTGGTGCT	952
5	Junction	J1F_7762: TCAGCACGTGAAACCTACGA	J4R_10214: GCTTGGACAGCTGCCAGATA	2,453
6	Junction	J1F_7762: TCAGCACGTGAAACCTACGA	J3R_9587: TCAGAAGAGACAAGCTAGGCC	1,825
7	Junction	J3F_8594: TCTTTTGATTCGGCCTAGGGA	J4R_10214: GCTTGGACAGCTGCCAGATA	1,621
8	Junction	J3F_8594: TCTTTTGATTCGGCCTAGGGA	J1R_10870: GGACATACTGTGCTGGTGCT	2,276

Results

Viral sequences detection in Amphibian

Nearly 789 thousand contigs were analysed from amphibian transcriptomes. Prodigal software predicted about 287 thousand Open Reading Frames. The results of the first homology investigation, against a viral database, indicated that 6.02% of ORFs have the best Blast hit as a viral sequence, representing nearly 17.3 thousand contigs. The result of the reciprocal BLAST, equivalent to the second homology search and using the NCBI complete non-redundant protein database, confirmed the viral taxonomy assignment for 16,700 ORFs.

Finally, of the sequences having passed all the previous detector filters, slightly less than 0.3% (52 sequences) corresponded to homologs of retroviral protein. A more detailed study of each of these sequences then provided us with information to validate and interpret their presence in animal host species.

In all 52 sequences, as a first approximation, 3 have homology with Gag protein, 42 with Pol protein, 5 with Env protein and 2 have homology with accessory proteins. Among all these sequences, 1 has homology with a Env of *Betaretrovirus*, 1 with a Pol of *Epsilonretrovirus*, 1

with a Gag of *Gammaretrovirus*, 41 with *Spumavirus* (2 Gag, 34 Pol, 4 Env and 1 accessory) and finally 8 with unclassified retroviruses (7 Pol and 1 accessory). Among *Spumaretrovirus* sequences, 33 came from the *Salamandra salamandra* transcriptome and the remaining 8 from the *Pleurodeles waltl* transcriptome. The results of association between retroviral genus and animal host were examined to detect if among all these sequences there were undescribed host-virus association. Epsilon and Gammaretroviruses were previously found in some other Amphibians (Kambol et al., 2003; Sinzelle et al., 2011). Recently, spumaviruses were found in association with amphibians (Aiewsakun & Katzourakis, 2017) and as a high number of *Spumavirus* sequences were found in amphibian transcriptomes we chose to focus on these viruses.

Reconstruction and genomic organisation of a Spumavirus in *Salamandra salamandra*

Mapping on a 2,820 nt consensus made from 5 aligned contigs found in *S. salamandra* allowed to reconstruct a 14,850 nt contig containing 1.6 X the sequence of the genome of salamander *Spumavirus* (Figure 1A). The genome resulting could correspond more precisely to the succession of two genomes repeated in tandem, as evidenced by a dot plot (Figure 1B). This repetition is not identical; however, the homologous area between the two genomes was 4,199 nucleotides with a nucleotide identity of 78.2%. It is however not sure that the 2 genomes are direct repeats, as a chimeric assembly cannot be excluded, even though there is no evidence of misassemblies in the read alignments produced from the mapping step.

The first genome is nearly complete according to the organization of known exogenous Spumaviruses (King et al., 2012), and its coverage is 120X. Indeed, the first three ORFs corresponded to the genes GAG, POL and ENV, characteristic of the retroviruses, with a length of 1,566, 3,489 and 3,024 nucleotides respectively (Figure 1A). Annotation step by InterProScan and Cd-Search allowed to detect a capsid domain (CA) in the gag protein (E-value=4.49.E-25), a reverse transcriptase domain (RT), an RNAse-H domain (RH) and an integrase (IN) domain in the pol protein (E-value=3.14.E-36, 3.35.10E-3 and 2.E-3 respectively) and an envelope domain (TM) in the env protein (E-value=2.28.E-11) (Figure 1A).

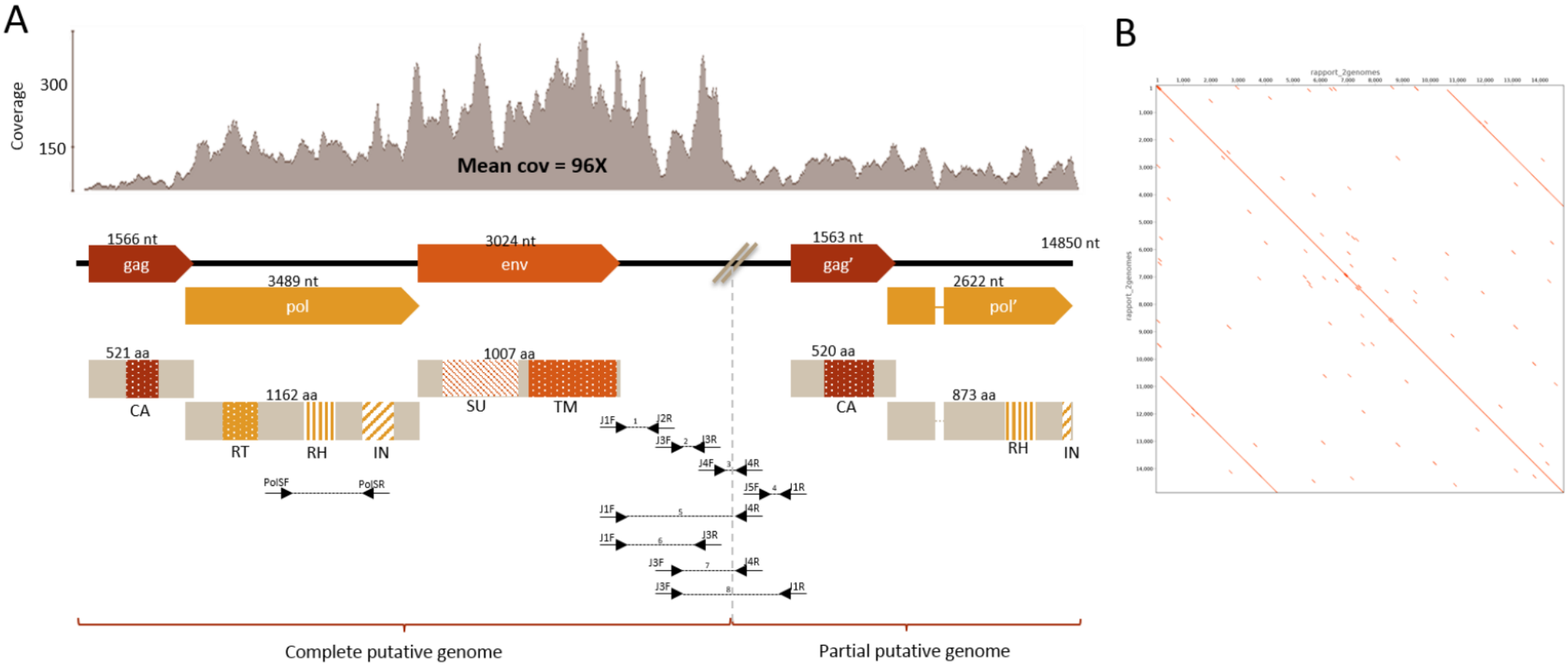


Figure 1: Genome characterization of the Spumavirus sequence found in *Salamandra salamandra*. A) Coverage and annotation of ORFs are represented, gag, pol and env. Conserved protein domains are indicated; CA: capsid, RT: retrotranscriptase, RN: Rnase-H, IN: integrase, SU: internal envelope subunit, TM: envelope transmembrane unit. Arrows correspond to PCR amplifications tested, see Table 1 for details. B) Self-self dotplot.

The conserved domain SU of the envelope could not be found by homology searches against known protein in the databases, although the size of the envelope is very similar to other endogenous or exogenous spumaviruses. A manual annotation step, by comparison with known functional domains of envelope of foamy viruses, revealed the SU domain sharing foamy features. The envelope protein indeed contained a signal peptide (position 123-160 of the env protein), a putative cleavage signal delimitating SU-TM (658-KKGR-651) followed by hydrophobic residues putatively corresponding to the fusion peptide; two pairs of Cysteine (824-825 and 843-844); and finally the detection of a transmembrane domain putatively corresponding to the membrane spanning domain (MSD) (Bénit et al., 2001; Wang & Mulligan, 1999).

The second genome has a 53X coverage and harbored retroviral protein domains forming a nearly complete genome (Figure 1). The pol gene was truncated (2,761 nucleotide whereas the Coelacanth FERV is 3,633) and this ORF contained a premature STOP codon in position 850, corresponding to pseudogenisation.

For the second amphibian transcriptome of *Pleurodeles waltl*, the development of a consensus followed by the reconstruction of a genome, similar to that used for the salamander, could not be used. The spumavirus-like contigs were not overlapping enough to reconstruct longer sequences. However, a consensus sequence of the integrase domain of the pol gene could be integrated for phylogenetic analyses.

Spumavirus similarity searches in other available Amphibian transcriptomes and genomes

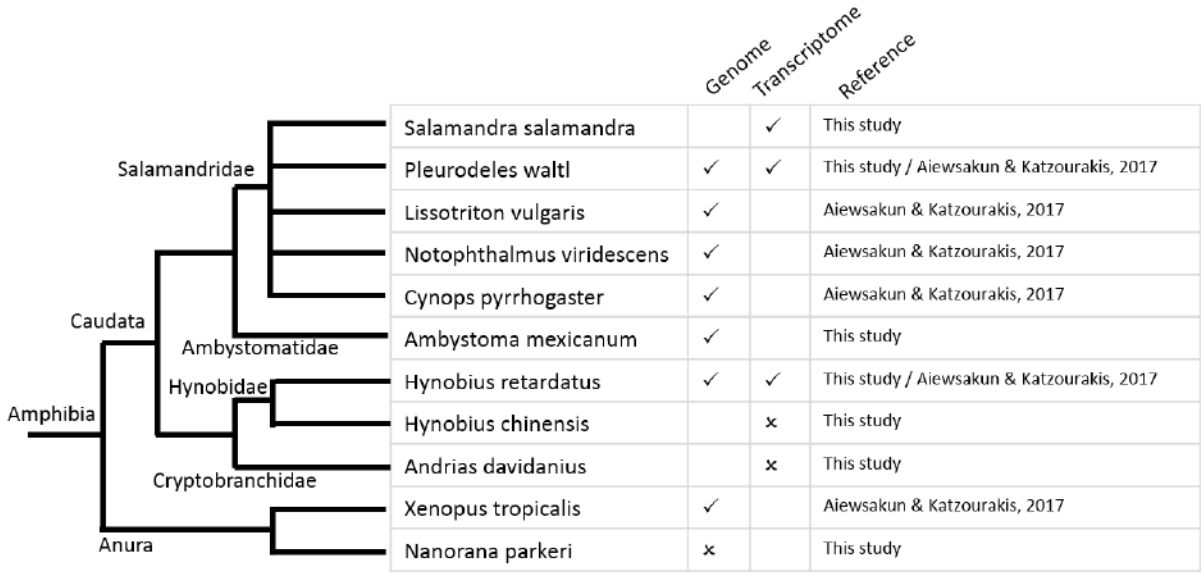
A broader search of *Spumavirus* elements was done in other transcriptome and genome of all amphibians available in databases. A BLAST search using the consensus sequence of the *Salamandra salamandra* spumavirus genome as a query against the axolotl (*Ambystoma mexicanum*, *Ambystomatidae*, Caudata) WGS database revealed the presence of several sequences with strong homology. The best hit possessed 39% of nucleotide identity and an E-value of 9.E-35. In the axolotl genome, the putative spumavirus sequence was truncated and pseudogenised but still a consensus protein of 240 aa was built (including 27 X in the amino acid translation) for phylogenetic analyses.

The BLAST search in the Hokkaido salamander (*Hynobius retardatus*, *Hynobiidae*, Caudata) transcriptome allowed to detect the presence of Spumavirus sequences. The best hit was for

a 709 nucleotide sequence (235 aa) with 46 % of identity (E-value=6.E-59). This sequence was also included in phylogenetic analyses.

Furthermore, similar search in the Chinese salamander (*Hynobius chinensis*), the Chinese giant salamander (*Andrias davidianus*), the Tibetan frog (*Nanorana parkeri*) and the western clawed frog (*Xenopus tropicalis*) did not reveal the presence of *Spumavirus* homologies (Figure 2).

Figure 2: Summary of the presence (ticks) and absence (crosses) of *Spumavirus* sequences in Amphibian genomes and transcriptomes.



Phylogenetic position of new Spumaviruses found in Amphibian

Two distinct ML phylogenies of the reverse transcriptase (Figure 3A) and integrase (Figure 3B) domains were built to analyse the phylogenetic position of the new fragmented FV-like sequences obtained from the fire salamander *Salamandra salamandra*, the Waltl newt *Pleurodeles walt*, the axolotl *Ambystoma mexicanum* and the Hokkaido salamander *Hynobius retardatus*.

The seven retrovirus genus already described in the literature clearly form distinct clades supported by important nodes values in both phylogenies outgrouped by Yoyo and Gypsy retrotransposons (Figure 3).

According to both phylogenies, the four amphibian retrovirus sequences formed a monophyletic clade within the Spumavirus genus well supported by very high aLRT values of 0.99 and 1 for reverse transcriptase and integrase, respectively. This clade therefore belongs very clearly to spumaviruses and constitutes a fully-fledged lineage within this genus and is

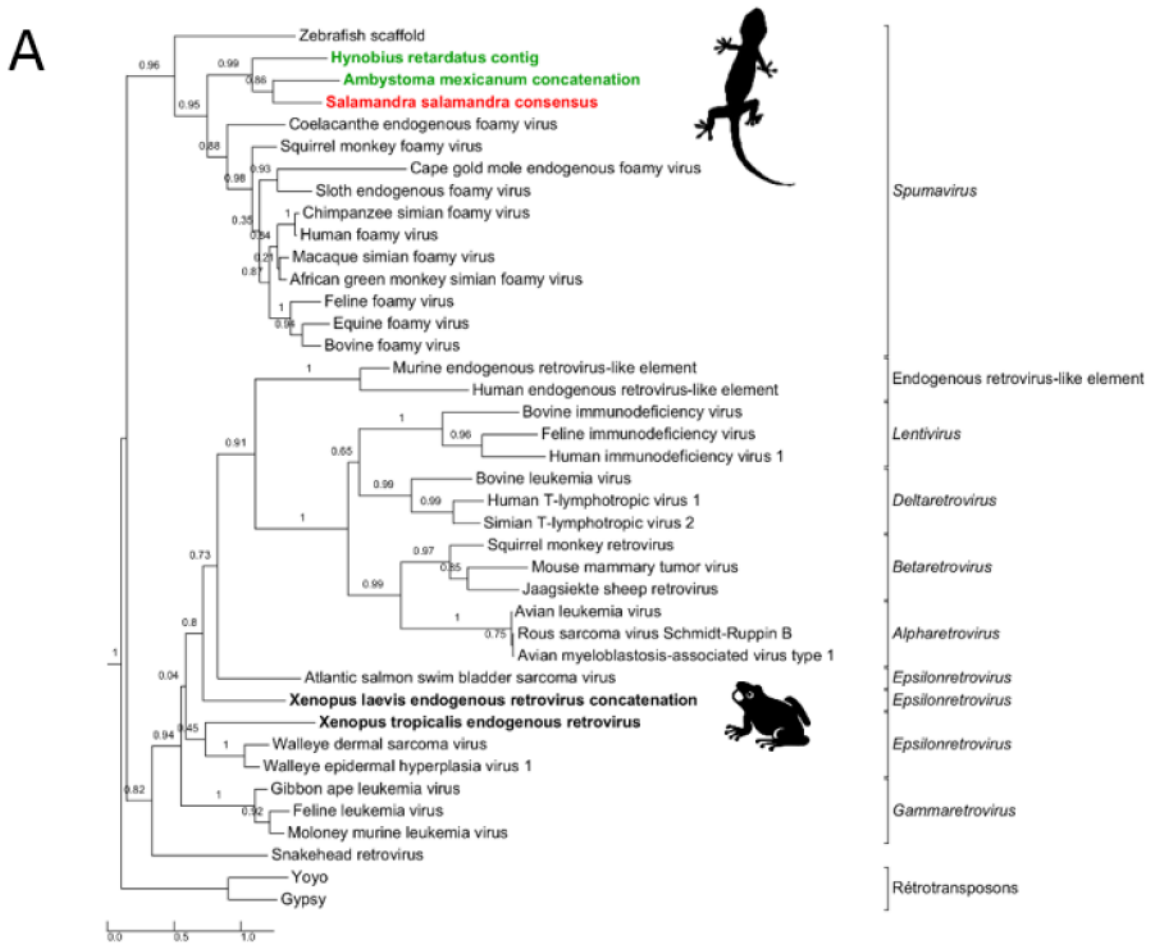
not close to any other known amphibian retrovirus sequence found in the anura *Xenopus* sp. (in bold in phylogenies) belonging to the genus *Epsilonretrovirus* and not to Spumavirus, as proposed in previous study (Herniou et al., 1998). Moreover, the Spumaviruses, whether endogenous or exogenous, formed a monophyletic group (aLRT=0.96 and 0.99 for reverse transcriptase and integrase, respectively) (Figure 3).

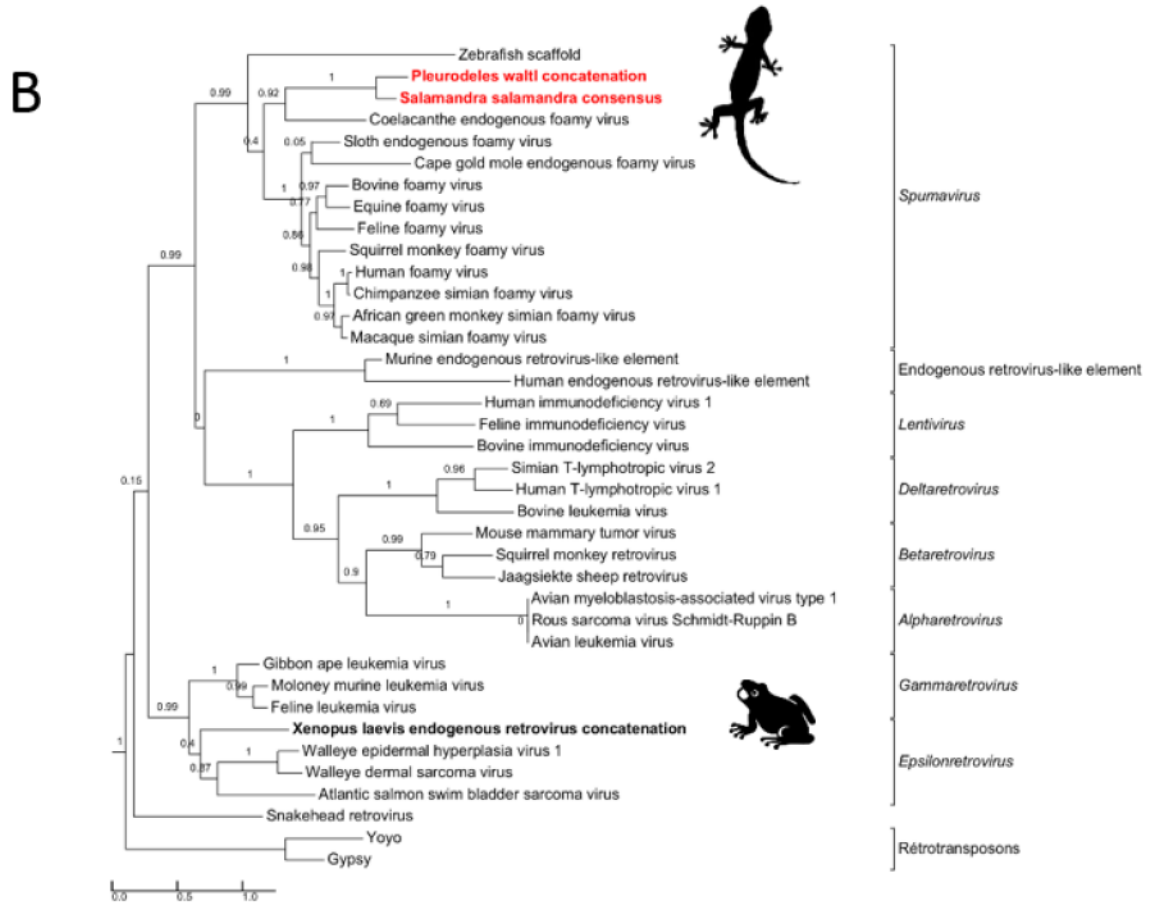
Sequences of *S. salamandra*, and *P. waltl* are sister taxa. The sequence of *A. mexicanum* would have diverged previously, and finally the sequence of *H. retardatus* is the most external to this clade (in green on phylogenies). The relationship of these four retroviral sequences is congruent with the host phylogeny: the Hynobiidae (*H. retardatus*) form with the Cryptobranchidae the group which diverged first among the urodeles and the Ambystomatidae (including *A. mexicanum*) are a sister group of Salamandridae (including *S. salamandra, and P. waltl*) (Figure 2) (Frost et al., 2006).

In order to confirm the presence of spumaviruses within the fire salamandra genome, PCR detection of distinct genomic regions were performed. Two samples of fire salamander and of palmate newt were collected. A sufficient quality of DNA extraction using non-invasive method was achieved for ST2 sample (*Salamandra salamandra*) only.

A first PCR allowed the amplification of the polymerase gene (Primers PolSF/PolSR, Figure 1, and Figure 4A). Sanger sequencing of PCR product was performed using both forward and reverse primers. The sequence obtained with the forward primer was 1,782bp size with 86% identity with the in silico salamandra *Spumavirus*. This first PCR amplification and sequencing confirmed the presence of a Spumavirus polymerase within the fire salamandra genome.

Figure 3: Amino acid phylogenies of retroviral sequences found in Amphibians. A) Retrotranscriptase domain (238 sites), B) Integrase domain (440 sites). Substitution model RtRev. Bold black sequences are amphibian known retroviruses. Sequences produced in this study are indicated in red (fire salamader and Waltl newt) and found from Blast searches in green. Accession number of publicly available sequences are indicated in Supplementary Table 1. The scale bars is substitution per sites. Node support values are aLRT (%) statistic.


A
Zebrafish scaffold
Hynobius retardatus contig
Ambystoma mexicanum concatenation
Salamandra salamandra consensus
Coelacanthe endogenous foamy virus
Squirrel monkey foamy virus
Cape gold mole endogenous foamy virus
Sloth endogenous foamy virus
Chimpanzee simian foamy virus
Human foamy virus
Macaque simian foamy virus
African green monkey simian foamy virus
Feline foamy virus
Equine foamy virus
Bovine foamy virus
Murine endogenous retrovirus-like element
Human endogenous retrovirus-like element
Bovine immunodeficiency virus
Feline immunodeficiency virus
Human immunodeficiency virus 1
Bovine leukemia virus
Human T-lymphotropic virus 1
Simian T-lymphotropic virus 2
Squirrel monkey retrovirus
Mouse mammary tumor virus
Jaagsiekte sheep retrovirus
Avian leukemia virus
Rous sarcoma virus Schmidt-Ruppin B
Avian myeloblastosis-associated virus type 1
Atlantic salmon swim bladder sarcoma virus
Xenopus laevis endogenous retrovirus concatenation
Xenopus tropicalis endogenous retrovirus
Walleye dermal sarcoma virus
Walleye epidermal hyperplasia virus 1
Gibbon ape leukemia virus
Feline leukemia virus
Moloney murine leukemia virus
Snakehead retrovirus
Yoyo
Gypsy
0.0
0.5
1.0
Spumavirus
Endogenous retrovirus-like element
Lentivirus
Deltaretrovirus
Betaretrovirus
Alpharetrovirus
Epsilonretrovirus
Epsilonretrovirus
Epsilonretrovirus
Gammaretrovirus
Rétrotransposons


B
Zebrafish scaffold
Pleurodeles waltl concatenation
Salamandra salamandra consensus
Coelacanthe endogenous foamy virus
Sloth endogenous foamy virus
Cape gold mole endogenous foamy virus
Bovine foamy virus
Equine foamy virus
Feline foamy virus
Squirrel monkey foamy virus
Human foamy virus
Chimpanzee simian foamy virus
African green monkey simian foamy virus
Macaque simian foamy virus
Murine endogenous retrovirus-like element
Human endogenous retrovirus-like element
Human immunodeficiency virus 1
Feline immunodeficiency virus
Bovine immunodeficiency virus
Simian T-lymphotropic virus 2
Human T-lymphotropic virus 1
Bovine leukemia virus
Mouse mammary tumor virus
Squirrel monkey retrovirus
Jaagsiekte sheep retrovirus
Avian myeloblastosis-associated virus type 1
Rous sarcoma virus Schmidt-Ruppin B
Avian leukemia virus
Gibbon ape leukemia virus
Moloney murine leukemia virus
Feline leukemia virus
Xenopus laevis endogenous retrovirus concatenation
Walleye epidermal hyperplasia virus 1
Walleye dermal sarcoma virus
Atlantic salmon swim bladder sarcoma virus
Snakehead retrovirus
Yoyo
Gypsy
0.0
0.5
1.0
Spumavirus
Endogenous retrovirus-like element
Lentivirus
Deltaretrovirus
Betaretrovirus
Alpharetrovirus
Gammaretrovirus
Epsilonretrovirus
Rétrotransposons

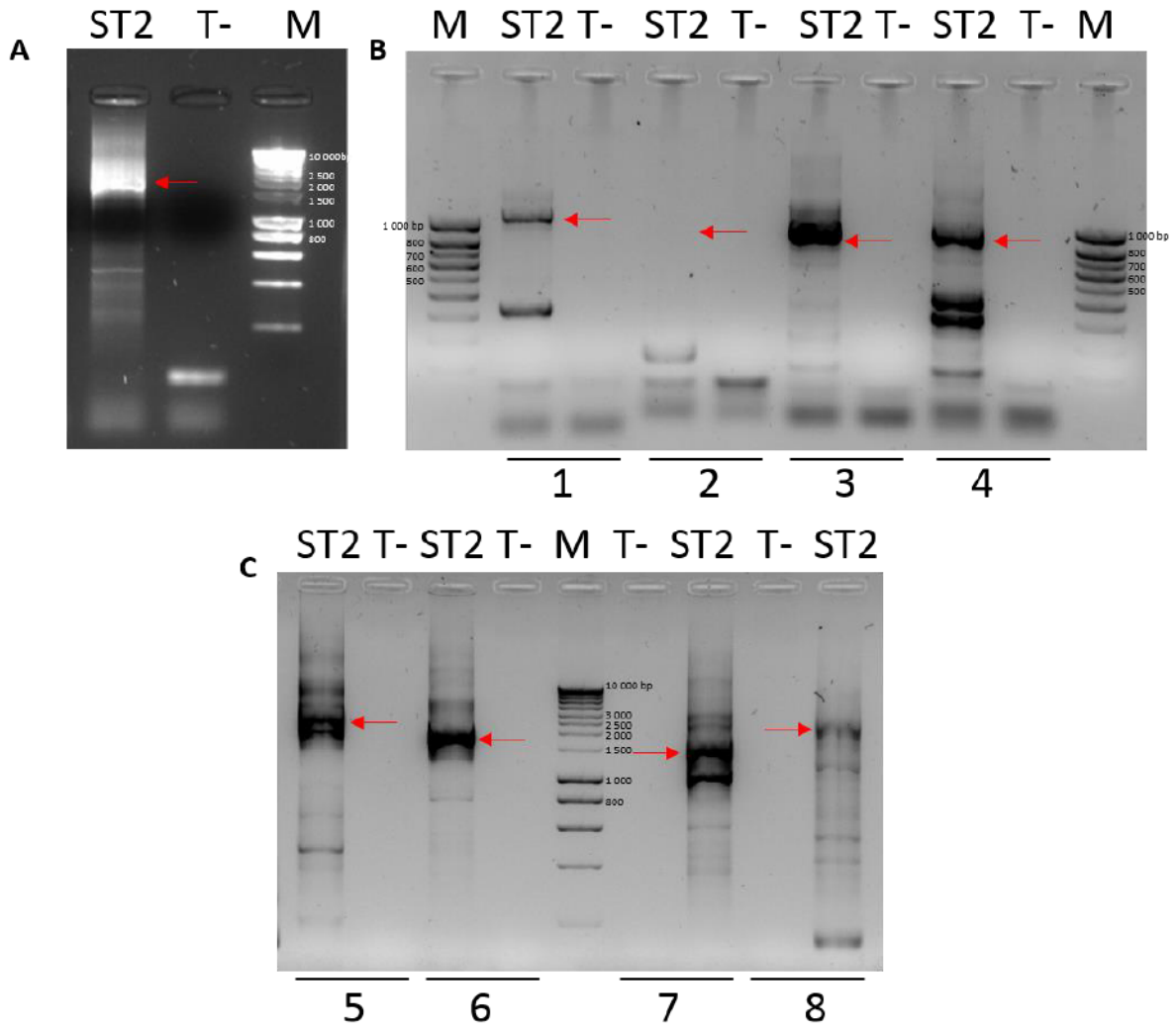


Figure 4: Biological validation of spumaviral integration in Caudata *Salamandra salamandra* (ST1&2) and *Lissotriton helveticus* (TP1&2). A) Polymerase PCR (agarose 1.5%) primers PolSF-PolSR (1,893 bp). B) and C) Junctions PCRs (agarose 1.5%), PCR 1: J1F-J2R (1,127 bp); 2: J3F-J3R (994 bp), 3: J4F-J4R (965 bp), 4: J5F-J1R (952 bp), 5: J1F-J4R (2,453 bp), 6: J1F-J3R (1,825 bp), 7: J3F-J4R (1,621 bp), 8: J3F-J1R (2,276 bp). See Figure 1 for PCR localization on genome. Red arrows indicate expected product waiting sizes.

Biological evidences of integrated *Spumavirus* viral genome in *Salamandra salamandra*

Eight other PCR amplifications were performed in order to confirm the presence of the end of the nearly complete genome and the putative junction between the two genomes (Figure 1, Table 1, Figure 4B and Figure 4C, PCR tested 1-8). Seven out of eight PCRs allowed amplification, the only one without positive result was the junction PCR 2 (Primers J3F-J3R, Figure 4B). Sanger sequencing was done only for the junction PCR 3 (Primers J4F-J4R) which was the most amplified without unspecific amplifications. The obtained sequence was 692bp size with 96.3% identity with the in silico salamandra *Spumavirus*.

Discussion

Amphibian infections and diseases have been studied extensively since the recognition of their global decline (Blaustein et al., 2012).

Our study allows the detection of a new Spumavirus in 4 species of urodeles: the fire salamander (*Salamandra salamandra*) and the Waltl newt (*Pleurodeles waltl*) using pipeline to detect retroviral sequences from transcriptomes, and in the Axolotl genome (*Ambystoma mexicanum*) and the transcriptome of the Hokkaido salamander (*Hynobius retardatus*) by targeted BLAST searches in the NCBI databases.

This result was new according to our knowledge at the beginning of this work and it was the first time we showed the existence of an association between Spumavirus and amphibians. However, since few month another group also identified spumaviruses in amphibian, including *Pleurodeles waltl, Hynobius retardatus* as in our work (Aiewsakun & Katzourakis, 2017). They also found gag of *Spumavirus* in *Xenopus tropicalis* where we did not found any similarities and some other genes (gag-pol-env) in three other species (*Cynops pyrrhogaster*, *Lissotriton vulgaris* and *Notophthalmus viridescens*).

To date, the only known exogenous Spumaviruses were found in the Laurasiatheria (bat, horse, cow) (Materniak et al., 2013; Tobaly-Tapiero et al., 2000; Wu et al., 2012) and in the Euarchontoglires (ex: primates) (Falcone et al., 1999), placental mammals which form the clade of the Boroeoeutherians. The endogenization of Spumaviruses has been shown in the coelacanth (*Latimeria chalumnae*) a lobe-finned fish of the Sarcopterygian group, very close to the bifurcation of the root of a tetrapod (Han & Worobey, 2012); in a ray-finned fish radiated from the Actinopterygii group, zebrafish (*Danio rerio*), in the sloth (*Choloepus hoffmanni*) of the basal mammalian Xenarthral group; as well as in a cape golden mole (*Chrysochloris asiatica*), an insectivorous mammal (Han & Worobey, 2014). The detection of spumavirus within the salamander's genome and transcriptome bring more knowledge on this group of viruses which have an ancient marine originate and originate more than 450 million years ago (Aiewsakun & Katzourakis, 2017).

Whatever the regions of the retroviral genome from which the alignments are derived, phylogenetic analyzes indicate that these four new viruses form a monophyletic taxonomic group within the Spumavirus clade, as also shown in (Aiewsakun & Katzourakis, 2017). The phylogenies performed with the reverse transcriptase and integrase domains of the polymerase, as well as that performed with the whole polymerase (result not included in the

study) show that the topology of the new urodel spumaviruses is well resolved. These results show that the method used for reconstruction by consensus and mapping does not introduce errors as to the phylogenetic position of the viruses described. Finally, these results also confirm the hypothesis that new retroviruses can be discovered in vertebrates and especially when studying non-model animals. This spumavirus lineage seemed specific to urodeles, since other spumaviral sequences from anura did not group in the salamander spumavirus group, and rather derived from *Epsilonretrovirus* (Kambol et al., 2003; Sinzelle et al., 2011).

We are able to reconstruct a nearly entire spumavirus genome with the fire salamander transcriptome analysis. The complete annotation of predicted proteins confirmed the presence of all specific retroviral elements, such as the capsid encoded by the gag gene; the reverse transcriptase, the RNase-H and the integrase encoded by the pol gene; and the envelope subunit and the transmembrane glycoproteins encoded by the env gene. The presence of complete, non-truncated genes with complete domains indicates that this genome is close to a functional state. The sizes of the ORF are very similar to those observed in the nearest Spumavirus integrated into the genome of the coelacanth (Han & Worobey, 2012) and the organization of genomic RNA and ORF sizes are therefore consistent with what is usually found in Spumaviruses (King et al., 2012). However, we can note the absence of the genes tas and bet, accessory genes found in almost all spumaviruses, except for the spumaviruses of the zebrafish and the golden taupe for which the information of the presence of these gene is absent from publications. These two genes appear to be key elements in the infectious success of exogenous Spumaviruses.

The fact that at least two spumavirus genomes - quite different genetically - exist in the salamander support the hypothesis that this sequence is probably integrated, and that this Spumavirus is endogenous. It is also verified by the PCR of some part of the salamandra spumavirus within genomic DNA. If this hypothesis is correct, we may wonder why this ERV is a potentially functional genome, and why it is always strongly expressed. It is possible that the lineage leading to this ERV has recently been endogenized sufficiently to maintain functional ORFs. Another hypothesis is that this line is active, and that this ERV would still be capable of producing virions and transmitting horizontally. Molecular biology tools (e.g. detection of virions by electron microscope) will help in the future to validate and verify theses assumptions.

Acknowledgements

Data used in this work were partly produced through molecular genetic analysis technical facilities of the SFR ‘‘Montpellier Environnement Biodiversite´’’, thanks to Dr. Philippe Clair (UM2-Montpellier GenomiX). We are grateful to the Genotoul Bioinformatics Platform Toulouse Midi-Pyrenees for providing computing and storage resources. We would like to thank Jean-Yves Sires and ANR Jaws for sharing unpublished data of *Pleurodeles waltl*. This work has been supported by a European Research Council (ERC) grant to Nicolas Galtier (ERC PopPhyl 232971).

The authors declare that they have no competing interests.

References

Abascal, F., Zardoya, R. & Posada, D. (2005). ProtTest: Selection of best-fit models of protein evolution. Bioinformatics 21, 2104–2105.

Aiewsakun, P. & Katzourakis, A. (2017). Marine origin of retroviruses in the early Palaeozoic Era. Nat Commun 8, 13954. Nature Publishing Group.

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.

Anisimova, M. & Gascuel, O. (2006). Approximate likelihood-ratio test for branches: a fast, accurate, and powerful alternative. Syst Biol 55, 539–52.

Bénit, L., Dessen, P. & Heidmann, T. (2001). Identification, Phylogeny , and Evolution of Retroviral Elements Based on Their Envelope Genes. J Virol 75, 11709–11719.

Birol, I., Jackman, S. D., Nielsen, C. B., Qian, J. Q., Varhol, R., Stazyk, G., Morin, R. D., Zhao, Y., Hirst, M. & other authors. (2009). De novo transcriptome assembly with ABySS. Bioinformatics 25, 2872–2877.

Blaustein, A. R., Gervasi, S. S., Johnson, P. T. J., Hoverman, J. T., Belden, L. K., Bradley, P. W. & Xie, G. Y. (2012). Ecophysiology meets conservation: understanding the role of disease in amphibian population declines. Philos Trans R Soc B Biol Sci 367, 1688–1707.

Cahais, V., Gayral, P., Tsagkogeorga, G., Melo-Ferreira, J., Ballenghien, M., Weinert, L., Chiari, Y., Belkhir, K., Ranwez, V. & Galtier, N. (2012). Reference-free transcriptome assembly in non-model animals from next-generation sequencing data. Mol Ecol Resour 12, 834–45.

Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K. & Madden, T. L. (2009). BLAST+: architecture and applications. BMC Bioinformatics 10, 421.

Castresana, J. (2000). Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. Mol Biol Evol 17, 540–52.

Delelis, O., Lehmann-Che, J. & Saïb, A. (2004). Foamy viruses — a world apart. Curr Opin Microbiol 7, 400–406.

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

Falcone, V., Leupold, J., Clotten, J., Urbanyi, E., Herchenröder, O., Spatz, W., Volk, B., Böhm, N., Toniolo, A. & other authors. (1999). Sites of simian foamy virus persistence in naturally infected African green monkeys: latent provirus is ubiquitous, whereas viral replication is restricted to the oral mucosa. Virology 257, 7–14.

Feschotte, C. & Gilbert, C. (2012). Endogenous viruses: insights into viral evolution and impact on host biology. Nat Rev Genet 13, 283–296. Nature Publishing Group.

Figuet, E., Ballenghien, M., Romiguier, J. & Galtier, N. (2015). Biased Gene Conversion and GC-Content Evolution in the Coding Sequences of Reptiles and Vertebrates. Genome Biol Evol 7, 240–250. Oxford University Press.

Frost, D. R., Grant, T., Faivovich, J., Bain, R. H., Haas, A., Haddad, C. F. B., Sá, R. O. D. E., Channing, A., Donnellan, S. C. & other authors. (2006). The amphibian tree of life. Bull Am Museum Nat Hist 297, 1–291. American Museum of Natural History.

Gifford, R. & Tristem, M. (2003). The Evolution, Distribution and Diversity of Endogenous Retroviruses. Virus Genes 26, 291–315. Kluwer Academic Publishers.

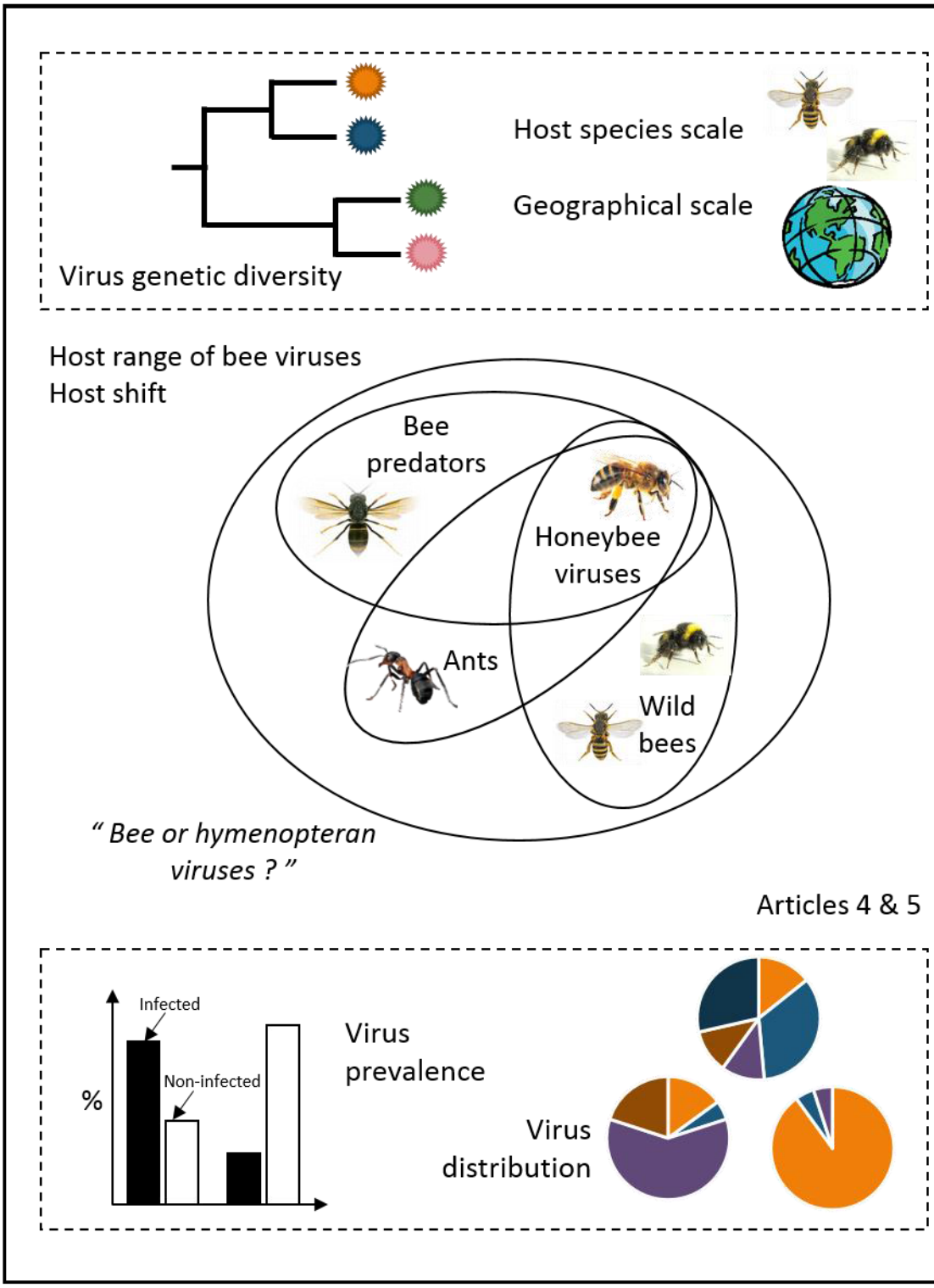
Gouy, M., Guindon, S. & Gascuel, O. (2010). SeaView Version 4: A Multiplatform Graphical User Interface for Sequence Alignment and Phylogenetic Tree Building. Mol Biol Evol 27, 221–224.
Guindon, S. & Gascuel, O. (2003). A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. Syst Biol 52, 696–704.
Han, G. Z. & Worobey, M. (2012). An endogenous foamy-like viral element in the coelacanth genome. PLoS Pathog 8, 1–7.
Han, G. Z. & Worobey, M. (2014). Endogenous viral sequences from the Cape golden mole (Chrysochloris asiatica) reveal the presence of foamy viruses in all major placental mammal clades. PLoS One 9, 3–6.
Herniou, E., Martin, J., Miller, K., Cook, J., Wilkinson, M. & Tristem, M. (1998). Retroviral diversity and distribution in vertebrates. J Virol 72, 5955–5966.
Huang, X. & Madan, A. (1999). CAP3: a DNA sequence assembly program. Genome Res 9, 868–877.
Hyatt, D., Chen, G.-L., Locascio, P. F., Land, M. L., Larimer, F. W. & Hauser, L. J. (2010). Prodigal: prokaryotic gene recognition and translation initiation site identification. BMC Bioinformatics 11, 119.
Hyatt, D., Locascio, P. F., Hauser, L. J. & Uberbacher, E. C. (2012). Gene and translation initiation site prediction in metagenomic sequences. Bioinformatics 28, 2223–2230.
Jones, P., Binns, D., Chang, H.-Y., Fraser, M., Li, W., McAnulla, C., McWilliam, H., Maslen, J., Mitchell, A. & other authors. (2014). InterProScan 5: genome-scale protein function classification. Bioinformatics 30, 1236–1240.
Kambol, R., Kabat, P. & Tristem, M. (2003). Complete nucleotide sequence of an endogenous retrovirus from the amphibian, Xenopus laevis. Virology 311, 1–6.
Katoh, K., Misawa, K., Kuma, K. & Miyata, T. (2002). MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. Nucleic Acids Res 30, 3059–3066.
Katzourakis, A., Gifford, R. J., Tristem, M., Gilbert, M. T. P. & Pybus, O. G. (2009). Macroevolution of Complex Retroviruses. Science 325, 1512–1512.
Katzourakis, A. & Gifford, R. J. (2010). Endogenous Viral Elements in Animal Genomes. PLoS Genet 6, e1001191.
King, A. M. Q., Adams, M. J., Carstens, E. B. & Lefkowitz, E. J. (2012). Virus taxonomy. Ninth report of the International Committee on Taxonomy of Viruses. Virus Taxon (A. M. Q. King, M. J. Adams, E. B. Carstens & E. J. Lefkowitz, Eds.). Academic Press.
Linial, M. (2000). Why aren't foamy viruses pathogenic? Trends Microbiol 8, 284–289.
Materniak, M., Hechler, T., Lochelt, M. & Kuzmak, J. (2013). Similar Patterns of Infection with Bovine Foamy Virus in Experimentally Inoculated Calves and Sheep. J Virol 87, 3516–3525.
Pais, F. S.-M., Ruy, P. de, Oliveira, G. & Coimbra, R. (2014). Assessing the efficiency of multiple sequence alignment programs. Algorithms Mol Biol 9, 4.
Patel, M. R., Emerman, M. & Malik, H. S. (2011). Paleovirology—ghosts and gifts of viruses past. Curr Opin Virol 1, 304–309.
Pichlmüller, F., Straub, C. & Helfer, V. (2013). Skin swabbing of amphibian larvae yields sufficient DNA for efficient sequencing and reliable microsatellite genotyping. Amphibia-Reptilia 34, 517–523.
Rost, B. (1996). PHD: predicting one-dimensional protein structure by profile-based neural networks. Methods Enzymol 266, 525–39.
Ruboyianes, R. & Worobey, M. (2016). Foamy-like endogenous retroviruses are extensive and abundant in teleosts. Virus Evol 2, vew032.
Simpson, J. T., Wong, K., Jackman, S. D., Schein, J. E., Jones, S. J. M. & Birol, I. (2009). ABySS: a parallel assembler for short read sequence data. Genome Res 19, 1117–1123.
Sinzelle, L., Carradec, Q., Paillard, E., Bronchain, O. J. & Pollet, N. (2011). Characterization of a Xenopus tropicalis Endogenous Retrovirus with Developmental and Stress-Dependent Expression. J Virol 85, 2167–2179.
Thompson, J. D., Higgins, D. G. & Gibson, T. J. (1994). CLUSTAL W: Improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res 22, 4673–4680.
Tobaly-Tapiero, J., Bittoun, P., Neves, M., Guillemin, M.-C., Lecellier, C.-H., Puvion-Dutilleul, F., Gicquel, B., Zientara, S., Giron, M.-L. & other authors. (2000). Isolation and characterization of an equine foamy virus. J Virol 74, 4064–4073.
Wang, G. & Mulligan, M. J. (1999). Comparative sequence analysis and predictions for the envelope glycoproteins of foamy viruses. J Gen Virol 80, 245–254.
Wu, Z., Ren, X., Yang, L., Hu, Y., Yang, J., He, G., Zhang, J., Dong, J., Sun, L. & other authors. (2012). Virome Analysis for Identification of Novel Mammalian Viruses in Bat Species from Chinese Provinces. J Virol 86, 10999–11012.

Supplementary table 1: Accession number list of sequences used in phylogenies.

Virus name	Retroviral genus	GenBank Accesion
Avian leukemia virus	*Alpharetrovirus*	YP_004222728
Avian myeloblastosis-associated virus type 1	*Alpharetrovirus*	AAA46304
Rous sarcoma virus Schmidt-Ruppin B	*Alpharetrovirus*	AAC08988
Jaagsiekte sheep retrovirus	*Betaretrovirus*	NP_041186
Mouse mammary tumor virus	*Betaretrovirus*	NP_955564
Squirrel monkey retrovirus	*Betaretrovirus*	NP_041261
Bovine leukemia virus	*Deltaretrovirus*	AAA42785
Human T-lymphotropic virus 1	*Deltaretrovirus*	NP_955619
Simian T-lymphotropic virus 2	*Deltaretrovirus*	CAA74901
Human endogenous retrovirus-like element	*Endogenous retroviral element*	X8221
Murine endogenous retrovirus-like element	*Endogenous retroviral element*	CAA73251
Atlantic salmon swim bladder sarcoma virus	*Epsilonretrovirus*	ABA54982
Walleye dermal sarcoma virus	*Epsilonretrovirus*	ABO25842
Walleye epidermal hyperplasia virus 1	*Epsilonretrovirus*	AAD30048
Xenopus laevis endogenous retrovirus	*Epsilonretrovirus*	AJ506107
Xenopus tropicalis endogenous retrovirus	*Epsilonretrovirus*	(Sinzelle et al., 2011)
Feline leukemia virus	*Gammaretrovirus*	NP_955577
Gibbon ape leukemia virus	*Gammaretrovirus*	NP_056790
Moloney murine leukemia virus	*Gammaretrovirus*	NP_057933
Bovine immunodeficiency virus	*Lentivirus*	NP_040563
Feline immunodeficiency virus	*Lentivirus*	NP_040973
Human immunodeficiency virus 1	*Lentivirus*	NP_789740
Gypsy	Retrotransposons	ABQ84946
Yoyo	Retrotransposons	AAC28743
African green monkey simian foamy virus	*Spumavirus*	YP_001956722
Bovine foamy virus	*Spumavirus*	NP_044929
Cape gold mole endogenous foamy virus	*Spumavirus*	(Han & Worobey, 2014)
Chimpanzee simian foamy virus	*Spumavirus*	Q87040
Coelacanthe endogenous foamy virus	*Spumavirus*	(Han & Worobey, 2012)
Equine foamy virus	*Spumavirus*	NP_054716
Feline foamy virus	*Spumavirus*	AGC11908
Human foamy virus	*Spumavirus*	CAA68997
Macaque simian foamy virus	*Spumavirus*	AGM61336
Sloth endogenous foamy virus	*Spumavirus*	(Katzourakis et al., 2009)
Snakehead retrovirus	*Spumavirus*	NP_043924
Squirrel monkey foamy virus	*Spumavirus*	ADE05995
Zebrafish scaffold	*Spumavirus*	AL935186

CHAPITRE 2

Biodiversité et prévalence de virus d'abeilles


Host species scale
Geographical scale
Virus genetic diversity
Host range of bee viruses
Host shift
Bee predators
Honeybee viruses
Ants
Wild bees
" Bee or hymenopteran viruses ? "
Articles 4 & 5
Infected
Non-infected
%
Virus prevalence
Virus distribution

4.1. Préambule au Chapitre 2

"With the new molecular tools capable of analysing not only a single microorganism, but also the larger community of microorganisms living in bees, there is no doubt that more honey bee viruses will be discovered."

***De Miranda et al, 2011**, Honey Bee Colony Health: Challenges and Sustainable Solutions. p. 71–102, CRC Press.*

A ce jour et depuis les années 1960, au moins 24 virus différents ont été décrits affectant l'abeille domestique *Apis mellifera.* En revanche, bien qu'il y ait eu plusieurs études qui ont décrit la présence de virus d'abeilles chez certains hyménoptères sauvages (bourdons, fourmis, abeilles solitaires), la diversité des virus d'hyménoptères sauvages reste largement inconnue. Si bien que les nouvelles techniques de séquençages permettent alors d'en découvrir, tel est le cas du Halictus scabiosae Adlikon Virus décrit dans le premier article de ce chapitre (voir **Article 4**). De plus, ce premier travail a permis de montrer la présence d'un nouveau cas de changement d'hôte, au sein des hyménoptères pour un virus récemment décrit chez *A. mellifera*, le *Lake Sinai virus*, découvert chez des fourmis du genre *Messor*. Le changement d'hôte est finalement très fréquent au sein des hyménoptères puisque la seconde étude de ce chapitre en fait également la preuve (voir **Article 5**). Ce second travail permet de montrer que de nombreux hôtes hyménoptères peuvent être infectés par les mêmes souches infectant les abeilles.

Ces deux travaux permettent ainsi de confirmer l'idée que de très nombreuses espèces d'hyménoptères sauvages peuvent habriter des virus découverts chez les abeilles domestiques. Il apparait donc nécessaire de découvrir de nouveaux virus chez les hyménoptères sauvages autant que chez les abeilles domestiques, afin de comprendre leur réelle gamme d'hôtes (voir **Article 4 & Article 5**).

4.2. Les hyménoptères sauvages habritent des virus d'abeille domestique (**Article 4**)

Ce travail se présente sous la forme d'un article accepté dans *Journal of General Virology* (doi : 10.1099/jgv.0.000957). Il présente la description d'un nouveau virus d'abeille sauvage halictes, la détection d'un virus d'abeille chez trois espèces de fourmis moissonneuse et l'étude de la diversité génétique d'un virus d'abeille domestique, le *Lake Sinai virus*.

The discovery of Halictivirus resolves the *Sinaivirus* phylogeny

Diane Bigot[1], Anne Dalmon[2], Bronwen Roy[3], Chunsheng Hou[4,5], Michèle Germain[1], Manon Romary[1], Shuai Deng[4,5], Qingyun Diao[4,5], Lucy A. Weinert[6,7], James M. Cook[3], Elisabeth A. Herniou[1§] and Philippe Gayral[1§]

[1]Institut de Recherche sur la Biologie de l'Insecte, UMR 7261, CNRS, Université François-Rabelais, 37200 Tours, France ; [2]INRA UR 406 Abeilles et environnement, Centre de recherche Provence-Alpes-Côte d'Azur, Site Agroparc, Domaine St Paul 228, Route de l'aérodrome CS40509 84914 Avignon, France ; [3]Hawkesbury Institute for the Environment, Western Sydney University, Locked Bag 1797, Penrith NSW 2751, Australia ; [4]Institute of Apicultural Research, Chinese Academy of Agricultural Sciences, Beijing 100093, China ; [5]Key Laboratory of Pollinating Insect Biology, Ministry of Agriculture, Beijing 100093, China; [6]Institut des Sciences de l'Evolution UMR5554, Université Montpellier–CNRS–IRD–EPHE, Montpellier, France; [7]Present address: University of Cambridge, Department of Veterinary Medicine, Madingley Road, Cambridge, CB3 0ES, United Kingdom

[§] Equal contribution

Abbreviations: AACV: Anopheline-associated C virus; ALPV: Aphid Lethal Paralysis virus; BQCV: Black queen cell virus; BSRV: Big Sioux River virus; CBPV: Chronic bee paralysis virus; CCD: Colony Collapse Disorder; Df: Degree of freedom; dN/dS: non-synonymous to synonymous substitution rates; DWV: Deformed wing virus; HsAV: Halictus scabiosae Adlikon Virus; IAPV: Israeli Acute Paralysis Virus; ICTV: International Committee on Taxonomy of Viruses; LRT: Likelihood Ratio Tests; LSV: Lake Sinai virus; MoNV: Mosinovirus; MTase-GTase: methyltransferase-guanylyltransferase; NGS: Next generation sequencing; NoV: Nodamura virus; RdRp: RNA-dependent RNA-polymerase; SNP: Single nucleotide polymorphism.

New sequence data: Sequence Read Archive (SRA) were deposited on accession numbers SRX2559194, SRX2188455-SRX2188457, SRX2188473, SRX2188475, and SRX2960331-SRX2960344. Sequence accessions number MF491478-MF491508 were deposited on Genbank.

Abstract

By providing pollination services, bees are among the most important insects, both in ecological and economical terms. Combined next generation and classical sequencing approaches were applied to discover and study new insect viruses potentially harmful to bees. A bioinformatics virus discovery pipeline was used on individual Illumina transcriptomes of 13 wild bees from 3 species from the genus *Halictus* and 30 ants from 6 species of the genera *Messor* and *Aphaenogaster*. This allowed the discovery and description of three sequences of a new virus termed Halictus scabiosae Adlikon Virus (HsAV). Phylogenetic analyses of ORF1, RdRp and capsid genes showed that HsAV is closely related to ssRNA+ viruses of the unassigned *Sinaivirus* genus but distant enough to belong to a different new genus we called Halictivirus. In addition, our study of ant transcriptomes revealed the first four sinaivirus sequences from ants (*Messor barbarus*, *M. capitatus* and *M. concolor*). Maximum likelihood phylogenetic analyses were performed on a 594 nt fragment of the ORF1/RdRp region from 84 sinaivirus sequences, including 31 new LSV from honey bees collected in five countries across the globe and the 4 ant viral sequences. The phylogeny revealed 4 main clades potentially representing different viral species infecting honey bees. Moreover, the ant viruses belonged to the LSV4 clade, suggesting a possible cross-species transmission between bees and ants. Lastly, wide honey bee screening showed that all four LSV clades have worldwide distributions with no obvious geographical segregation.

Key-words: Hymenoptera, Wild bee, Ants, RNA virus discovery, LSV, HsAV

Introduction

The worldwide economic value of pollination is about €153 billion [1], as 70 % of the main crops used for human consumption depend on insect pollinators [2]. Consequently, there is concern over the implications of recent declines in insect pollinators and raised awareness of the importance of honey bee (*Apis mellifera*) health. A combination of various elements, including pesticides, nutrition, management practices, environmental factors, parasites, and pathogens, including viruses, have been linked to the decline of managed honey bees [3–11]. Honey bee colonies affected by Colony Collapse Disorder (CCD) have been shown to host more pathogens than control hives [12]. However whether pathogens are causing or contributing factors of CCD, or spread through opportunistic infections, remains unknown. Recently, in order to understand the causes of honey bee decline, extensive efforts have been made to monitor viruses of insect pollinators [13–15], reviewed by [16]. However, these efforts are hampered by limited knowledge of the true biodiversity of viruses infecting insect pollinators. Since the discovery of the first honey bee viruses in the 1960s [17], 24 honey bee viruses and satellites have been described, reviewed in [18], and this number is increasing [19]. Several bee viruses are associated with CCD (reviewed in [14, 20]), but pathogenic effects per se are known for only a few of these viruses. One double-stranded DNA virus (*Apis mellifera filamentous virus*) has been described for honey bees [21, 22], but most bee viruses have positive single-stranded RNA (ssRNA+) genomes and belong to *Dicistroviridae* and *Iflaviridae* families (*Picornavirales*). Other unclassified ssRNA+ viruses have also been described, such as *Chronic bee paralysis virus* (CBPV) and *Lake Sinai virus* (LSV), both showing similarities with the members of the *Nodaviridae* family [23].

Between 2013 and 2015, 21 % to 33 % of surveyed honey bee colonies were positive for LSV-2 in the USA [24]. This high prevalence was further observed in 2013-2014 as over 34 % of colonies tested positive for pathogens in the Western US bore LSV infections [25]. Although LSV abundance is correlated with weak colonies, its pathology remains unknown and no visible symptoms have been attributed to LSV infection in honey bees [25]. Moreover, LSV has been detected in the *Varroa destructor* mites [25, 26] and a positive correlation with the presence of LSV and *Nosema* microsporidia has been demonstrated [24]. Furthermore, there have been few studies on LSV diversity and distribution and they were mainly based on American, Belgium and Spanish samples [12, 23, 25–32]. LSV has also been detected by PCR in Africa (Benin and Algeria) and South America (Colombia) but no sequences are available [9, 33, 34].

The limited geographic screening of LSV to date may well underestimate the true diversity of this virus. Currently, the International Committee on Taxonomy of Viruses (ICTV) recognizes only two LSV species, but other species or strains [25] have been described (Table S1).

Replicative forms of LSV, as demonstrated by detection of the negative-strand RNA intermediate by strand-specific PCR, have so far been found in only three bee species: in *Apis mellifera* where it was first discovered [23, 25, 26, 32], in the bumble bee *Bombus pascuorum* [30] and the solitary mason bee *Osmia cornuta* [26]. The presence of the replicative form of the virus in these species indicates that wild bees are probably natural hosts for LSV.

Here we report the discovery and description of a virus closely related to Sinaiviruses in the sweat bee *Halictus*, as well as the first detection of LSV in ants based on meta-transcriptomic analyses of wild Hymenoptera. We further collected honey bees from five countries and sequenced the ORF1/RdRp region of LSV to study the genetic diversity and geographical distribution of the different *Sinaivirus* clades.

Results

Genome reconstruction of *Sinaivirus* and new Halictivirus in wild Hymenoptera

A total of 580 million reads, 1.5 million assembled contigs and 1.2 million ORFs were analyzed in this work (Table S2). Overall 7 new viral sequences (4 complete genomes, one nearly-complete and two partial) were found in 6 transcriptomes, all showing significant homology with ssRNA+ *Lake Sinai virus*, a discovered honey bee virus [23] (File S1).

The first three viral sequences labelled Halictus scabiosae Adlikon Virus (HsAV) strains D, E and H were found in three individual sweat bees (sample ID GA16D, GA16E and GA16H) sampled in Switzerland (Table S3). The HsAV genomes contain three ORFs: ORF1 of unknown function, ORF2 encoding the RNA-dependent RNA-polymerase (RdRp) and ORF3 encoding a capsid protein (Fig. 1). Full-length genomes were obtained for HsAV_D and E (detected in the individual transcriptomes of GA16D and GA16E). Their genomes were 5,203 nt and 5,238 nt in size respectively and were both highly covered by 34,837 reads (mean coverage 640.7 X) and by 14,108 reads (256.6 X), respectively (Fig. S1a and b). The 5,201 nt genome of HsAV_H was nearly complete, covered by 533 reads (9.5 X) and filled with 207 undetermined nucleotides (Fig. S1c). All three HsAV sequences share over 96% nucleotide identity (Fig. 2d). The genome of HsAV is smaller than that of LSV2 the type species of Sinaivirus. It lacks ORF4 and harbors a repetition of 50 Adenine at the 5' end of the genome (Fig. 1).

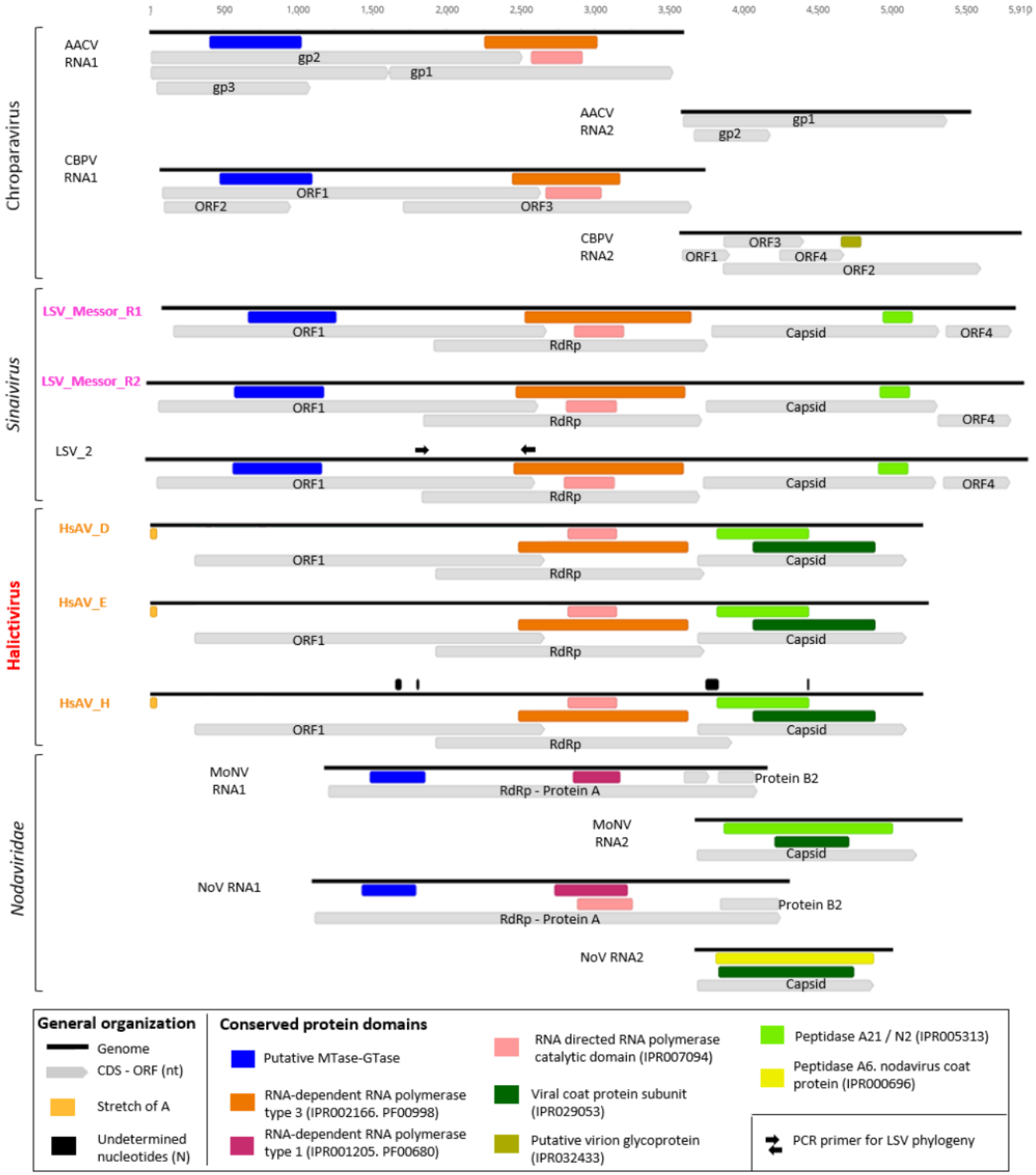


Fig. 1: Schematic representation of Halictus scabiosae Adlikon Virus genomes and new LSV sequences from *Messor* ants. *Chroparavirus*: AACV: *Anopheline associated C virus* (Genbank accession RNA1 NC_023682; RNA2 NC_023683), CBPV: *Chronic Bee Paralysis virus* (RNA1: NC_010711; RNA2: NC_010712); *Nodaviridae*: MoNV: *Mosinovirus* (RNA1: KJ632942; RNA2: KJ632943); NoV: *Nodamura virus* (RNA1: AF174533; RNA2: AF174534); *Sinaivirus*: LSV 2: *Lake Sinai virus 2* (HQ888865).

Four viral sequences were detected in ants. Two full-length genomes (LSV_Messor_R1 and LSV_Messor_R2) were found in a single *Messor concolor* harvester ant sampled in Crete, Greece (individual ID GA09R). Genomic organization was typical of LSV with 4 ORFs: ORF1 and ORF4 of unknown function, ORF2 encoding the RdRp and ORF3 the Capsid (Fig. 1). The genomes were respectively 5,816 nt and 5,877 nt in size and covered by 21,642 (mean coverage 329.6 X) and 48,503 (741.8 X) reads for LSV_Messor_R1 and LSV_Messor_R2 respectively (Fig. S1d and e). Two additional partial sequences with homology to LSV were found in the ants *M. barbarus* (individual ID GA09J) and *M. capitatus* (individual ID GA09P). Both sequences were too small to be fully annotated (1,613 and 553 nucleotides, respectively) but could be included in the ORF1/RdRp LSV phylogeny (part 3.3).

HsAV has a specific genomic organization

The five new ant LSV and HsAV genomes were annotated and compared to the genomes of related viruses including *Lake Sinai virus 2* (LSV2; *Sinaivirus*), *Anopheline-associated C virus* (AACV; Chroparavirus), *Chronic bee paralysis virus* (CBPV; Chroparavirus), and *Mosinovirus* (MoNV; *Nodaviridae*), *Nodamura virus* (NoV; *Nodaviridae*) (Fig. 1). The ant LSV_Messor sequences had the typical genomic organization of LSV2 (Fig. 1), in contrast with the HsAV sequences from sweat bees. HsAV has a type 3 RdRp domain (IPR002166) with a conserved catalytic domain (IPR007094) similar to those of Chroparavirus, *Sinaivirus*, and some *Nodaviridae* (NoV) (Fig. 1). This suggests that HsAV has conserved the function of RNA virus replicase.

The ORF1 of LSVs contained a putative MTase-GTase domain, also detected in Chroparavirus (AACV and CBPV), and in the N-terminal position of the protein A/RdRp of *Nodaviridae* (MoNV and NoV) [35, 36]. This MTase-GTase domain, with all conserved sites [35], could be identified by sequence homology in LSV_Messor_R1 and LSV_Messor_R2, but was lacking from all three HsAV sequences.

The capsid found in the HsAV genome was markedly different from to those of LSVs (16% identity, 30% similarity (Blosum62) at the protein level). Sinaiviruses possessed a single short peptidase A21/N2 (IPR005313) domain at the 3' end of ORF3. A significantly longer peptidase A21/N2 domain was found in the HsAV capsid ORF, but in an N terminal position. This peculiarity was also observed in the MoNV capsid. In addition, the HsAV and MoNV capsid ORF displayed a second overlapping viral coat domain (IPR029053) in a C-terminal position.

MoNV, which is a recombinant virus with a nodavirus-like RdRp, is the only other virus known to have such capsid domain organization. Other nodaviruses instead possess a viral coat domain (IPR029053) embedded with a peptidase A6 nodavirus coat domain (IPR000696) (Fig. 1).

Finally, the monopartite genomic organization of Sinaivirus and HsAV differs to the bipartite genomes of Chroparavirus and *Nodaviridae*, in which RNA1 encode ORF1 and RdRp and RNA2 the capsid. This segmentation could explain some evolutionary dissimilarities observed between Chroparavirus/*Nodaviridae* and *Sinaivirus*/HsAV. Furthermore, segmentation could favor gene exchanges, possibly explaining the shared origin of the MoNV and HsAV capsids.

Genome-scale phylogenies revealed the relationships of Sinaivirus, Halictivirus and Chroparavirus

Phylogenetic analyses were performed on each of the three ORFs (ORF1, RdRp and capsid) to explore the evolutionary history of the 5 new HsAV and LSV_Messor genomes.

First, phylogenetic analyses for ORF1 were performed from an alignment of 603 amino acid sites using the LG+G+I evolutionary model (Fig. 2a). The phylogeny showed that the three HsAV strains formed a strongly supported monophyletic group (posterior probabilities=1), clearly distinct from chroparavirus and sinaivirus clades. The two ant LSV sequences, LSV_Messor_R1 and LSV_Messor_R2 discovered together in a single ant, both belonged to Sinaivirus, which formed a monophyletic group (posterior probabilities=1).

Second, the RdRp phylogeny was built from a 559 amino acid alignment using the LG+G+I evolutionary model (Fig. 2b). Since the RdRp gene is present in many RNA viruses, *Nodaviridae* sequences could be added as outgroups to root the LSV and HsAV clades. The RdRp phylogeny showed that HsAVs, which form a well-supported monophyletic group (posterior probabilities=1), and not the chroparaviruses, were the sister group of Sinaivirus (posterior probabilities=1). The RdRp tree also confirmed that ant LSV genomes belonged to the Sinaivirus clade and did not form an independent lineage.

Third, the capsid phylogeny was built from an alignment of 884 amino acid sites using the Blosum62+G+I evolutionary model (Fig. 2c). The monophyly of HsAV, the place of ant LSVs within sinaivirus clade and the chroparavirus as outgroups were consistent with analyses of the other genes. Interestingly, the phylogeny showed that the capsid of *Mosinovirus* (MoNV)

derived from a common ancestor of the Halictivirus, and not *Sinaivirus*. This evolutionary scenario was strongly supported by posterior probabilities of 0.99.

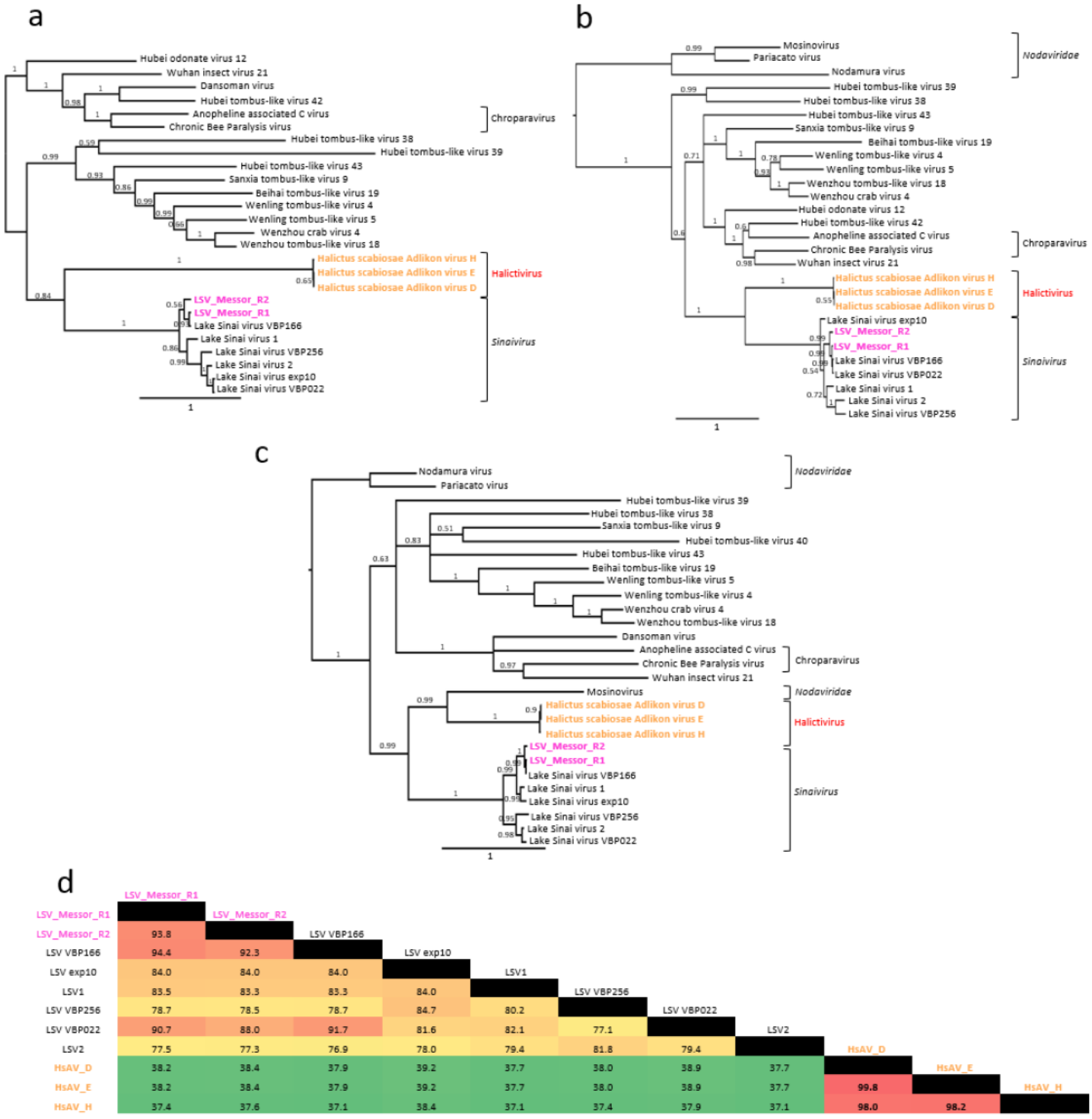


Fig. 2: Bayesian phylogenetic trees of ORF1, RdRp and capsid proteins of Halictus scabiosae Adlikon Virus and new *Lake Sinai virus* sequences found in *Messor* ants. (a) ORF1 (603 amino acid sites), (b) RNA-dependant RNA-polymerase phylogeny (559 amino acid sites). (c) Capsid phylogeny (884 amino acid sites). HsAV sequences are indicated in orange and LSV sequences from ants in pink. GenBank accessions are indicated in the supplementary Table S6. Scale bar represents substitutions rate per site and node values are posterior probabilities. (d) Matrix of protein identities of RdRp (%) between shared sequences of full-length genomes.

Analysis of protein identity of RdRp showed that while within-strain protein identity was high for HsAV (98.7%) and Sinaivirus (83.2%), there was only 38% nucleotide identity between shared sequences from the Sinaiviruses and HsAV genomes (Fig. 2d), suggesting they could belong to different genera. This is a proposal that should be examined by the relevant ICTV committee to determine the appropriate species/ genus demarcation criteria.

HsAV and LSV strain divergence

Nucleotide comparisons between full-length genomes of HsAV strains D, E and H revealed that the three sequences had accumulated 42 strain-specific SNPs widespread along the entire genomes. In HsAV_D, 14 synonymous and 4 non-synonymous SNPs were observed relative to the consensus sequence. The HsAV_E genome revealed 2 synonymous SNPs. HsAV_H contained 4 synonymous and 18 non-synonymous SNPs. This result suggests that the three assembled genomes are not contamination artifacts; for which nucleotide identity would be expected to approach 100%. The low level of polymorphism observed here confirmed that three different strains of a single virus species have been sequenced.

Molecular evolution suggests HsAV and ant LSVs are functional

Selective pressures acting on ORF1, RdRp and capsid of the newly discovered HsAV and ant LSV were estimated to verify if their evolutionary rates reflected those of functional infectious viruses.

As comparison, we first estimated the dN/dS ratio typical of *Sinaivirus*, Chroparavirus and *Nodaviridae* to identify reference selective pressures acting on infectious viruses. On average, dN/dS ratios of the latter viruses were 0.17 for ORF1, 0.09 for RdRp and 0.01 for capsid gene. ORFs of ant LSVs were on average more constrained than in other infectious viruses: dN/dS = 0.08 for ORF1 (Likelihood ratio test: $\chi2=2\Delta LnL=12.200$, p=4.8E-4), = 0.03 for RdRp (LRT, $\chi2=29.173$, p=6.6E-8), and = 0.01 for capsid (LRT, $\chi2=0.014$, p=0.9). RdRp of HsAV displayed similar selective constraints compared to infectious viruses: dN/dS = 0.03 for RdRp (LRT, $\chi2=2.541$, p=0.11), whereas ORF1 and capsid of HsAV seemed to evolve under slightly more relaxed selection: dN/dS = 0.3 (LRT, $\chi2=1.082$, p=0.3) and = 0.06 (LRT, $\chi2=4.008$, p=0.045) for ORF1 and capsid, respectively. Altogether, molecular evolution analyses of ant LSV and HsAV shows that dN/dS values were below 1 suggesting a selective regime of purifying evolution, as expected in functional infectious viruses.

Worldwide *Sinaivirus* genetic diversity

There are currently 58 LSV sequences available in public database, including 6 complete viral genomes, 41 sequences of the region overlapping ORF1/RdRp, 3 partial ORF1, 2 partial RdRp and 6 partial capsid sequences (Table S1). Most sequences were obtained from *Apis mellifera*, but a few come from the wild bees *Andrena vaga*, *Bombus lapidarius* and *B. pascuorum*. So far LSV sequences have been produced from only 3 countries: USA [12, 23, 25], Belgium [30–32, 37] and Spain [28, 29]. To increase both geographical and taxon sampling to improve phylogenetic resolution > 650 honey bees sampled worldwide were screened for ORF1/RdRp region. Thirty-six *A. mellifera* honey bees (pool or individual samples) were positive for LSV and were sequenced (Table S4). No LSV sequences from pooled honey bees produced electropherograms displaying double peaks that would indicate a mixture of different strains or species. LSV sequences were obtained from 5 new countries: 11 from France, 3 from Italy, 5 from Canada, 5 from China and 7 from Australia, confirming that LSVs have a very wide geographic distribution across continents. In total 81 LSV sequences, as well as the 3 HsAV sequences were collated into a 594 nucleotide alignment of the ORF1/RdRp region. Bayesian phylogenetic analyses were performed using the GTR+G substitution model. As shown above, HsAV is the most closely related virus to the genus Sinaivirus and thus was used for outgroup rooting of the LSV phylogeny. The phylogeny distinguished at least 4 LSV lineages, which we named Clades A to D. The ICTV currently recognizes only two LSV species: LSV1, which belongs to Clade C and LSV2 from Clade A. All four LSV clades were strongly supported by high posterior probabilities of 0.76, 0.94, 0.92 and 1, respectively (Fig. 3a).

Interestingly, each country contained LSVs from multiple clades and no evidence of geographical pattern could be associated with the four clades. The same pattern was observed at the continent scale as different LSV strains from 3 to 4 clades co-circulate in Europe, North America, Asia and Oceania (Fig. 3b). It should be noted that European LSVs from Belgium were overrepresented in Clades A, C and D, reflecting a higher sampling effort in this country [26].

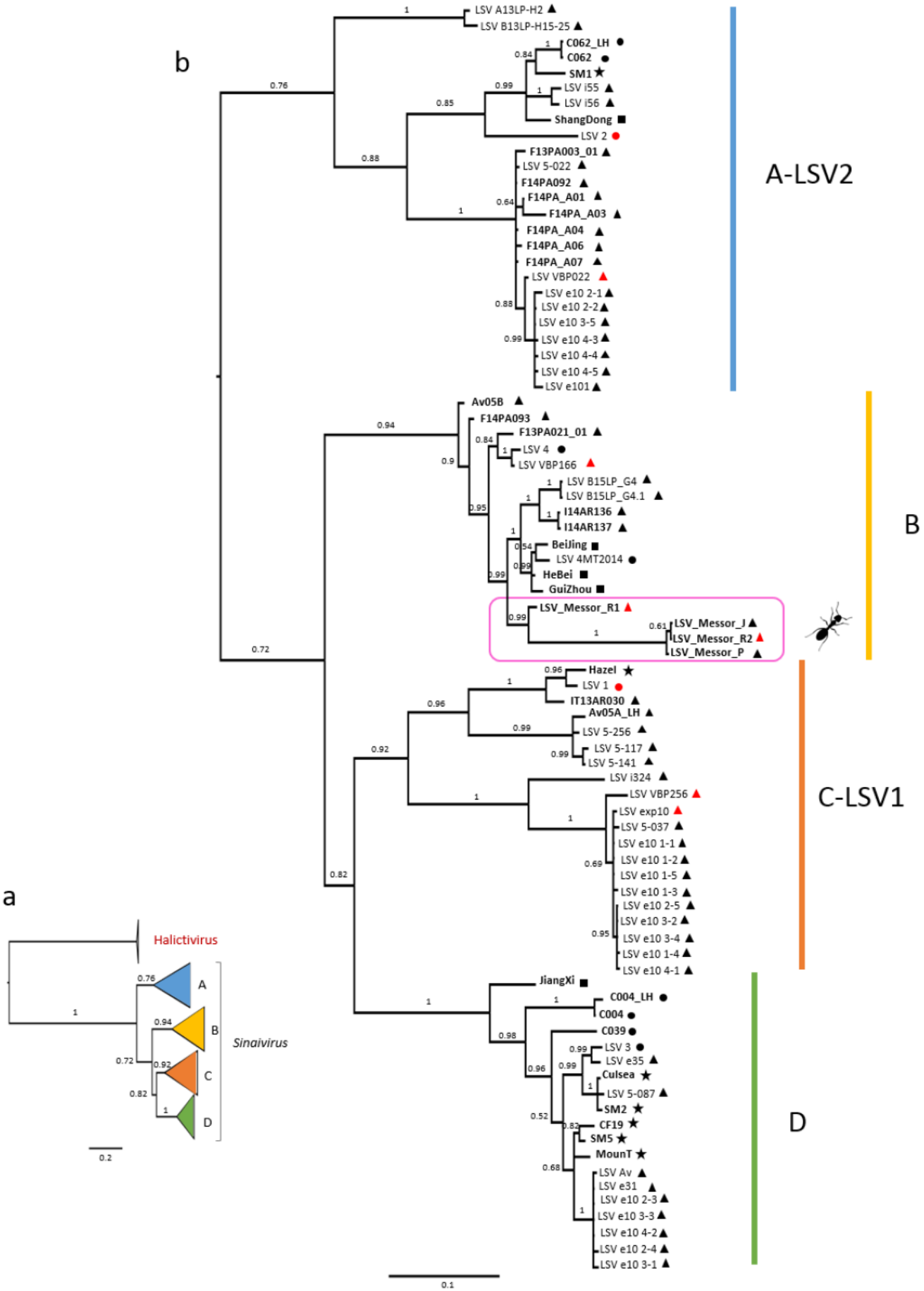

a
Halictivirus
A
B
C
D
Sinaivirus
0.2
b
LSV A13LP-H2
LSV B13LP-H15-25
C062_LH
C062
SM1
LSV i55
LSV i56
ShangDong
LSV 2
F13PA003_01
LSV 5-022
F14PA092
F14PA_A01
F14PA_A03
F14PA_A04
F14PA_A06
F14PA_A07
LSV VBP022
LSV e10 2-1
LSV e10 2-2
LSV e10 3-5
LSV e10 4-3
LSV e10 4-4
LSV e10 4-5
LSV e101
A-LSV2
Av05B
F14PA093
F13PA021_01
LSV 4
LSV VBP166
LSV B15LP_G4
LSV B15LP_G4.1
I14AR136
I14AR137
BeiJing
LSV 4MT2014
HeBei
GuiZhou
LSV_Messor_R1
LSV_Messor_J
LSV_Messor_R2
LSV_Messor_P
B
Hazel
LSV 1
IT13AR030
Av05A_LH
LSV 5-256
LSV 5-117
LSV 5-141
LSV i324
LSV VBP256
LSV exp10
LSV 5-037
LSV e10 1-1
LSV e10 1-2
LSV e10 1-5
LSV e10 1-3
LSV e10 2-5
LSV e10 3-2
LSV e10 3-4
LSV e10 1-4
LSV e10 4-1
C-LSV1
JiangXi
C004_LH
C004
C039
LSV 3
LSV e35
Culsea
LSV 5-087
SM2
CF19
SM5
MounT
LSV Av
LSV e31
LSV e10 2-3
LSV e10 3-3
LSV e10 4-2
LSV e10 2-4
LSV e10 3-1
D
0.1

Fig. 3: Bayesian phylogenetic tree of the ORF1/RdRp nucleotide region of all known and new *Lake Sinai virus* sequences. (a) Zoom of the LSV clade (594 nucleotide sites). New sequences from ants are indicated by pink box. Samples from Europe (France, Italy, Belgium, Spain) are indicated by triangles, from North America (Canada, USA) by circles, from Asia (China) by squares and from Oceania (Australia) by stars. Taxons in bold were sequenced in this study. Red symbols are full length LSV genomes. Taxon information for LSV sequences are in Table S1. (b) Collapsed phylogeny with HsAV as outgroup. Scale bar represents substitutions rate per site. Node values are posterior probabilities.

The 4 LSV sequences associated with *Messor* ants all belong to clade B. The two virus genomes discovered within the same ants, LSV_Messor_R1 and R2, were phylogenetically distinct (posterior probabilities=0.99), while the two partial sequences from two other *Messor* ants, LSV_Messor_J and P, were closely related to LSV_Messor_R2 (posterior probabilities=1) (Fig. 3b). In ants, the two LSV clades formed two strains named LSV Messor 1 (comprising LSV_Messor_R1 sequence) and LSV Messor 2 (comprising LSV_Messor_R2, _J and _P sequences).

In this study, two bee samples (C004, C062) produced PCR amplicons using two different primer pairs (LSV and LSV-HsAV noted -LH in the phylogeny). For both C004 and C062 samples, overlapping sequences from both primer pairs were nearly identical (100% and 99.2% nucleotide identity) and therefore clustered in the phylogeny, showing they result from a single virus population circulating in the bees.

Discussion

Halictivirus: a new viral genus

The generation of metagenomic data via the development of Next Generation Sequencing technologies has fueled the discovery of many new viruses [38–40]. However, to date few honey bee viruses have been discovered in this way. A new *Iflavirus* (ssRNA+), the *Moku virus*, was recently found in *A. mellifera*, in the mite *Varroa destructor* and in the wasp *Vespula pensylvanica* [41]. In addition, NGS allowed the discovery of four new RNA viruses: *Aphid Lethal Paralysis virus* strain Brookings (ALPV-Brookings), *Big Sioux River virus* (BSRV), *Lake Sinai virus* 1 and 2 (LSV1 and LSV2) in honey bees [23]. In the era of metagenomics where genomic and phylogenetic analyses are powerful and efficient, new viral genomes deserve attribution of genus and species names, even in the absence of additional biological information, microscopic descriptions or pathology [42].

Our current study allowed the discovery and the description of three isolates of Halictus scabiosae Adlikon Virus (HsAV). The genomic reconstructions, annotation and phylogenies permit complete description of these new viruses, phylogenetically closely related to Chroparavirus, *Sinaivirus* and *Nodaviridae*. HsAV is distinguishable from other closely related viruses by the absence of the MTase-GTase domain within the ORF1, putatively implicated in 5' cap formation [35, 36], and suggesting that it has another mechanism of initiation of translation. The stretch of Adenine at the 5' end of the HsAV genome might form a non-conventional poly(A) head initiating virus translation, similar to the poly(A) head, which significantly enhances cap-independent translation of mRNAs in some poxviruses [43]. Besides their specific genomic organisation, several evolutionary features distinguish HsAVs. HsAVs form a monophyletic group, are genetically homogenous and clearly divergent from their closest relatives the sinaiviruses, from which they are separated by long branches of equivalent length to those defining the Chroparavirus and *Sinaivirus* genera. Altogether this supports the proposal that HsAVs belong to a distinct and new viral genus, which we call Halictivirus. Molecular evolution analyses revealed that all HsAV proteins are subjected to strong purifying selection, suggesting that this virus is functional and infectious. This is also suggested by a high transcriptome coverage (Fig. S1). However, symptoms associated to HsAV remain to be elucidated.

***Lake Sinai virus* infect multiple and diverse hosts**

LSVs were discovered in three independently collected harvester ants: LSV_Messor_1 in *Messor concolor* and LSV_Messor_2 in *M. barbarus*, *M. capitatus* and *M. concolor*. This is the first time this virus has been reported from insects outside the superfamily Apoidea. LSV was discovered in the honey bee *Apis mellifera* in North America [12, 23, 25], in Europe [26, 29, 32] and in Africa [33]. Moreover, LSVs have been detected in wild solitary bees of the Andrenidae family (*Andrena vaga* and *A. ventralis*), in Megachilidae (*Osmia bicornis* and *O. cornuta*) in Belgium [31], and in Apidae bumble bees in Colombia (*Bombus atratus*) [34] and in Belgium (*B. lapidarius*, *B. pratorum* and *B. pascuorum*) [30]. Ant LSVs formed a monophyletic group and were all unequivocally incorporated within LSV clade D. Given that all other known LSVs are from bee hosts, this result suggests that host jump events between bees and ants may have occurred. Interestingly, one ant harboured two viral strains, showing that co-infection might also occur in ants.

Since its discovery in 2011 [23], LSV screening in non-bee insects is lacking. However other honey bee viruses have also been detected successfully in several other hymenopteran hosts, mostly non-*Apis* bees [16, 44]. *Israeli Acute Paralysis Virus* (IAPV) has also been reported in the wasp *Vespula vulgaris* [45, 46] and the replicative form of the virus was found in *Vespa velutina* [47]. The invasive hornet *V. velutina* mainly feeds on honey bees [48], but detection of replicative viral genomes excludes a simple trophic contamination. Honey bee virus detection in ants is also scarce, and to our knowledge, CBPV and DWV are the only honey bee viruses detected in ants. CBPV was found in *Formica rufa* (viral genome) and in *Camponotus vagus* (replicative genome) ants living close to apiaries [49]. The genome equivalent copy numbers of CBPV were comparable between ants and bees (103 to 1011 copy per individual) [49]. DWV was found in invasive Argentine ants *Linepithema humile* (replicative genome) [50]. DWV was found in New Zealand ants and a replicative form was found by strand-specific PCR in 7% of tested ants. Although no symptoms were observed in ants, the high copy number of the virus and the presence of viral replication suggest that honey bee viruses can infect ants. Additional studies, in which more samples should be analyzed, are needed to determine if LSV infections in ants are dead-ends or could participate in spreading the viruses in bees or other insects.

The discovery of ant LSV would clearly benefit from further wider sampling and detection of the replicative form of the virus using specific detection of the minus-strand RNA genome [51]. However, several lines of evidence already suggest that *Messor* ants are not simple passive trophic carriers of LSV. First, *Messor* ants are mainly granivorous, and dead bees would not be major foraging targets. Second, the *Messor* used for the transcriptome sequencing were not collected near apiaries. Third, ant LSVs were detected 3 times independently i) in 3 ant species, ii) sampled up to 2,000 km apart, and iii) displaying high between-strains polymorphism exceeding Illumina sequencing errors; thereby excluding cross contaminations during the experiment. Altogether, these findings provide strong arguments in favor of a genuine LSV infection in ants.

Cross-species transmissions of viruses have been shown to occur more frequently than previously thought and play a major role in evolution compared to rare co-divergence events [52]. Adaptations of RNA viruses to a new host in a new environment are enhanced by high mutation rates and fast viral replication by RNA polymerases [53]. In addition, close phylogenetic relatedness between hosts may also facilitate cell entry via similar receptors

[54]. Transmission vectors shared between host species can also mediate host switches. For instance the *Deformed wing virus* (DWV) is transmitted by the *Varroa destructor* mite [51, 55–57] but also via the environment through contaminated pollen [58]. Both transmission routes mediate DWV inoculation in non-*Apis* hymenopteran species [46]. In the case of LSV, both pollen pellets and *Varroa* mite can carry LSV particles, but LSV replicative forms were absent from these vectors. In addition, LSV presence in the honey bee gut could indicate that a potential food-associated (i.e. through pollen) and/or fecal-oral horizontal transmission route can occur for LSV [25]. This kind of transmission appears more random in the case of the harvester ants, which principally eat seeds. Carnivorous ants (*Camponotus vagus*) were found to be potential hosts of CBPV, as replicating forms of the virus were found in ants living near infected apiaries [49].

Resolution of *Sinaivirus* phylogeny and characterization of LSV diversity

By combining 47 separately published LSV sequences with 35 new LSV sequences from this work, the ORF1/RdRp phylogeny represents the most exhaustive characterization of LSV diversity so far. Furthermore, the use of the new Halictivirus as outgroup allowed better resolution of the sinaivirus tree topology. No recombination was detected in this dataset, legitimating inferences drawn from this *Sinaivirus* phylogeny. The phylogeny showed 4 main LSV clades, 3 of which correspond to the previously described clades A, C and D [26]. Clade B, which includes virus sequences from bees collected in the Northern hemisphere as well as in ants, is novel (Fig. 3). The ICTV currently recognizes two species within the new *Sinaivirus* genus: LSV1 and LSV2, respectively belonging to clades C and A. Our results suggest there are at least 2 additional LSV species corresponding to clades B and D, depending on the sequence divergence cut-off applied (Table S5). Previous work named some LSV sequences as LSV 5, but here multiple LSV 5 were dispersed in multiple clades and not corresponding to a distinct species. Altogether the phylogenetic analyses revealed the great diversity of sinaiviruses both in terms of species and strains, based on which taxonomic revision could be undertaken (Table S5).

Coinfections of a single host insect by distinct LSV strains or species was observed with the identification of both LSV Messor 1 and 2 strains (Clade B) in a single *M. concolor* ant. Occurrence of LSV coinfections from clades A, C and D in single honey bees have also been recently reported from Belgium [26]. This shows that LSV coinfections are relatively frequent,

whatever the level of relatedness between the viruses and whatever the hosts. As each species and each strain might have different pathology and virulence, this may complicate identification of symptoms associated with specific LSVs.

Strikingly, all four LSV clades have wide geographic distributions, revealed by our screening from several new countries. Moreover, all of the main clades were distributed across several continents. This confirms on a far wider geographic scale, the observations based on LSV sequences from Belgium and USA [26]. Recent honey bee trade such as import and export of queens or recurrent hive transports could explain the lack of geographical segregation of virus species. Notably, DWV also displays a global distribution of genotypes, reflecting a worldwide spread of viruses driven by *Varroa* mites [57, 59]. Interestingly, this heterogeneity in LSV and DWV genetic distribution contrasts with other bee viruses such as IAPV [60, 61], SBV [62], or *Black queen cell virus* (BQCV) [63], for which genetic diversity shows clear biogeographic structure. As no symptoms have yet been associated with LSVs, which was only discovered in 2011, there is no regulation yet to manage LSV spread. Further research is required on the pathology of *Sinaivirus* and Halictivirus to determine their impacts on honey bee and wild pollinator health.

Materials and Methods

Virus detection in bees and ants transcriptomes

The 43 transcriptomes used in this study were obtained from single adult insects (i.e. each individual was treated separately, without pooling) including 13 wild bees belonging to three *Halictus* species (Apoidea, Halictidae): *H. scabiosae*, *H. sexcinctus*, *H. simplex*, and 30 ants from 6 species: *Messor barbarus*, *M. concolor*, *M. structor*, *M. bouvieri*, *M. capitatus* and *Aphaenogaster subterrannea* (Formicidae). Twenty new transcriptomes were produced for this work to complement 23 previously published transcriptomes [64] (Table S3). Total RNA isolation of whole individual bees and ants was performed using standard protocols [65]. Succinctly, 50 nt single-end reads were produced by an Illumina Hiseq 2000 sequencer after cDNA synthesis using the SMART cDNA library Construction kit (Clontech, Mountain View, USA) from 5 µg of total RNA [64]. The 20 new transcriptomes were de novo assembled using the same method as previously [64] that is assembly with ABYSS V1.2.0 [66, 67] with Kmer set at 40 [68] and contig re-assembly with CAP3 program [69]. Open Reading Frames (ORFs) were predicted on assembled contigs of the 43 transcriptomes using Prodigal V2_60 software for

metagenomic data [70, 71] using the standard genetic code. Translated ORFs were annotated based on protein homology using the HHblits program implemented in the HHSuite package [72, 73]. To minimize false-positive results only ORFs displaying homology e-values <10-5 and probability >95% were kept. Significant positive homology hits were then parsed to retrieve their NCBI taxonomic identifiers (TaxID; ftp://ftp.ncbi.nih.gov/pub/taxonomy) using the BLAST+ program [74], and the corresponding taxonomic identification was assigned to the predicted ORFs. Viral ORFs were kept for further analyses. When multiple hits of the same viral family occurred in a single transcriptome, full-length viral genomes were reconstructed by assembly of the corresponding contigs into scaffolds (Geneious assembler program) and extension by successive mappings of initial reads (Geneious mapper program) using default parameters of Geneious® 8.1.7 software [75]. A final mapping of all Illumina reads of initial transcriptomes was performed using the previously extended viral genome as a reference sequence to validate the accuracy of genome reconstruction and correct for mapping errors.

Genome annotation, phylogeny and molecular diversity of new viruses

In order to annotate new full-length viral genomes, conserved protein domains of all predicted genes were searched against the 14 protein domain databases available in the InterPro consortium [76] using InterProScan version 5 [77]. Multiple protein alignments were performed with MAFFT [78] using default parameters on ORF1, RdRp and Capsid ORFs (Table S6). The best amino-acid substitution model was predicted using ProtTest [79]. Bayesian phylogenetic trees were inferred using MrBayes version 3.2.6 [80], by running four Markov chains for 106 generations. Branch support values indicate posterior probabilities estimated from trees sampled every 20 generations once the Markov chains had become stationary (determined by empirical checking of likelihood values).

By comparing complete aligned viral genomes, polymorphisms between HsAV strains were analyzed. In order to assess the functionality of the three ORFs of HsAV and LSV associated with ants, non-synonymous to synonymous substitution rates (dN/dS) were estimated to quantify selection pressures. The PAL2NAL program [81] was first used to guide codon alignments using protein alignments. dN/dS was then estimated from codon alignments using branch-models [82, 83] of the CodeML program [84, 85] implemented in the PAML software version 4.9c [86]. Finally, different nested models were used to compare dN/dS of the branches of interest (newly discovered HsAV and LSV found in ants) to those of reference

viruses (LSVs and the closely-related Chroparavirus and *Nodaviridae*). Model comparisons were performed using Likelihood Ratio Tests (LRTs), using $\chi 2$ tests with type I error = 0.05, df= 1 (i.e. the difference of number of parameters between two models) and the test statistics $\chi 2 = 2\Delta LnL$ (i.e. twice the difference of the Log-Likelihood of each model).

Large-scale *de novo* detection and phylogeny of *Lake Sinai virus* in honey bees

We screened for LSV sequences from 569 Apis mellifera honey bees sampled in France, Italy, Canada, China, and Australia. Individual or pooled bees were sampled in the summers 2013 to 2016 (Table S4). Due to the worldwide sampling effort, biological material underwent distinct processes, summarized in Table S7 for simplicity. Briefly, individual or pooled bees were mechanically disrupted and homogenized in lysis buffer and total RNA was isolated according to kit manufacturer instructions using phenol/chlorophorm or guanidinium thiocyanate protocols. Total RNA was quantified using the Qubit Fluorometer or Nanodrop and 1-4 µg of total RNA was reverse-transcribed using random hexamer primers following the reverse transcription kit manuals.

PCR detection of multiple strains of LSV and/or Sinaiviruses (LSV-HsAV) was performed using custom degenerate primers (Fig. 1, Table S8) targeting the region overlapping ORF1/RdRp, commonly used in LSV genetic studies [32]. PCR reaction mixes and cycling conditions (identical for LSV or LSV-HsAV primers) are detailed in Table S7. PCR products were analyzed by electrophoresis in 1.5% agarose gels, stained with GelRed and visualized under UV light. All positive PCRs were Sanger sequenced by GATC Biotech (Germany), Sangon Biotech (China) or the Hawkesbury Institute for the Environment (Australia) using forward M13FP and reverse M13-RP primers. Sequences from both strands were assembled using DNAman software package, version 6.0.3 (Lynnon BioSoft, http://www.lynnon.com). Electropherograms were manually corrected and ambiguities were replaced by N using Geneious R9 [75].

Multiple nucleotide alignments were performed with MAFFT [78] using default parameters on ORF1/RdRp sequences. In order to draw proper conclusions from the phylogeny, recombination was detected using the GARD program [87] implemented in the Datamonkey web server [88]. The best evolutionary model was predicted using JModelTest v2 [89]. A Bayesian phylogenetic tree was inferred using MrBayes version 3.2.6 [80] as described above.

Funding information

This work was supported by two European Research Council (ERC) grants to Nicolas Galtier (ERC PopPhyl 232971) and to EAH (ERC GENOVIR 205206), to HCS (NSFC 31572471), and to DQY (CAAS-ASTIP-2017-IAR). The funders played no role in the study or in the preparation of the article or decision to publish. This work was supported by a Jean & Marie-Louise Dufrenoy grant to DB from the French Agriculture Academy.

Acknowledgements

We would like to thank Nicolas Galtier for data acquisition, Jonathan Romiguier for help in data accessibility and Marion Ballenghien for technical assistance. Analyses largely benefited from the ISEM computing cluster platform. We are also grateful to the Genotoul Bioinformatics Platform Toulouse Midi-Pyrenees for providing computing and storage resources. We would thank the ARNIA company for Italian honey bee samples and ADAPI (Association pour le Développement de l'Apiculture) for honey bee samples in south of France.

References

1. Gallai N, Salles JM, Settele J, Vaissière BE: Economic valuation of the vulnerability of world agriculture confronted with pollinator decline. Ecol Econ 2009, 68:810–821.
2. Klein A-M, Vaissiere BE, Cane JH, Steffan-Dewenter I, Cunningham SA, Kremen C, Tscharntke T: Importance of pollinators in changing landscapes for world crops. Proc R Soc B Biol Sci 2007, 274:303–313.
3. Core A, Runckel C, Ivers J, Quock C, Siapno T, DeNault S, Brown B, DeRisi J, Smith CD, Hafernik J: A new threat to honey bees, the parasitic phorid fly Apocephalus borealis. PLoS One 2012, 7:1–9.
4. Doublet V, Labarussias M, de Miranda JR, Moritz RFA, Paxton RJ: Bees under stress: Sublethal doses of a neonicotinoid pesticide and pathogens interact to elevate honey bee mortality across the life cycle. Environ Microbiol 2015, 17:969–983.
5. Evans JD, Schwarz RS: Bees brought to their knees: Microbes affecting honey bee health. Trends Microbiol 2011, 19:614–620.
6. Fairbrother A, Purdy J, Anderson T, Fell R: Risks of neonicotinoid insecticides to honeybees. Environ Toxicol Chem 2014, 33:719–31.
7. Goulson D, Nicholls E, Botias C, Rotheray EL: Bee declines driven by combined stress from parasites, pesticides, and lack of flowers. Science 2015, 347:1255957–1255957.
8. Kielmanowicz MG, Inberg A, Lerner IM, Golani Y, Brown N, Turner CL, Hayes GJR, Ballam JM: Prospective large-scale field study generates predictive model identifying major contributors to colony losses. PLOS Pathog 2015, 11:e1004816.
9. Menail AH, Piot N, Meeus I, Smagghe G, Loucif-Ayad W: Large pathogen screening reveals first report of Megaselia scalaris (Diptera: Phoridae) parasitizing Apis mellifera intermissa (Hymenoptera: Apidae). J Invertebr Pathol 2016, 137:33–37.
10. Sánchez-Bayo F, Goulson D, Pennacchio F, Nazzi F, Goka K, Desneux N: Are bee diseases linked to pesticides? - A brief review. Environ Int 2016, 89–90:7–11.
11. VanEngelsdorp D, Meixner MD: A historical review of managed honey bee populations in Europe and the United States and the factors that may affect them. J Invertebr Pathol 2010, 103(SUPPL. 1):S80–S95.
12. Cornman RS, Tarpy DR, Chen Y, Jeffreys L, Lopez D, Pettis JS, VanEngelsdorp D, Evans JD: Pathogen webs in collapsing honey bee colonies. PLoS One 2012, 7:e43562.
13. Genersch E, Aubert M: Emerging and re-emerging viruses of the honey bee (Apis mellifera L.). Vet Res 2010, 41:54.
14. McMenamin AJ, Genersch E: Honey bee colony losses and associated viruses. Curr Opin Insect Sci 2015, 8:121–129.
15. Potts SG, Biesmeijer JC, Kremen C, Neumann P, Schweiger O, Kunin WE: Global pollinator declines: trends, impacts and drivers. Trends Ecol Evol 2010, 25:345–353.

16. Tehel A, Brown MJF, Paxton RJ: Impact of managed honey bee viruses on wild bees. Curr Opin Virol 2016, 19:16–22.
17. Bailey L, Gibbs A, Woods R: Two viruses from adult honey bees (Apis mellifera Linnaeus). Virology 1963, 21:390–395.
18. de Miranda JR, Bailey L, Ball B V, Blanchard P, Budge GE, Chejanovsky N, Chen Y-P, Gauthier L, Genersch E, de Graaf, DC, Ribière M, Ryabov E, De Smet L, van der Steen JJM: Standard methods for virus research in Apis mellifera. J Apic Res 2013, 52:1–56.
19. Remnant EJ, Shi M, Buchmann G, Blacquière T, Holmes EC, Beekman M, Ashe A: A Diverse Range of Novel RNA Viruses in Geographically Distinct Honey Bee Populations. J Virol 2017(May):JVI.00158-17.
20. Brutscher LM, McMenamin AJ, Flenniken ML, Goulson D, Nicholls E, Botías C, Rotheray E, Klein A, Vaissière B, Cane J, Steffan-Dewenter I, Cunningham S, Kremen C, Gallai N, Salles J-M, Settele J, Vaissière B, Lee K, Steinhauer N, Rennich K, Wilson M, Tarpy D, Caron D, Traynor K, Rennich K, Forsgren E, Rose R, Pettis J, VanEngelsdorp D, Evans J, et al.: The buzz about honey bee viruses. PLOS Pathog 2016, 12:e1005757.
21. Clark TB: A filamentous virus of the honey bee. J Invertebr Pathol 1978, 32:332–340.
22. Gauthier L, Cornman S, Hartmann U, Cousserans F, Evans JD, De Miranda JR, Neumann P: The Apis mellifera filamentous virus genome. Viruses 2015, 7:3798–3815.
23. Runckel C, Flenniken ML, Engel JC, Ruby JG, Ganem D, Andino R, DeRisi JL: Temporal analysis of the honey bee microbiome reveals four novel viruses and seasonal prevalence of known viruses, Nosema, and Crithidia. PLoS One 2011, 6:e20656.
24. Traynor KS, Rennich K, Forsgren E, Rose R, Pettis J, Kunkel G, Madella S, Evans J, Lopez D, VanEngelsdorp D: Multiyear survey targeting disease incidence in US honey bees. Apidologie 2016, 47:325–347.
25. Daughenbaugh KF, Martin M, Brutscher LM, Cavigli I, Garcia E, Lavin M, Flenniken ML: Honey bee infecting Lake Sinai Viruses. Viruses 2015, 7:3285–309.
26. Ravoet J, De Smet L, Wenseleers T, de Graaf DC: Genome sequence heterogeneity of Lake Sinai Virus found in honey bees and Orf1/RdRP-based polymorphisms in a single host. Virus Res 2015, 201:67–72.
27. Cavigli I, Daughenbaugh KF, Martin M, Lerch M, Banner K, Garcia E, Brutscher LM, Flenniken ML: Pathogen prevalence and abundance in honey bee colonies involved in almond pollination. Apidologie 2016, 47:251–266.
28. Cepero A, Ravoet J, Gómez-Moracho T, Bernal J, Del Nozal MJ, Bartolomé C, Maside X, Meana A, González-Porto A V, de Graaf DC, Martín-Hernández R, Higes M: Holistic screening of collapsing honey bee colonies in Spain: a case study. BMC Res Notes 2014, 7:649.
29. Granberg F, Vicente-Rubiano M, Rubio-Guerri C, Karlsson OE, Kukielka D, Belák S, Sánchez-Vizcaíno JM: Metagenomic detection of viral pathogens in Spanish honeybees: Co-infection by Aphid Lethal Paralysis, Israel Acute Paralysis and Lake Sinai Viruses. PLoS One 2013, 8:e57459.
30. Parmentier L, Smagghe G, de Graaf DC, Meeus I: Varroa destructor Macula-like virus, Lake Sinai virus and other new RNA viruses in wild bumblebee hosts (Bombus pascuorum, Bombus lapidarius and Bombus pratorum). J Invertebr Pathol 2016, 134:6–11.
31. Ravoet J, De Smet L, Meeus I, Smagghe G, Wenseleers T, de Graaf DC: Widespread occurrence of honey bee pathogens in solitary bees. J Invertebr Pathol 2014, 122:55–58.
32. Ravoet J, Maharramov J, Meeus I, De Smet L, Wenseleers T, Smagghe G, de Graaf DC: Comprehensive bee pathogen screening in Belgium reveals Crithidia mellificae as a new contributory factor to winter mortality. PLoS One 2013, 8:e72443.
33. Amakpe F, De Smet L, Brunain M, Ravoet J, Jacobs FJ, Reybroeck W, Sinsin B, de Graaf DC: Discovery of Lake Sinai virus and an unusual strain of Acute Bee Paralysis virus in West African apiaries. Apidologie 2015:35–47.
34. Gamboa V, Ravoet J, Brunain M, Smagghe G, Meeus I, Figueroa J, Riaño D, de Graaf DC: Bee pathogens found in Bombus atratus from Colombia: A case study. J Invertebr Pathol 2015, 129:36–39.
35. Ahola T, Karlin DG: Sequence analysis reveals a conserved extension in the capping enzyme of the alphavirus supergroup, and a homologous domain in nodaviruses. Biol Direct 2015, 10:16.
36. Kuchibhatla DB, Sherman WA, Chung BYW, Cook S, Schneider G, Eisenhaber B, Karlin DG: Powerful sequence similarity search methods and in-depth manual analyses can identify remote homologs in many apparently “orphan” viral proteins. J Virol 2014, 88:10–20.
37. Ravoet J, De Smet L, Wenseleers T, de Graaf DC: Vertical transmission of honey bee viruses in a Belgian queen breeding program. BMC Vet Res 2015, 11:61.
38. Edwards RA, Rohwer F: Opinion: Viral metagenomics. Nat Rev Microbiol 2005, 3:504–510.
39. Mokili JL, Rohwer F, Dutilh BE: Metagenomics and future perspectives in virus discovery. Curr Opin Virol 2012, 2:63–77.
40. Rosario K, Breitbart M: Exploring the viral world through metagenomics. Curr Opin Virol 2011, 1:289–297.

41. Mordecai GJ, Brettell LE, Pachori P, Villalobos EM, Martin SJ, Jones IM, Schroeder DC: Moku virus; a new Iflavirus found in wasps, honey bees and Varroa. Sci Rep 2016, 6(October):34983.
42. Simmonds P, Adams MJ, Benkő M, Breitbart M, Brister JR, Carstens EB, Davison AJ, Delwart E, Gorbalenya AE, Harrach B, Hull R, King AMQ, Koonin E V., Krupovic M, Kuhn JH, Lefkowitz EJ, Nibert ML, Orton R, Roossinck MJ, Sabanadzovic S, Sullivan MB, Suttle CA, Tesh RB, van der Vlugt RA, Varsani A, Zerbini FM: Consensus statement: Virus taxonomy in the age of metagenomics. Nat Rev Microbiol 2017, 15:161–168.
43. Shirokikh NE, Spirin AS: Poly(A) leader of eukaryotic mRNA bypasses the dependence of translation on initiation factors. Proc Natl Acad Sci U S A 2008, 105:10738–43.
44. Manley R, Boots M, Wilfert L: Emerging viral disease risk to pollinating insects: ecological, evolutionary and anthropogenic factors. J Appl Ecol 2015, 52:331–340.
45. Evison SEF, Roberts KE, Laurenson L, Pietravalle S, Hui J, Biesmeijer JC, Smith JE, Budge G, Hughes WOH: Pervasiveness of parasites in pollinators. PLoS One 2012, 7:e30641.
46. Singh R, Levitt AL, Rajotte EG, Holmes EC, Ostiguy N, VanEngelsdorp D, Lipkin WI, DePamphilis CW, Toth AL, Cox-Foster DL: RNA Viruses in Hymenopteran pollinators: Evidence of inter-taxa virus transmission via pollen and potential impact on non-Apis Hymenopteran species. PLoS One 2010, 5:e14357.
47. Yañez O, Zheng HQ, Hu FL, Neumann P, Dietemann V: A scientific note on Israeli acute paralysis virus infection of Eastern honeybee Apis cerana and vespine predator Vespa velutina. Apidologie 2012, 43:587–589.
48. Monceau K, Bonnard O, Thiéry D: Vespa velutina: a new invasive predator of honeybees in Europe. J Pest Sci (2004) 2014, 87:1–16.
49. Celle O, Blanchard P, Olivier V, Schurr F, Cougoule N, Faucon J-P, Ribière M: Detection of Chronic bee paralysis virus (CBPV) genome and its replicative RNA form in various hosts and possible ways of spread. Virus Res 2008, 133:280–4.
50. Sébastien A, Lester PJ, Hall RJ, Wang J, Moore NE, Gruber MAM: Invasive ants carry novel viruses in their new range and form reservoirs for a honeybee pathogen. Biol Lett 2015, 11:20150610.
51. Yue C, Genersch E: RT-PCR analysis of Deformed wing virus in honeybees (Apis mellifera) and mites (Varroa destructor). J Gen Virol 2005, 86:3419–3424.
52. Geoghegan JL, Duchêne S, Holmes EC: Comparative analysis estimates the relative frequencies of co-divergence and cross-species transmission within viral families. PLOS Pathog 2017, 13:e1006215.
53. Drake JW, Holland JJ: Mutation rates among RNA viruses. Proc Natl Acad Sci 1999, 96:13910–13913.
54. Parrish CR, Holmes EC, Morens DM, Park E-C, Burke DS, Calisher CH, Laughlin CA, Saif LJ, Daszak P: Cross-species virus transmission and the emergence of new epidemic diseases. Microbiol Mol Biol Rev 2008, 72:457–470.
55. Boncristiani HF, Di Prisco G, Pettis JS, Hamilton M, Chen YP: Molecular approaches to the analysis of deformed wing virus replication and pathogenesis in the honey bee, Apis mellifera. Virol J 2009, 6:221.
56. Gisder S, Aumeier P, Genersch E: Deformed wing virus: Replication and viral load in mites (Varroa destructor). J Gen Virol 2009, 90:463–467.
57. Wilfert L, Long G, Leggett HC, Schmid-Hempel P, Butlin R, Martin SJM, Boots M: Deformed wing virus is a recent global epidemic in honeybees driven by Varroa mites. Science 2016, 351:594–597.
58. Mazzei M, Carrozza ML, Luisi E, Forzan M, Giusti M, Sagona S, Tolari F, Felicioli A: Infectivity of DWV associated to flower pollen: Experimental evidence of a horizontal transmission route. PLoS One 2014, 9:e113448.
59. Berenyi O, Bakonyi T, Derakhshifar I, Köglberger H, Topolska G, Ritter W, Pechhacker H, Nowotny N: Phylogenetic analysis of Deformed Wing Virus genotypes from diverse geographic origins indicates recent global distribution of the virus. Appl Environ Microbiol 2007, 73:3605–3611.
60. Chen YP, Pettis JS, Corona M, Chen WP, Li CJ, Spivak M, Visscher PK, DeGrandi-Hoffman G, Boncristiani H, Zhao Y, VanEngelsdorp D, Delaplane K, Solter L, Drummond F, Kramer M, Lipkin WI, Palacios G, Hamilton MC, Smith B, Huang SK, Zheng HQ, Li JL, Zhang X, Zhou AF, Wu LY, Zhou JZ, Lee M-L, Teixeira EW, Li ZG, Evans JD: Israeli Acute Paralysis Virus: Epidemiology, Pathogenesis and Implications for Honey Bee Health. PLoS Pathog 2014, 10:e1004261.
61. Palacios G, Hui J, Quan PL, Kalkstein A, Honkavuori KS, Bussetti A V, Conlan S, Evans J, Chen YP, VanEngelsdorp D, Efrat H, Pettis J, Cox-Foster D, Holmes EC, Briese T, Lipkin WI: Genetic analysis of Israel acute paralysis virus: distinct clusters are circulating in the United States. J Virol 2008, 82:6209–17.
62. Roberts JMK, Anderson DL: A novel strain of sacbrood virus of interest to world apiculture. J Invertebr Pathol 2014, 118:71–74.
63. Reddy KE, Noh JH, Choe SE, Kweon CH, Yoo MS, Doan HTT, Ramya M, Yoon B-S, Nguyen LTK, Nguyen TTD, Van Quyen D, Jung S-C, Chang K-Y, Kang SW: Analysis of the complete genome sequence and capsid region of black queen cell viruses from infected honeybees (Apis mellifera) in Korea. Virus Genes 2013, 47:126–132.

64. Romiguier J, Lourenco J, Gayral P, Faivre N, Weinert LA, Ravel S, Ballenghien M, Cahais V, Bernard A, Loire E, Keller L, Galtier N: Population genomics of eusocial insects: the costs of a vertebrate-like effective population size. J Evol Biol 2014, 27:593–603.
65. Gayral P, Weinert L, Chiari Y, Tsagkogeorga G, Ballenghien M, Galtier N: Next-generation sequencing of transcriptomes: a guide to RNA isolation in nonmodel animals. Mol Ecol Resour 2011, 11:650–661.
66. Birol I, Jackman SD, Nielsen CB, Qian JQ, Varhol R, Stazyk G, Morin RD, Zhao Y, Hirst M, Schein JE, Horsman DE, Connors JM, Gascoyne RD, Marra M a., Jones SJM: De novo transcriptome assembly with ABySS. Bioinformatics 2009, 25:2872–2877.
67. Simpson JT, Wong K, Jackman SD, Schein JE, Jones SJM, Birol I: ABySS: a parallel assembler for short read sequence data. Genome Res 2009, 19:1117–1123.
68. Cahais V, Gayral P, Tsagkogeorga G, Melo-Ferreira J, Ballenghien M, Weinert L, Chiari Y, Belkhir K, Ranwez V, Galtier N: Reference-free transcriptome assembly in non-model animals from next-generation sequencing data. Mol Ecol Resour 2012, 12:834–45.
69. Huang X, Madan A: CAP3: a DNA sequence assembly program. Genome Res 1999, 9:868–877.
70. Hyatt D, Chen G-L, Locascio PF, Land ML, Larimer FW, Hauser LJ: Prodigal: prokaryotic gene recognition and translation initiation site identification. BMC Bioinformatics 2010, 11:119.
71. Hyatt D, Locascio PF, Hauser LJ, Uberbacher EC: Gene and translation initiation site prediction in metagenomic sequences. Bioinformatics 2012, 28:2223–2230.
72. Remmert M, Biegert A, Hauser A, Söding J: HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment. Nat Methods 2011, 9:173–175.
73. Söding J: Protein homology detection by HMM-HMM comparison. Bioinformatics 2005, 21:951–960.
74. Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL: BLAST+: architecture and applications. BMC Bioinformatics 2009, 10:421.
75. 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, Mentjies P, Drummond A: Geneious Basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data. Bioinformatics 2012, 28:1647–1649.
76. Finn RD, Attwood TK, Babbitt PC, Bateman A, Bork P, Bridge AJ, Chang HY, Dosztanyi Z, El-Gebali S, Fraser M, Gough J, Haft D, Holliday GL, Huang H, Huang X, Letunic I, Lopez R, Lu S, Marchler-Bauer A, Mi H, Mistry J, Natale DA, Necci M, Nuka G, Orengo CA, Park Y, Pesseat S, Piovesan D, Potter SC, Rawlings ND, et al.: InterPro in 2017-beyond protein family and domain annotations. Nucleic Acids Res 2017, 45:D190–D199.
77. Jones P, Binns D, Chang H-Y, Fraser M, Li W, McAnulla C, McWilliam H, Maslen J, Mitchell A, Nuka G, Pesseat S, Quinn AF, Sangrador-Vegas A, Scheremetjew M, Yong S-Y, Lopez R, Hunter S: InterProScan 5: genome-scale protein function classification. Bioinformatics 2014, 30:1236–1240.
78. Katoh K, Misawa K, Kuma K, Miyata T: MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. Nucleic Acids Res 2002, 30:3059–3066.
79. Abascal F, Zardoya R, Posada D: ProtTest: Selection of best-fit models of protein evolution. Bioinformatics 2005, 21:2104–2105.
80. Huelsenbeck JP, Ronquist F: MRBAYES: Bayesian inference of phylogenetic trees. Bioinformatics 2001, 17:754–755.
81. Suyama M, Torrents D, Bork P: PAL2NAL: robust conversion of protein sequence alignments into the corresponding codon alignments. Nucleic Acids Res 2006, 34(Web Server issue):W609-12.
82. Kryazhimskiy S, Plotkin JB: The population genetics of dN/dS. PLoS Genet 2008, 4:e1000304.
83. Nei M, Gojobori T: Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. Mol Biol Evol 1986, 3:418–426.
84. Yang Z: Likelihood ratio tests for detecting positive selection and application to primate lysozyme evolution. Mol Biol Evol 1998, 15:568–573.
85. Yang Z, Bielawski JP: Statistical methods for detecting molecular adaptation. Trends Ecol Evol 2000, 15:496–503.
86. Yang Z: PAML 4: Phylogenetic Analysis by Maximum Likelihood. Mol Biol Evol 2007, 24:1586–1591.
87. Kosakovsky Pond SL, Posada D, Gravenor MB, Woelk CH, Frost SDW: GARD: A genetic algorithm for recombination detection. Bioinformatics 2006, 22:3096–3098.
88. Delport W, Poon AFY, Frost SDW, Kosakovsky Pond SL: Datamonkey 2010: a suite of phylogenetic analysis tools for evolutionary biology. Bioinformatics 2010, 26:2455–7.

89. Darriba D, Taboada GL, Doallo R, Posada D: jModelTest 2: more models, new heuristics and parallel computing. Nat Methods 2012, 9:772.

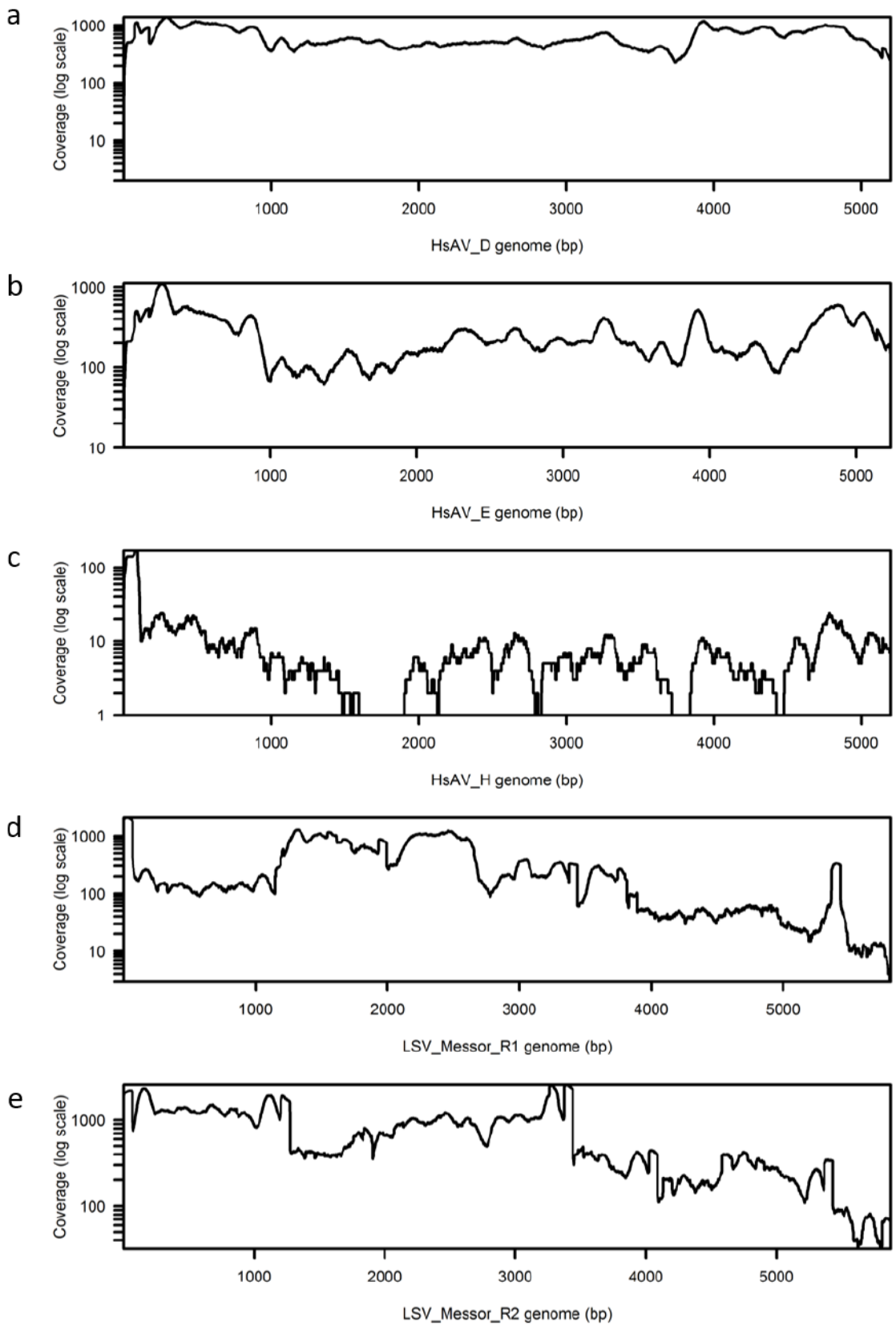


Fig. S1: Log-transformed Illumina read coverage of all new complete sequences. (a) HsAV_D, (b) HsAV_E, (c) HsAV_H, (d) LSV_Messor_R1, and (e) LSV_Messor_R2. Of note, graph C is not on the same scale as the others.

Table S1: Summary of all published LSV sequences since its discovery.

Acronyme	Full name	Part of the genome	Genbank Accession	Country	Host	Reference
LSV 1	*Lake Sinai Virus 1*	Complete	HQ871931	USA	*Apis mellifera*	(Runckel *et al.*, 2011)
LSV 2	*Lake Sinai Virus 2 strain BruceSD_T17E01*	Complete	HQ888865			
LSV 3	*Lake Sinai Virus 3 clone BRL-1-15-12*	Partial (ORF1/RdRp)	JQ480620	USA	*Apis mellifera*	(Comman *et al.*, 2012)
LSV 4	*Lake Sinai Virus 4*	Partial (ORF1/RdRp)	JX878492	Belgium	*Apis mellifera*	(Ravoet *et al.*, 2013)
LSV 5-022	*Lake Sinai Virus 5 JR 2013 isolate LSV022*	Partial (ORF1/RdRp)	KC880121			
LSV 5-037	*Lake Sinai Virus 5 JR 2013 isolate LSV037*	Partial (ORF1/RdRp)	KC880122			
LSV 5-087	*Lake Sinai Virus 5 JR 2013 isolate LSV087*	Partial (ORF1/RdRp)	KC880123			
LSV 5-117	*Lake Sinai Virus 5 JR 2013 isolate LSV117*	Partial (ORF1/RdRp)	KC880124			
LSV 5-141	*Lake Sinai Virus 5 JR 2013 isolate LSV141*	Partial (ORF1/RdRp)	KC880125			
LSV 5-256	*Lake Sinai Virus 5 JR 2013 isolate LSV256*	Partial (ORF1/RdRp)	KC880126			
LSV Navarra	*Lake Sinai Virus strain Navarra isolate 4782*	Partial (ORF1)	JX045859	Spain	*Apis mellifera*	(Granberg *et al.*, 2013)
LSV e31	*Lake Sinai Virus isolate e31*	Partial (ORF1/RdRp)	KF768348	Belgium	*Apis mellifera*	(Ravoet *et al.*, 2014)
LSV e35	*Lake Sinai Virus isolate e35*	Partial (ORF1/RdRp)	KF768349			
LSV e101	*Lake Sinai Virus isolate e101*	Partial (ORF1/RdRp)	KF768351			
LSV Av	*Lake Sinai Virus isolate Av*	Partial (ORF1/RdRp)	KF768350		*Andrena vaga*	
LSV i324	*Lake Sinai Virus isolate 324*	Partial (ORF1/RdRp)	KJ561227	Spain	*Apis mellifera*	(Cepero *et al.*, 2014)
LSV i55	*Lake Sinai Virus isolate 55*	Partial (ORF1/RdRp)	KJ561228			
LSV i56	*Lake Sinai Virus isolate 56*	Partial (ORF1/RdRp)	KJ561229			
LSV exp10	*Lake Sinai Virus strain exp10*	Complete	KM886905	Belgium	*Apis mellifera*	(Ravoet *et al.*, 2015)
LSV VBP166	*Lake Sinai Virus strain VBP166*	Complete	KM886903			
LSV VBP256	*Lake Sinai Virus strain VBP256*	Complete	KM886904			
LSV VBP022	*Lake Sinai Virus strain VBP022*	Complete	KM886902			
LSV e10_1-1	*Lake Sinai Virus strain LSVe10 clone Bee1-1*	Partial (ORF1/RdRp)	KM886906			
LSV e10_1-2	*Lake Sinai Virus strain LSVe10 clone Bee1-2*	Partial (ORF1/RdRp)	KM886907			
LSV e10_1-3	*Lake Sinai Virus strain LSVe10 clone Bee1-3*	Partial (ORF1/RdRp)	KM886908			
LSV e10_1-4	*Lake Sinai Virus strain LSVe10 clone Bee1-4*	Partial (ORF1/RdRp)	KM886909			
LSV e10_1-5	*Lake Sinai Virus strain LSVe10 clone Bee1-5*	Partial (ORF1/RdRp)	KM886910			
LSV e10_2-1	*Lake Sinai Virus strain LSVe10 clone Bee2-1*	Partial (ORF1/RdRp)	KM886911			
LSV e10_2-2	*Lake Sinai Virus strain LSVe10 clone Bee2-2*	Partial (ORF1/RdRp)	KM886912			
LSV e10_2-3	*Lake Sinai Virus strain LSVe10 clone Bee2-3*	Partial (ORF1/RdRp)	KM886913			
LSV e10_2-4	*Lake Sinai Virus strain LSVe10 clone Bee2-4*	Partial (ORF1/RdRp)	KM886914			
LSV e10_2-5	*Lake Sinai Virus strain LSVe10 clone Bee2-5*	Partial (ORF1/RdRp)	KM886915			
LSV e10_3-1	*Lake Sinai Virus strain LSVe10 clone Bee3-1*	Partial (ORF1/RdRp)	KM886916			

LSV e10_3-2	*Lake Sinai Virus strain LSVe10 clone Bee3-2*	Partial (ORF1/RdRp)	KM886917			
LSV e10_3-3	*Lake Sinai Virus strain LSVe10 clone Bee3-3*	Partial (ORF1/RdRp)	KM886918			
LSV e10_3-4	*Lake Sinai Virus strain LSVe10 clone Bee3-4*	Partial (ORF1/RdRp)	KM886919			
LSV e10_3-5	*Lake Sinai Virus strain LSVe10 clone Bee3-5*	Partial (ORF1/RdRp)	KM886920			
LSV e10_4-1	*Lake Sinai Virus strain LSVe10 clone Bee4-1*	Partial (ORF1/RdRp)	KM886921			
LSV e10_4-2	*Lake Sinai Virus strain LSVe10 clone Bee4-2*	Partial (ORF1/RdRp)	KM886922			
LSV e10_4-3	*Lake Sinai Virus strain LSVe10 clone Bee4-3*	Partial (ORF1/RdRp)	KM886923			
LSV e10_4-4	*Lake Sinai Virus strain LSVe10 clone Bee4-4*	Partial (ORF1/RdRp)	KM886924			
LSV e10_4-5	*Lake Sinai Virus strain LSVe10 clone Bee4-5*	Partial (ORF1/RdRp)	KM886925			
LSV 4MT2014	*Lake Sinai Virus 4 clone MT2014*	Partial (ORF1/RdRp)	KP892556	USA	*Apis mellifera*	(Daughenbaugh *et al.*, 2015)
LSV 7MT2014	*Lake Sinai Virus 7 clone MT2014*	Partial (ORF1)	KR021355			
LSV 1MT2014	*Lake Sinai Virus 1 clone MT2014*	Partial (RdRp)	KR021356			
LSV 6MT2014	*Lake Sinai Virus 6 clone MT2014*	Partial (RdRp)	KR021357			
LSV 2MT2014	*Lake Sinai Virus 2 clone MT2014*	Partial (ORF1)	KR022002			
LSV 1MT2014cap	*Lake Sinai Virus 1 clone MT2014 capsid*	Partial (capsid)	KR022003			
LSV 2MT2014cap	*Lake Sinai Virus 2 clone MT2014 capsid*	Partial (capsid)	KR022004			
LSV A13LP_H2	*Lake Sinai Virus isolate Apis2013LP_H2*	Partial (ORF1/RdRp)	KT956845	Belgium	*Apis mellifera*	(Parmentier *et al.*, 2016)
LSV B13LP_H15-25	*Lake Sinai Virus isolate Bombus2013LP_H15-25*	Partial (ORF1/RdRp)	KT956846		*Bombus lapidarius*	
LSV B15LP_G4	*Lake Sinai Virus isolate Bombus2015LP_G4_fat*	Partial (ORF1/RdRp)	KT956847		*Bombus pascuorum*	
LSV B15LP_G4.2	*Lake Sinai Virus isolate Bombus2015LP_G4.2_body*	Partial (ORF1/RdRp)	KT956848			
LSV B15LP_G4.1	*Lake Sinai Virus isolate Bombus2015LP_G4.1_body*	Partial (ORF1/RdRp)	KT956849			
LSV1_NI1	*Lake Sinai virus 1 strain Norfolk Island 1*	Partial (capsid)	KT380002	Norfolk Island (Australia)	*Apis mellifera*	(Malfroy *et al.*, 2016)
LSV1_NI2	*Lake Sinai virus 1 strain Norfolk Island 2*	Partial (capsid)	KT380003			
LSV1_C1	*Lake Sinai virus 1 strain Cairns 1*	Partial (capsid)	KT380004			
LSV1_C2	*Lake Sinai virus 1 strain Cairns 2*	Partial (capsid)	KT380005			

Supp References:

Malfroy, S. F., Roberts, J. M. K., Perrone, S., Maynard, G. & Chapman, N. (2016). A pest and disease survey of the isolated Norfolk Island honey bee (Apis mellifera) population. J Apic Res 55, 202–211.

Table S2: Summary statistics (means for each species) of transcriptome assembly quality, ORF prediction and viral homology search.

	Sweat bees				**Ants**				
Genus	*Halictus*				*Messor*				*Aphaenogaster*
Species	*scabiosae*	*sexcinctus*	*simplex*	*barbarus*	*structor*	*capitatus*	*bouvieri*	*concolor*	*subterranea*
No. of transcriptomes analyzed (equal to # individual)	11	1	1	20	4	3	1	1	1
Initial assembly (ABYSS)									
No. million reads	6.8	6.7	7.9	15.9	26.2	14.1	5.2	19.9	3.2
No. contigs (x 1000)	121	132	148	230	352	129	70	238	47
Median length	75	74	73	88	82	77	101	75	108
N50	158	151	157	159	159	190	194	153	182
Final assembly (ABYSS-CAP3)									
No. contigs (x 1000)	30	34	37	35	70	24	24	55	18
Median length	185	181	185	203	201	207	183	181	177
N50	447	393	444	581	636	627	417	540	315
Virus detection									
No. of ORF predicted per species (x1,000)	23	26	28	26	50	18	19	38	13
No. of transcriptomes with full-length viral genome	**3**	0	0	0	0	0	0	**1**	0
No. of transcriptomes with partial viral genome	0	0	0	**1**	0	**1**	0	0	0

Table S3: Origins and characteristics of insects samples used for transcriptomic analysis.

	Host species	Individual name	Country	Locality	Year	SRA Accession number	Refs
Sweat bees	*Halictus scabiosae*	GA16A	Switzerland	Weiach	2010	SRX565141	(Romiguier *et al.*, 2014b)
		GA16B	Germany	Essen	2010	SRX565142	
		GA16C	France	Montpellier	2010	SRX565143	
		*GA16D	Switzerland	Adlikon	2010	SRX565144	
		*GA16E	Switzerland	Adlikon	2010	SRX565145	
		GA16F	Switzerland	Adlikon	2010	SRX565146	
		GA16G	Switzerland	Lausanne	2010	SRX565147	
		*GA16H	Switzerland	Lausanne	2010	SRX565148	
		GA16I	Switzerland	Lausanne	2010	SRX565149	
		GA16J	Switzerland	Lausanne	2010	SRX565150	
		GA16M	Switzerland	Lausanne	2010	SRX565151	
	Halictus simplex	GA16K	Switzerland	Adlikon	2010	SRX1470188	
	Halictus sexcinctus	GA16L	Switzerland	Weiach	2010	SRX2559194	**This study**
Ants	*Messor barbarus*	GA09A	France	Montpellier	2010	SRX2960337	
		GA09B	France	Lac Salagou	2010	SRX2960338	
		GA09E	Morocco	Soualem	2010	SRX2960339	
		GA09F	France	Puget sur Argens	2010	SRX2960340	
		GA09H	Spain	Calahorra	2010	SRX2960333	
		GA09I	Spain	Montblanc	2010	SRX2960334	
		§GA09J	France	La Cladiere	2010	SRX2960335	
		GA09K	Spain	Andalousia	2010	SRX2960336	
		GA09L	Spain	Granada	2010	SRX2960331	
		GA09M	Spain	Ventosa, Salamanca	2010	SRX2960332	
		GA40A	Spain	Andalousia	2010	SRX565202	(Romiguier *et al.*, 2014b)
		GA40B	Spain	Calahorra	2010	SRX565203	
		GA40C	France	Corneilla-la-Rivière	2010	SRX565204	

		GA40D	France	Montpellier	2010	SRX565205	
		GA40E	Spain	Ventosa Salamanca	2010	SRX565206	
		GA40F	Spain	Vilajoan	2010	SRX565207	
		GA40G	Spain	Grenade	2010	SRX565208	
		GA40H	Morocco	Soualem	2010	SRX565209	
		GA40I	France	La Cadière d'Azur	2010	SRX565210	
		GA40J	France	La Cadière d'Azur	2010	SRX565211	
	Messor structor	GA40M	France	Nimes	2012	SRX1470199	
		GA09S	France	Nimes	2010	SRX2960343	
		GA40L	France	Saint Guilhem le Désert	2014	SRX2188473	
		GA40N	France	La Doua, Lyon	2014	SRX2188475	
	Messor bouvieri	GA09T	France	Saint Guilhem le Desert	2010	SRX2960344	
	Messor capitatus	§GA09P	France	St Jean de Cuculles,	2014	SRX2188455	**This study**
		GA09Q	Spain	Villoria, Salamanca	2014	SRX2188456	
		GA40K	France	Saint Jean de Cuculles	2014	SRX2188457	
	Messor concolor	*GA09R	Crete	Kakopetros, Hamia,	2010	SRX2960341	
	Aphaenogaster subterranea	GA09N	France	Bois Montmaur	2010	SRX2960342	

*Individuals with a full-length viral genome
§Individuals with partial sequences

Table S4: Origins and characteristics of LSV positive honeybees *Apis mellifera*.

Pool or Individual name	Nb of bees	Country	Location	Year	Latitude	Longitude	LSV GenBank accession
F13PA003-01	10	France	Montfavet	2013	43.9160583	4.8758333	MF491488
F13PA021-01	5	France	Mazan	2013	44.056376	5.127605	MF491489
IT13AR030	1 (trembling)	Italy	Bagni di Lucca	2013	44.010924499	10.59157730	MF491502
F14PA092	40 (trembling)	France	Lambesc	2014	43.653995	5.261712	MF491490
F14PA093	40	France	Lambesc	2014	43.653995	5.261712	MF491491
I14AR136	4 (CBPV infection)	Italy	Bagni di Lucca	2014	44.010924499	10.59157730	MF491500
I14AR137	7 (CBPV infection)	Italy	Bagni di Lucca	2014	44.010924499	10.59157730	MF491501
F14PA-A01	40	France	Montfavet	2014	43.9160583	4.8758333	MF491492
F14PA-A03	40	France	Montfavet	2014	43.9160583	4.8758333	MF491493
F14PA-A04	40	France	Montfavet	2014	43.9160583	4.8758333	MF491494
F14PA-A06	40	France	Montfavet	2014	43.9160583	4.8758333	MF491495
F14PA-A07	40	France	Montfavet	2014	43.9160583	4.8758333	MF491496
Av5A	1	France	Avoine	2015	47.205697	0.1819820	MF491478
Av5B	1	France	Avoine	2015	47.205697	0.1819820	MF491479
C004	1	Canada	Vancouver	2015	49.274273	-123.099224	MF491481-MF491482
C039	1	Canada	Victoria	2015	48.424197	-123.376052	MF491483
C062	1	Canada	Squamish	2015	49.697827	-123.155240	MF491484-MF491485
BeiJing	50	China	Xiangshan, HaiDian District, BeiJing	2016	40.02	116.20	MF491480
HeBei	50	China	LuanPing city, HeBei province	2015	40.95	117.34	MF491499
GuiZhou	50	China	HuaXi district, GuiYang city, GuiZhou province	2016	26.42	106.68	MF491497
JiangXi	50	China	JingAn city, JiangXi province	2016	28.87	115.37	MF491503
ShanDong	50	China	QingDao City, ShanDong province	2016	35.88	119.79	MF491505
Culsea	1	Australia	Culburra Beach NSW	2016	-34.930556	150.757222	MF491487
SM1	1	Australia	Bryon Bay Area NSW	2016	-28.779722	153.478056	MF491506
CF19	1	Australia	Kin Kin QLD	2016	-26.252778	152.854722	MF491486
SM2	1	Australia	Bryon Bay Area NSW	2016	-28.779722	153.478056	MF491507
SM5	1	Australia	Bryon Bay Area NSW	2016	-28.779722	153.478056	MF491508
Hazel	1	Australia	Hazelbrook NSW	2016	-33.7225	150.459167	MF491498
MounT	1	Australia	Mount Tomah, NSW	2016	-33.539444	150.421389	MF491504

Table S5: Patristic distances for the ORF1/RdRp sequences between 81 LSV isolates. (not supplied)

Table S6: GenBank accessions of viruses used for ORF1, RdRp and capsid phylogenies.

Virus name	GenBank accession
Anopheline associated C virus	NC_023682; NC_023683
Beihai tombus-like virus 19	NC_032726
Chronic Bee Paralysis virus	NC_010711; NC_010712
Dansoman virus	KP714086; KP714087
Hubei odonate virus 12	NC_032846
Hubei tombus-like virus 38	NC_032984
Hubei tombus-like virus 39	NC_032741
Hubei tombus-like virus 40	NC_032767
Hubei tombus-like virus 42	NC_033207
Hubei tombus-like virus 43	NC_033263
Lake Sinai virus 1	HQ871931
Lake Sinai virus 2	HQ888865
Lake Sinai virus exp10	KM886905
Lake Sinai virus VBP022	KM886902
Lake Sinai virus VBP166	KM886903
Lake Sinai virus VBP256	KM886904
Mosinovirus	KJ632942; KJ632943
Nodamura virus	AF174533; AF174534
Pariacato virus	NC_003691; NC_003692
Sanxia tombus-like virus 9	NC_033149
Wenling tombus-like virus 4	NC_033024
Wenling tombus-like virus 5	NC_033090
Wenzhou crab virus 4	NC_033240
Wenzhou tombus-like virus 18	NC_033431
Wuhan insect virus 21	NC_033481;NC_033491

Table S7: LSV detection methods.

Sample collection of honeybees *Apis mellifera*				RNA extraction		Reverse Transcription		PCR reaction			Sequencing	
Country	**Year**	**Method**	**Individuals**	**Lysis Method**	**RNA Extraction kit**	**Kit**	**Amount of RNA**	**Taq Polymerase**	**Amplification cycle**	**Mix composition**	**Company**	**Primers**
France / **Italy**	2013-2014	Random or symptomatic	Individual or Pool (5 to 40 bees pooled per morpho-types)	TissueLyser® II (4 pulses 30Hz-30sec) 500 to 800µL phenol ; 0,8cm bead	RNAeasy® minikit, (Qiagen) 100µL chlorophorm per 500µL phenol	High capacity RNA to cDNA, (Life Technologies)	1 µg	Diamond® Taq DNA polymerase (Eurogentec)	95°C for 3 min, 4 cycles [1 min at 95°C, 1 min at 48°C, 1 min at 72°C], 35 cycles at 54°C for the annealing temperature, 10 min at 72°C	1.25 µL each primer (20 µM), 3 µL of MgCl2 (25 mM), 0.25 µL dNTPs mix (20 mM), 2.5 µL PCR buffer 10X, 0.15 µL Diamond Taq®, 2 µL cDNA (40 ng/µL), H2O to total volume 25 µL	GATC Biotech (Germany)	Forward M13-FP
France / **Canada**	2015	Random	Individual bees	TissueLyser® II (3 pulses 30Hz-90sec) 1,500µL RA1 buffer, one 5mm stainless bead	NucleoSpin® RNA (Macherey-Nagel)	RevertAid® First Strand cDNA synthesis kit (Thermo Scientific)	1 µg					
China	2015-2016	Targeted (3 colonies per apiaries ; 3 apiaries per location ; 3 or 4 location per province; 5 provinces)	Pool (50 bees randomly selected per province)	Homogenizer 10mL Trizol® reagent	Trizol® (Invitrogen)	GoScript® Reverse Transcription system (Promega)	4 µg	Es Taq DNA polymerase (CWBio, Beijing, China)	2 min at 94 °C; 4 cycles [30 s at 94°C, 30s at 48°C, 30s at 72°C], 30 cycles at 54°C for the annealing temperature, 10 min at 72°C.	25 µL 2× Es Taq MasterMix, 1 µL each primer (10µM), 1 µL cDNA (200 ng/µL), H2O to total volume 50µL	Sangon Biotech (China)	Forward M13-FP & Reverse M13-RP (Assembly DNAman software)
Australia	2016	Random	Individual bees	TissueLyser® II (3 pulses 30Hz-30sec) 1,500µL RLY buffer	ISOLATE II RNA Mini Kit (Bioline)	SensiFAST cDNA Synthesis Kit (Bioline)	2 µg	My Taq™ polymerase (Bioline)	1 min at 95°C: 40 cycles (95°c for 15 s, 55°C for 15 s, 72°C for 15 s), 2 min at 72°C	12µl My Taq HS red Mix 2x, 0.5µl each primer (20µM each),2µl cDNA (200ng/µl), 7µl H2O	Hawkesbury Institute for the Environment (Australia)	Forward M13-FP and Reverse M13-RP

Table S8: Primers used for PCR detection of LSV and Sinaiviruses.

Primer name	Target	Primer sequence	Expected product size
LSV-F-1791	LSV	**tgtaaaacgacggccagt**GCCWCGRYTGYTRGTDCCYCC	616 bp
LSV-R-2368		**caggaaacagctatgacc**GAVGTGGNGGNGCNAGATARAGT	
LSV-HsAV_deg_F	Sinaiviruses	**tgtaaaacgacggccagt**GARCGNTTNCSNGCNGAGGCC	747 or 812 bp
LSV-HsAV_deg_R		**caggaaacagctatgacc**TGWCKKKYBWGHGGGTACCGMGA	
M13-FP	Universal sequençing primers	**tgtaaaacgacggccagt**	
M13-RP		**caggaaacagctatgacc**	

4.3. Etude épidémiologique et de diversité génétique de virus d'abeilles (Article 5)

Ce travail se présente sous la forme d'un article en cours de rédaction et qui sera soumis dans *Journal of Invertebrate Pathology*. Il représente une étude de prévalence de virus décrits chez l'abeille domestique dans un grand nombre d'espèces d'hyménoptères sauvages et chez l'abeille domestique, récoltés sur trois continents Europe (France, Italie, Espagne), Amérique du Nord (Canada, Hawaii) et Amérique du Sud (Guyane française). Cette étude confirme la forte prévalence de ces virus, la présence de virus d'abeilles dans de nouvelles espèces d'hôtes, et permet également de quantifier la diversité génétique des virus détectés.

Overview of the prevalence and diversity of honey bee-infecting viruses in evolutionary diverse wild hymenoptera

Diane Bigot[1], Michèle Germain[1], Manon Romary[1], Philippe Gayral[1] and Elisabeth A. Herniou[1].

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

Abstract

Honey bee decline is observed since more than a decade and affects *Apis mellifera* bees all around the world. This decline is a multifactorial syndrome where interactions between pesticides and pathogens play an important role. More than twenty viruses have been identified in honey bees. Some of these viruses were also detected in other hymenopteran like ants and wild bees, which could serve as honey bee virus reservoir. In order to evaluate the prevalence and the viral diversity of honey bee viruses in a large range of wild hymenopteran, 350 ants and 175 wild bees occurring in sympatry were sampled in 60 localities in Western & Eastern Canada, French Guyana, France, Italy, Spain and Hawaii. All samples were analysed individually using total RNA extraction and RT-PCR virus detection. Taxonomic determination of the hymenopteran host species was performed using cDNA barcoding using COI sequencing. This allowed to detect phorid and conopid endoparasites in some *Bombus*. Virus detection was performed using a multiplex PCR for ssRNA+ CBPV, ABPV, DWV, SBV, IAPV, and BQCV and degenerated PCR for the detection of *Lake Sinai virus* (LSV) and a new closely related virus tentatively named Halictus scabiosae Adlikon virus (HsAV) (Bigot et al, in press). Nine new putative hosts were found for honey bee viruses such as four bumble bees, one hornet and four ant species. The genetic diversity of SBV, BQCV and DWV remained unchanged despite host shift demonstrating virus host flexibility and the role of hymenoptera as virus reservoir. Altogether, our study establishes a first step in the understanding of the ecological network in which bee viruses might evolve.

Key-words: Honey bee viruses, Prevalence, Virus diversity, *Deformed wing virus*, *Sacbrood virus*, *Black Queen Cell Virus*, *Formica exsecta virus 1*.

Introduction

Bees are keystones for ecosystem by providing pollination services evaluated worldwide about €153 billion (Gallai et al., 2009). Agricultural systems for human consumption are very dependent of bee pollination and 70% main crops are clearly pollinator-dependent (Klein et al., 2007).

Since over a decade, the honey bee *Apis mellifera* is impacted by a global decline (Potts et al., 2010). This decline has multiple and interactives causes, including environmental or abiotic factors as pesticides, nutrition, climate change, management practices and biotic factors as genetic diversity, micro and macro parasites, bacteria, fungi and viruses (Core et al., 2012; Doublet et al., 2015; Evans & Schwarz, 2011; Fairbrother et al., 2014; Goulson et al., 2015; Kielmanowicz et al., 2015; Menail et al., 2016; Sánchez-Bayo et al., 2016; VanEngelsdorp & Meixner, 2010).

To date, at least 24 viruses have been described to infect honey bees (Chen & Siede, 2007; Li et al., 2014; Runckel et al., 2011). The majority are RNA viruses belonging to the order *Picornavirales* (*Iflavirividae* and *Dicistroviridae* families) or unclassified viruses and, *Apis mellifera filamentous virus* is the only DNA virus (Clark, 1978; Gauthier et al., 2015). Extensive efforts were recently done to survey honey bee virus expansion in order to evaluate viral infections acting on pollinator declines (Genersch & Aubert, 2010; McMenamin & Genersch, 2015). As the impact of pathogens was proved to be correlated to honey bee decline (Cornman et al., 2012), studies of viruses became an important key for understanding and safeguarding honey bees and wild pollinators (Potts et al., 2016).

Honey bee RNA viruses can be transmitted by pollen or mite vectors such as *Varroa destructor* mite, and have potential impact on non-*Apis* hymenopteran (Fürst et al., 2014; Singh et al., 2010). Several honey bee viruses: *Deformed wing virus* (DWV, *Iflaviridae*, *Iflavirus*), *Black queen cell virus* (BQCV, *Dicistroviridae*, *Triatovirus*), *Israeli acute paralysis virus* (IAPV, *Dicistroviridae*, *Aparavirus*), *Acute bee paralysis virus* (ABPV, *Dicistroviridae*, *Aparavirus*), *Kashmir bee virus* (KBV, *Iflaviridae*, *Iflavirus*), *Chronic bee paralysis virus* (CBPV, Chroparavirus), *Sacbrood virus* (SBV, *Iflaviridae*, *Iflavirus*) and *Slow bee paralysis virus* (SBPV, *Iflaviridae*, *Iflavirus*) have been described in various hosts species based on RT-PCR detection, detection of replicative genomes, and pathogenicity, reviewed in (Manley et al., 2015; Tehel et al., 2016).

Bumble bees *Bombus* spp. (Hymenoptera, Apidae) are known to harbour various honey bee virus species (Fürst et al., 2014; Genersch et al., 2006; Graystock et al., 2013, 2014; Parmentier et al., 2016; Singh et al., 2010). DWV replicates in 7 bumble bee species: *B. impatiens*, *B. huntii*, *B. monticola*, *B. vagans*, *B. lapidarius*, *B. terrestris* and *B. lucorum* (Fürst et al., 2014; Levitt et al., 2013; Li et al., 2011); BQCV was replicative in *B. huntii* (Peng et al., 2011) and IAPV replicates in *B. impatiens*. *B terrestris* was particularly sensitive to bee virus since DWV, IAPV, ABPV and KBV can infect this species (Bailey & Gibbs, 1964; Evison et al., 2012; Fürst et al., 2014; Genersch et al., 2006; McMahon et al., 2015; Meeus et al., 2014). Finally, ABPV was proven to be pathogenic for 5 species: *B. agrorum, B. horturum, B. lucurum, B. ruderarius*, and *B. terrestris* (Bailey & Gibbs, 1964; McMahon et al., 2015).

Moreover, it has been observed that wild bees other than bumble bees also carry honey bee viruses. DWV was found in the sweat bee *Augochlora pura*, the carpenter bees *Ceratina dupla*, *Xylocopa latreille* and *X. virginica*, the masson bees *Osmia bicornis* and *O. cornuta* and the stingless bee *Scaptotrigona mexicana*. BQCV was found in the mining bee *Andrena vaga*, the masson bee *Heriades truncorum*, the stingless bee *Scaptotrigona mexicana*, and the carpenter bee *Xylocopa virginica*. IAPV was also found in the mining bee *Andrena vaga* and finally SBPV was also found in the carpenter bee Xylocopa virginica (Tehel et al., 2016). However, replicative forms of bee viruses were only proven in the DWV-*Osmia cornuta* pathosystem (Mazzei et al., 2014).

Some wasps (Hymenoptera, Vespidae) have been found positive for honey bee viruses. *Vespula vulgaris* was positive for DWV, BQCV, SBV and IAPV (Evison et al., 2012; Singh et al., 2010) and IAPV was found as a replicative form in the invasive hornet *Vespa velutina* (Yañez et al., 2012). Finally, *Formica rufa* and *Camponotus vagus* ants have been found positive for CBPV, and the replicative form of its genome was detected in *C. vagus* (Celle et al., 2008) as well as DWV found in *Linepithema humile* ant in New Zealand (Sébastien et al., 2015).

Usually, detection of honey bee viruses relies on simple or multiplex RT-PCR amplification of a small region of the viral genome or using ELISA techniques (Evans et al., 2013; de Miranda, 2008; de Miranda et al., 2013). More recently, metagenomic approaches have been used for bee virus surveys in bee hives (Cox-Foster et al., 2007; Granberg et al., 2013; Tozkar et al., 2015), or for the discovery of new bee viruses (Mordecai et al., 2016a; Runckel et al., 2011).

In our study 557 individual insects, regrouping hymenopteran, mostly honey bees, wild solitary bees and ants were sampled worldwide to study the prevalence of the 6 most pathogenic honey bee viruses SBV, ABPV, BQCV, CBPV, IAPV, DWV and two recently discovered viruses, *Lake Sinai virus* (LSV) and Halictus scabiosae Adlikon Virus (HsAV, Bigot et al, in press). This work shows the presence of honey bee viruses in some four species of ants, two hornets, at least four bumble bees and phylogenies of SBV, DWV and BQCV allows study of viral diversity of new found viruses.

Materials and Methods

Insects Field sampling

A large sampling of 557 individual insects was done in 6 countries: France, Canada, Spain, Italy, Hawaii (USA) and French Guyana, in summers 2013 to 2016 (Fig. 1a and Table S1). Mainly hymenoptera, sampled insects were classified as honey bees (*Apis mellifera*), wild bees (other non-*Apis mellifera* bees), ants (Formicidae), and other insects including Vespidae (hornets), Coleoptera, Diptera and Hemiptera formed the last category (Fig. 1a). All samples were collected alive in the field using visual random capture. All insects were brought back alive to the laboratory, and either immediately preserved at -80°C or cut using sterile blades, individually stored in RNALater (Ambion) following manufacturer's instructions.

RNA extraction and cDNA synthesis

Individual insects were disrupted mechanically and homogenized by a TissueLyser II (Qiagen) using a 5mm stainless bead during three 90 sec repetitions at 30Hz. Volume of lysis buffer RA1 (NucleoSpin RNA isolation kit, Macherey-Nagel) was adapted to insect weight according to the manufacturer's instructions. Lysates were centrifuged 3 min at 14,000 x g and 350µL of the supernatant was used for total RNA extraction using the NucleoSpin RNA isolation kit following manufacturer's instructions. A final elution was performed using 40µL of RNAse free water, repeated 3 time with the same eluate to increase final RNA concentration.

For some bees, lysate was too viscous and clung to adsorption columns. To circumvent this, a guanidium thiocyanate-phenol extraction was performed from 40µL of the NucleoSpin lysate and 100µL of NucleoZol® solution (Macherey-Nagel) following manufacturer's instructions, and eluted in 40µL of RNAse free water.

Quality of RNA was visually controlled by a 1% agarose gel electrophoresis and quantity was measured using a Qubit fluorimeter.

Reverse Transcription was performed using RevertAid First Strand cDNA Synthesis Kit (Thermo Scientific) with random hexamer primers following manufacturer's instructions usually from 1µg of total RNA (or between 100ng-1µg for very small ants).

Molecular taxonomic determination of hosts

Cytochrome Oxidase I gene was directly amplified on cDNA using universal primers targeting invertebrates LCO1490 and HCO2198 for a 709bp expected product size (Folmer et al., 1994). PCR mix contained 0.5 U of Diamond Taq Polymerase (Eurogentec), 2.5 µL buffer 10x (Eurogentec), 1.5 mM MgCl2, 0.2 mM each dNTP, 0.4 µM of forward and reverse primers, 100 ng of cDNA (10 ng for very small ants). Mix was completed by nuclease-free water for a total volume 25 µL. The amplification cycle was optimized to reduce aspecific signal observed in honey bees and increase PCR yield: 2 min at 95°C, 4 cycles [1 min at 95°C, 1 min at 54°C, 1 min at 72°C], 30 cycles [1 min at 95°C, 1 min at 49°C, 1 min at 72°C], 10 min at 72°C. All PCR products were visualized under UV light after a 40 min/100 V electrophoresis in gelRED-stained 1.5% agarose gel.

PCR products were Sanger-sequenced using LCO1490 primer by GATC Biotech (Germany), and electropherograms were manually corrected using Geneious version 9. Curated sequences were submitted to the Barcode Of Life Database (BOLD Sytem Version 3, http://www.boldsystems.org/) (Ratnasingham & Hebert, 2007) in order to retrieve taxonomic assignment of species or at least genus of the sampled insects.

Pathogen detection and sequencing

Detection of the 6 most common honey bee viruses: ABPV, IAPV, SBV, DWV, BQCV and CBPV was done by a multiplex PCR for simultaneous amplification, using previous described method (Sguazza et al., 2013). PCR mix for multiplex PCR was identical to the one previously described for COI amplification. PCR products were visualised as described above however using a 2.5% agarose gel and a 50 min/50V electrophoresis.

Detection of LSV was performed using LSV and LSV-HsAV degenerate primers pairs targeting the overlapping ORF1/RdRp region, described in (Bigot et al, in press). PCR mix was: 0.75 U of Diamond Taq Polymerase (Eurogentec), 2.5 µL buffer 10x (Eurogentec), 1.5 mM MgCl2, 0.2

mM dNTP , 0.4 µM of each forward and reverse primers, and 100 ng of cDNA, (10 ng for very small ants). Mix was completed by nuclease-free water for a total volume 25 µL. Amplification cycle was: 3 min at 95°C, 4 cycles [1 min at 95°C, 1 min at 48°C, 1 min at 72°C], 30 cycles [1 min at 95°C, 1 min at 54°C, 1 min at 72°C], 10 min at 72°C. PCR products were visualised as described for COI PCR.

Positives PCR products were Sanger-sequenced using corresponding forward primer by GATC Biotech (Germany) and electropherograms were manually corrected using Geneious version 9. When the multiplex PCR revealed coinfections (multiple bands of the sizes expected for the detected viruses), simplex PCR were performed on the same insect host cDNA again before sequencing.

Virus taxonomy and phylogenies

Homology search was performed on all sequenced PCR products using BLASTN online (https://blast.ncbi.nlm.nih.gov/Blast.cgi) against the nucleotide (nt) database and allowed species assignment. The new viral sequences were then aligned with all available sequences of the same species and genomic region found in NCBI Nucleotide database, as well as appropriate outgroups from more divergent species using MAFFT program (Katoh et al., 2002) implemented in Geneious.

The best evolutionary model of each alignment was determined by JModelTest (Darriba et al., 2012). Maximum Likelihood phylogenies were build using PhyML (Guindon & Gascuel, 2003), implemented in SeaView (Gouy et al., 2010). Node support values were estimated using approximate likelihood-ratio test (aLRT) (Anisimova & Gascuel, 2006).

Statistical analysis

Statistical analysis were performed using R software (R Development Core Team, 2008) using threshold for statistical significance of 0.05. Fisher's exact tests were used to test independence between taxonomic categories of hosts ("honey bees", "wild bees", "ants" and "other insects") and the proportion of infected individuals. Pairwise comparison was also performed using Bonferroni-corrected p-values to compare each pairs independently. 95% confidence intervals for virus prevalence were calculated using the Wilson procedure with a correction for continuity described in (Newcombe, 1998).

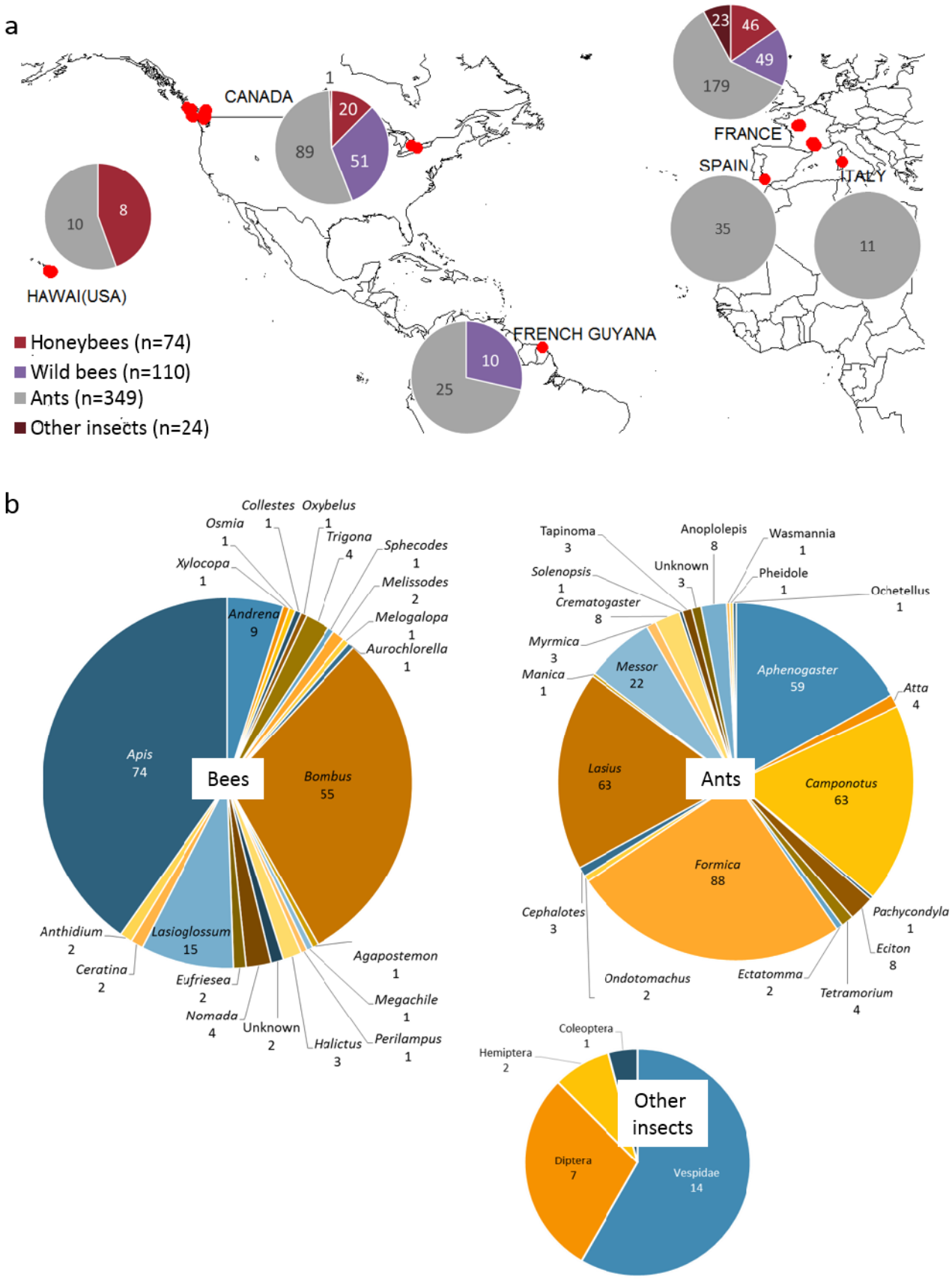


Figure 1: Origin and taxonomy of insect hosts. (a) Geographical repartition of honey bees, wild bees, ants and other insects sampled. (b) Taxonomic repartition of insects at the genus level determined by COI (Cytochrome Oxidase subunit I) barcode (BOLD System).

Results

Host barcoding

Among all the 557 individual insects, the COI identification using the BOLD Database allowed taxonomic identification of insect samples at the genus or species level. As the COI PCR was performed on cDNA, positive amplifications were used as a positive control of RNA extraction and RT, and negative COI PCRs were excluded from further analysis and calculations. The 349 individuals of "ants" category were belonged to 21 genera of the family Formicidae (Fig. 1b). Among them, the most represented genera were *Formica* (88 individuals), *Camponotus* (63 indiv.), *Lasius* (63 indiv.), *Aphaenogaster* (59 indiv.) and *Messor* (22 indiv.) followed by 16 genera less represented by 1 to 8 individuals (Fig. 1b). Finally, 3 individuals were attributed to the Formicidae family but the genus could not be undetermined using COI barcoding. Within the category honey bees, 74 individual *Apis mellifera* were collected (Fig. 1b). In the wild bees category, 20 genera were represented and the three most represented genera were *Bombus* (55 individuals), *Lasioglossum* (15 indiv.), and *Andrena* (9 indiv.). The remaining 16 genera were represented by 1 to 4 individuals (Fig. 1b). Finally, 2 individuals were attributed to the Apoidae (bee) family but the genus was undetermined. For nine *Bombus* samples, the COI barcode identification corresponded to various Diptera species, although collected insects were unequivocally bumble bees. Four individuals were attributed to Phoridae (n°C001, C034, C037, C045). Four other individuals were attributed to Conopidae; two *Physocephala sp*. (n°F001, F020) and two *Physocephala burgessi* (n°C069, C070). The last individual was attributed to *Cladotanytarsus cyrylae* (Chironomidae) (n°F017). As the COI barcode was unable to attribute the correct species and for avoid misattribution of taxonomy, all nine samples were attributed to Bombus sp. species. The last category other insects, is represented by the Vespoidae family (hornets, 14 indiv.) and by others order, Diptera (7 indiv.), Hemiptera (2 indiv.), Coleoptera (1 indiv.) (Fig. 1b), which we considered to be negative controls.

Virus prevalence in wild insects

Sanger sequencing of PCR products was achieved for 23 DWV, 9 BQCV, 9 SBV, 3 ABPV, 7 LSV positive PCRs. Since extensive phylogenetic analysis of Sinaivirus sequences from LSV primers has already been analysed previously (Bigot et al, in press), we will focus here on the 44 remaining sequences.

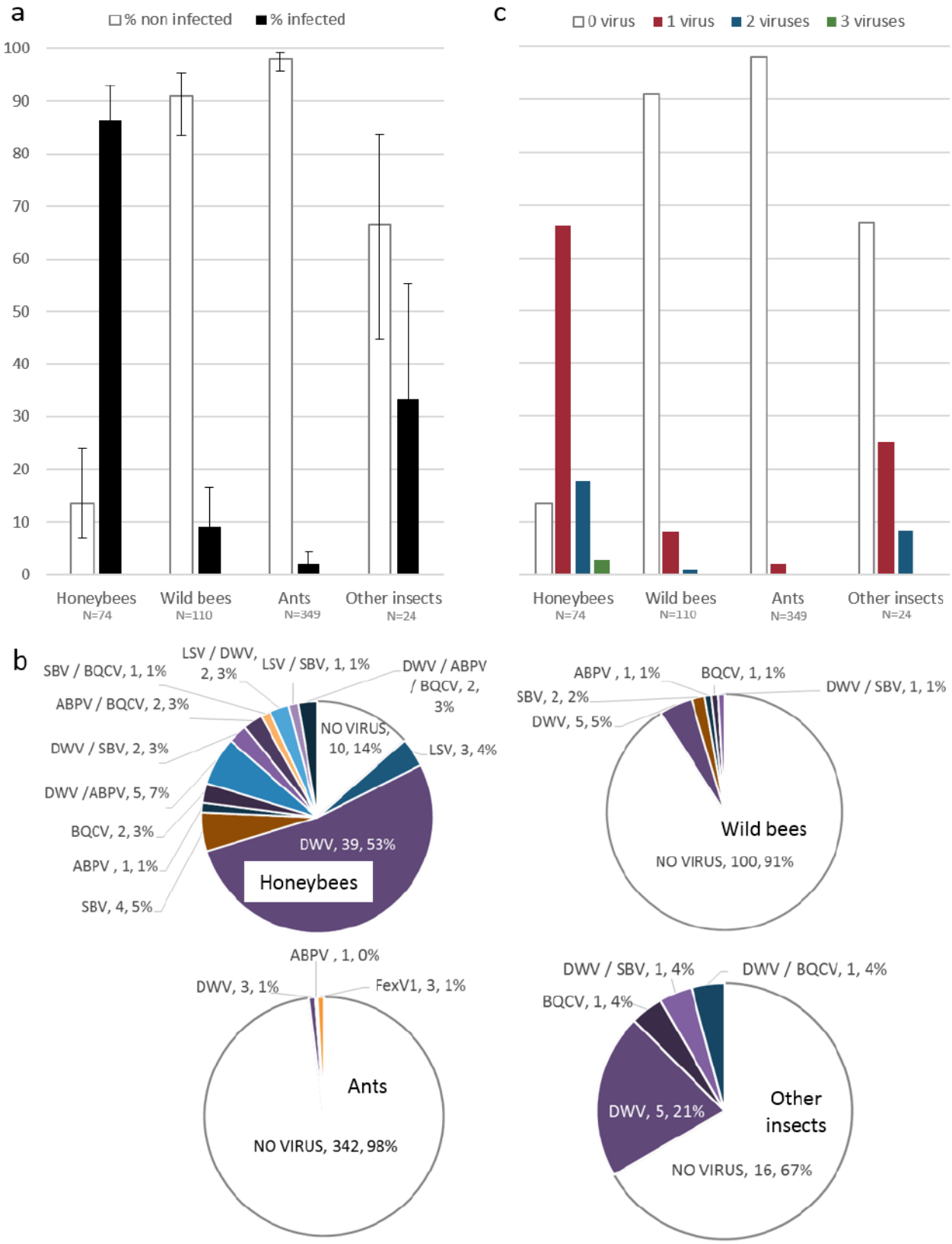


Figure 2: Virus prevalence. (a) Prevalence of infections, (b) Detailed virus prevalence, with name of virus or group of viruses found, followed by number of infected individual and prevalence in percentage and (c) Distribution of coinfections. All these graphics are detailed four each taxonomic groups, honey bee, wild bees, ants and other insects.

First of all, BLAST homology identification was performed to address the validity of the virus detection results. We confirmed that all sequences obtained from DWV, SBV and BQCV primers corresponded to the targeted virus (percentage identity varied from 92% - 100%, query coverages from 96% - 100% and E-values from 1E-70 to 0). For DWV, sequences were furthermore identified as DWV and a VDV-1/DWV recombinant with E-values between 1E-70 and 6E-103 and nucleotide identity 92-100 % (Table S2). ABPV detection was the only one giving unexpected BLAST results. The three sequences obtained from *Formica aserva* ants using ABPV primer pairs corresponded to Formica exsecta virus 1 (FexV1) (E-values between 1E-173 and 7E-172 and nucleotide identity of 95%) (Table S2). FexV1 and ABPV are closely related virus both belonging to the *Dicistroviridae* family and this genetic similarity can explain why ABPV primers also amplified FexV1.

Within the 557 barcoded hosts, 5 viruses over the 7 screened were detected in this study: ABPV (global prevalence without taking account for coinfections: 2.15%), LSV (1.08%), BQCV (1.80%), DWV (12.03%) and SBV (2.15%). CBPV and IAPV were screened but not found in any of our sample.

The global percentage of viral infection was measured for each host categories, and Fisher's exact test indicated that the prevalence significantly dependent from the host category ($p= 2.2E-16$). Honey bees had the highest infection rate (86.5% of infected insects), followed by wild bees (9.1%) and "ants" (2%) (Fig. 2a). Other insects, mainly represented by hornets had high virus prevalence (33.3%). Pairwise comparisons and Bonferroni-correction indicated that all prevalence rates differed significantly from one another ($p< 0.0045$).

Among the 7 distinct virus species screened in this study, 5 were detected in honey bees, 4 in wild bees, 3 in the other Insect category and 2 in ants (Fig. 2b and Table S1). In all cases DWV was the most prevalent in 39 individuals (53%) *Apis mellifera*, in 5 individuals (5%) in "wild bees" (three in *Bombus* sp., n°C010, C011, C020, one in *Bombus flavifrons*, n°C025, one in *Bombus humilis*, n°F016), in 3 individuals (1%) in "ants" (one in *Formica neorufibarbis* n°A062; one in *Crematogaster scutellaris*, n°SCR-09; and one in *Pheidole megacephala*, n°Hilo23), and in 5 individuals (21%) in "other insects" (four *Vespa crabro* n°VC-F1, VC-S1, VC-S2 and VC-S3 and one in *Vespa velutina* n°VV-L2).

The presence of simultaneous infection (or co-infections) was found in all this dataset. Ants were only infected by single viruses (DWV, FexV1 or ABPV), no co-infections were found (Fig. 2c). Co-occurrence of 2 and 3 viruses was frequently observed in the remaining host

categories. 18.92% of honey bees were infected by two viruses and the most prevalent species pair was DWV/ABPV found in 7% of samples. Coinfections of 3 virus species was also found in 3% of samples, two honey bees carried DWV/ABPV/BQCV (Fig. 2b and 2c). In wild bees a DWV/SBV coinfection (4%) was found in one *Bombus* sp. n°C001. In the group other insects, the presence of 2 viruses were found in 8.33% of individuals: DWV/SBV (4%) was found in one *Vespa velutina* hornet n°VV-L1; and DWV/BQCV (4%) was found in one *Vespa crabro* hornet n°VC-F2 (Fig. 2b and 2c and Table S1). The distribution of co-infection between groups was tested by the Kolmogorov-Smirnov test and no significant differences between groups were found.

Virus phylogenies

A maximum Likelihood phylogeny was first built in order to investigate the taxonomy of the FexV1-related sequences found in *Formica aserva* ants. This phylogeny focused on members of the *Aparavirus* genus within the *Dicistroviridae* family, and included the single homologous FexV1 sequence available thus far (Table S3). The tree shows that the three new sequences (F2, A079 and A080) formed a robust monophyletic group (node support aLRT=0.93) closely related to Fexv1 (node support aLRT=1) (Fig. 3).

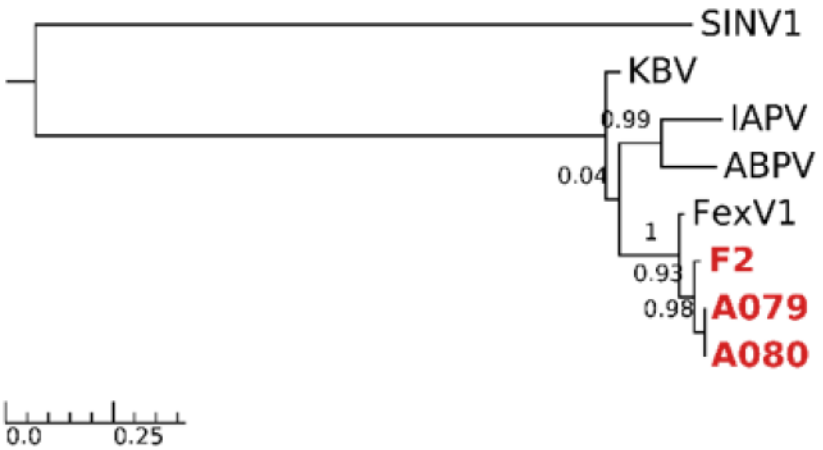


Figure 3: Maximum Likelihood phylogenetic tree of the capsid of *Formica exsecta virus 1* and new sequences. (430 nucleotide sites, model GTR+G). Taxa in red were produced for this study. ABPV: *Acute bee paralysis virus* (NC_002548), FexV1: *Formica exsecta virus 1* strain FexV1 (KF500001), IAPV: *Israeli acute paralysis virus* (NC_009025), KBV: *Kashmir bee virus* (NC_004807), SINV1: *Solenopsis invicta virus 1* (NC_006559). Details information are in Table S3 and Table S2 for new sequences. The scale bars represents the substitutions rate per site. The values correspond to the values of nodes supporting by aLRT statistics.

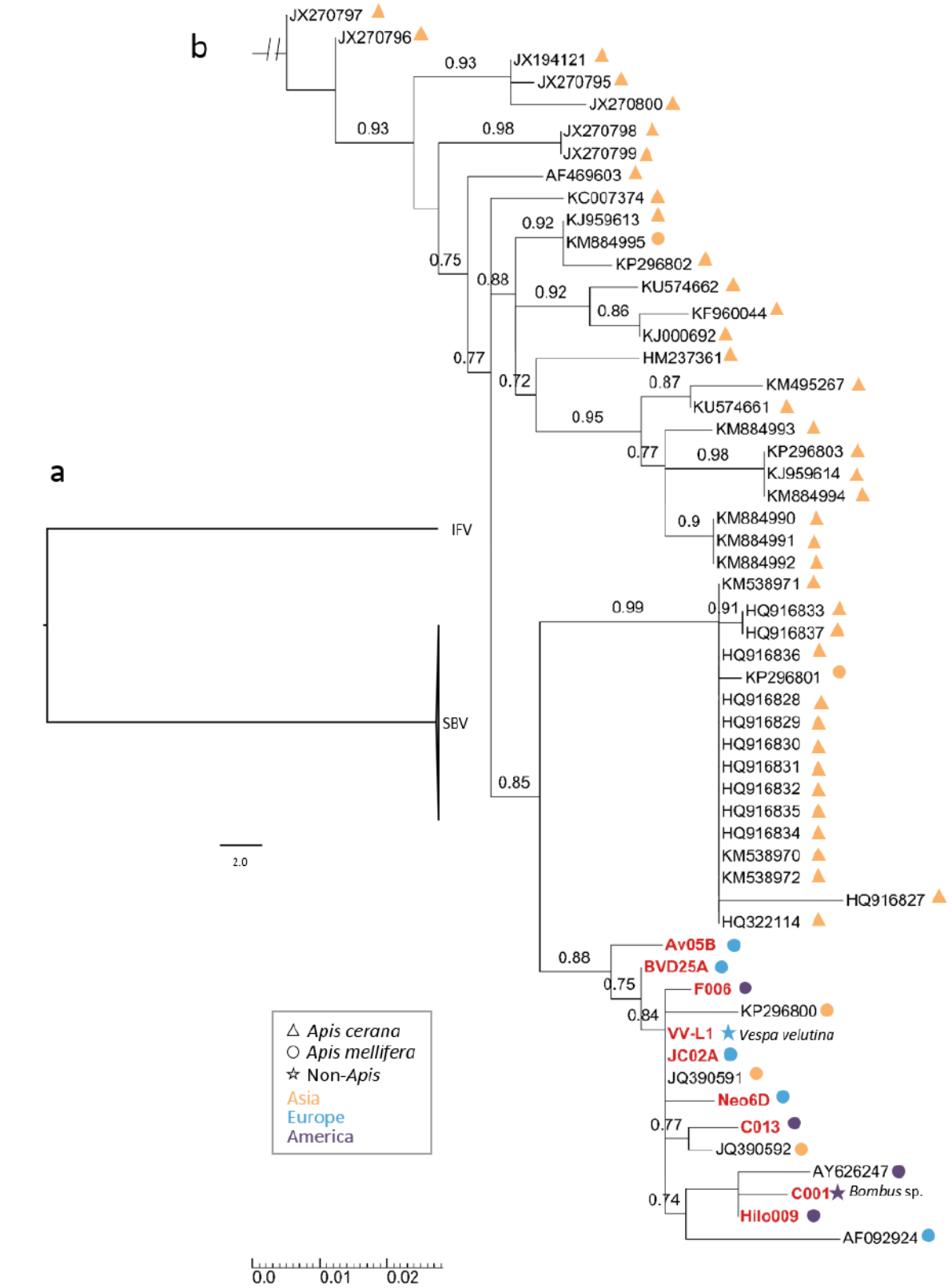


Figure 4: Maximum Likelihood phylogenetic tree of the capsid of all known and new *Sacbrood virus* sequences. (a) Collapsed phylogeny with *Infectious flacherie virus* (IFV) as outgroup (AB000906). (b) Zoom of the SBV clade (335 nucleotide sites, model GTR+G). Taxa in red were produced for this study. Host and geographical origins of all sequences are indicated by shape and colour of symbols next to GenBank accession and by host name in parentheses when the host was non-*Apis* (see the figure legend). Taxon information for known SBV sequences are in Table S3 and Table S2 for new sequences. The scale bars represents the substitutions rate per site. The values correspond to the values of nodes supporting by aLRT statistics.

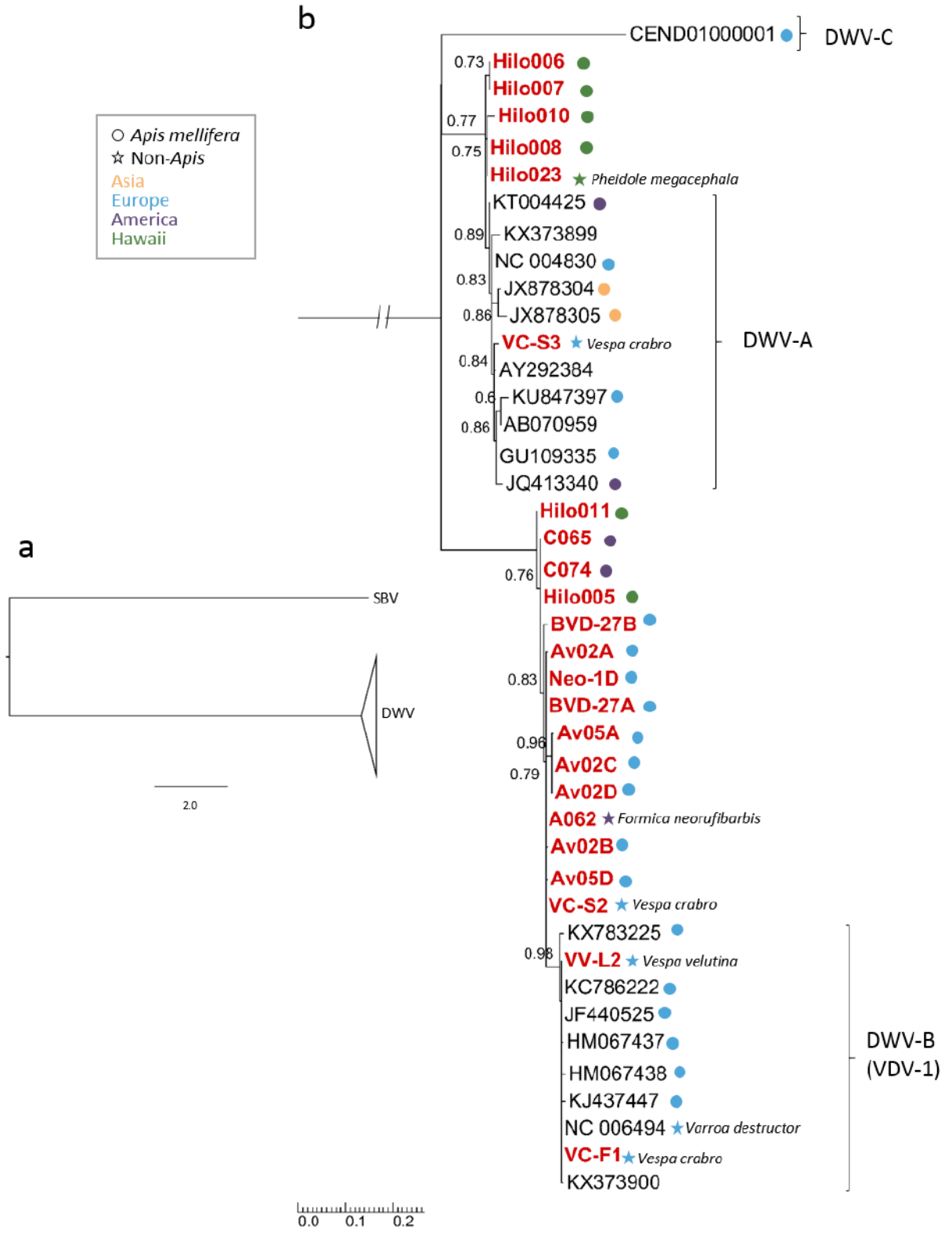


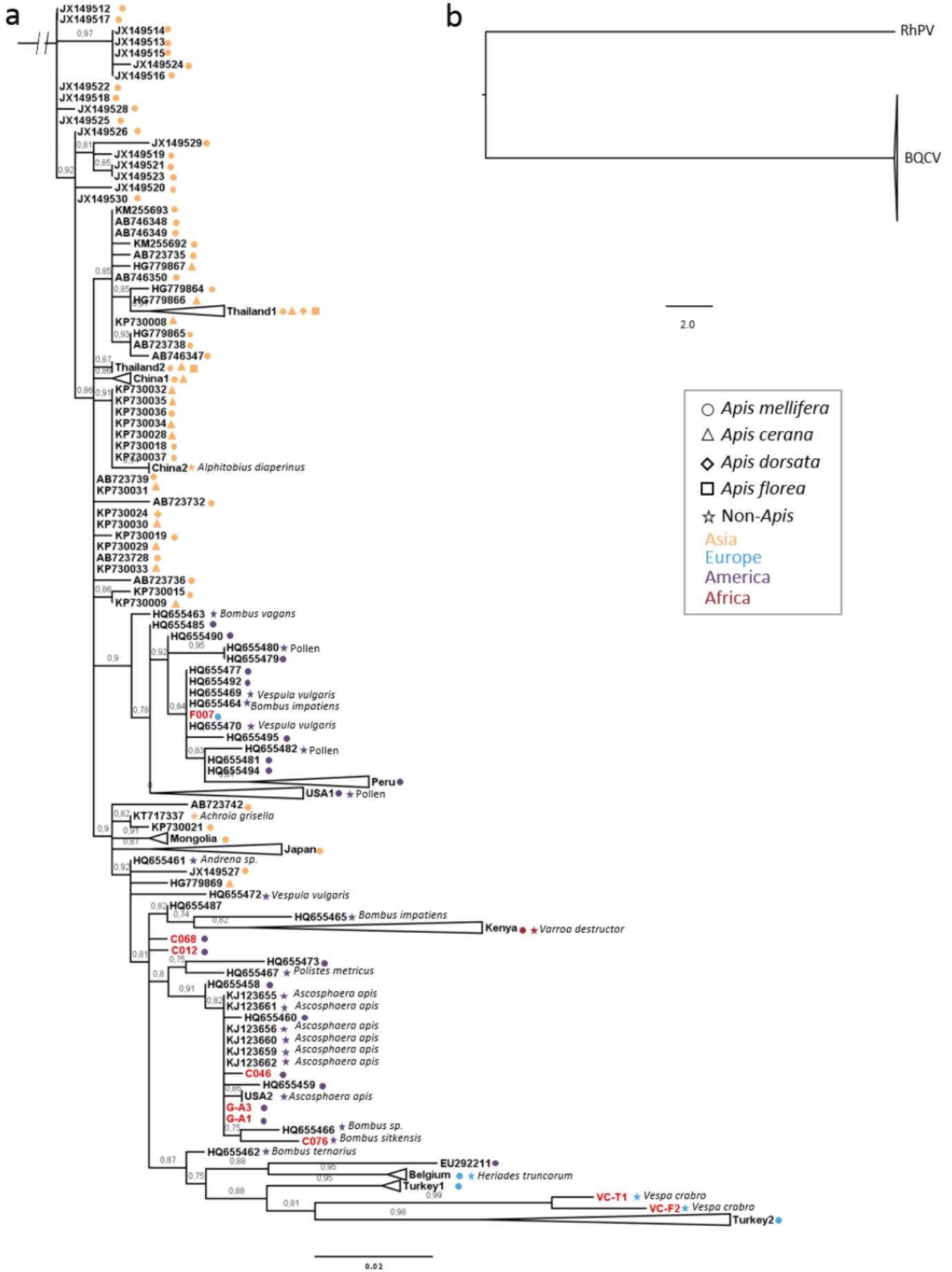
Figure 5: Maximum Likelihood phylogenetic tree of the capsid of all known and new *Deformed wing virus* sequences. (a) Collapsed phylogeny with *Sacbrood virus* (SBV) as outgroup (AF092924). (b) Zoom of the DWV clade (253 nucleotide sites, model HKY85+G). Taxan in red were produced for this study. Host and geographical origins of all sequences are indicated by shape and color of symbols next to GenBank accession and by host name in parentheses when the host was non-*Apis* (see the figure legend). Taxon information for known DWV sequences are in Table S3 and Table S2 for new sequences. The scale bars represents the substitutions rate per site. The values correspond to the values of nodes supporting by aLRT statistics.

The remaining SBV, DWV and BQCV phylogenies were built to investigate genetic diversity at the intraspecific level. Especially, we wanted to know if targeting non-*Apis* insects could enhance our knowledge of virus biodiversity. To answer this, we built SBV, DWV and BQCV nucleotide phylogenies using all publicly available sequences of the capsid conserved region used for multiplex PCRs.

The SBV phylogeny was performed using *Infectious flacherie virus* (IFV) as outgroup (Fig. 4a). In this phylogeny, all SBV sequences formed a monophyletic group, likely corresponding to the same viral species. The majority of sub-clades of SBV phylogeny are all composed of viruses collected in Asia, mainly on *Apis cerana* bees (Fig. 4b). All 9 new SBV sequences grouped within SBV, and all belonged to a single well-supported sub-clade (node support = 0.88). This clade comprised sequences from Europe, North America and Asia (see Table S3 for accession and origins details). These 9 new sequences collected from invasive Asian hornet *Vespa velutina* collected in France, in *Bombus* sp. from Canada, and from *A. mellifera* in Europe (France), North America (Canada) and Hawaii enriched the biodiversity of this sub-clade.

The DWV phylogeny was built with SBV as outgroup (Fig. 5a). DWV tree is shaped by three major sub-clades, corresponding to the three known genetic groups named DWV-A, DWV-B (previously found in the *Varroa destructor* mite and named *Varroa destructor virus 1*, VDV-1) and DWV-C (Fig. 5b). This phylogeny confirms that all 23 new sequences belonged to DWV-A (6 sequences) and DWV-B groups (17 sequences). DWV sequences from hornets were found in both DWV-A and DWV-B. Interestingly, several occurrence of DWV in non-Apis hymenoptera were reported: DWV-A in *Pheidole megacephala* ant from Hawaii, DWV-B in *Formica neorufibarbis* ant in Canada, DW-A and –B in hornet *Vespa crabro* and DWV-B in *Vespa velutina* from France.

Figure 6: Maximum Likelihood phylogenetic tree of the capsid of all known and new *Black Queen Cell Virus* sequences. (a) Zoom of the BQCV clade (557 nucleotide sites, model GTR+G). Taxon in red were produced for this study. Host and geographical origins of all sequences are indicated by shape and color of symbols next to GenBank accession and by host name in parentheses when the host was non-Apis (see the figure legend). Taxon information for known BQCV sequences are in Table S3 and Table S2 for new sequences. (b) Collapsed phylogeny with Rhopalosiphum padi virus (RhPV) as outgroup (AF022937). The scale bars represents the substitutions rate per site. The values correspond to the values of nodes supporting by aLRT statistics.


a
b
RhPV
BQCV
2.0
0.02
Apis mellifera
Apis cerana
Apis dorsata
Apis florea
Non-Apis
Asia
Europe
America
Africa
JX149512
JX149517
JX149514
JX149513
JX149515
JX149524
JX149516
JX149522
JX149518
JX149528
JX149525
JX149526
JX149529
JX149519
JX149521
JX149523
JX149520
JX149530
KM255693
AB746348
AB746349
KM255692
AB723735
HG779867
AB746350
HG779864
HG779866
Thailand1
KP730008
HG779865
AB723738
AB746347
Thailand2
China1
KP730032
KP730035
KP730036
KP730034
KP730028
KP730018
KP730037
China2 Alphitobius diaperinus
AB723739
KP730031
AB723732
KP730024
KP730030
KP730019
KP730029
AB723728
KP730033
AB723736
KP730015
KP730009
HQ655463 Bombus vagans
HQ655485
HQ655490
HQ655480 Pollen
HQ655479
HQ655477
HQ655492
HQ655469 Vespula vulgaris
HQ655464 Bombus impatiens
F007
HQ655470 Vespula vulgaris
HQ655495
HQ655482 Pollen
HQ655481
HQ655494
Peru
USA1 Pollen
AB723742
KT717337 Achroia grisella
KP730021
Mongolia
Japan
HQ655461 Andrena sp.
JX149527
HG779869
HQ655472 Vespula vulgaris
HQ655487
HQ655465 Bombus impatiens
Kenya Varroa destructor
C068
C012
HQ655473
HQ655467 Polistes metricus
HQ655458
KJ123655 Ascosphaera apis
KJ123661 Ascosphaera apis
HQ655460
KJ123656 Ascosphaera apis
KJ123660 Ascosphaera apis
KJ123659 Ascosphaera apis
KJ123662 Ascosphaera apis
C046
HQ655459
USA2 Ascosphaera apis
G-A3
G-A1
HQ655466 Bombus sp.
C076 Bombus sitkensis
HQ655462 Bombus ternarius
EU292211
Belgium Heriades truncorum
Turkey1
VC-T1 Vespa crabro
VC-F2 Vespa crabro
Turkey2

The BQCV phylogeny showed that, as observed for SBV and DWV that the clade formed by DWV sequences was very phylogenetically distant from the *Rhopalosiphum padi virus* (RhPV) closest outgroup, with no intermediate lineage (Fig. 6a) The tree shows that BQCV is very polymorphic (Fig. 6b). The phylogeny seemed to be geographically structured in clades from same geographical origin at the continent scale (Asia, North America, Africa, and Europe).

All published sequences in this tree came from various hosts, mainly from *A. mellifera* but also from *A. cerana*, *A. dorsata* and *A. florea* bees, from *Bombus impatiens*, *B. ternarius*, *Bombus spp*. and *B. vagans* bumble bees, from *Heriades truncorum* and *Andrena sp*. wild bees, from *Vespula vulgaris* and *Polistes metricus* wasps, from *Alphitobius diaperinus* coleopteran, from hive parasites such as *Varroa destructor* mite, *Ascosphaera apis* fungi, and *Achroia grisella* lepidoptera or from pollen pellets.

The nine new BQCV sequences belonged to several sub-clades of the tree but not from the Asian sub-clades. This work brought six new sequences from *A. mellifera* from France, Canada and French Guyana, one sequence from *Bombus sitkensis* bumble bee from Canada (n°C076) and two sequences from European hornet Vespa crabro (n° VC-T1 and VC-F) from France. Sequences from Canada and French Guyana grouped within a clade comprising other North American sequences. The sequence from *A. mellifera* from France (n°F007) grouped within another clade comprising American sequences. Finally, the two hornet sequences grouped together (node support = 0.81) with a very long branches.

Discussion

Virus prevalence in *Apis mellifera*

Within the honey bees 86 % are infected by at least one virus in this study. Several studies have already shown that honey bee viruses are often in coinfections in *Apis mellifera*, with two viruses in around 50% of cases (Berényi et al., 2006; Chen et al., 2004), three viruses at 7 % (Chen et al., 2004), 27 % (Berényi et al., 2006) or 31 % (Tentcheva et al., 2004). Multiple infections are not rare and may cause inapparent symptom in seemingly healthy bee colonies compare to weak colonies with similar virus distribution where the viral load determine the status of colonies (Berényi et al., 2006). Multiple infections can also be detected in queens

suggesting a putative role of infected queens in the vertical transmission of viruses (Chen et al., 2005).

DWV was predominant in single or multiple infection, 53 % of single infections, 13 % in dual infection with ABPV, SBV or LSV and 3 % in triple infection with ABPV and BQCV. Several studies point out that DWV is often the most prevalent virus in 90 % of honey bees (Ai et al., 2012; Baker & Schroeder, 2008; Berényi et al., 2006; Tentcheva et al., 2004) but in some other works DWV prevalence is very low under 10 % (Cavigli et al., 2016; Daughenbaugh et al., 2015). DWV is mainly vectored by the *Varroa destructor* mite and variation in DWV prevalence could be directly linked with the presence or absence of *Varroa* (Wilfert et al., 2016). Variation in prevalence can also be driven by others factors such as *Varroa* mite or *Nosema* fungi exposure (Berényi et al., 2006; Traynor et al., 2016). As for DWV, for all other viruses prevalence is variable and it has been previously shown that variation of prevalence is common through seasons (Runckel et al., 2011; Traynor et al., 2016).

Viral presence in non-Apis insects

Within the "wild bees" group, *Bombus* were the only infected clade. It was already shown by multiple studies that various bumble bee species can carry honey bee viruses (Fürst et al., 2014; Genersch et al., 2006; Graystock et al., 2013, 2014; Parmentier et al., 2016; Singh et al., 2010). The main interest of this study is that the presence of bee viruses was shown in non-recorded *Bombus* species as *Bombus flavifrons* and *B. humilis* infected by DWV; *B. lapidarius* by SBV and *B. sitkensis* by BQCV, as confirmed by the phylogenetic analysis.

Our work also reported that hornets frequently carried honey bee viruses. This work shows the presence of three honey bee viruses, DWV, SBV and BQCV, alone or in co-infections, in two species *Vespa crabro* (the European hornet) and *Vespa velutina* (the Asian hornet, invasive species). Presence of IAPV has previously been found in *V. velutina* with the presence of a putative replicative form of the virus (Yañez et al., 2012). In the latter and the present study, whole insects were used for virus detection, and we cannot invalidate the hypothesis of trophic contamination as hornets, and especially *V. velutina*, are honey bee predators (Monceau et al., 2014). A similar study, realized without the entire digestive tract or targeting specific infected tissues, would help to know if honey bee viruses can truly infect hornets.

Honey bee viruses were also reported here several times in ants. DWV was found and confirmed phylogenetically in one *Pheidole megacephala* and one *Formica neorufibarbis* and was also found in one *Crematogaster scutellaris*. DWV is known to be vectored by the *Varroa destructor* mite (Wilfert et al., 2016) or via pollen (Mazzei et al., 2014). Nevertheless, the presence of honey bee viruses in ants has already been shown for CBPV in two species *Formica rufa* and *Camponotus vagus*, and DWV in the invasive Argentine ants *Linepithema humile* in New Zealand and a replicative form was found by strand-specific PCR in *C. vagus* (Celle et al., 2008) and in *L. humile* (Sébastien et al., 2015), showing that ants could be reservoir of bee viruses. Here again, the hypothesis that environmental contamination from dead bees to ants cannot be discarded, and the ability of bee virus replications in ants should be specifically tested.

Phylogenetic diversity of honey bee viruses

The SBV phylogeny shows a strong geographic distribution of SBV strains following the two main SBV serotypes (the 'European serotype" infecting *A. mellifera* worldwide and the 'Asian serotype' infecting *A. cerana*) and as previously demonstrated using an RdRp phylogeny (Roberts & Anderson, 2014; Xia et al., 2015). Our phylogeny correlates this geographical and species distribution of SBV serotypes, previously confirmed using capsid sequences (Hu et al., 2016).

The DWV phylogenetic tree was done on capsid nucleotides and was composed of few sequences but the tree showed the presence of the three major clades of this virus, DWV-A, DWV-B and DWV-C (Mordecai et al., 2016b), were DWV-A and DWV-B are known to differ in term of virulence (McMahon et al., 2016). A previous work, using the RdRp sequence, showed that these three major variants of DWV were ancient; DWV-A and DWV-B diverged 181 years ago and DWV-C was anterior and diverged form DWV-A/B 319 years ago (Mordecai et al., 2016b). Our tree, based on the capsid gene, shows a different signal, where the DWV-C and DWV-A clade were grouped with DWV-B as external clade. DWV was previously found in Hawaii by another work and the authors hypothesised that would be a new strain (Martin et al., 2012), probably corresponded to DWV-C (Mordecai et al., 2016b). It was not the case for our sequences from Hawaii, were five may correspond to DWV-A and two to DWV-B. Nevertheless, among them one was found in an ant, *Pheidole megacephala*, which is invasive

in Hawaii (Wetterer, 2007). With the few number of sequences, no visible geographical link was showed by the tree but new putative hosts were found for DWV, such as two ant species (*Pheidole megacephala* and *Formica neorufibarbis*) and two hornets species (*Vespa crabro* and *V. velutina*). As DWV is known to be principally vectored by *Varroa destructor* mites (Wilfert et al., 2016) and could be vectored via pollen (Mazzei et al., 2014), we can hypothesized that the contamination of these new species can be also driven by these factors.

The BQCV phylogenetic tree showed a strong geographic distribution of Asian sequences, previously reported in Korea (Reddy et al., 2013) or in Thailand and Japan (Mookhploy et al., 2015). Our new sequences were generally grouped within clades possessing the same geographic origins, confirming this geographical pattern. BQCV was previously reported to be present in various hosts as bumble bees, wasps, wild bees, bee parasites (mite or fungi) and also pollen (Singh et al., 2010). The virus diversity do not reflect a host structuration of strains as previously reported on few species and RdRp phylogeny (Zhang et al., 2012). Nevertheless, the description of BQCV in the hornet *Vespa crabro* is new and phylogenetically may correspond to a specific BQCV variant as the branch size showed.

Conclusion

Various hosts are known to be infected by bee viruses, as previously demonstrated but this work brings a new list of potential honey bee virus hosts, such as four species of bumble bees, one species of hornet and three species of ants. This work also showed that the virus genetic diversity was not host dependant in most cases. This study demonstrate that honey bee viruses can be found at large scale in various hymenoptera suggesting the role of wild hymenoptera as reservoir of honey bee viruses.

Availability of supporting data

Sanger sequences are available in GenBank, see Table S2 and (Bigot et al, JGV).

Authors' contributions

EAH and PG performed the design and the coordination of the study. DB, PG and EAH carried out the sample collection. DB, MG and MR performed RNA isolation, PCR amplifications and

sequencing. DB carried out the bioinformatics analysis. DB, EAH and PG analysed the result and wrote the manuscript. All authors read and approved the final manuscript.

Funding information

This work was supported by two European Research Council (ERC) grants to NG (ERC PopPhyl 232971) and to EAH (ERC GENOVIR 205206). The funders had no role in study design, data collection and interpretation, or the decision to submit the work for publication. This work was supported by a Jean & Marie-Louise Dufrenoy grant to DB form the French Agriculture Academy.

Acknowledgements

Multiple collaborators from our laboratory were implicated in some insect's collections: Carlos Lopez-Vaamonde for sampling in French Guyana and Hawaii. Alain Lenoir and Raphael Boulay for sampling of ants in Spain and France. Jeremy Gauthier for sampling of ants in Italy. Jérémy Gevar for hornets sampling in France. Sébastien Moreau for help and sampling in Tours. Jean-Christophe Lenoir for sampling of bees in France. Magali Chabert-Ribière and Eric Dubois (ANSES, French Agency for Food, Environmental and Occupational Health & Safety) for providing CBPV positive control.

The authors declare that they have no competing interests.

References

Ai, H., Yan, X. & Han, R. (2012). Occurrence and prevalence of seven bee viruses in Apis mellifera and Apis cerana apiaries in China. J Invertebr Pathol 109, 160–164.

Anisimova, M. & Gascuel, O. (2006). Approximate likelihood-ratio test for branches: a fast, accurate, and powerful alternative. Syst Biol 55, 539–52.

Bailey, L. & Gibbs, A. J. (1964). Acute infection of bees with Paralysis Virus. J Insect Pathol 6, 395–407.

Baker, A. & Schroeder, D. (2008). Occurrence and genetic analysis of picorna-like viruses infecting worker bees of Apis mellifera L. populations in Devon, South West England. J Invertebr Pathol 98, 239–242.

Berényi, O., Bakonyi, T., Derakhshifar, I., Köglberger, H. & Nowotny, N. (2006). Occurrence of Six Honey bee Viruses in Diseased Austrian Apiaries. Appl Environ Microbiol 72, 2414–2420.

Cavigli, I., Daughenbaugh, K. F., Martin, M., Lerch, M., Banner, K., Garcia, E., Brutscher, L. M. & Flenniken, M. L. (2016). Pathogen prevalence and abundance in honey bee colonies involved in almond pollination. Apidologie 47, 251–266.

Celle, O., Blanchard, P., Olivier, V., Schurr, F., Cougoule, N., Faucon, J.-P. & Ribière, M. (2008). Detection of Chronic bee paralysis virus (CBPV) genome and its replicative RNA form in various hosts and possible ways of spread. Virus Res 133, 280–4.

Chen, Y. P. & Siede, R. (2007). Honey Bee Viruses. In Adv Virus Res, pp. 33–80. Academic Press.

Chen, Y., Zhao, Y., Hammond, J., Hsu, H. T., Evans, J. & Feldlaufer, M. (2004). Multiple virus infections in the honey bee and genome divergence of honey bee viruses. J Invertebr Pathol 87, 84–93.

Chen, Y., Pettis, J. S. & Feldlaufer, M. F. (2005). Detection of multiple viruses in queens of the honey bee Apis mellifera L. J Invertebr Pathol 90, 118–121.
Clark, T. B. (1978). A filamentous virus of the honey bee. J Invertebr Pathol 32, 332–340.
Core, A., Runckel, C., Ivers, J., Quock, C., Siapno, T., DeNault, S., Brown, B., DeRisi, J., Smith, C. D. & Hafernik, J. (2012). A new threat to honey bees, the parasitic phorid fly Apocephalus borealis. PLoS One 7, 1–9.
Cornman, R. S., Tarpy, D. R., Chen, Y., Jeffreys, L., Lopez, D., Pettis, J. S., VanEngelsdorp, D. & Evans, J. D. (2012). Pathogen webs in collapsing honey bee colonies. PLoS One 7, e43562.
Cox-Foster, D. L., Conlan, S., Holmes, E. C., Palacios, G., Evans, J. D., Moran, N. A., Quan, P.-L., Briese, T., Hornig, M. & other authors. (2007). A metagenomic survey of microbes in honey bee Colony Collapse Disorder. Science 318, 283–287.
Darriba, D., Taboada, G. L., Doallo, R. & Posada, D. (2012). jModelTest 2: more models, new heuristics and parallel computing. Nat Methods 9, 772.
Daughenbaugh, K. F., Martin, M., Brutscher, L. M., Cavigli, I., Garcia, E., Lavin, M. & Flenniken, M. L. (2015). Honey bee infecting Lake Sinai Viruses. Viruses 7, 3285–309.
Doublet, V., Labarussias, M., de Miranda, J. R., Moritz, R. F. A. & Paxton, R. J. (2015). Bees under stress: Sublethal doses of a neonicotinoid pesticide and pathogens interact to elevate honey bee mortality across the life cycle. Environ Microbiol 17, 969–983.
Evans, J. D., Schwarz, R. S., Chen, Y. P., Budge, G., Cornman, R. S., De la Rua, P., de Miranda, J. R., Foret, S., Foster, L. & other authors. (2013). Standard methods for molecular research in Apis mellifera. J Apic Res 52, 1–54.
Evans, J. D. & Schwarz, R. S. (2011). Bees brought to their knees: Microbes affecting honey bee health. Trends Microbiol 19, 614–620.
Evison, S. E. F., Roberts, K. E., Laurenson, L., Pietravalle, S., Hui, J., Biesmeijer, J. C., Smith, J. E., Budge, G. & Hughes, W. O. H. (2012). Pervasiveness of parasites in pollinators. PLoS One 7, e30641.
Fairbrother, A., Purdy, J., Anderson, T. & Fell, R. (2014). Risks of neonicotinoid insecticides to honey bees. Environ Toxicol Chem 33, 719–31.
Folmer, 0., Black, M., Hoeh, W., Lutz, R. & Vrijenhoek, R. (1994). DNA primers for amplification of mitochondrial cytochrome c oxidase subunit I from diverse metazoan invertebrates. Mol Mar Biol Biotechnol 3, 294–299.
Fürst, M. A., McMahon, D. P., Osborne, J. L., Paxton, R. J. & Brown, M. J. F. (2014). Disease associations between honey bees and bumble bees as a threat to wild pollinators. Nature 506, 364–6. Nature Publishing Group.
Gallai, N., Salles, J. M., Settele, J. & Vaissière, B. E. (2009). Economic valuation of the vulnerability of world agriculture confronted with pollinator decline. Ecol Econ 68, 810–821.
Gauthier, L., Cornman, S., Hartmann, U., Cousserans, F., Evans, J. D., De Miranda, J. R. & Neumann, P. (2015). The Apis mellifera filamentous virus genome. Viruses 7, 3798–3815.
Genersch, E. & Aubert, M. (2010). Emerging and re-emerging viruses of the honey bee (Apis mellifera L.). Vet Res 41, 54.
Genersch, E., Yue, C., Fries, I. & de Miranda, J. R. (2006). Detection of Deformed wing virus, a honey bee viral pathogen, in bumble bees (Bombus terrestris and Bombus pascuorum) with wing deformities. J Invertebr Pathol 91, 61–63.
Goulson, D., Nicholls, E., Botias, C. & Rotheray, E. L. (2015). Bee declines driven by combined stress from parasites, pesticides, and lack of flowers. Science 347, 1255957–1255957.
Gouy, M., Guindon, S. & Gascuel, O. (2010). SeaView Version 4: A Multiplatform Graphical User Interface for Sequence Alignment and Phylogenetic Tree Building. Mol Biol Evol 27, 221–224.
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 honey bees: Co-infection by Aphid Lethal Paralysis, Israel Acute Paralysis and Lake Sinai Viruses. PLoS One 8, e57459.
Graystock, P., Yates, K., Evison, S. E. F., Darvill, B., Goulson, D. & Hughes, W. O. H. (2013). The Trojan hives: Pollinator pathogens, imported and distributed in bumble bee colonies. J Appl Ecol 50, 1207–1215.
Graystock, P., Goulson, D. & Hughes, W. O. H. (2014). The relationship between managed bees and the prevalence of parasites in bumble bees. PeerJ 2, e522. PeerJ Inc.
Guindon, S. & Gascuel, O. (2003). A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. Syst Biol 52, 696–704.
Hu, Y., Fei, D., Jiang, L., Wei, D., Li, F., Diao, Q. & Ma, M. (2016). A comparison of biological characteristics of three strains of Chinese sacbrood virus in Apis cerana. Sci Rep 6, 37424. Nature Publishing Group.
Katoh, K., Misawa, K., Kuma, K. & Miyata, T. (2002). MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. Nucleic Acids Res 30, 3059–3066.

Kielmanowicz, M. G., Inberg, A., Lerner, I. M., Golani, Y., Brown, N., Turner, C. L., Hayes, G. J. R. & Ballam, J. M. (2015). Prospective large-scale field study generates predictive model identifying major contributors to colony losses. PLOS Pathog 11, e1004816.

Klein, A.-M., Vaissiere, B. E., Cane, J. H., Steffan-Dewenter, I., Cunningham, S. A., Kremen, C. & Tscharntke, T. (2007). Importance of pollinators in changing landscapes for world crops. Proc R Soc B Biol Sci 274, 303–313.

Levitt, A. L., Singh, R., Cox-Foster, D. L., Rajotte, E., Hoover, K., Ostiguy, N. & Holmes, E. C. (2013). Cross-species transmission of honey bee viruses in associated arthropods. Virus Res 176, 232–240.

Li, J. L., Cornman, R. S., Evans, J. D., Pettis, J. S., Zhao, Y., Murphy, C., Peng, W. J., Wu, J., Hamilton, M. & other authors. (2014). Systemic Spread and Propagation of a Plant-Pathogenic Virus in European Honey bees, Apis mellifera. MBio 5, e00898-13.

Li, J., Peng, W., Wu, J., Strange, J. P., Boncristiani, H. & Chen, Y. (2011). Cross-species infection of deformed wing virus poses a new threat to pollinator conservation. J Econ Entomol 104, 732–739.

Manley, R., Boots, M. & Wilfert, L. (2015). Emerging viral disease risk to pollinating insects: ecological, evolutionary and anthropogenic factors. J Appl Ecol 52, 331–340.

Martin, S. J., Highfield, A. C., Brettell, L., Villalobos, E. M., Budge, G. E., Powell, M., Nikaido, S. & Schroeder, D. C. (2012). Global Honey Bee Viral Landscape Altered by a Parasitic Mite. Science 336, 1304–1306.

Mazzei, M., Carrozza, M. L., Luisi, E., Forzan, M., Giusti, M., Sagona, S., Tolari, F. & Felicioli, A. (2014). Infectivity of DWV associated to flower pollen: Experimental evidence of a horizontal transmission route. PLoS One 9, e113448.

McMahon, D. P., Fürst, M. A., Caspar, J., Theodorou, P., Brown, M. J. F. & Paxton, R. J. (2015). A sting in the spit: widespread cross-infection of multiple RNA viruses across wild and managed bees. J Anim Ecol 84, 615–624 (S. Altizer, Ed.).

McMahon, D. P., Natsopoulou, M. E., Doublet, V., Fürst, M., Weging, S., Brown, M. J. F., Gogol-Döring, A. & Paxton, R. J. (2016). Elevated virulence of an emerging viral genotype as a driver of honey bee loss. Proc R Soc B Biol Sci 283, 20160811.

McMenamin, A. J. & Genersch, E. (2015). Honey bee colony losses and associated viruses. Curr Opin Insect Sci 8, 121–129.

Meeus, I., de Miranda, J. R., de Graaf, D. C., Wäckers, F. & Smagghe, G. (2014). Effect of oral infection with Kashmir bee virus and Israeli acute paralysis virus on bumble bee (Bombus terrestris) reproductive success. J Invertebr Pathol 121, 64–69. Elsevier Inc.

Menail, A. H., Piot, N., Meeus, I., Smagghe, G. & Loucif-Ayad, W. (2016). Large pathogen screening reveals first report of Megaselia scalaris (Diptera: Phoridae) parasitizing Apis mellifera intermissa (Hymenoptera: Apidae). J Invertebr Pathol 137, 33–37.

de Miranda, J. R. (2008). Diagnostic techniques for virus detection in honey bees. In Virol Honey bee, pp. 121–232. Edited by M. Aubert, B. Ball, I. Fries, R. Moritz, N. Milani & I. Bernardinelli. Brussels, Belgium: EEC Publications.

de Miranda, J. R., Bailey, L., Ball, B. V, Blanchard, P., Budge, G. E., Chejanovsky, N., Chen, Y.-P., Gauthier, L., Genersch, E. & other authors. (2013). Standard methods for virus research in Apis mellifera. J Apic Res 52, 1–56.

Monceau, K., Bonnard, O. & Thiéry, D. (2014). Vespa velutina: a new invasive predator of honey bees in Europe. J Pest Sci (2004) 87, 1–16.

Mookhploy, W., Kimura, K., Disayathanoowat, T., Yoshiyama, M., Hondo, K. & Chantawannakul, P. (2015). Capsid Gene Divergence of Black Queen Cell Virus Isolates in Thailand and Japan Honey Bee Species. J Econ Entomol 108, 1460–1464.

Mordecai, G. J., Brettell, L. E., Pachori, P., Villalobos, E. M., Martin, S. J., Jones, I. M. & Schroeder, D. C. (2016a). Moku virus; a new Iflavirus found in wasps, honey bees and Varroa. Sci Rep 6, 34983.

Mordecai, G. J., Wilfert, L., Martin, S. J., Jones, I. M. & Schroeder, D. C. (2016b). Diversity in a honey bee pathogen: first report of a third master variant of the Deformed Wing Virus quasispecies. ISME J 10, 1–10. Nature Publishing Group.

Newcombe, R. G. (1998). Two-sided confidence intervals for the single proportion: comparison of seven methods. Stat Med 17, 857–72.

Parmentier, L., Smagghe, G., de Graaf, D. C. & Meeus, I. (2016). Varroa destructor Macula-like virus, Lake Sinai virus and other new RNA viruses in wild bumble bee hosts (Bombus pascuorum, Bombus lapidarius and Bombus pratorum). J Invertebr Pathol 134, 6–11.

Peng, W., Li, J., Boncristiani, H., Strange, J. P., Hamilton, M. & Chen, Y. (2011). Host range expansion of honey bee Black Queen Cell Virus in the bumble bee, Bombus huntii. Apidologie 42, 650–658.

Potts, S. G., Biesmeijer, J. C., Kremen, C., Neumann, P., Schweiger, O. & Kunin, W. E. (2010). Global pollinator declines: trends, impacts and drivers. Trends Ecol Evol 25, 345–353.

Potts, S. G., Imperatriz-Fonseca, V., Ngo, H. T., Aizen, M. A., Biesmeijer, J. C., Breeze, T. D., Dicks, L. V., Garibaldi, L. A., Hill, R. & other authors. (2016). Safeguarding pollinators and their values to human well-being. Nature. Nature Research.

R Development Core Team. (2008). R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing.

Ratnasingham, S. & Hebert, P. D. N. (2007). BARCODING: bold: The Barcode of Life Data System (http://www.barcodinglife.org). Mol Ecol Notes 7, 355–364.

Reddy, K. E., Noh, J. H., Choe, S. E., Kweon, C. H., Yoo, M. S., Doan, H. T. T., Ramya, M., Yoon, B.-S., Nguyen, L. T. K. & other authors. (2013). Analysis of the complete genome sequence and capsid region of black queen cell viruses from infected honey bees (Apis mellifera) in Korea. Virus Genes 47, 126–132.

Roberts, J. M. K. & Anderson, D. L. (2014). A novel strain of sacbrood virus of interest to world apiculture. J Invertebr Pathol 118, 71–74. Elsevier Inc.

Runckel, C., Flenniken, M. L., Engel, J. C., Ruby, J. G., Ganem, D., Andino, R. & DeRisi, J. L. (2011). Temporal analysis of the honey bee microbiome reveals four novel viruses and seasonal prevalence of known viruses, Nosema, and Crithidia. PLoS One 6, e20656.

Sánchez-Bayo, F., Goulson, D., Pennacchio, F., Nazzi, F., Goka, K. & Desneux, N. (2016). Are bee diseases linked to pesticides? - A brief review. Environ Int 89–90, 7–11.

Sébastien, A., Lester, P. J., Hall, R. J., Wang, J., Moore, N. E. & Gruber, M. A. M. (2015). Invasive ants carry novel viruses in their new range and form reservoirs for a honey bee pathogen. Biol Lett 11, 20150610.

Sguazza, G. H., Reynaldi, F. J., Galosi, C. M. & Pecoraro, M. R. (2013). Simultaneous detection of bee viruses by multiplex PCR. J Virol Methods 194, 102–106.

Singh, R., Levitt, A. L., Rajotte, E. G., Holmes, E. C., Ostiguy, N., VanEngelsdorp, D., Lipkin, W. I., DePamphilis, C. W., Toth, A. L. & Cox-Foster, D. L. (2010). RNA Viruses in Hymenopteran pollinators: Evidence of inter-taxa virus transmission via pollen and potential impact on non-Apis Hymenopteran species. PLoS One 5, e14357.

Tehel, A., Brown, M. J. F. & Paxton, R. J. (2016). Impact of managed honey bee viruses on wild bees. Curr Opin Virol 19, 16–22.

Tentcheva, D., Gauthier, L., Zappulla, N., Dainat, B., Cousserans, F., Colin, M. E. & Bergoin, M. (2004). Prevalence and seasonal variations of six bee viruses in Apis mellifera L. and Varroa destructor mite populations in France. Appl Environ Microbiol 70, 7185–7191.

Tozkar, C. Ö., Kence, M., Kence, A., Huang, Q. & Evans, J. D. (2015). Metatranscriptomic analyses of honey bee colonies. Front Genet 6, 100. Frontiers Media SA.

Traynor, K. S., Rennich, K., Forsgren, E., Rose, R., Pettis, J., Kunkel, G., Madella, S., Evans, J., Lopez, D. & VanEngelsdorp, D. (2016). Multiyear survey targeting disease incidence in US honey bees. Apidologie 47, 325–347.

VanEngelsdorp, D. & Meixner, M. D. (2010). A historical review of managed honey bee populations in Europe and the United States and the factors that may affect them. J Invertebr Pathol 103, S80–S95.

Wetterer, J. K. (2007). Biology and impacts of Pacific Island invasive species. 3. The African big-headed ant, Pheidole megacephala (Hymenoptera : Formicidae). Pacific Sci 61, 437–456.

Wilfert, L., Long, G., Leggett, H. C., Schmid-Hempel, P., Butlin, R., Martin, S. J. M. & Boots, M. (2016). Deformed wing virus is a recent global epidemic in honey bees driven by Varroa mites. Science 351, 594–597.

Xia, X., Zhou, B. & Wei, T. (2015). Complete genome of Chinese sacbrood virus from Apis cerana and analysis of the 3C-like cysteine protease. Virus Genes 50, 277–285.

Yañez, O., Zheng, H. Q., Hu, F. L., Neumann, P. & Dietemann, V. (2012). A scientific note on Israeli acute paralysis virus infection of Eastern honey bee Apis cerana and vespine predator Vespa velutina. Apidologie 43, 587–589.

Zhang, X., He, S. Y., Evans, J. D., Pettis, J. S., Yin, G. F. & Chen, Y. P. (2012). New evidence that deformed wing virus and black queen cell virus are multi-host pathogens. J Invertebr Pathol 109, 156–159.

Table S1: Detailed information of all 523 individual samples used in this study. (not supplied)

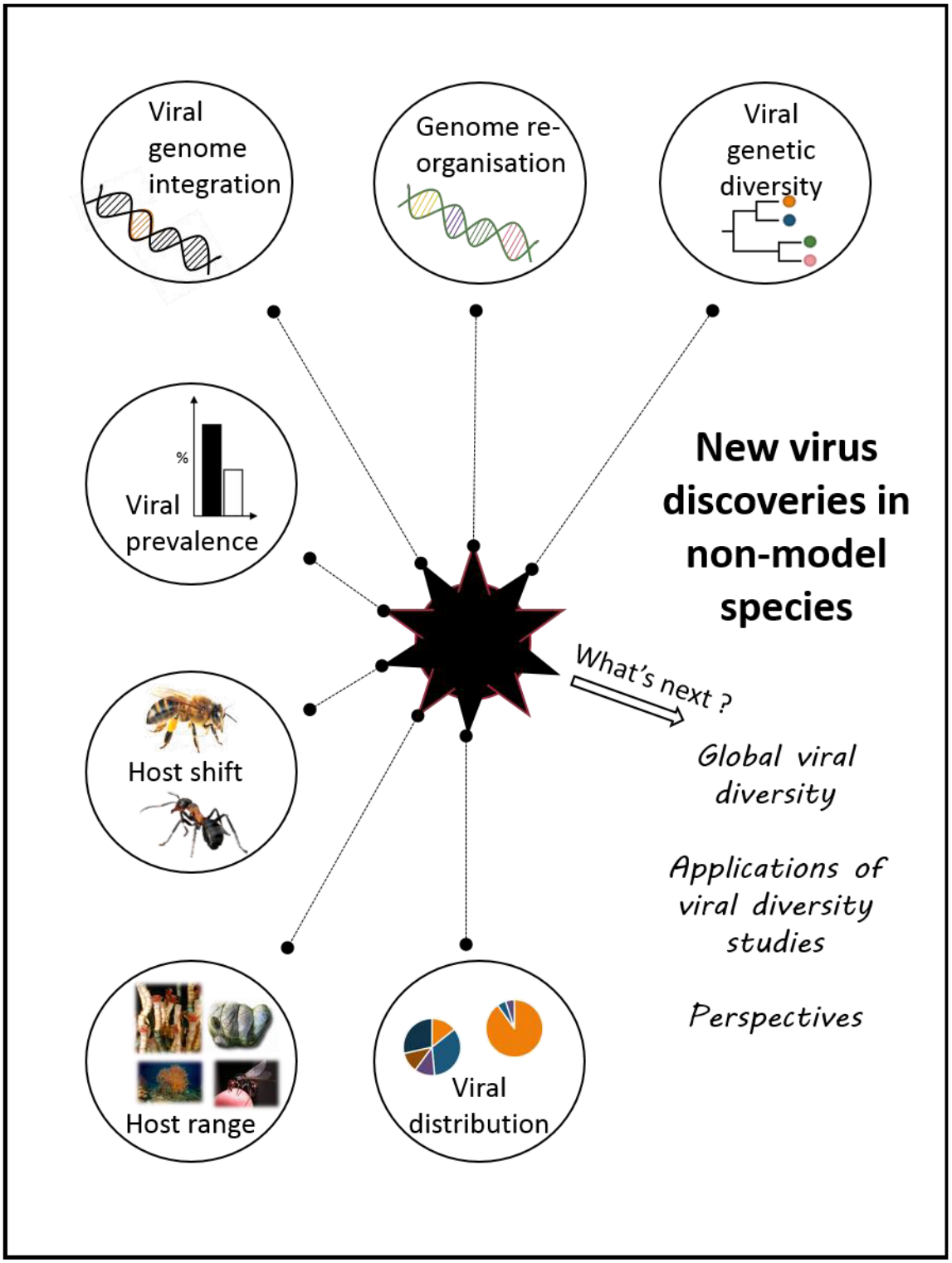
Table S2: Detailed information of all new viral sequences produced in this study.

Samples informations							PCR screening	BLASTN results of amplified sequences					Taxonomic assignement confirmed by phylogeny	
Individual	Country	Location	Order	Family	Genus	Specie	Targeted virus species	Sequence first BLAST hit	Target GenBank Accession	Query coverage	E-value	Identity	Phylogenetic virus	Sequence GenBank accession
A079	Canada	Whistler	Hymenoptera	Formicidae	Formica	aserva	Acute bee paralysis virus (ABPV)	*Formica exsecta virus 1*	KF500001	97%	7,00E-172	95%	*Formica exsecta virus 1*	**Pending**
A080	Canada	Whistler	Hymenoptera	Formicidae	Formica	aserva		*Formica exsecta virus 1*	KF500001	96%	7,00E-172	95%	*Formica exsecta virus 1*	**Pending**
F2	Canada	Guelph	Hymenoptera	Formicidae	Formica	aserva		*Formica exsecta virus 1*	KF500001	97%	1,00E-173	95%	*Formica exsecta virus 1*	**Pending**
A062	Canada	Whistler	Hymenoptera	Formicidae	Formica	neorufibarbis	Deformed wing virus (DWV)	*Deformed wing virus*	KX783225	100%	6,00E-88	99%	*Deformed wing virus*	**Pending**
Av02A	France	Avoine	Hymenoptera	Apidae	Apis	mellifera		*Deformed wing virus*	KX580899	98%	3,00E-101	98%	*Deformed wing virus*	**Pending**
Av02B	France	Avoine	Hymenoptera	Apidae	Apis	mellifera		*Deformed wing virus*	KX580899	98%	3,00E-96	98%	*Deformed wing virus*	**Pending**
Av02C	France	Avoine	Hymenoptera	Apidae	Apis	mellifera		*VDV-1/DWV recombinant*	KX373900	98%	3,00E-96	97%	*Deformed wing virus*	**Pending**
Av02D	France	Avoine	Hymenoptera	Apidae	Apis	mellifera		*VDV-1/DWV recombinant*	KX373900	98%	1,00E-99	97%	*Deformed wing virus*	**Pending**
Av05A	France	Avoine	Hymenoptera	Apidae	Apis	mellifera		*VDV-1/DWV recombinant*	KX373900	98%	6,00E-98	97%	*Deformed wing virus*	**Pending**
Av05D	France	Avoine	Hymenoptera	Apidae	Apis	mellifera		*Deformed wing virus*	KX580899	98%	3,00E-101	98%	*Deformed wing virus*	**Pending**
BVD27A	France	Beaumont-en-Véron	Hymenoptera	Apidae	Apis	mellifera		*Deformed wing virus*	KX580899	97%	6,00E-103	98%	*Deformed wing virus*	**Pending**
BVD27B	France	Beaumont-en-Véron	Hymenoptera	Apidae	Apis	mellifera		*VDV-1/DWV recombinant*	KX373900	97%	3,00E-100	98%	*Deformed wing virus*	**Pending**
C065	Canada	Squamish	Hymenoptera	Apidae	Apis	mellifera		*Deformed wing virus*	KX783225	100%	1,00E-89	98%	*Deformed wing virus*	**Pending**
C074	Canada	Whistler	Hymenoptera	Apidae	Apis	mellifera		*Deformed wing virus*	KX783225	100%	1,00E-89	98%	*Deformed wing virus*	**Pending**
Hilo005	Hawaï	Hilo	Hymenoptera	Apidae	Apis	mellifera		*Deformed wing virus*	KX783225	100%	1,00E-89	98%	*Deformed wing virus*	**Pending**
Hilo006	Hawaï	Hilo	Hymenoptera	Apidae	Apis	mellifera		*Deformed wing virus*	KT004425	100%	7,00E-87	98%	*Deformed wing virus*	**Pending**
Hilo007	Hawaï	Hilo	Hymenoptera	Apidae	Apis	mellifera		*Deformed wing virus*	KT004425	99%	2,00E-93	99%	*Deformed wing virus*	**Pending**
Hilo008	Hawaï	Hilo	Hymenoptera	Apidae	Apis	mellifera		*Deformed wing virus*	KT004425	99%	3,00E-95	99%	*Deformed wing virus*	**Pending**
Hilo010	Hawaï	Hilo	Hymenoptera	Apidae	Apis	mellifera		*Deformed wing virus*	KT004425	100%	7,00E-87	97%	*Deformed wing virus*	**Pending**
Hilo011	Hawaï	Hilo	Hymenoptera	Apidae	Apis	mellifera		*Deformed wing virus*	KX580899	99%	3,00E-90	98%	*Deformed wing virus*	**Pending**
Hilo023	Hawaï	Hilo	Hymenoptera	Formicidae	Pheidole	megacephala		*Deformed wing virus*	KT004425	100%	1,00E-70	92%	*Deformed wing virus*	**Pending**
Neo1D	France	Saint-Branchs	Hymenoptera	Apidae	Apis	mellifera		*Deformed wing virus*	KX580899	98%	6,00E-103	98%	*Deformed wing virus*	**Pending**
VC-F1	France	Fondettes	Hymenoptera	Vespidae	Vespa	crabro		*Deformed wing virus*	KX580899	100%	3,00E-91	100%	*Deformed wing virus*	**Pending**
VC-S2	France	Saché	Hymenoptera	Vespidae	Vespa	crabro		*Deformed wing virus*	KX783225	100%	6,00E-93	99%	*Deformed wing virus*	**Pending**

VC-S3	France	Saché	Hymenoptera	Vespidae	Vespa	crabro		*Deformed wing virus*	GU903475	99%	7,00E-97	99%	*Deformed wing virus*	**Pending**
VV-L2	France	Lussault-sur-Loire	Hymenoptera	Vespidae	Vespa	velutina		*VDV-1/DWV recombinant*	KX373900	100%	2,00E-87	98%	*Deformed wing virus*	**Pending**
Av05B	France	Avoine	Hymenoptera	Apidae	Apis	mellifera	Sacbrood virus (SBV)	*Sacbrood virus*	JQ390591	99%	3,00E-140	99%	*Sacbrood virus*	**Pending**
BVD25A	France	Beaumont-en-Véron	Hymenoptera	Apidae	Apis	mellifera		*Sacbrood virus*	JQ390591	100%	2,00E-147	99%	*Sacbrood virus*	**Pending**
C001	Canada	Vancouver	Hymenoptera	Apidae	Bombus	sp.		*Sacbrood virus*	KM001901	100%	7,00E-142	99%	*Sacbrood virus*	**Pending**
C013	Canada	Tofino	Hymenoptera	Apidae	Apis	mellifera		*Sacbrood virus*	JQ390592	98%	7,00E-147	99%	*Sacbrood virus*	**Pending**
F006	France	Saint-Avertin	Hymenoptera	Apidae	Apis	mellifera		*Sacbrood virus*	JQ390591	100%	2,00E-132	99%	*Sacbrood virus*	**Pending**
Hilo009	Hawaï	Hilo	Hymenoptera	Apidae	Apis	mellifera		*Sacbrood virus*	KC513760	100%	4,00E-134	100%	*Sacbrood virus*	**Pending**
Jc02A	France	Saint-Branchs	Hymenoptera	Apidae	Apis	mellifera		*Sacbrood virus*	JQ390591	100%	1,00E-143	100%	*Sacbrood virus*	**Pending**
Neo6D	France	Saint-Branchs	Hymenoptera	Apidae	Apis	mellifera		*Sacbrood virus*	JQ390591	100%	4,00E-144	99%	*Sacbrood virus*	**Pending**
VV-L1	France	Lussault-sur-Loire	Hymenoptera	Vespidae	Vespa	velutina		*Sacbrood virus*	JQ390591	100%	5,00E-138	99%	*Sacbrood virus*	**Pending**
C012	Canada	Tofino	Hymenoptera	Apidae	Apis	mellifera	Black queen cell virus (BQCV)	*Black queen cell virus*	KJ123661	100%	0.0	98%	*Black queen cell virus*	**Pending**
C046	Canada	Victoria	Hymenoptera	Apidae	Apis	mellifera		*Black queen cell virus*	KJ123661	99%	0.0	99%	*Black queen cell virus*	**Pending**
C068	Canada	Whistler	Hymenoptera	Apidae	Apis	mellifera		*Black queen cell virus*	HQ655487	99%	0.0	99%	*Black queen cell virus*	**Pending**
C076	Canada	Whistler	Hymenoptera	Apidae	Bombus	sitkensis		*Black queen cell virus*	KJ123661	100%	0.0	99%	*Black queen cell virus*	**Pending**
F007	France	Saint-Avertin	Hymenoptera	Apidae	Apis	mellifera		*Black queen cell virus*	KP223795	100%	0.0	100%	*Black queen cell virus*	**Pending**
G-A1	Canada	Guelph	Hymenoptera	Apidae	Apis	mellifera		*Black queen cell virus*	KJ123661	100%	0.0	99%	*Black queen cell virus*	**Pending**
G-A3	Canada	Guelph	Hymenoptera	Apidae	Apis	mellifera		*Black queen cell virus*	KJ123661	100%	0.0	99%	*Black queen cell virus*	**Pending**
VC-F2	France	Fondettes	Hymenoptera	Vespidae	Vespa	crabro		*Black queen cell virus*	KP223793	100%	0.0	97%	*Black queen cell virus*	**Pending**
VC-T1	France	Tours	Hymenoptera	Vespidae	Vespa	crabro		*Black queen cell virus*	KP223792	99%	0.0	97%	*Black queen cell virus*	**Pending**

Table S3: Accession numbers of all published sequences used in phylogenies. (not supplied)

DISCUSSION GENERALE


Viral genome integration
Genome re-organisation
Viral genetic diversity
%
Viral prevalence
New virus discoveries in non-model species
What's next ?
Global viral diversity
Applications of viral diversity studies
Perspectives
Host shift
Host range
Viral distribution

5.1. Préambule à la discussion générale

Après avoir énoncé dans un premier chapitre trois exemples de découverte de nouveaux virus, par une approche centrée sur les virus, puis après avoir énoncé l'exemple des virus d'abeilles, se plaçant d'un point vue plus écologique et plus du côté de l'étude des hôtes, cette partie permettra l'intégration de ces deux approches.

Dans un premier temps, je vais rappeler en quoi l'arrivée des NGS a permi d'avoir une nouvelle perspective sur la taxonomie virale. Les génomes viraux étant au cœur de l'affiliation taxonomique ; les études de caractérisation génomiques et phylogénétiques sont essentielles. Sera ensuite abordé de manière succincte le bilan général des résultats viraux.

Puis, dans un deuxième temps, je discuterai de l'importance de l'étude des hôtes impliqués dans les interactions hôtes-virus et de l'étude de la gamme d'hôte sur la transmission et distribution des virus.

Dans un troisième temps, je ferai le bilan des virus découverts dans l'ensemble des transcriptomes animaux à disposition dans cette étude et montrerai en quoi la diversité virale reste encore à découvrir par l'étude de plus en plus d'animaux non-modèles.

Enfin, dans une conclusion générale, j'aborderai les limites de mon approche tout en soulignant les avancées que cette étude a permises et les perspectives que cela induira.

5.2. Importance de l'étude des virus dans les animaux non-modèles

L'étude d'animaux non-modèles par l'utilisation des nouvelles technologies de séquençage permet d'avoir accès à de nouvelles informations utiles à la compréhension de l'évolution des virus. Ainsi, des informations relatives i) à la diversité virale au sein de ces hôtes, ii) à la taxonomie des virus par leurs caractéristiques génomiques propres ou iii) aux liens qui les unissent à leur(s) hôte(s), deviennent accessibles.

5.2.1. Nécessité des NGS pour étudier la diversité virale

Depuis plus d'une décennie, la production de données de métagénomique par le développement des nouvelles génération de séquençage a permi la découverte de très nombreux nouveaux virus (Edwards & Rohwer, 2005; Mokili et al., 2012; Rosario & Breitbart, 2011). En effet, les découvertes de virus se sont accélérées encore récemment lors d'une étude de grande ampleur permettant de mettre à jour plus de 125 000 génomes de virus à ADN au sein de différents environnements tels que le milieu aquatique, terrestre ou au sein d'hôtes (Paez-Espino *et al.*, 2016). D'autres études, se basant sur des mélanges d'espèces invertébrées, ont montré que la diversité virale est finalement beaucoup plus vaste que celle qui est actuellement décrite par la taxonomie, agrandissant encore plus la vision que l'on a de cette diversité (Li *et al.*, 2015; Shi *et al.*, 2016; Webster *et al.*, 2015).

Bien que mon approche ne soit pas semblable à celles utilisées dans les études précédentes puisqu'il ne s'agit pas de méta-viromique sensu stricto, mon travail a toutefois permis de montrer encore un peu plus l'étendue de la diversité virale par l'étude individuelle d'animaux non-modèles. Dans un premier temps, elle est la preuve de concept que l'étude des transcriptomes, bien qu'elle ait des limites par rapport à la découverte de virus (comme le biais de l'étude des ARN uniquement et la reconstruction de génomes incomplets aux extrémités) s'avère néanmoins fructueuse. Ainsi, mon étude montre que l'utilisation des transcriptome est une voie qui gagne à être approfondie et requière bien d'autres études au vu du très grand nombre de transcriptomes publiés à ce jour dans les bases de données.

L'un des exemples développé dans ma thèse est celui de la diversité des parvovirus. Des recherches dans les bases de données publiques réalisées par mes collaborateurs ainsi que celles détectées par mon approche, permettent de souligner que cette famille virale possède une gamme d'hôte beaucoup plus étendue qu'initialement décrite (**Article 2**). Les *Parvoviridae,* classés en deux sous-familles, étaient connus pour infecter les vertébrés principalement les oiseaux et les mammifères pour les *Parvovirinae* et les arthropodes principalement les insectes et les crustacés pour les *Densovirinae*. Cette nouvelle étude a ainsi permi de montrer que les *Parvoviridae* sont finalement capable d'infecter plus d'hôtes, incluant des mollusques, annélides, nématodes et cnidaires, supportant l'hypothèse que les parvovirus sont distribués dans tous le règne animal (**Article 2**). Ainsi, comme les génomes de plus en plus d'animaux sont séquencés à l'heure actuelle, de grandes quantité de données seront disponibles dans les bases de données dans les années qui viennent et permettront d'augmenter considérablement les connaissances sur la diversité des parvovirus par exemple, mais de tous les virus de manière plus générale.

L'étude des virus est très liée aux avancées technologiques et l'apport du séquençage permet ainsi de résoudre des questions liées à l'origine des virus, leur évolution et surtout leur classification taxonomique. Jusqu'à maintenant, afin que les espèces virales soient intégrées à la classification officielle des virus acceptée par l'ICTV, de nombreuse preuves biologiques sont nécessaires telles que la visualisation par microscopie ou la présence d'une pathologie induite par le(s) virus considéré(s). Il a ainsi été proposé récemment d'intégrer de manière plus simple les virus découverts par des approches de métagénomique au sein de la taxonomie virale. Avec l'arrivée de la métagénomique, et le fait que les études des génomes et des phylogénies soient plus pertinentes, il est proposé d'attribuer des noms de genre et d'espèces aux nouveaux génomes viraux pour que la taxonomie virale puisse suivre le rythme des découvertes. En effet, toutes les découvertes ne pouvaient bénéficier d'un statut taxonomique validé par l'ICTV sans preuve biologique (pathogénicité, gamme d'hôte, épidémiologie, structure du virion, propriétés antigéniques) et donc se voyaient dans l'impossibilité d'être réutilisés facilement et de manière rigoureuse dans des études ultérieures, ce qui est un frein à l'avancée de la taxonomie virale (Simmonds *et al.*, 2017).

L'un des biais rencontré par l'étude de métagénomes/-transcriptomes est la présence de nombreuses séquences virales qui ne peuvent pas être assemblés de manière complète et demeurent alors partielles et fragmentées. Cependant, il est possible de réaliser des assignations taxonomiques même pour les petits fragments par phylogénie (de la réplicase notamment), des études de génomique comparative jusqu'à des études de la variabilité intra-population pour les séquences les plus complètes lorsque la couverture le permet (Ladner *et al.*, 2014). A titre d'exemple, Wang et ses collaborateurs ont déterminé que pour du séquençage haut-débit une couverture d'environ 400X est nécessaire pour identifier des variant mineurs présents à une fréquence de 1 % avec 99.999 % de confiance, et environ 1000X pour des variant de fréquence 0,5 % et que dans ces cas précis des amplifications ciblées sont nécessaires pour obtenir de telles couvertures (Wang *et al.*, 2007a).

Selon Ladner et collaborateurs, la définition de génome viraux "coding complete" (pour lesquels tous les cadres ouverts de lectures sont disponibles mais dont on ne possède pas l'entièreté du génome incluant les extrémités) correspondent à la définition de nouveaux virus et sont suffisants pour permettre une identification et des analyses phylogénétiques pertinentes (Ladner *et al.*, 2014).

5.2.2. Apporter des connaissances sur les génomes pour établir la taxonomie virale

La classification actuelle des virus à ARN est très largement basée sur des phylogénies de la réplicase (ADN-dépendant ADN polymérase chez les virus à ADN classes I-II ; ARN-dépendant ARN polymérase chez les virus à ARN classes III-IV-V ; ARN-dépendant ADN polymérase chez les rétrovirus classes VI-VII) qui possède des domaines protéiques très conservés liés à la fonction importante et maintenue qu'elle occupe pour tous les virus (Koonin, 1991). Néanmoins, l'analyse complète des génomes reste nécessaire pour valider l'existence et étudier les origines évolutives de ces virus.

Deux exemples issus de mon travail de thèse montrent à quel point l'analyse approfondie des génomes est essentielle et permet la description de nouvelles organisations génomiques illustrant la présence de nouvelles espèces, de nouveaux genres voire de nouvelles familles de virus.

Ainsi le CpATV, avec un ensemble d'autres séquences incluses dans le même clade (Clade 8, **Article 1**), ne présente pas les caractéristiques d'un groupe viral actuellement reconnu par l'ICTV et les analyses de génomique comparative et de phylogénie permettent de montrer qu'ils appartiennent à une nouvelle famille virale. En effet, le génome du CpATV contient des ORFs proches de *Virgaviridae* et Negevirus mais possède aussi un ORF particulier qui possède des similarités avec les *Plasmodium* ce qui montre son originalité génomique. Le CpATV est proche de trois autres séquences virales, le Boutonnet virus, TSA *Musca domestica* et *Adelphocoris suturalis virus 1* (ASV1) qui forment avec lui un clade monophylétique représentant au moins un nouveau genre viral présent au sein d'une nouvelle famille virale. Enfin, les analyses phylogénétiques ont permis de retracer l'évolution du CpATV parmi d'autres nouveau virus d'insectes découverts récemment (Webster *et al.*, 2016) et montrent que les analyses de génomique comparative permettent de mieux comprendre l'évolution des génomes viraux et de leur attribuer une taxonomie plus précise. Il reste cependant des analyses complémentaires à réaliser sur ce virus afin d'acquérir des connaissances sur sa gamme d'hôtes (plantes ou insectes, le moustique est-il l'hôte ou le vecteur de ce virus ?), sa prévalence et sur la biodiversité de ce virus.

C'est également le cas pour le HsAV (**Article 4**), où l'analyse de l'organisation génomique de ce virus, couplée à l'analyse phylogénétique, permettent elles aussi de montrer que plusieurs caractéristiques évolutives distinguent les HsAV des virus qui lui sont proches tels les *Chroparavirus*, *Sinaivirus* et *Nodaviridae*. Un nouveau genre, Halictivirus, a ainsi été proposé pour ce virus qui phylogénétiquement forme un groupe monophylétique ancestral aux *Sinaivirus* avec qui il partage 30 % d'identité de séquence.

Dans les deux cas, des analyses d'évolution moléculaire et de couverture génomique suggèrent que ces virus, soumis à une forte sélection négative, sont potentiellement fonctionnels et infectieux bien qu'aucun symptôme associé à l'infection par ces deux virus n'ait été observé. Grâce à cette première étude de découverte et de caractérisation génomique, étape utile à la définition taxonomique des virus, des études plus approfondies sur la pathogénicité, la réplication et la transmission de ces virus sont maintenant possibles et seront nécessaires à leur complète caractérisation.

5.2.3. Aperçu des virus présents dans les métagénomes animaux

Ce travail de thèse s'inscrit dans une approche globale de description de la diversité virale présente dans un ensemble de 523 transcriptomes d'animaux non-modèles, échantillonnés de manière à représenter au mieux la diversité animale. En effet, l'échantillonnage tentait de prendre en compte la présence des grands phyla animaux mais il existe cependant des phyla absents tels que les Platyhelminthes (29 285 espèces répertoriées), les Bryozoa (5 486 esp.), les Rotifera (2 049 esp.) ou encore les Hemichordata (103 esp.) (Zhang, 2013). Il existe aussi un biais en termes de nombre d'espèces échantillonnées, notamment par la surreprésentation des arthropodes et des vertébrés représentant à eux seuls près de 61 % des données (respectivement 38,4 % et 22,4 % des transcriptomes) et une sous-représentation de tous les autres phyla. Parmi les 523 transcriptomes animaux étudiés, cela représente néanmoins 135 espèces répartis dans les phyla avec 47 espèces chez les arthropodes, 35 chez les vertébrés, 21 chez les mollusques, 11 chez les annélides, 6 chez les nématodes, 5 chez les échinodermes et les némertes et 3 chez les cnidaires (Annexe 1). Le jeu de donnée initial était composé d'environ 11,3 milliard de lectures de séquençage Illumina, qui ont été assemblées en 35 millions de contigs, sur lesquels 20 millions de protéines ont été prédites (Figure 16).

Sur l'ensemble des données, 4720 protéines ayant des homologies virales ont été détectées par BLAST et BLAST réciproque, représentant environ **0,03 %** des protéines totales. Une analyse plus approfondie des résultats des hits viraux a permis d'identifier la présence potentielle de 200 génomes viraux complets représentants potentiellement environ 50 espèces de virus distinctes (Figure 16).

Ce qui est intéressant dans ce type d'analyse à grande échelle est de pouvoir observer sous deux angles différents la diversité virale suivant si l'on observe les hôtes ou les virus.

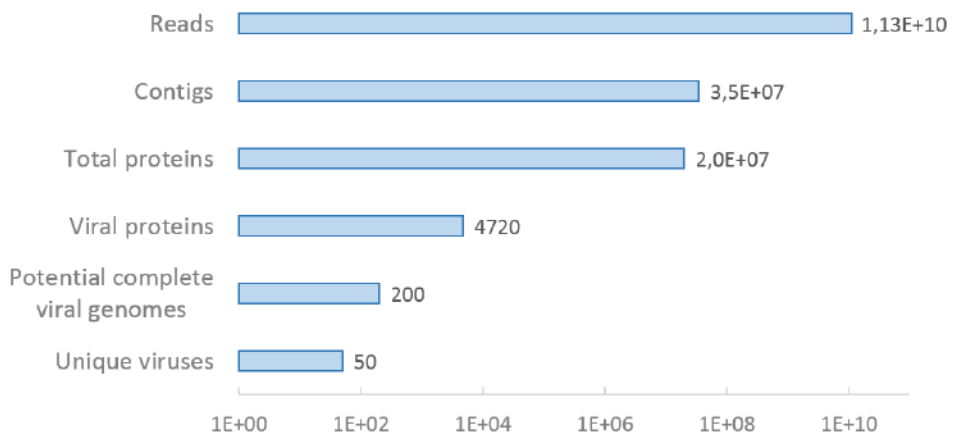


Figure 16 : Statistiques générales de l'analyse des 523 transcriptomes d'animaux non modèles.

5.3. Les hôtes animaux

5.3.1. Les connaissances actuelles des associations hôte-virus chez les animaux

A l'heure actuelle, 1 525 728 espèces animales sont répertoriées parmi lesquelles se trouvent les grands phyla animaux étudiés durant ce travail tels que les Annelida (17 388 espèces), Arthropoda (1 257 040 esp.), Chordata (68 295 esp.), Cnidaria (10 183 esp.), Echinodermata (7 550 esp.), Mollusca (84 977 esp.), Nematoda (25 033 esp.) et Nemerta (1 358 esp.) (Zhang, 2013).

Une base de données récente **Virus-Host db** (http://www.genome.jp/virushostdb/) répertorie les associations hôtes-virus présentes actuellement dans les bases de données de séquences génomiques (Mihara *et al.*, 2016). Virus-Host db utilise les bases de données de séquences telles que RefSeq et GenBank (NCBI) mais s'appuie également sur la base de données protéiques UniProt (EMBL-EBI) et la base de données ViralZone (Hulo *et al.*, 2011) pour établir les liens taxonomiques entre les virus et leurs hôtes (naturels et de laboratoire). Les associations répertoriées sont également vérifiées par une surveillance de la littérature et sont manuellement corrigées (Mihara *et al.*, 2016).

De nombreux virus sont connus pour infecter des animaux de ces différents phyla et pour lesquels des associations hôtes-virus sont répertoriées dans les bases, cependant nos connaissances sur la diversité virale se limitent souvent aux hôtes d'intérêts médicaux ou

agronomiques (arthropodes, vertébrés et mollusques), alors que pour les autres phyla nos connaissances semblent beaucoup plus rares (Tableau 4).

Tableau 4 : Nombre des associations hôtes-virus répertoriées dans les bases de données.

	dsDNA	dsRNA	ssDNA	ssRNA+	ssRNA-	RT	Non assignés	TOTAL
Annelida (17 388 esp.)	-	1	-	-	-	-	36	**37**
Arthropoda (1 257 040 esp.)	170	55	113	292	173	-	667	**1 470**
Chordata (68 295 esp.)	978	115	333	1 360	505	149	88	**3 528**
Cnidaria (10 183 esp.)	3	-	-	-	-	-	9	**12**
Echinodermata (7 550 esp.)	-	-	1	-	-	-	9	**10**
Mollusca (84 977 esp.)	6	1	38	7	-	-	217	**269**
Nematoda (25 033 esp.)	-	-	-	4	1	-	17	**22**
Nemerta (1 358 esp.)	-	-	-	-	-	-	-	**0**

Données des associations issues de la base de données *Virus-Host db* (http://www.genome.jp/virushostdb/) version GenBank (15/06/2017) version RefSeq (17/07/2017) (Mihara *et al.*, 2016).

Les arthropodes, les mollusques et les vertébrés comptent le plus grand nombre d'espèces ayant une association virale connue. Les mollusques possèdent 5 à 13 fois moins d'associations connues avec des virus que les arthropodes et les vertébrés respectivement, alors qu'ils ont 1,3 fois plus d'espèce que les vertébrés (Tableau 4). Les vertébrés possèdent 2,4 fois plus d'associations avec des virus répertoriées alors qu'ils ont 15 fois moins d'espèces que les arthropodes. Cependant, même au sein de ces 3 phyla bien connus, la biodiversité virale des hôtes animaux modèles est sûrement bien mieux caractérisée que celle des animaux non modèles. En effet, ce qui ressort de ces chiffres montre que les bases de données possèdent beaucoup de séquences virales dont les hôtes sont des vertébrés parmi lesquels le plus connu et le plus étudié est l'Homme, suivi des animaux d'élevages et domestiques. Les arthropodes, parmi lesquels on retrouve les insectes (majoritaires avec 1 020 007 espèces) (Junglen & Drosten, 2013) sont connus pour être vecteurs (moustiques, tiques, acariens) de maladies virales transmissible à l'Homme et à d'autres animaux (Vasilakis & Tesh, 2015) ou employés comme modèles d'études associés à la lutte biologique des ravageurs de cultures (Szewczyk *et al.*, 2006).

Ainsi il existe un biais des connaissances acquises sur les interactions hôte-virus directement en lien avec les connaissances acquises sur les hôtes. En effet, à côté des grands groupes

animaux, les annélides, les nématodes, les némertes, les échinodermes et les cnidaires sont très peu connus et ne possèdent pas beaucoup (voire pas pour les némertes) d'association avec des virus répertoriées dans les bases de données. Pour toutes ces catégories, les associations actuellement répertoriées comme non assignées à une classe virale proviennent d'une seule et même étude récente réalisée sur un grand échantillonnage et séquençage de mélanges d'espèces (Shi *et al.*, 2016). On se rend compte alors que l'inclusion de nouvelles espèces, dites non-modèles, permet d'avoir de plus en plus de connaissances dans l'étude de la diversité virale qui s'agrandit alors continuellement. En effet, cette étude a permis de combler des lacunes taxonomiques qui existaient entre familles et genre viraux (Shi *et al.*, 2016). D'autres études de ce type ont permi elles aussi d'apporter de nouvelles connaissances notamment sur les virus à ARN en observant un grand ensemble d'hôtes arthropodes (Li *et al.*, 2015) ou encore en réalisant des surveillances d'animaux susceptibles de contenir des virus pathogènes pour les humains tels les chauves-souris ou les rongeurs en découvrant une grande diversité de nouveaux virus dans leurs fèces (Li *et al.*, 2010; Phan *et al.*, 2011).

Par rapport au nombre d'espèces échantillonnées et au nombre d'associations hôte-virus connus dans ces phyla, l'attendu majeur de notre étude était de pourvoir détecter plus de hits viraux chez les arthropodes, les vertébrés et les mollusques de par leur surreprésentation dans l'échantillonnage et que ces hits restent proches de ce qui est connu dans les bases de données puisque les recherches d'homologies sont biaisées par les séquences virales déjà présentes dans ces bases.

Une limite à l'analyse des transcriptomes pour les associations hôte-virus est que les associations observées ne sont pas forcément corrélées à une réelle interaction entre l'hôte échantillonné et le virus. Il peut en effet s'agir de contamination trophique liée à la nourriture ingérée par les hôtes et il est nécessaire de vérifier la réplication au sein de l'hôte pour éliminer cette hypothèse. Néanmoins, l'étude des associations reste une première étape vers la caractérisation des interactions hôtes-virus, et cela ne remet pas en cause la réalité biologique de la détection virale. L'ultime preuve de l'interaction hôte-virus pourrait être apportée par une analyse de paléovirologie en détectant *in silico* des virus (entiers ou fragmentés) intégrés dans les génomes hôtes, reflets d'interactions passées (Gallot-Lavallée & Blanc, 2017; Katzourakis, 2013; Patel *et al.*, 2011).

5.3.2. L'observation de la diversité virale au sein des transcriptomes animaux

Tableau 5 : Répartition des différentes classes de hits viraux* détectés dans les transcriptomes animaux de cette étude.

Phylum animaux		Arthropoda	Chordata	Mollusca	Annelida	Nematoda	Echinodermata	Cnidaria	Nemerta	Haptophyta
# transcriptomes		201	117	79	43	27	24	12	8	12
Virus ADN	ADNdb	747	164	384	196	18	78	13	57	329
	ADNsb	113	32	41	31	24	4	8	1	0
Retro virus	ARNsb(RT)	32	213	1	8	0	4	0	1	2
	ADNdb(RT)	0	2	1	4	0	0	0	0	0
Virus ARN	ARNdb	59	5	5	0	0	1	0	0	0
	ARNsb(+)	967	94	124	77	2	14	3	2	0
	ARNsb(-)	228	164	37	33	5	1	17	0	0
N/A	Unclassified	322	8	20	14	0	1	0	0	9

* Ensemble de tous les hits ayant une homologie virale détectée par BLAST et BLAST réciproque dans les 523 transcriptomes. Le détail des familles virales détectées dans ces classes de virus sont accessible en Annexe 2.

Les arthropodes et les vertébrés

Les vertébrés et les arthropodes sont les deux phyla animaux pour lesquels de nombreuses associations hôte-virus sont répertoriées, dans toutes les classes de virus (Tableau 4). Dans cette étude, les hits viraux détectés parmi les 117 transcriptomes de vertébrés représentes pour la plupart de potentiels virus complets à ARNsb(+), ARNsb(-) ou encore chez les rétrovirus (Tableau 5, Annexe 2 et **Article 3**). C'est également le cas pour les arthropodes étudiés, pour lesquels toutes les catégories virales sont représentées, y compris les rétrovirus pourtant très majoritairement décrits chez les vertébrés à l'exception de rétrotransposons (*Pseudoviridae* et *Metaviridae*) connus chez les invertébrés (Terzian *et al.*, 2000; Zhou & Haymer, 1997).

Pour les arthropodes, de nombreux hits viraux à ADNdb sont catégorisés dans des familles virales connues pour infecter les arthropodes (Tableau 5, Annexe 2) tels que les *Polydnaviridae*, *Baculoviridae*, *Nudiviridae*, *Iridoviridae* ou *Poxviridae*. Cependant quelques familles virales elles aussi très bien caractérisées chez les arthropodes font défaut à cause d'un échantillonnage non exhaustif, tels que les *Hytrosaviridae* décrits chez certains diptères ou les *Ascovirus* décrits chez les lépidoptères Noctuidae (Asgari & Johnson, 2010; King *et al.*, 2012).

Parmi les hits de virus à ARNsb(+), de nombreux hits correspondent à de potentiels nouveaux virus, également appartenant à des familles virales très présentes chez les arthropodes tels que les *Picornavirales* (dont font partie les *Dicistroviridae* et *Iflaviridae)*, ou encore des virus non classés pour lesquels certains nouveaux virus complets ont été décrits précédemment dans les **Articles 1 et 4** (section 5.4., Tableau 6).

Ces résultats montrent que les arthropodes et vertébrés, étant les animaux les plus étudiés, possèdent plus de hits viraux détectés, ce qui était attendu au vu des nombreuses associations virales décrites pour ces deux groupes. Cette étude permet néanmoins de découvrir de nouveaux virus bien plus distants phylogénétiquement des virus déjà connus (**Article 1, 3 et 4**). Ceci est probablement lié à la conservation des domaines protéiques des réplicases virales qui permettent en première détection des homologies protéiques même quand les virus sont très éloignés phylogénétiquement.

Autres animaux vivants en milieu terrestre

Les nématodes, et certains annélides échantillonnés vivent en milieu terrestre. Pour ces phyla animaux, très peu d'associations hôtes-virus sont répertoriées (Tableau 4). Très peu de hits viraux ont été détectés dans ces phyla probablement due à une taille et représentativité sous-estimée dans notre échantillonnage (Tableau 5, Annexe 2). La faible connaissance des virus de ces hôtes couplée à notre méthodologie qui se base sur les références de génomes viraux connus, augmente le risque de sous-estimer la biodiversité virale. Néanmoins, comme évoqué précédemment, les réplicases étant très conservées et partagées par tous les virus pourraient permettre des détections virales et des reconstructions de génomes *de novo* non observées dans ces phyla particuliers.

Pour les nématodes, quelques hits de virus à ADNdb ont été détectés, ainsi que des hits ADNsb, mais ne correspondent pas à de potentiels virus complets (Tableau 5, Annexe 2). Les associations découvertes récemment chez les nématodes sont celles de virus à ARNsb(-) avec le *Soybean cyst nematode socyvirus* (*Mononegavirales*, *Nyamiviridae*) chez *Heterodera glycines* (Bekal *et al.*, 2011) et à ARNsb(+) avec *Santeuil nodavirus* et *Le Blanc nodavirus* (*Nodaviridae*) chez *Caenorhabditis briggsae* et *Orsay virus* (*Nodaviridae*) chez *C. elegans* (Félix *et al.*, 2011; Franz *et al.*, 2012), ou encore le *Soybean cyst nematode virus 5* (*Flaviviridae*) chez *H. glycines* (Bekal *et al.*, 2014).

En ce qui concerne les annélides, un virus ARNdb (*Beihai sipunculid worm virus 6*, *Hypoviridae*) a été détecté chez des *Sipuncula* (Polychaeta) ainsi que d'autres non classés (Shi *et al.*, 2016). Dans la présente étude, chez les annélides terrestres, deux virus potentiellement complets à ARNsb(+) ont été détectés chez *Allolobophora*, appartenant à l'ordre des *Picornavirales* et des *Tymovirales*, *Betaflexiviridae* (section 5.4., Tableau 6).

On se rend compte ici, que moins les bases de données sont fournies en séquences virales de

référence infectant des animaux de phyla peu étudié, moins de hits viraux proches sont détectés. Cependant, quelques hits viraux voire quelques virus sont détectables ce qui n'enlève en rien de la pertinence de la méthode.

Autres animaux et microorganismes vivants en milieu aquatique

Des *Malacoherpesviridae* (ADNdb) sont connus pour infecter des mollusques marins et notamment des mollusques comestibles tels que les huitres avec *Ostreid herpesvirus 1* chez *Crassostrea gigas* (Burioli *et al.*, 2017)*, l'Abalone herpesvirus Victoria/AUS/2009* chez *Haliotis rubra* un escargot de mer comestible (Savin *et al.*, 2010) ou encore le *Chlamys acute necrobiotic virus* chez *Chlamys farreri* un coquillage dont la production asiatique a été décimée par ce virus (Ren *et al.*, 2013). Un virus à ADNsb est connu pour infecter les moules du genre *Mytilus*, le *Mytilus sp. clam associated circular virus* (*Circoviridae*) (Rosario *et al.*, 2015). Dans cette étude, chez les mollusques différents virus détectés sont potentiellement complets et sont des virus à ARNsb(+) tels qu'un *Bacillarnavirus* (*Picornavirales*), un *Tombusviridae*, un *Flaviviridae* ou un virus à ARNsb(-) *Mononegavirales* (section 5.4., Tableau 6).

Très peu de hits viraux ont été détectés chez les échinodermes échantillonnés ne représentant aucun potentiel nouveau virus complet (Tableau 5, Annexe 2). A l'heure actuelle, un seul virus connu à ADNsb infecte un oursin, le *Lytechinus variegatus variable sea urchin associated circular virus* (*Circoviridae*) (Rosario *et al.*, 2015). Pour les cnidaires, également très peu de hits viraux ont été détectés par cette étude, dont un virus à ARNsb(-) potentiellement complet de l'ordre des *Bunyavirales* a été détecté chez un individu *Eunicella* (Tableau 6). Seuls trois Circoviridae (ADNdb) infectent des cnidaires, *Paramuricea placomus associated circular virus*, *Aiptasia sp. sea anemone associated circular virus*, et *Primnoa pacifica coral associated circular virus* (Rosario *et al.*, 2015). Enfin, pour les némertes seuls quelques hits de virus à ADNdb ont été détectés majoritairement, mais ne correspondent pas à de potentiels virus complets (Annexe 2).

Un cas qui est à part dans le milieu marin est celui des individus Haptophyta de l'échantillonnage, représenté majoritairement par l'espèce focale *Emiliania huxleyi* qui est une algue unicellulaire. Les résultats indiquent principalement des hits viraux à ADNdb parmi lesquels des *Phycodnaviridae* connus pour infecter ces algues et se transmettant par diffusion passive dans l'eau (exemple *Coccolithovirus : Emiliania huxleyi virus 203* ; *Chlorovirus* : *Paramecium bursaria Chlorella virus 1* ; non classé : *Emiliania huxleyi virus PS401*) (Nagasaki,

2008) et des *Marseilleviridae* (*Marseillevirus marseillevirus*, *Port-miou virus*) des virus d'amibes (La Scola, 2003).

De manière générale, pour les animaux marins (mollusques, cnidaire et échinodermes) et les algues unicellulaire, une partie des hits viraux détectés est représentée par des virus dont les plus proches parents connus infectent d'autres types d'organismes vivant dans le même milieu que l'hôte séquencé. En effet parmi les hits, certains sont des hits de *Mimiviridae* et *Marseilleviridae* infectant les amibes (La Scola, 2003) ou de *Bacillarnavirus* infectant les diatomées (Tomaru *et al.*, 2013) (Tableau 5, Annexe 2).

L'échantillonnage n'est pas assez important pour généraliser à l'ensemble du milieu marin mais d'une manière générale comme cela a été montré précédemment (Breitbart *et al.*, 2002; Suttle, 2007) il existe une grande quantité de virus à ADN dans ce milieu, majoritairement des bactériophages, mais également des virus à ARN (Steward *et al.*, 2012), beaucoup moins retrouvés dans cette étude.

5.3.3. La transmission des virus au sein des hôtes

La gamme d'hôte des virus

Les virus sont capables d'infecter certains hôtes particuliers, ce qui correspond à leur gamme d'hôtes. Ils peuvent ainsi être spécialistes d'un seul type d'hôte ou alors plus généralistes et infecter une large gamme d'hôtes potentiels. Cependant le type de gamme d'hôte préexistante n'est pas un facteur favorisant l'expansion de la gamme d'hôtes, les deux (spécialistes et généralistes) pouvant d'adapter à de nouveaux hôtes (Parrish *et al.*, 2008). Comme cela a été montré précédemment dans l'**Article 2** chez les *Parvoviridae*, la gamme d'hôtes des différents genres et espèces de cette famille virale, au fil de l'accumulation des connaissances peut alors se révéler plus élargie qu'attendue et inclure de nouveaux phyla animaux.

Les capsides virales sont les principales protéines qui permettent aux virus d'élargir leur gamme d'hôte car elles sont en lien direct avec les cellules hôtes cibles. En effet, les capsides virales ont de multiples fonctions dans l'infection, la pathogénicité, le mouvement ou encore la transmission des virus (Weber & Bujarski, 2015).

Les sauts d'hôtes

Le saut d'hôte concerne la transmission d'un virus d'un hôte à un autre hôte d'une espèce

différente. La transmission entre espèces de virus a été montrée comme étant plus fréquente que précédemment connue et peut alors jouer un rôle important dans l'évolution des hôtes et des virus (Geoghegan et al., 2017). Pour les virus à ARN plus précisément, l'adaptation à un nouvel hôte dans un nouvel environnement est favorisée par le fort taux de mutation produites par les ARN polymérases (Drake & Holland, 1999). Les relations phylogénétiques des espèces hôtes, comme évoquées dans l'**Article 4** entre les abeilles et fourmis hyménoptères, peuvent favoriser également le phénomène de saut d'hôte et l'adaptation des virus aux cellules via des récepteurs cellulaires plus similaires qu'entre des espèces plus éloignées (Parrish et al., 2008). Le LSV a été découvert infectant l'abeille domestique *Apis mellifera* à travers le monde (Amakpe *et al.*, 2015; Cornman *et al.*, 2012; Daughenbaugh *et al.*, 2015; Granberg *et al.*, 2013; Ravoet *et al.*, 2013, 2015; Runckel *et al.*, 2011) et infectant également diverses abeilles sauvages solitaires (Ravoet *et al.*, 2014) et des bourdons (Gamboa *et al.*, 2015; Parmentier *et al.*, 2016). Notre étude a ajouté à cette liste d'hôtes du LSV les fourmis moissonneuses *Messor concolor, M. barbarus* et *M. capitatus* (**Article 4**) suggérant que ce virus a réalisé un saut d'hôte au sein des hyménoptères. Cependant, le sens du changement d'hôte n'est pas connu, et nécessiterait des études plus approfondies.

La transmission et la distribution des virus dans les hôtes

La transmission des virus entre hôtes, de même espèce peut être verticale (transmission à la descendance), mais aussi horizontale entre espèces différentes (transmission entre organismes de la population), cette dernière pouvant être facilitée par des vecteurs. Chez les abeilles, la vectorisation virale peut se faire par l'intermédiaire d'autres animaux, comme l'acarien *Varroa destructor* pour le DWV (Wilfert et al., 2016) ou via des facteurs environnementaux tel que le pollen (Mazzei et al., 2014) ce qui favorise aussi la transmission à de nouvelles espèces pollinisatrices d'hyménoptères non-*Apis* (Singh et al., 2010). Dans le cas du LSV évoqué (**Article 4)**, le pollen et *Varroa* peuvent contenir des particules virales de LSV, sans preuve de réplication dans ces vecteurs mais la présence de LSV dans le tube digestif des abeilles domestiques suggère que la transmission de ce virus peut se faire par la nourriture (pollen) ou par trophallaxie avec ses congénères (Daughenbaugh *et al.*, 2015).

Il reste nécessaire de vérifier que le virus, détecté dans un nouvel hôte peut se répliquer dans celui-ci, afin d'apporter la démonstration qu'il s'agit d'une réelle nouvelle infection et montrant que l'hôte est définitif ou s'il n'est qu'un hôte intermédiaire (ou qu'il s'agit d'une

contamination trophique), devenant alors un possible nouveau vecteur au virus. Pour les virus à ARNsb(+), principalement détectés dans cette étude, la détection du brin réplicatif (sens négatif) reste le moyen le plus utilisé pour distinguer ces deux phénomènes (Yue & Genersch, 2005).

Enfin, l'un des facteurs essentiel à la distribution des virus au sein de différents hôtes est l'environnement dans lequel évoluent les hôtes. En effet, des études de prévalence et d'épidémiologie permettent d'avoir un meilleur aperçu de la distribution des virus dans différents hôtes partageant le même milieu (Lambin *et al.*, 2010; Ostfeld *et al.*, 2005). Il a ainsi été montré dans l'**Article 5** que différents virus d'abeilles sont capables capables de se retrouver en association avec différents hyménoptères partageant le même milieu.
C'est ainsi les cas de nouvelles espèces de bourdons (*Bombus flavifrons* et *B. humilis* infectés par DWV, *B. lapidarius* par SBV et *B. sitkensis* par BQCV), où déjà de nombreuses espèces sont connues pour être infectées par des virus d'abeilles (Fürst *et al.*, 2014; Genersch *et al.*, 2006; Graystock *et al.*, 2013, 2014; Parmentier *et al.*, 2016; Singh *et al.*, 2010). De nouvelles espèces de fourmis (*Pheidole megacephala, Formica neorufibarbis* et *Crematogaster scutellaris*) ont également été retrouvée infectées par le DWV, déjà détecté dans la fourmi argentine *Linepithema humile* (Sébastien *et al.*, 2015). Enfin, dans les frelons Européen (*Vespa crabro*) et Asiatique (*Vespa velutina*) des détections de virus d'abeilles ont été réalisées dans notre étude montrant que les virus d'abeilles peuvent être trouvés dans une large gamme d'hôte hyménoptères servant de potentiels réservoirs de virus. Des études sont encore nécessaires pour infirmer l'hypothèse de possible contamination trophique (ingestion de pollen contaminé et/ou ingestion d'abeille infectée pour les prédateurs) et qu'il existe une réplication virale dans l'ensemble de ces nouveaux hôtes montrant l'impact du partage de la même niche écologique sur la transmission des virus dans divers hôtes.

5.4. Les virus présents dans les métagénomes animaux

5.4.1. De nouvelles associations hôtes-virus

Parmi les tous les hits viraux détectés dans les transcriptomes, certains correspondent à des hits « uniques » et d'autres correspondent à des potentiels virus complets qui pourront être

reconstruits entièrement suivant le protocole décrit précédemment et utilisé dans les **Articles 1, 3 et 4**. Ainsi, si l'on regroupe l'ensemble des associations hôte-virus détectées par cette approche, de nouvelles associations ont été découvertes, de nouveau(x) hôte(s) ont été trouvés pour des virus connus et finalement de nouveaux virus ont été décrits (Tableau 6). Des nouvelles associations jamais décrites telles que la présence d'un virus de plante ou de champignon chez un animal (*Betaflexiviridae/Allolobophora* ; *Benyviridae/Basiliscus*) ou celle d'un virus de vertébré chez un invertébré (*Flaviviridae/Parus*, *Physa* ; *Nodaviridae/Melitaea*) permettent de montrer que les barrières d'espèces, à l'échelle de la famille virale, peuvent ne pas être imperméables (Tableau 6). Néanmoins, il est nécessaire d'aller plus loin en terme d'utilisation de l'outil phylogénétique pour savoir exactement de quel virus il s'agit et d'utiliser les outils de biologie moléculaire pour vérifier la présence de ces nouveaux virus et infirmer de possibles contaminations environnementales ou trophiques.

Tableau 6 : Liste des potentiels virus complets détectés dans les 523 transcriptomes animaux

Classe	Taxonomie	Nouvel hôte potentiel	Représentativité (nb indiv/nb total indiv)	Phyla	Hôtes connus (ICTV)	Association Hôte-virus	Validation phylogénétique
II. ADNsb	*Parvoviridae*	*Lamellibrachia*	1/18	Annelida	Invertébrés	*	**Article 2**
III. ARNdb	*Partitiviridae*	*Messor*	3/30	Arthropoda	Plantes, Champignons	!	x
	Reoviridae	*Thymelicus* *Myrmica* *Halictus*	3/12 1/5 2/13	Arthropoda	Invertébrés	✓	Nouveaux virus
	Totiviridae	*Formica*	5/26	Arthropoda	Champignons, Protozoaires	!	x
IV. ARNsb(+)	*Arteriviridae*	*Microtus*	3/9	Chordata	Vertébrés	✓	x
	Benyviridae	*Basilicus*	1/1	Chordata	Plantes	!	x
	Flaviviridae	*Physa* *Parus*	1/13 1/12	Mollusca	Vertébrés	!	*Mauvaise affiliation ?*
	Hepeviridae	*Cystodites*	1/8	Chordata	Vertébrés	*	*Mauvaise affiliation ?*
	Luteoviridae-Sobemovirus	*Camponotus* *Formica* *Necora* *Polyergus* *Armadilidium*	1/7 10/26 2/10 3/4 1/14	Arthropoda	Invertébrés	✓	Nouveaux virus/ hôtes
	Negevirus	*Culex*	2/22	Arthropoda	Invertébrés	*	**Article 1**
	Nidovirales (Coronaviridae)	*Hippocampus*	1/8	Chordata	Vertébrés	*	Nouveaux virus/ hôte
	Nodaviridae	*Melitaea*	1/16	Arthropoda	Vertébrés	!	Nouveaux virus/ hôtes
	Picornavirales	*Melitaea* *Messor*	1/16 23/30	Arthropoda		✓	Nouveaux virus/ hôtes

		Halictus *Formica* *Myrmica* *Pheidole* *Polyergus*	1/13 6/26 3/5 4/4 2/4		Invertébrés, Vertébrés, Microorganismes		
		Crepidula *Mytilus* *Galba*	1/12 1/14 2/15	Mollusca		!	Nouveaux virus/ hôtes
		Allolobophora	1/16	Annelida		*	Nouveaux virus
		Cystodites	1/8	Chordata		*	×
	Tombusviridae	*Physa*	1/13	Mollusca	Plantes	!	*Mauvaise affiliation /* nouveau virus ?
	Tymovirales (Betaflexiviridae)	*Allolobophora*	2/16	Annelida	Plantes	!	Nouvel hôte
	Virgaviridae	*Ciona*	2/20	Chordata	Plantes	!	*Mauvaise affiliation ?*
		Armadilidium *Formica*	1/14 5/26	Arthropoda		!	
	Unassigned	*Halictus* *Messor*	3/13 3/30	Arthropoda	Invertébrés	*	**Article 4**
V. ARNsb(-)	*Arenaviridae*	*Basiliscus*	1/1	Chordata	Vertébrés	*	Nouveau virus
	Bunyavirales	*Eunicella*	1/12	Cnidaria	Invertébrés, Vertébrés	*	×
		Messor *Formica* *Melitaea* *Necora*	1/30 4/26 1/16 1/10	Arthropoda		*	Nouveaux virus/ hôtes
		Ciona	2/20	Chordata	Champignons	!	Nouveau virus
	Mononegavirales (Filoviridae, Rhabdoviridae)	*Microtus* *Hippocampus* *Ciona*	6/9 1/8 3/20	Chordata	Vertébrés	*	Nouveau virus
		Myrmica *Formica* *Camponotus* *Armadilidium* *Thymelicus*	2/5 10/26 5/7 6/14 1/12	Arthropoda	Invertébrés, Vertébrés	✓	Nouveaux virus/ hôtes
		Sepia	1/11	Mollusca	Invertébrés	*	Nouveau virus
	Orthomyxoviridae	*Hippocampus*	1/8	Chordata	Vertébrés	*	Nouvel hôte
VI. ARNsb(RT)	*Retroviridae*	*Crocodylus* *Emys* *Lepus* *Microtus* *Tupinambis*	1/1 8/13 12/12 9/9 1/1	Chordata	Vertébrés	✓	Nouvel hôte
		Pleurodeles *Salamandra*	1/1 1/1			*	**Article 3**

! Association inattendue (nouveau genre viral) ; * Nouvel hôte potentiel (nouvelle espèce virale) ; ✓ Association hôte-virus connue ; × Pas de phylogénie réplicase

La représentativité (nombre d'individu animaux infectés par rapport au nombre total d'individus échantillonnés) des virus complets détectés dans les différents genres animaux est variable puisque cela concerne en majorité qu'un ou peu d'individus d'un même genre animal

mais dans un cas particulier, celui des rétrovirus, quasiment tous les individus sont concernés (Tableau 6). La détection dans des individus uniques est ici courante, relatant des infections virales uniques ou peu fréquentes et d'un échantillonnage restreint de la population. En revanche cela reste l'avantage de l'utilisation des transcriptomes individuels en comparaison de pool d'individus pour lesquels l'information aurait pu alors être diluée et les virus être non détectables.

5.4.2. Une grande diversité de nouveaux virus

En complément de cette synthèse sur les nouvelles associations hôtes-virus que ce travail a permis, quatorze phylogénies préliminaires, réalisées à l'échelle de l'ordre ou de la famille virale présumées, ont été construites par Maximum de vraisemblance à partir des alignements de la protéine de réplicase des virus ARN et des rétrovirus. (Figure 17). Ce travail de phylogénie permet alors de manière plus précise de définir si les affiliations virales détectées sont correctes et si le virus détecté est proche d'un virus connu, ou s'il s'agit d'un nouveau virus (Figure 17, Annexe 3).

Ces résultats préliminaires sur l'ensemble des données permettent de montrer que dans quelques cas l'affiliation virale, attribuée par l'homologie BLAST, ne semble pas correcte (Figure 17, Annexe 3). Chez le mollusque *Physa* (Annexe 3G) le nouveau virus semble proche des *Tombusviridae*, cependant les supports de nœuds ne permettent pas de confirmer l'affiliation avec certitude. Chez ce même hôte encore, le nouveau virus proche des *Flaviviridae* (Annexe 3M) qui semble alors faire partie d'une troisième famille virale phylogénétiquement proche des *Tombusviridae* et des *Flaviviridae*. Chez *Cystodites*, le virus détecté a comme famille connue la plus proche les *Hepeviridae* (Annexe 3D), mais la phylogénie indique qu'il fait partie d'une lignée toutefois fort divergente, et donc potentiellement d'une nouvelle famille virale. Des études plus approfondies, notamment en élargissant le nombre de séquences externes, et en reconstruisant les génomes les plus complets possibles par *mapping*, pourraient, à la manière de l'étude sur le CpATV (**Article 1**), corriger ces incertitudes d'affiliation virale.

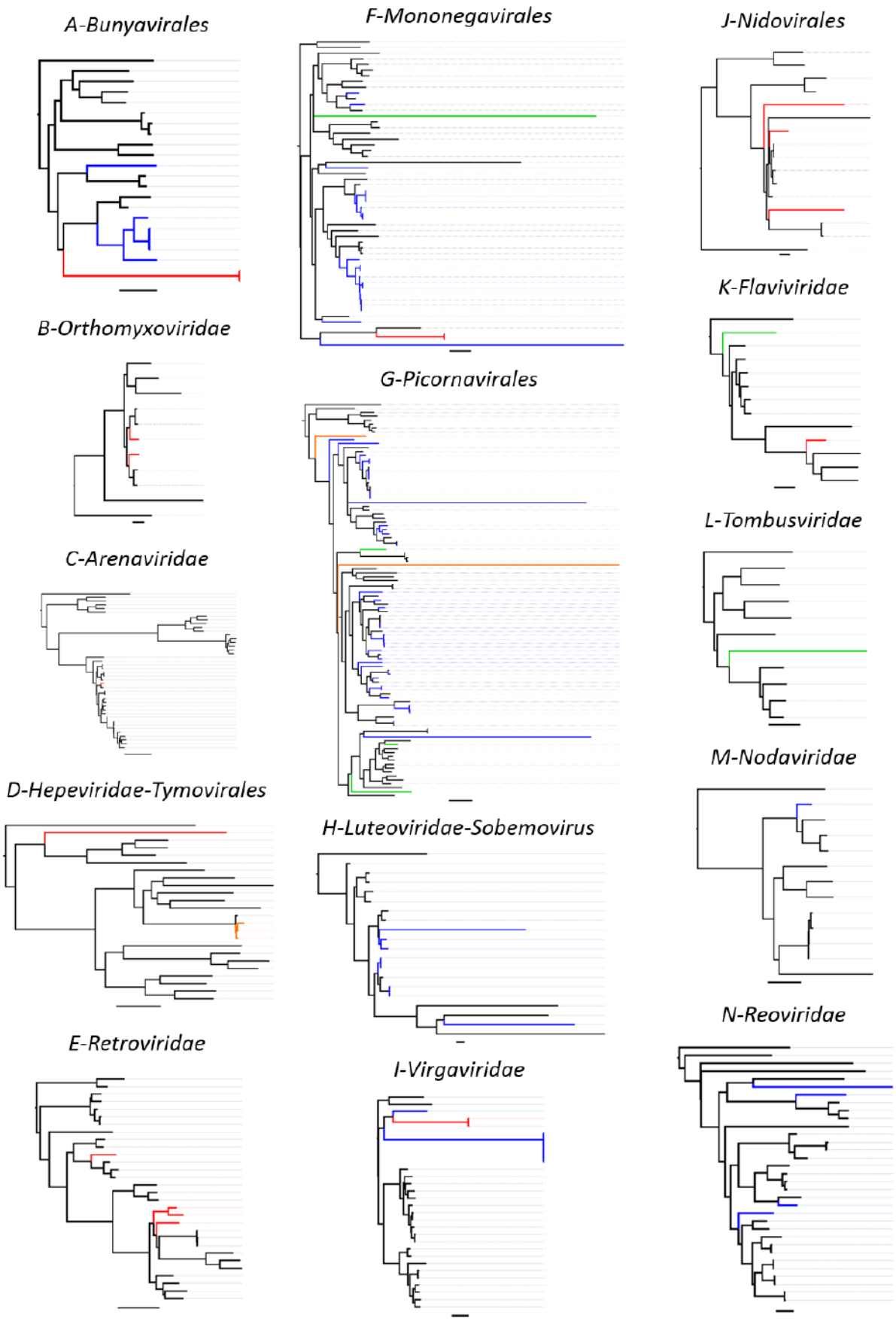


Figure 17 : Phylogénies de la RdRp des différents potentiels virus complets détectés.

En couleur sont indiquées les nouvelles séquences détectées dans cette étude (Bleu: Arthropoda, Rouge: Chordata, Vert: Mollusca, Orange: Annelida). Les barres d'échelles sont toutes calibrées à 1 substitution/sites. Les phylogénies détaillées sont en Annexe 3, le tableau des accessions Annexe 4.

Dans tous les autres cas étudiés grâce à ces phylogénies, les différents hits viraux correspondent à la détection de virus proche de virus connus mais i) potentiellement présents dans de nouveaux hôtes tels que le virus proche de *Orthomyxoviridae* chez *Hippocampus* (Annexe 3B), proche *d'Arenaviridae* chez *Basiliscus* (Annexe 3C) ou proche de *Tymovirales* chez *Allolobophora* (Annexe 3D) ou ii) de nouveaux virus très différents de ceux déjà décrits, généralement représentés par des longues branches tels que les virus de *Formica paralugubris* et *Sepia* chez les *Mononegavirales* (Annexe 3F) ou les virus de *F. fusca* et *Myrmica rubra* chez les *Picornavirales* (Annexe 3G). En perspective à ce travail, des études plus poussées seront nécessaires, au cas par cas et sur chaque virus, pour établir le statut de chacun des nouveaux génomes viraux et pour reconstruire si possible des génomes complets et comprendre leur histoire évolutive.

5.4.3. Des virus intégrés et des virus libres

Au cours de cette étude, la présence de virus libres (exogènes) et intégrés (endogènes) a pu être détectée. Les phénomènes d'intégrations étudiés ici concernent les parvovirus (**Article 2**) et les rétrovirus (**Article 3**) et pourrait concerner également d'autres hits rétrovirus détectés chez des sauriens (la tortue *Emys*, le tegu *Tupinambis* et le crocodile *Crocodylus*, Annexe 3E).

Des endogénisation de *Spumavirus* ont été récemment découvertes dans des poissons (*Latimeria chalumnae* et *Danio rerio*), dans le paresseux (*Choloepus hoffmanni*), dans une taupe dorée (*Chrysochloris asiatica*) et un mammifère insectivore (Han & Worobey, 2012, 2014). Notre étude sur la salamandre tachetée *Salamandra salamandra* dans l'**Article 3** a permis de montrer la détection d'un nouveau *Spumavirus* dans le transcriptome par une analyse de bioinformatique et dans le génome par détection PCR, ce qui tend à confirmer l'hypothèse d'une intégration ancienne de *Spumavirus* dans les amphibiens déjà observée récemment (Aiewsakun & Katzourakis, 2017). En effet, l'étude de divers animaux permet d'avoir des aperçus de la date d'intégration de ces génomes viraux au sein des génomes animaux, ayant une origine ici datée à plus de 450 millions d'année pour les rétrovirus (Aiewsakun & Katzourakis, 2017). L'utilisation de la divergence des séquences répétées aux extrémités des génomes rétroviraux (LTR) identiques au moment de l'intégration permet de dater approximativement l'intégration en se basant sur le taux d'évolution neutre du génome

hôte (Aiewsakun & Katzourakis, 2017). L'identification de séquences virales orthologues entre des hôtes divergents permet aussi d'estimer des dates d'intégrations à partir de la date estimée de spéciation de ces hôtes ; tel qu'utilisé dans l'étude de l'élément RELIK chez le lapin européen, datant l'intégration de lentivirus il y a 12 millions d'années chez les lagomorphes (Keckesova et al., 2009).

L'intégration ne concerne pas seulement les génomes complets et peut tout aussi bien être observée par les hits « uniques » détectés par cette étude, pouvant correspondre alors à des éléments viraux endogènes (EVE). Les EVE peuvent provenir soit de l'intégration du génome viral complet au cours de la réplication virale par une intégrase active, par exemple des rétrovirus (Hayward *et al.*, 2013), soit par un transfert aléatoire de matériel génomique complet ou partiel transféré au génome de l'hôte (Chabannes *et al.*, 2013; Cui & Holmes, 2012). Les EVE ne sont pas limités aux virus à ARN et aux virus à ADN se répliquant dans le noyau (Feschotte & Gilbert, 2012), suggérant qu'une variété de mécanismes génétiques sont impliqués dans les processus d'endogénisation. Décrit dans les plantes, les animaux, les champignons, les protistes et le génome bactérien (Katzourakis & Gifford, 2010), les EVE pourraient avoir des conséquences évolutives majeures.

Premièrement, les EVE participent à l'accumulation d'ADN dans le génome de l'hôte, tels les génomes vertébrés qui portent plusieurs centaines à des milliers de copies virales (Gifford & Tristem, 2003) avec par exemple, 8% du génome humain qui est constitué de rétrovirus endogènes (ERV) (Lander *et al.*, 2001; Paces, 2002).

Deuxièmement, les EVE peuvent être un réservoir d'agents pathogènes latents avec des capacités infectieuses provocatrices de maladies virales, déclenchées par des stress génomiques ou environnementaux ; tels l'exemple du *Ectocarpus siliculosus virus 1* (*Phycodnaviridae, Phaeovirus*) qui est intégré dans le génome de l'algue brune *Ectocarpus siliculosus* (Delaroque et al., 1999) ou le *Banana streak virus* (*Caulimoviridae*, *Badnavirus*) qui est intégré au génome du bananier (genre *Musa*) (Gayral et al., 2008; Iskra-Caruana et al., 2010).

Troisièmement, de nombreux EVE peuvent être bénéfiques pour les génomes, permettant d'avoir un réservoir plus riche et plus diversifié de nouveautés génétiques en fournissant à la machinerie cellulaire hôte des promoteurs de gènes hétérologues ou de nouveaux sites de liaison pour les facteurs de transcription (Kunarso *et al.*, 2010; Wang *et al.*, 2007b). L'ORF viral

endogène peut également servir à produire de nouveaux gènes cellulaires dans un processus appelé domestication, c'est-à-dire l'utilisation de la fonction virale préexistante de l'ORF (exaptation) ou l'évolution d'une nouvelle fonction du matériel génomique viral (néo-fonctionnalisation). Plus d'une douzaine de cas de domestication d'EVE ont été signalés dans plusieurs génomes hôtes jusqu'à présent (Drezen *et al.*, 2017b; Feschotte & Gilbert, 2012; Herniou *et al.*, 2013; Roossinck, 2011) et l'étendue de ce phénomène n'est en fait pas entièrement connue.

Les analyses d'évolution moléculaire et les estimations de la force de la sélection purifiante sont souvent la seule source de données (indirecte) disponible pour déduire les fonctions actuelles ou passées putatives associées à ces ORF dans le génome de l'hôte (Fort *et al.*, 2012). De plus, la caractérisation génomique des EVE (leur taille complète) et la compréhension des conséquences évolutives sur le génome de l'hôte dépendent encore largement de la disponibilité de génomes complets, encore rares pour les organismes non modèles. Les cas d'EVE bien décrits sont donc limités aux organismes modèles et la répartition des EVE à travers la taxonomie hôte restent largement inexplorée.

En contraste, divers virus libres sont également détectés dans cette étude tels que le CpATV (**Article 1**), le HsAV (**Article 4**) et d'autres sont encore à analyser. L'étude des couvertures de nouveaux génomes viraux en comparaison à la couverture moyenne du transcriptome dont il est issu, montre qu'il ne s'agit pas d'un bruit de fond de transcription ou de contamination d'ADN génomique permis par la profondeur de séquençage Illumina. Ici cela permet d'avoir une vision des virus circulant dans les différents hôtes animaux et d'observer des variations génétiques de ces virus comme analysés avec les virus d'abeilles tels que DWV, SBV, BQCV ou LSV (**Article 4 et 5**).

5.5. Conclusion générale

En résumé, cette étude décrit une méthodologie puissante pour la découverte de nouveaux virus à partir de transcriptomes, qui complétée par des analyses de génomique comparative et de phylogénie permet d'apporter de nouvelles connaissances liées à la diversité virale au sein des animaux non-modèles.

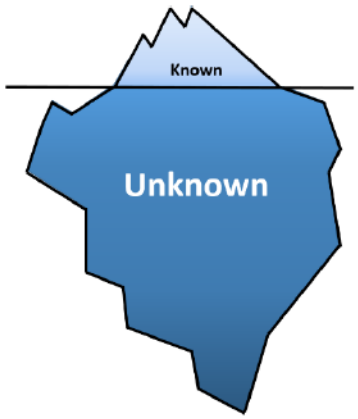
Une des principales limites de cette approche est son échantillonnage qui n'est pas réellement représentatif des espèces animales en termes de nombre d'espèces et d'individus étudiés, et qui se cantonne aux animaux excluant l'immense majorité de la diversité du vivant (Simpson & Roger, 2004). Ainsi, des généralisations ne sont pas faisables à toutes les espèces d'un phylum, les virus étant en plus souvent détectés dans des individus isolés. De plus, parmi tous ces potentiels virus complets, quasi tous appartiennent à des ordre/familles virales à génome ARN (classes III à VI). C'est l'un des biais évident de l'analyse des transcriptomes qui, par le séquençage des ARN, enrichi par conséquence la détection des virus à ARN et diminue les capacités de détection de séquences de virus à ADN. Cependant, les virus à ADN dont le génome s'exprime peuvent aussi être séquencés par cette approche de transcriptomique, de même que les virus ADN à fort titre dans les cellules hôtes (infection en cours) car la méthode utilisée est un enrichissement en ARNm et non l'utilisation d'un traitement DNAse systématique laissant place à la présence d'ADN. Bien que de nombreux hits viraux à ADN aient alors été détectés par cette approche, aucune vérification (autre que le BLAST réciproque) ne prouve qu'il s'agisse de virus ou au contraire de gènes cellulaires ; tel l'exemple du gène inibiteur d'apoptose (IAP) présent chez les grands virus à ADN mais aussi ayant des homologues cellulaires (Clem, 2015).

Cette étude a ainsi pris le parti d'utiliser ces données transcriptomiques pour l'originalité que cela apporte par rapport à toutes les méthodes de détections virales existantes. De plus, les transcriptomes sont des ressources plus facilement accessibles que les génomes. En effet, les transcriptomes sont moins chers d'acquisition, plus rapide en termes d'assemblage et d'analyse des données que les génomes. L'utilisation des transcriptomes permet le séquençage d'animaux non-modèles sauvages et donc permet de découvrir de nouveaux virus ; le fait de ne pas avoir le génome de l'hôte n'est pas rédhibitoire.

Enfin, comme les bases de données ne sont pas très riches en séquences virales de nombreuses espèces animales variées, les virus détectés sont proches de virus connus et nous n'avons pas suffisamment de connaissances pour découvrir des virus plus éloignés. Cela a pu être le cas pour les annélides ou encore les némertes, pour lesquels très peu voire pas de virus sont répertoriés, indiquant peut-être un manque de détection d'une réelle infection virale alors non observée.

Malgré toutes ces limites, l'utilisation de ce type d'étude à grande échelle est tout à fait pertinente. Notamment par l'utilisation d'animaux non-modèles de manière individuelle ce qui permet d'avoir accès à des informations non « noyées » dans des pools d'individus de même espèces ou d'espèces différentes. Cela permet alors d'avoir accès à une connaissance plus fine des associations hôtes/virus, bien que cela ait néanmoins un coût puisque il est plus difficile de séquencer un transcriptome pour les hôtes de petites tailles tels que les insectes par exemple. Néanmoins, bien que cela représente une faible quantité des protéines séquencées (environ 0,03 %), le taux de découverte - c'est-à-dire la quantité de nouveaux virus par transcriptome – est non négligeable, ici mesurée à 40 % (200 potentiels virus complets dans 523 transcriptomes).

L'étude de la diversité virale est un champ disciplinaire qui n'as pas fini d'attirer les scientifiques et promet des découvertes passionnantes dans les années à venir par la mise en place de nouvelles stratégies (telle que celle présentée ici), de plus en plus performantes et accessibles à une communauté de plus en plus vaste (Rose *et al.*, 2016). Au vue de la masse importante de données non explorées dans les bases de données et au vue de l'immensité du monde vivant (dont on ne connait pas encore les limites), la diversité virale actuellement observée n'est en réalité que « *la partie émergée de l'iceberg* », et représente qu'une partie infime de ce qui peut réellement exister.



BIBLIOGRAPHIE

Abascal, F., Zardoya, R. & Posada, D. (2005). ProtTest: Selection of best-fit models of protein evolution. *Bioinformatics* **21**, 2104–2105.

Ackermann, H. W. (2003). Bacteriophage observations and evolution. *Res Microbiol* **154**, 245–251.

Adams, I. P., Glover, R. H., Monger, W. A., Mumford, R., Jackeviciene, E., Navalinskiene, M., Samuitiene, M. & Boonham, N. (2009). Next-generation sequencing and metagenomic analysis: a universal diagnostic tool in plant virology. *Mol Plant Pathol* **10**, 537–545.

Ahmed, F., Kumar, M. & Raghava, G. P. S. (2009). Prediction of polyadenylation signals in human DNA sequences using nucleotide frequencies. *In Silico Biol*.

Aiewsakun, P. & Katzourakis, A. (2017). Marine origin of retroviruses in the early Palaeozoic Era. *Nat Commun* **8**, 13954.

Alavandi, S. V. & Poornima, M. (2012). Viral metagenomics: A tool for virus discovery and diversity in aquaculture. *Indian J Virol* **23**, 88–98.

Amakpe, F., De Smet, L., Brunain, M., Ravoet, J., Jacobs, F. J., Reybroeck, W., Sinsin, B. & de Graaf, D. C. (2015). Discovery of *Lake Sinai virus* and an unusual strain of *Acute Bee Paralysis virus* in West African apiaries. *Apidologie* 35–47.

Anantharaman, K., Duhaime, M. B., Breier, J. a, Wendt, K. a, Toner, B. M. & Dick, G. J. (2014). Sulfur oxidation genes in diverse deep-sea viruses. *Science* **344**, 757–60.

Angly, F., Rodriguez-Brito, B., Bangor, D., McNairnie, P., Breitbart, M., Salamon, P., Felts, B., Nulton, J., Mahaffy, J. & Rohwer, F. (2005). PHACCS, an online tool for estimating the structure and diversity of uncultured viral communities using metagenomic information. *BMC Bioinformatics* **6**, 41.

Anisimova, M. & Gascuel, O. (2006). Approximate likelihood-ratio test for branches: a fast, accurate, and powerful alternative. *Syst Biol* **55**, 539–52.

Asgari, S. & Johnson, K. (2010). *Insect virology* (S. Asgari & K. Johnson, Eds.). Norfolk, United Kingdom: Caister Academic Press.

Baker, A. C. & Schroeder, D. C. (2008). The use of RNA-dependent RNA polymerase for the taxonomic assignment of Picorna-like viruses (order *Picornavirales*) infecting *Apis mellifera* L. populations. *Virol J* **5**, 10.

Baltimore, D. (1971). Expression of animal virus genomes. *Bacteriol Rev* **35**, 235–241.

Bamford, D. H. (2003). Do viruses form lineages across different domains of life? *Res Microbiol* **154**, 231–236.

Barba, M., Czosnek, H. & Hadidi, A. (2014). Historical perspective, development and applications of next-generation sequencing in plant virology. *Viruses* **6**, 106–36.

Barré-Sinoussi, F., Chermann, J. C., Rey, F., Nugeyre, M. T., Chamaret, S., Gruest, J., Dauguet, C., Axler-Blin, C., Vézinet-Brun, F. & other authors. (1983). Isolation of a T-lymphotropic retrovirus from a patient at risk for acquired immune deficiency syndrome (AIDS). *Science* **220**, 868–871.

Bekal, S., Domier, L. L., Niblack, T. L. & Lambert, K. N. (2011). Discovery and initial analysis of novel viral genomes in the soybean cyst nematode. *J Gen Virol* **92**, 1870–1879.

Bekal, S., Domier, L. L., Gonfa, B., McCoppin, N. K., Lambert, K. N. & Bhalerao, K. (2014). A novel flavivirus in the soybean cyst nematode. *J Gen Virol* **95**, 1272–1280.

Bergh, Ø., BØrsheim, K. Y., Bratbak, G. & Heldal, M. (1989). High abundance of viruses found in aquatic environments. *Nature* **340**, 467–468.

Bexfield, N. & Kellam, P. (2011). Metagenomics and the molecular identification of novel viruses. *Vet J* **190**, 191–198.

Bichaud, L., de Lamballerie, X., Alkan, C., Izri, A., Gould, E. A. & Charrel, R. N. (2014). Arthropods as a source of new RNA viruses. *Microb Pathog* **77**, 136–41.

Birol, I., Jackman, S. D., Nielsen, C. B., Qian, J. Q., Varhol, R., Stazyk, G., Morin, R. D., Zhao, Y., Hirst, M. & other authors. (2009). De novo transcriptome assembly with ABySS. *Bioinformatics* **25**, 2872–2877.

Boujelben, I., Yarza, P., Almansa, C., Villamor, J., Maalej, S., Anton, J., Santos, F. & Antón, J. (2012). Virioplankton community structure in Tunisian solar salterns. *Appl Environ Microbiol* **78**, 7429–7437.

Breitbart, M., Felts, B., Kelley, S., Mahaffy, J. M., Nulton, J., Salamon, P. & Rohwer, F. (2004). Diversity and population structure of a near-shore marine-sediment viral community. *Proc R Soc B Biol Sci* **271**, 565–574.

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

Breitbart, M., Salamon, P., Andresen, B., Mahaffy, J. M., Segall, A. M., Mead, D., Azam, F. & Rohwer, F. (2002). Genomic analysis of uncultured marine viral communities. *Proc Natl Acad Sci U S A* **99**, 14250–5.

Breitbart, M., Hewson, I., Felts, B., Mahaffy, J. M., Nulton, J., Salamon, P. & Rohwer, F. (2003). Metagenomic analyses of an uncultured viral community from human feces. *J Bacteriol* **185**, 6220–6223.

Brum, J. R., Ignacio-Espinoza, J. C., Roux, S., Doulcier, G., Acinas, S. G., Alberti, A., Chaffron, S., Cruaud, C., de Vargas, C. & other authors. (2015). Patterns and ecological drivers of ocean viral communities. *Science* **348**, 1261498–1261498.

Burioli, E. A. V., Prearo, M. & Houssin, M. (2017). Complete genome sequence of *Ostreid herpesvirus* type 1 µVar isolated during mortality events in the Pacific oyster *Crassostrea gigas* in France and Ireland. *Virology* **509**, 239–251.

Cahais, V., Gayral, P., Tsagkogeorga, G., Melo-Ferreira, J., Ballenghien, M., Weinert, L., Chiari, Y., Belkhir, K., Ranwez, V. & Galtier, N. (2012). Reference-free transcriptome assembly in non-model animals from next-generation sequencing data. *Mol Ecol Resour* **12**, 834–45.

Calisher, C. H., Childs, J. E., Field, H. E., Holmes, K. V. & Schountz, T. (2006). Bats: Important reservoir hosts of emerging viruses. *Clin Microbiol Rev* **19**, 531–545.

Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K. & Madden, T. L. (2009). BLAST+: architecture and applications. *BMC Bioinformatics* **10**, 421.

Cann, A. J., Fandrich, S. E. & Heaphy, S. (2005). Analysis of the virus population present in equine faeces indicates the presence of hundreds of uncharacterized virus genomes. *Virus Genes* **30**, 151–156.

Castresana, J. (2000). Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. *Mol Biol Evol* **17**, 540–52.

Cepero, A., Ravoet, J., Gómez-Moracho, T., Bernal, J., Del Nozal, M. J., Bartolomé, C., Maside, X., Meana, A., González-Porto, A. V & other authors. (2014). Holistic screening of collapsing honey bee colonies in Spain: a case study. *BMC Res Notes* **7**, 649.

Chabannes, M., Baurens, F.-C., Duroy, P.-O., Bocs, S., Vernerey, M.-S., Rodier-Goud, M., Barbe, V., Gayral, P. & Iskra-Caruana, M.-L. (2013). Three infectious viral species lying in wait in the banana genome. *J Virol* **87**, 8624–8637.

Chang, T.-H., Huang, H.-Y., Hsu, J. B.-K., Weng, S.-L., Horng, J.-T. & Huang, H.-D. (2013). An enhanced computational platform for investigating the roles of regulatory RNA and for identifying functional RNA motifs. *BMC Bioinformatics* **14 Suppl 2**, S4.

Chapman, A. D. (2009). *Numbers of living species in Australia and the world*. Australian Biodiversity Information Services, Toowoomba, Australia.

Claverie, J.-M. (2006). Viruses take center stage in cellular evolution. *Genome Biol* **7**, 110.

Clem, R. J. (2015). Viral IAPs, then and now. *Semin Cell Dev Biol* **39**, 72–79.

Conway, M. J., Colpitts, T. M. & Fikrig, E. (2014). Role of the vector in arbovirus transmission. *Annu Rev Virol* **1**, 71–88.

Cornman, R. S., Tarpy, D. R., Chen, Y., Jeffreys, L., Lopez, D., Pettis, J. S., VanEngelsdorp, D. & Evans, J. D. (2012). Pathogen webs in collapsing honey bee colonies. *PLoS One* **7**, e43562.

Crick, F. H. & Watson, J. D. (1956). Structure of small viruses. *Nature* **177**, 473–475.

Cui, J. & Holmes, E. C. (2012). Endogenous RNA viruses of plants in insect genomes. *Virology* **427**, 77–79.

Darriba, D., Taboada, G. L., Doallo, R. & Posada, D. (2012). jModelTest 2: more models, new heuristics and parallel computing. *Nat Methods* **9**, 772.

Daughenbaugh, K. F., Martin, M., Brutscher, L. M., Cavigli, I., Garcia, E., Lavin, M. & Flenniken, M. L. (2015). Honey bee infecting Lake Sinai Viruses. *Viruses* **7**, 3285–309.

Dedeine, F., Weinert, L. A., Bigot, D., Josse, T., Ballenghien, M., Cahais, V., Galtier, N. & Gayral, P. (2015). Comparative analysis of transcriptomes from secondary reproductives of three *Reticulitermes* termite species. *PLoS One* **10**, e0145596.

Delaroque, N., Maier, L., Knippers, R. & Müller, D. G. (1999). Persistent virus integration into the genome of its algal host, *Ectocarpus siliculosus* (Phaeophyceae). *J Gen Virol* **80**, 1367–1370.

Delwart, E. L. (2007). Viral metagenomics. *Rev Med Virol* **17**, 115–131.

Dodd, M. S., Papineau, D., Grenne, T., Slack, J. F., Rittner, M., Pirajno, F., O'Neil, J. & Little, C. T. S. (2017). Evidence for early life in Earth's oldest hydrothermal vent precipitates. *Nature* **543**, 60–64.

Drake, J. W. & Holland, J. J. (1999). Mutation rates among RNA viruses. *Proc Natl Acad Sci* **96**, 13910–13913.

Drezen, J.-M., Gauthier, J., Josse, T., Bézier, A., Herniou, E. & Huguet, E. (2017a). Foreign DNA acquisition by invertebrate genomes. *J Invertebr Pathol* **147**, 157–168.

Drezen, J.-M., Leobold, M., Bézier, A., Huguet, E., Volkoff, A.-N. & Herniou, E. A. (2017b). Endogenous viruses of parasitic wasps: variations on a common theme. *Curr Opin Virol* **25**, 41–48.

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.

Durzyńska, J. & Goździcka-Józefiak, A. (2015). Viruses and cells intertwined since the dawn of evolution. *Virol J* **12**, 169.

Edwards, R. A. & Rohwer, F. (2005). Opinion: Viral metagenomics. *Nat Rev Microbiol* **3**, 504–510.

Emanuelsson, O., Nielsen, H., Brunak, S. & von Heijne, G. (2000). Predicting subcellular localization of proteins based on their N-terminal amino acid sequence. *J Mol Biol* **300**, 1005–1016.

Félix, M. A., Ashe, A., Piffaretti, J., Wu, G., Nuez, I., Bélicard, T., Jiang, Y., Zhao, G., Franz, C. J. & other authors. (2011). Natural and experimental infection of *Caenorhabditis* nematodes by novel viruses related to nodaviruses. *PLoS Biol* **9**.

Feschotte, C. & Gilbert, C. (2012). Endogenous viruses: insights into viral evolution and impact on host biology. *Nat Rev Genet* **13**, 283–296.

Flint, J., Racaniello, V. R., Rall, G. F., Skalka, A. M. & Enquist, L. W. (2015a). *Principles of Virology, 4th Edition, Volume I: Molecular Biology*. American Society of Microbiology.

Flint, J., Racaniello, V. R., Rall, G. F., Skalka, A. M. & Enquist, L. W. (2015b). *Principles of Virology, 4th Edition, Volume II: Pathogenesis and Control*. American Society of Microbiology.

Flygare, S., Simmon, K., Miller, C., Qiao, Y., Kennedy, B., Di Sera, T., Graf, E. H., Tardif, K. D., Kapusta, A. & other authors. (2016). Taxonomer: an interactive metagenomics analysis portal for universal pathogen detection and host mRNA expression profiling. *Genome Biol* **17**, 111.

Fort, P., Albertini, A., Van-Hua, A., Berthomieu, A., Roche, S., Delsuc, F., Pasteur, N., Capy, P., Gaudin, Y. & Weill, M. (2012). Fossil rhabdoviral sequences integrated into arthropod genomes: ontogeny, evolution, and potential functionality. *Mol Biol Evol* **29**, 381–90.

Forterre, P. (2006). The origin of viruses and their possible roles in major evolutionary transitions. *Virus Res* **117**, 5–16.

Forterre, P. & Prangishvili, D. (2009). The origin of viruses. *Res Microbiol* **160**, 466–72.

Fosso, B., Santamaria, M., D'Antonio, M., Lovero, D., Corrado, G., Vizza, E., Passaro, N., Garbuglia, A. R., Capobianchi, M. R. & other authors. (2017). MetaShot: An accurate workflow for taxon classification of host-associated microbiome from shotgun metagenomic data. *Bioinformatics* **33**, 1730–1732.

Fraenkel-Conrat, H., Singer, B. & Williams, R. C. (1957). Infectivity of viral nucleic acid. *Biochim Biophys Acta* **25**, 87–96.

François, S., Filloux, D., Roumagnac, P., Bigot, D., Gayral, P., Martin, D. P., Froissart, R. & Ogliastro, M. (2016). Discovery of parvovirus-related sequences in an unexpected broad range of animals. *Sci Rep* **6**, 30880.

Franz, C. J., Zhao, G., Felix, M.-A. & Wang, D. (2012). Complete genome sequence of Le Blanc virus, a third *Caenorhabditis* nematode-infecting virus. *J Virol* **86**, 11940–11940.

Fredslund, J. (2006). PHY·FI: fast and easy online creation and manipulation of phylogeny color figures. *BMC Bioinformatics* **7**, 315.

Fürst, M. A., McMahon, D. P., Osborne, J. L., Paxton, R. J. & Brown, M. J. F. (2014). Disease associations between honeybees and bumblebees as a threat to wild pollinators. *Nature* **506**, 364–6.

Galassi, F. M., Habicht, M. E. & Rühli, F. J. (2017). Poliomyelitis in Ancient Egypt ? *Neurol Sci* **38**, 375.

Gallot-Lavallée, L. & Blanc, G. (2017). A glimpse of nucleo-cytoplasmic large DNA virus biodiversity through the eukaryotic genomicswindow. *Viruses* **9**.

Gamboa, V., Ravoet, J., Brunain, M., Smagghe, G., Meeus, I., Figueroa, J., Riaño, D. & de Graaf, D. C. (2015). Bee pathogens found in *Bombus atratus* from Colombia: A case study. *J Invertebr Pathol* **129**, 36–39.

Gayral, P., Noa-Carrazana, J.-C., Lescot, M., Lheureux, F., Lockhart, B. E. L., Matsumoto, T., Piffanelli, P. & Iskra-Caruana, M.-L. (2008). A single *Banana streak virus* integration event in the banana genome as the origin of infectious endogenous pararetrovirus. *J Virol* **82**, 6697–710.

Gayral, P., Weinert, L., Chiari, Y., Tsagkogeorga, G., Ballenghien, M. & Galtier, N. (2011). Next-generation sequencing of transcriptomes: a guide to RNA isolation in nonmodel animals. *Mol Ecol Resour* **11**, 650–661.

Gayral, P., Melo-Ferreira, J., Glémin, S., Bierne, N., Carneiro, M., Nabholz, B., Lourenco, J. M., Alves, P. C., Ballenghien, M. & other authors. (2013). Reference-free population genomics from Next-Generation transcriptome data and the vertebrate–invertebrate gap. *PLoS Genet* **9**, e1003457.

Genersch, E., Yue, C., Fries, I. & de Miranda, J. R. (2006). Detection of *Deformed wing virus*, a honey bee viral pathogen, in bumble bees (*Bombus terrestris* and *Bombus pascuorum*) with wing deformities. *J Invertebr Pathol* **91**, 61–63.

Geoghegan, J. L., Duchêne, S. & Holmes, E. C. (2017). Comparative analysis estimates the relative frequencies of co-divergence and cross-species transmission within viral families. *PLOS Pathog* **13**, e1006215.

Ghosh, T. S., Mohammed, M. H., Komanduri, D. & Mande, S. S. (2011). ProViDE: A software tool for accurate estimation of viral diversity in metagenomic samples. *Bioinformation* **6**, 91–94.

Gifford, R. & Tristem, M. (2003). The evolution, distribution and diversity of endogenous retroviruses. *Virus Genes* **26**, 291–315.

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**, 3348.

Gouy, M., Guindon, S. & Gascuel, O. (2010). SeaView Version 4: A multiplatform graphical user interface for sequence alignment and phylogenetic tree building. *Mol Biol Evol* **27**, 221–224.

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.

Graystock, P., Yates, K., Evison, S. E. F., Darvill, B., Goulson, D. & Hughes, W. O. H. (2013). The Trojan hives: Pollinator pathogens, imported and distributed in bumblebee colonies. *J Appl Ecol* **50**, 1207–1215.

Graystock, P., Goulson, D. & Hughes, W. O. H. (2014). The relationship between managed bees and the prevalence of parasites in bumblebees. *PeerJ* **2**, e522.

Guindon, S. & Gascuel, O. (2003). A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Syst Biol* **52**, 696–704.

Han, G. Z. & Worobey, M. (2012). An endogenous foamy-like viral element in the coelacanth genome. *PLoS Pathog* **8**, 1–7.

Han, G. Z. & Worobey, M. (2014). Endogenous viral sequences from the Cape golden mole (*Chrysochloris asiatica*) reveal the presence of foamy viruses in all major placental mammal clades. *PLoS One* **9**, 3–6.

Hayward, A., Grabherr, M. & Jern, P. (2013). Broad-scale phylogenomics provides insights into retrovirus-host evolution. *Proc Natl Acad Sci U S A* **110**, 20146–51.

Hebert, P. D. N., Cywinska, A., Ball, S. L. & DeWaard, J. R. (2003a). Biological identifications through DNA barcodes. *Proc R Soc Biol Sci* **270**, 313–321.

Hebert, P. D. N., Ratnasingham, S. & deWaard, J. R. (2003b). Barcoding animal life: cytochrome c oxidase subunit 1 divergences among closely related species. *Proc Biol Sci* **270 Suppl**, S96-9.

Hendrix, R. W., Hatfull, G. F. & Smith, M. C. M. (2003). Bacteriophages with tails: Chasing their origins and evolution. *Res Microbiol* **154**, 253–257.

Herniou, E. A., Huguet, E., Thézé, J., Bézier, A., Périquet, G. & Drezen, J.-M. (2013). When parasitic wasps hijacked viruses : genomic and functional evolution of polydnaviruses. *Philos Trans R Soc B Biol Sci* **368**, 20130051.

Hershey, A. D. & Chase, M. (1952). Independent functions of viral protein and nucleic acid in growth of bacteriophage. *J Gen Physiol* **36**, 39–56.

Ho, T. & Tzanetakis, I. E. (2014). Development of a virus detection and discovery pipeline using next generation sequencing. *Virology* **471–473C**, 54–60.

Holmes, E. C. & Moya, A. (2002). Is the quasispecies concept relevant to RNA viruses? *J Virol* **76**, 460–462.

Holmes, E. C. (2010). The RNA virus quasispecies: Fact or fiction? *J Mol Biol* **400**, 271–273.

Hong, J.-J., Wu, T.-Y., Chang, T.-Y. & Chen, C.-Y. (2013). Viral IRES prediction system - a web server for prediction of the IRES secondary structure in silico. *PLoS One* **8**, e79288.

Huang, X. & Madan, A. (1999). CAP3: a DNA sequence assembly program. *Genome Res* **9**, 868–877.

Hueffer, K., Parker, J. S. L., Weichert, W. S., Geisel, R. E., Sgro, J.-Y. & Parrish, C. R. (2003). The natural host range shift and subsequent evolution of *Canine parvovirus* resulted from virus-specific binding to the canine transferrin receptor. *J Virol* **77**, 1718–1726.

Huelsenbeck, J. P. & Ronquist, F. (2001). MRBAYES: Bayesian inference of phylogenetic trees. *Bioinformatics* **17**, 754–755.

Hulo, C., de Castro, E., Masson, P., Bougueleret, L., Bairoch, A., Xenarios, I. & Le Mercier, P. (2011). ViralZone: a knowledge resource to understand virus diversity. *Nucleic Acids Res* **39**, D576-582.

Hyatt, D., Chen, G.-L., Locascio, P. F., Land, M. L., Larimer, F. W. & Hauser, L. J. (2010). Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics* **11**, 119.

Hyatt, D., Locascio, P. F., Hauser, L. J. & Uberbacher, E. C. (2012). Gene and translation initiation site prediction in metagenomic sequences. *Bioinformatics* **28**, 2223–2230.

Iskra-Caruana, M.-L., Baurens, F.-C., Gayral, P. & Chabannes, M. (2010). A four-partner plant–virus interaction: enemies can also come from within. *Mol Plant Microbe Interact* **23**, 1394–402.

Jiang, S., Steward, G., Jellison, R., Chu, W. & Choi, S. (2004). Abundance, distribution, and diversity of viruses in alkaline, hypersaline Mono Lake, California. *Microb Ecol* **47**, 9–17.

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.

Jones, P., Binns, D., Chang, H.-Y., Fraser, M., Li, W., McAnulla, C., McWilliam, H., Maslen, J., Mitchell, A. & other authors. (2014). InterProScan 5: genome-scale protein function classification. *Bioinformatics* **30**, 1236–1240.

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

Katoh, K., Misawa, K., Kuma, K. & Miyata, T. (2002). MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Res* **30**, 3059–3066.

Katzourakis, A. (2013). Paleovirology: inferring viral evolution from host genome sequence data. *Philos Trans R Soc B Biol Sci* **368**, 20120493–20120493.

Katzourakis, A. & Gifford, R. J. (2010). Endogenous viral elements in animal genomes. *PLoS Genet* **6**, e1001191.

Katzourakis, A., Tristem, M., Pybus, O. G. & Gifford, R. J. (2007). Discovery and analysis of the first endogenous lentivirus. *Proc Natl Acad Sci* **104**, 6261–6265.

Kearse, M., Moir, R., Wilson, A., Stones-Havas, S., Cheung, M., Sturrock, S., Buxton, S., Cooper, A., Markowitz, S. & other authors. (2012). Geneious Basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data. *Bioinformatics* **28**, 1647–1649.

Keckesova, Z., Ylinen, L. M. J., Towers, G. J., Gifford, R. J. & Katzourakis, A. (2009). Identification of a RELIK orthologue in the European hare (*Lepus europaeus*) reveals a minimum age of 12 million years for the lagomorph lentiviruses. *Virology* **384**, 7–11.

Kepner, R. L., Wharton, R. A. & Suttle, C. A. (1998). Viruses in Antarctic lakes. *Limnol Oceanogr* **43**, 1754–1761.

King, A. M. Q., Adams, M. J., Carstens, E. B. & Lefkowitz, E. J. (2012). *Virus taxonomy: classification and nomenclature of viruses: Ninth report of the International Committee on Taxonomy of Viruses. Virus Taxon*, Elsevier A. (A. M. Q. King, M. J. Adams, E. B. Carstens & E. J. Lefkowitz, Eds.). Academic Press.

Kishino, H. & Hasegawa, M. (1989). Evaluation of the maximum likelihood estimate of the evolutionary tree topologies from DNA sequence data, and the branching order in hominoidea. *J Mol Evol* **29**, 170–179.

Kofler, R., Orozco-terWengel, P., De Maio, N., Pandey, R. V., Nolte, V., Futschik, A., Kosiol, C. & Schlötterer, C. (2011). PoPoolation: a toolbox for population genetic analysis of next generation sequencing data from pooled individuals. *PLoS One* **6**, e15925.

Koonin, E. V & Dolja, V. V. (1993). Evolution and taxonomy of positive-strand RNA viruses: implications of comparative analysis of amino acid sequences. *Crit Rev Biochem Mol Biol* **28**, 375–430.

Koonin, E. V & Dolja, V. V. (2013). A virocentric perspective on the evolution of life. *Curr Opin Virol* **3**, 546–557.

Koonin, E. V, Senkevich, T. G. & Dolja, V. V. (2006). The ancient Virus World and evolution of cells. *Biol Direct* **1**, 29.

Koonin, E. V. (1991). The phylogeny of RNA-dependent RNA polymerases of positive-strand RNA viruses. *J Gen Virol* **72**, 2197–2206.

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

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.

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**, 11–9.

Kryazhimskiy, S. & Plotkin, J. B. (2008). The population genetics of dN/dS. *PLoS Genet* **4**, e1000304.

Kuhn, J. H. & Jahrling, P. B. (2010). Clarification and guidance on the proper usage of virus and virus species names. *Arch Virol* **155**, 445–453.

Kunarso, G., Chia, N.-Y., Jeyakani, J., Hwang, C., Lu, X., Chan, Y.-S., Ng, H.-H. & Bourque, G. (2010). Transposable elements have rewired the core regulatory network of human embryonic stem cells. *Nat Genet* **42**, 631–4.

Kuno, G. & Chang, G.-J. J. (2005). Biological transmission of arboviruses: reexamination of and new insights into components, mechanisms, and unique traits as well as their evolutionary trends. *Clin Microbiol Rev* **18**, 608–37.

Ladner, J. T., Beitzel, B., Chain, P. S. G., Davenport, M. G., Donaldson, E., Frieman, M., Kugelman, J., Kuhn, J. H., O'Rear, J. & other authors. (2014). Standards for sequencing viral genomes in the era of high-throughput sequencing. *MBio* **5**, e01360-14-e01360-14.

Laffy, P. W., Wood-Charlson, E. M., Turaev, D., Weynberg, K. D., Botté, E. S., Van Oppen, M. J. H., Webster, N. S. & Rattei, T. (2016). HoloVir: A workflow for investigating the diversity and function of viruses in invertebrate holobionts. *Front Microbiol* **7**, 1–15.

Lambin, E. F., Tran, A., Vanwambeke, S. O., Linard, C. & Soti, V. (2010). Pathogenic landscapes: Interactions between land, people, disease vectors, and their animal hosts. *Int J Health Geogr* **9**, 54.

Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J., Devon, K., Dewar, K., Doyle, M. & other authors. (2001). Initial sequencing and analysis of the human genome. *Nature* **409**, 860–921.

Lane, D. J., Pace, B., Olsen, G. J., Stahl, D. A., Sogin, M. L. & Pace, N. R. (1985). Rapid determination of 16S ribosomal RNA sequences for phylogenetic analyses. *Proc Natl Acad Sci U S A* **82**, 6955–9.

Lauring, A. S. & Andino, R. (2010). Quasispecies theory and the behavior of RNA viruses. *PLoS Pathog* **6**, 1–8.

Lesnaw, J. A. & Ghabrial, S. A. (2000). Tulip breaking : Past, present, and future. *Am Phytopathol Soc* **84**.

Letunic, I., Doerks, T. & Bork, P. (2012). SMART 7: recent updates to the protein domain annotation resource. *Nucleic Acids Res* **40**, D302–D305.

Letunic, I. & Bork, P. (2007). Interactive Tree Of Life (iTOL): an online tool for phylogenetic tree display and annotation. *Bioinformatics* **23**, 127–128.

Letunic, I. & Bork, P. (2011). Interactive Tree Of Life v2: online annotation and display of phylogenetic trees made easy. *Nucleic Acids Res* **39**, W475–W478.

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, 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, Y., Wang, H., Nie, K., Zhang, C., Zhang, Y., Wang, J., Niu, P. & Ma, X. (2016). VIP: an integrated pipeline for metagenomics of virus identification and discovery. *Sci Rep* **6**, 23774.

Lin, H.-H. & Liao, Y.-C. (2017). drVM: a new tool for efficient genome assembly of known eukaryotic viruses from metagenomes. *Gigascience* **6**, 1–10.

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**, 378.

Lipkin, W. I. & Firth, C. (2013). Viral surveillance and discovery. *Curr Opin Virol* **3**, 199–204.

Liu, H., Fu, Y., Jiang, D., Li, G., Xie, J., Cheng, J., Peng, Y., Ghabrial, S. a & Yi, X. (2010). Widespread horizontal gene transfer from double-stranded RNA viruses to eukaryotic nuclear genomes. *J Virol* **84**, 11876–11887.

Liu, S., Vijayendran, D. & Bonning, B. C. (2011). Next generation sequencing technologies for insect virus discovery. *Viruses* **3**, 1849–1869.

Lorenzi, H. A., Hoover, J., Inman, J., Safford, T., Murphy, S., Kagan, L. & Williamson, S. J. (2011). The Viral MetaGenome Annotation Pipeline (VMGAP):an automated tool for the functional annotation of viral Metagenomic shotgun sequencing data. *Stand Genomic Sci* **4**, 418–429.

Lwoff, A. (1957). The concept of virus. *J Gen Microbiol* **17**, 239–253.

Lwoff, A., Horne, R. & Tournier, P. (1962). A System of Viruses. *Cold Spring Harb Symp Quant Biol* **27**, 51–55.

Mackenzie, J. S. & Jeggo, M. (2013). Reservoirs and vectors of emerging viruses. *Curr Opin Virol* **3**, 170–9.

Malfroy, S. F., Roberts, J. M. K., Perrone, S., Maynard, G. & Chapman, N. (2016). A pest and disease survey of the isolated Norfolk Island honey bee (*Apis mellifera*) population. *J Apic Res* **55**, 202–211.

Mandl, J. N., Ahmed, R., Barreiro, L. B., Daszak, P., Epstein, J. H., Virgin, H. W. & Feinberg, M. B. (2015). Reservoir host immune responses to emerging zoonotic viruses. *Cell* **160**, 20–35.

Marchler-Bauer, A., Lu, S., Anderson, J. B., Chitsaz, F., Derbyshire, M. K., DeWeese-Scott, C., Fong, J. H., Geer, L. Y., Geer, R. C. & other authors. (2011). CDD: a Conserved

Domain Database for the functional annotation of proteins. *Nucleic Acids Res* **39**, D225-229.

Marchler-Bauer, A., Derbyshire, M. K., Gonzales, N. R., Lu, S., Chitsaz, F., Geer, L. Y., Geer, R. C., He, J., Gwadz, M. & other authors. (2015). CDD: NCBI's conserved domain database. *Nucleic Acids Res* **43**, D222-6.

Marshall, N., Priyamvada, L., Ende, Z., Steel, J. & Lowen, A. C. (2013). *Influenza virus* reassortment occurs with high frequency in the absence of segment mismatch. *PLoS Pathog* **9**, 1–11.

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.

Marz, M., Beerenwinkel, N., Drosten, C., Fricke, M., Frishman, D., Hofacker, I. L., Hoffmann, D., Middendorf, M., Rattei, T. & other authors. (2014). Challenges in RNA virus bioinformatics. *Bioinformatics* **30**, 1793–1799.

Mazzei, M., Carrozza, M. L., Luisi, E., Forzan, M., Giusti, M., Sagona, S., Tolari, F. & Felicioli, A. (2014). Infectivity of DWV associated to flower pollen: Experimental evidence of a horizontal transmission route. *PLoS One* **9**, e113448.

McDonald, S. M., Nelson, M. I., Turner, P. E. & Patton, J. T. (2016). Reassortment in segmented RNA viruses: mechanisms and outcomes. *Nat Rev Microbiol* **14**, 448–460.

Mihara, T., Nishimura, Y., Shimizu, Y., Nishiyama, H., Yoshikawa, G., Uehara, H., Hingamp, P., Goto, S. & Ogata, H. (2016). Linking virus genomes with host taxonomy. *Viruses* **8**, 10–15.

Misof, B., Liu, S., Meusemann, K., Peters, R. S., Donath, A., Mayer, C., Frandsen, P. B., Ware, J., Flouri, T. & other authors. (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.

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.

Moreira, D. & López-García, P. (2009). Ten reasons to exclude viruses from the tree of life. *Nat Rev Microbiol* **7**, 306–11.

Nagasaki, K. (2008). Dinoflagellates, diatoms, and their viruses. *J Microbiol* **46**, 235–43.

Nasir, A. & Caetano-Anolles, G. (2015). A phylogenomic data-driven exploration of viral origins and evolution. *Sci Adv* **1**, e1500527–e1500527.

Nei, M. & Gojobori, T. (1986). Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. *Mol Biol Evol* **3**, 418–426.

Onafuwa-Nuga, A. & Telesnitsky, A. (2009). The remarkable frequency of human immunodeficiency virus type 1 genetic recombination. *Microbiol Mol Biol Rev* **73**, 451–80.

Ostfeld, R. S., Glass, G. E. & Keesing, F. (2005). Spatial epidemiology: An emerging (or re-

emerging) discipline. *Trends Ecol Evol* **20**, 328–336.

Paces, J. (2002). HERVd: database of human endogenous retroviruses. *Nucleic Acids Res* **30**, 205–206.

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.

Parmentier, L., Smagghe, G., de Graaf, D. C. & Meeus, I. (2016). Varroa destructor Macula-like virus, Lake Sinai virus and other new RNA viruses in wild bumblebee hosts (*Bombus pascuorum*, *Bombus lapidarius* and *Bombus pratorum*). *J Invertebr Pathol* **134**, 6–11.

Parrish, C. R., Holmes, E. C., Morens, D. M., Park, E.-C., Burke, D. S., Calisher, C. H., Laughlin, C. A., Saif, L. J. & Daszak, P. (2008). Cross-species virus transmission and the emergence of new epidemic diseases. *Microbiol Mol Biol Rev* **72**, 457–470.

Patel, M. R., Emerman, M. & Malik, H. S. (2011). Paleovirology—ghosts and gifts of viruses past. *Curr Opin Virol* **1**, 304–309.

Paul, J. H. & Sullivan, M. B. (2005). Marine phage genomics: what have we learned? *Curr Opin Biotechnol* **16**, 299–307.

Petersen, T. N., Brunak, S., von Heijne, G. & Nielsen, H. (2011). SignalP 4.0: discriminating signal peptides from transmembrane regions. *Nat Methods* **8**, 785–786.

Phan, T. G., Kapusinszky, B., Wang, C., Rose, R. K., Lipton, H. L. & Delwart, E. L. (2011). The fecal viral flora of wild rodents. *PLoS Pathog* **7**, e1002218.

Postler, T. S., Clawson, A. N., Amarasinghe, G. K., Basler, C. F., Bavari, S., Benkő, M., Blasdell, K. R., Briese, T., Buchmeier, M. J. & other authors. (2016). Possibility and Challenges of Conversion of Current Virus Species Names to Linnaean Binomials. *Syst Biol* syw096.

Prangishvili, D. (2003). Evolutionary insights from studies on viruses of hyperthermophilic archaea. *Res Microbiol* **154**, 289–294.

Prangishvili, D. (2013). The wonderful world of archaeal viruses. *Annu Rev Microbiol* **67**, 565–85.

Quan, P.-L., Firth, C., Conte, J. M., Williams, S. H., Zambrana-Torrelio, C. M., Anthony, S. J., Ellison, J. A., Gilbert, A. T., Kuzmin, I. V. & other authors. (2013). Bats are a major natural reservoir for hepaciviruses and pegiviruses. *Proc Natl Acad Sci* **110**, 8194–8199.

Quiñones-Mateu, M. E., Avila, S., Reyes-Teran, G. & Martinez, M. A. (2014). Deep sequencing: Becoming a critical tool in clinical virology. *J Clin Virol* **61**, 9–19.

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.

Rampelli, S., Soverini, M., Turroni, S., Quercia, S., Biagi, E., Brigidi, P. & Candela, M. (2016). ViromeScan: a new tool for metagenomic viral community profiling. *BMC Genomics* **17**, 165.

Raoult, D. & Forterre, P. (2008). Redefining viruses: lessons from Mimivirus. *Nat Rev Microbiol* **6**, 315–319.

Ratnasingham, S. & Hebert, P. D. N. (2007). BARCODING: bold: The Barcode of Life Data System (http://www.barcodinglife.org). *Mol Ecol Notes* **7**, 355–364.

Ravoet, J., Maharramov, J., Meeus, I., De Smet, L., Wenseleers, T., Smagghe, G. & de Graaf, D. C. (2013). Comprehensive bee pathogen screening in Belgium reveals Crithidia mellificae as a new contributory factor to winter mortality. *PLoS One* **8**, e72443.

Ravoet, J., De Smet, L., Meeus, I., Smagghe, G., Wenseleers, T. & de Graaf, D. C. (2014). Widespread occurrence of honey bee pathogens in solitary bees. *J Invertebr Pathol* **122**, 55–58.

Ravoet, J., De Smet, L., Wenseleers, T. & de Graaf, D. C. (2015). Genome sequence heterogeneity of *Lake Sinai Virus* found in honey bees and Orf1/RdRP-based polymorphisms in a single host. *Virus Res* **201**, 67–72.

Van Regenmortel, M. H. V. (1990). Virus species, a much overlooked but essential concept in virus classification. *Intervirology* **31**, 241–254.

Remmert, M., Biegert, A., Hauser, A. & Söding, J. (2011). HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment. *Nat Methods* **9**, 173–175.

Ren, W., Chen, H., Renault, T., Cai, Y., Bai, C., Wang, C. & Huang, J. (2013). Complete genome sequence of acute viral necrosis virus associated with massive mortality outbreaks in the Chinese scallop, *Chlamys farreri*. *Virol J* **10**, 110.

Rice, G., Stedman, K., Snyder, J., Wiedenheft, B., Willits, D., Brumfield, S., McDermott, T. & Young, M. J. (2001). Viruses from extreme thermal environments. *Proc Natl Acad Sci* **98**, 13341–13345.

Robertson, G., Schein, J., Chiu, R., Corbett, R., Field, M., Jackman, S. D., Mungall, K., Lee, S., Okada, H. M. & other authors. (2010). De novo assembly and analysis of RNA-seq data. *Nat Methods* **7**, 909–912.

Rohwer, F. & Edwards, R. (2002). The phage proteomic tree : a genome-based taxonomy for phage **184**, 4529–4535.

Romiguier, J., Gayral, P., Ballenghien, M., Bernard, A., Cahais, V., Chenuil, A., Chiari, Y., Dernat, R., Duret, L. & other authors. (2014a). Comparative population genomics in animals uncovers the determinants of genetic diversity. *Nature* **515**, 261–263.

Romiguier, J., Lourenco, J., Gayral, P., Faivre, N., Weinert, L. A., Ravel, S., Ballenghien, M., Cahais, V., Bernard, A. & other authors. (2014b). Population genomics of eusocial insects: the costs of a vertebrate-like effective population size. *J Evol Biol* **27**, 593–603.

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

Rosario, K. & Breitbart, M. (2011). Exploring the viral world through metagenomics. *Curr Opin Virol* **1**, 289–297.

Rosario, K., Schenck, R. O., Harbeitner, R. C., Lawler, S. N. & Breitbart, M. (2015). Novel

circular single-stranded DNA viruses identified in marine invertebrates reveal high sequence diversity and consistent predicted intrinsic disorder patterns within putative structural proteins. *Front Microbiol* **6**, 1–13.

Rose, R., Constantinides, B., Tapinos, A., Robertson, D. L. & Prosperi, M. (2016). Challenges in the analysis of viral metagenomes. *Virus Evol* **2**, vew022.

Roux, S., Faubladier, M., Mahul, A., Paulhe, N., Bernard, A., Debroas, D. & Enault, F. (2011). Metavir: a web server dedicated to virome analysis. *Bioinformatics* **27**, 3074–3075.

Roux, S., Tournayre, J., Mahul, A., Debroas, D. & Enault, F. (2014). Metavir 2: new tools for viral metagenome comparison and assembled virome analysis. *BMC Bioinformatics* **15**, 76.

Roux, S., Enault, F., Hurwitz, B. L. & Sullivan, M. B. (2015). VirSorter: mining viral signal from microbial genomic data. *PeerJ* **3**, e985.

Runckel, C., Flenniken, M. L., Engel, J. C., Ruby, J. G., Ganem, D., Andino, R. & DeRisi, J. L. (2011). Temporal analysis of the honey bee microbiome reveals four novel viruses and seasonal prevalence of known viruses, *Nosema*, and *Crithidia*. *PLoS One* **6**, e20656.

Ruska, H., Borries, B. v. & Ruska, E. (1939). Die Bedeutung der übermikroskopie für die Virusforschung. *Arch Gesamte Virusforsch* **1**, 155–169.

Santos, F., Yarza, P., Parro, V. V., Briones, C. & Anton, J. (2010). The metavirome of a hypersaline environment. *Environ Microbiol* **12**, 2965–2976.

Savin, K. W., Cocks, B. G., Wong, F., Sawbridge, T., Cogan, N., Savage, D. & Warner, S. (2010). A neurotropic herpesvirus infecting the gastropod, abalone, shares ancestry with oyster herpesvirus and a herpesvirus associated with the amphioxus genome. *Virol J* **7**, 308.

Scheuch, M., Höper, D. & Beer, M. (2015). RIEMS: a software pipeline for sensitive and comprehensive taxonomic classification of reads from metagenomics datasets. *BMC Bioinformatics* **16**, 69.

Schmidt, H. A., Strimmer, K., Vingron, M. & von Haeseler, A. (2002). TREE-PUZZLE: maximum likelihood phylogenetic analysis using quartets and parallel computing. *Bioinformatics* **18**, 502–504.

Schultz, J., Milpetz, F., Bork, P. & Ponting, C. P. (1998). SMART, a simple modular architecture research tool: Identification of signaling domains. *Proc Natl Acad Sci* **95**, 5857–5864.

La Scola, B. (2003). A giant virus in Amoebae. *Science* **299**, 2033–2033.

La Scola, B., Desnues, C., Pagnier, I., Robert, C., Barrassi, L., Fournous, G., Merchat, M., Suzan-Monti, M., Forterre, P. & other authors. (2008). The virophage as a unique parasite of the giant mimivirus. *Nature* **455**, 100–104.

Sébastien, A., Lester, P. J., Hall, R. J., Wang, J., Moore, N. E. & Gruber, M. A. M. (2015). Invasive ants carry novel viruses in their new range and form reservoirs for a honeybee pathogen. *Biol Lett* **11**, 20150610.

Sharma, D., Priyadarshini, P. & Vrati, S. (2015). Unraveling the web of viroinformatics: computational tools and databases in virus research. *J Virol* **89**, 1489–501.

Shi, M., Lin, X.-D., Vasilakis, N., Tian, J.-H., Li, C.-X., Chen, L.-J., Eastwood, G., Diao, X.-N., Chen, M.-H. & other authors. (2015). Divergent viruses discovered in arthropods and vertebrates revise the evolutionary history of the Flaviviridae and related viruses. *J Virol* **90**, 659–69. American Society for Microbiology (ASM).

Shi, M., Lin, X.-D., Tian, J.-H., Chen, L.-J., Chen, X., Li, C.-X., Qin, X.-C., Li, J., Cao, J.-P. & other authors. (2016). Redefining the invertebrate RNA virosphere. *Nature* 1–12.

Shimodaira, H. & Hasegawa, M. (1999). Multiple comparisons of log-likelihoods with applications to phylogenetic inference. *Mol Biol Evol* **16**, 1114–1116.

Simmonds, P., Adams, M. J., Benkő, M., Breitbart, M., Brister, J. R., Carstens, E. B., Davison, A. J., Delwart, E., Gorbalenya, A. E. & other authors. (2017). Consensus statement: Virus taxonomy in the age of metagenomics. *Nat Rev Microbiol* **15**, 161–168.

Simpson, A. G. B. & Roger, A. J. (2004). The real 'kingdoms' of eukaryotes. *Curr Biol* **14**, R693–R696.

Simpson, J. T., Wong, K., Jackman, S. D., Schein, J. E., Jones, S. J. M. & Birol, I. (2009). ABySS: a parallel assembler for short read sequence data. *Genome Res* **19**, 1117–1123.

Singh, R., Levitt, A. L., Rajotte, E. G., Holmes, E. C., Ostiguy, N., VanEngelsdorp, D., Lipkin, W. I., DePamphilis, C. W., Toth, A. L. & Cox-Foster, D. L. (2010). RNA viruses in hymenopteran pollinators: Evidence of inter-taxa virus transmission via pollen and potential impact on non-Apis hymenopteran species. *PLoS One* **5**, e14357.

Smits, S. L., Bodewes, R., Ruiz-González, A., Baumgärtner, W., Koopmans, M. P., Osterhaus, A. D. M. E. & Schürch, A. C. (2015). Recovering full-length viral genomes from metagenomes. *Front Microbiol* **6**.

Söding, J. (2005). Protein homology detection by HMM–HMM comparison. *Bioinformatics* **21**, 951–960.

Söding, J., Biegert, A. & Lupas, A. N. (2005). The HHpred interactive server for protein homology detection and structure prediction. *Nucleic Acids Res* **33**, W244-248.

Stanley, W. M. (1935). Isolation of a crystalline protein possessing the properties of Tobacco-mosaic virus. *Science* **81**, 644–645.

Steward, G. F., Culley, A. I., Mueller, J. A., Wood-Charlson, E. M., Belcaid, M. & Poisson, G. (2012). Are we missing half of the viruses in the ocean? *ISME J* **7**, 672–679.

Sun, S., La Scola, B., Bowman, V. D., Ryan, C. M., Whitelegge, J. P., Raoult, D. & Rossmann, M. G. (2010). Structural Studies of the Sputnik Virophage. *J Virol* **84**, 894–897.

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

Suttle, C. A. (2016). Environmental microbiology: Viral diversity on the global stage. *Nat Microbiol* **1**, 16205.

Suyama, M., Torrents, D. & Bork, P. (2006). PAL2NAL: robust conversion of protein

sequence alignments into the corresponding codon alignments. *Nucleic Acids Res* **34**, W609-12.

Szewczyk, B., Hoyos-Carvajal, L., Paluszek, M., Skrzecz, I. & Lobo De Souza, M. (2006). Baculoviruses - re-emerging biopesticides. *Biotechnol Adv* **24**, 143–160.

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.

Terzian, C., Ferraz, C., Demaille, J. & Bucheton, A. (2000). Evolution of the gypsy endogenous retrovirus in the *Drosophila melanogaster* subgroup. *Mol Biol Evol* **17**, 908–914.

Theis, K. R., Dheilly, N. M., Klassen, J. L., Brucker, R. M., Baines, J. F., Bosch, T. C. G., Cryan, J. F., Gilbert, S. F., Goodnight, C. J. & other authors. (2016). Getting the Hologenome Concept Right: an Eco-Evolutionary Framework for Hosts and Their Microbiomes. *mSystems* **1**, e00028-16 (J. A. Gilbert, Ed.).

Tilman, D. (2000). Causes, consequences and ethics of biodiversity. *Nature* **405**, 208–211.

Tomaru, Y., Toyoda, K., Kimura, K., Takao, Y., Sakurada, K., Nakayama, N. & Nagasaki, K. (2013). Isolation and characterization of a single-stranded RNA virus that infects the marine planktonic diatom *Chaetoceros* sp. (SS08-C03). *Phycol Res* **61**, 27–36.

Ungerer, M. C., Johnson, L. C. & Herman, M. A. (2008). Ecological genomics: understanding gene and genome function in the natural environment. *Heredity (Edinb)* **100**, 178–183.

van Valen, L. (1973). A new evolutionary law. *Evol theory* 1–30.

Valsecchi, J., Marmontel, M., Franco, C. L. B., Cavalcante, D. P., Cobra, I. V. D., Lima, I. J., Lanna, J. M., Ferreira, M. T. M., Nassar, P. M. & other authors. (2017). *Atualização e composição da lista – Novas Espécies de Vertebrados e Plantas na Amazônia 2014-2015*, Iniciativa. Brasília: WWF e Instituto de Desenvolvimento Sustentável Mamirauá 2017.

Vasilakis, N. & Tesh, R. B. (2015). Insect-specific viruses and their potential impact on arbovirus transmission. *Curr Opin Virol* **15**, 69–74.

Verneau, J., Levasseur, A., Raoult, D., La Scola, B. & Colson, P. (2016). MG-digger: An automated pipeline to search for giant virus-related sequences in metagenomes. *Front Microbiol* **7**, 1–11.

Vijaykrishna, D., Mukerji, R. & Smith, G. J. D. (2015). RNA virus reassortment: an evolutionary mechanism for host jumps and immune evasion. *PLoS Pathog* **11**, 1–6.

Villarreal, L. P. (2004). Are Viruses Alive? *Sci Am* 101–105.

Wang, C., Mitsuya, Y., Gharizadeh, B., Ronaghi, M. & Shafer, R. W. (2007a). Characterization of mutation spectra with ultra-deep pyrosequencing: Application to HIV-1 drug resistance. *Genome Res* **17**, 1195–1201.

Wang, T., Zeng, J., Lowe, C. B., Sellers, R. G., Salama, S. R., Yang, M., Burgess, S. M., Brachmann, R. K. & Haussler, D. (2007b). Species-specific endogenous retroviruses shape the transcriptional network of the human tumor suppressor protein p53. *Proc*

Natl Acad Sci **104**, 18613–18618.

Wasik, B. R. & Turner, P. E. (2013). On the biological success of viruses. *Annu Rev Microbiol* **67**, 519–41.

Weber, P. H. & Bujarski, J. J. (2015). Multiple functions of capsid proteins in (+) stranded RNA viruses during plant-virus interactions. *Virus Res* **196**, 140–149.

Webster, C., Longdon, B., Lewis, S. & Obbard, D. (2016). Twenty-five new viruses associated with the Drosophilidae (Diptera). *Evol Bioinforma* 13.

Webster, C. L., Waldron, F. M., Robertson, S., Crowson, D., Ferrari, G., Quintana, J. F., Brouqui, J.-M., Bayne, E. H., Longdon, B. & other authors. (2015). The discovery, distribution, and evolution of viruses associated with *Drosophila melanogaster*. *PLoS Biol* **13**, e1002210.

Wilfert, L., Long, G., Leggett, H. C., Schmid-Hempel, P., Butlin, R., Martin, S. J. M. & Boots, M. (2016). *Deformed wing virus* is a recent global epidemic in honeybees driven by *Varroa* mites. *Science* **351**, 594–597.

Woese, C. R. (1987). Bacterial Evolution. *Microbiology* **51**, 221–271.

Wommack, K. E., Bhavsar, J., Polson, S. W., Chen, J., Dumas, M., Srinivasiah, S., Furman, M., Jamindar, S. & Nasko, D. J. (2012). VIROME: a standard operating procedure for analysis of viral metagenome sequences. *Stand Genomic Sci* **6**, 427–439.

Yamashita, A., Sekizuka, T. & Kuroda, M. (2016). VirusTAP: Viral genome-targeted assembly pipeline. *Front Microbiol* **7**, 1–5.

Yang, Z. (1998). Likelihood ratio tests for detecting positive selection and application to primate lysozyme evolution. *Mol Biol Evol* **15**, 568–573.

Yang, Z. (2007). PAML 4: Phylogenetic Analysis by Maximum Likelihood. *Mol Biol Evol* **24**, 1586–1591.

Yang, Z. & Bielawski, J. P. (2000). Statistical methods for detecting molecular adaptation. *Trends Ecol Evol* **15**, 496–503.

Yue, C. & Genersch, E. (2005). RT-PCR analysis of *Deformed wing virus* in honeybees (*Apis mellifera*) and mites (*Varroa destructor*). *J Gen Virol* **86**, 3419–3424.

Zhang, Z. Q. (2013). Animal biodiversity: An outline of higher-level classification and survey of taxonomic richness (Addenda 2013). *Zootaxa* **3703**, 1–82.

Zhou, Q. & Haymer, D. S. (1997). Molecular structure of yoyo, a gypsy-like retrotransposon from the mediterranean fruit fly, *Ceratitis capitata*. *Genetica* **101**, 167–178.

Zwart, M. P. & Elena, S. F. (2015). Matters of size: genetic bottlenecks in virus infection and their potential impact on evolution. *Annu Rev Virol* **2**, 161–179.

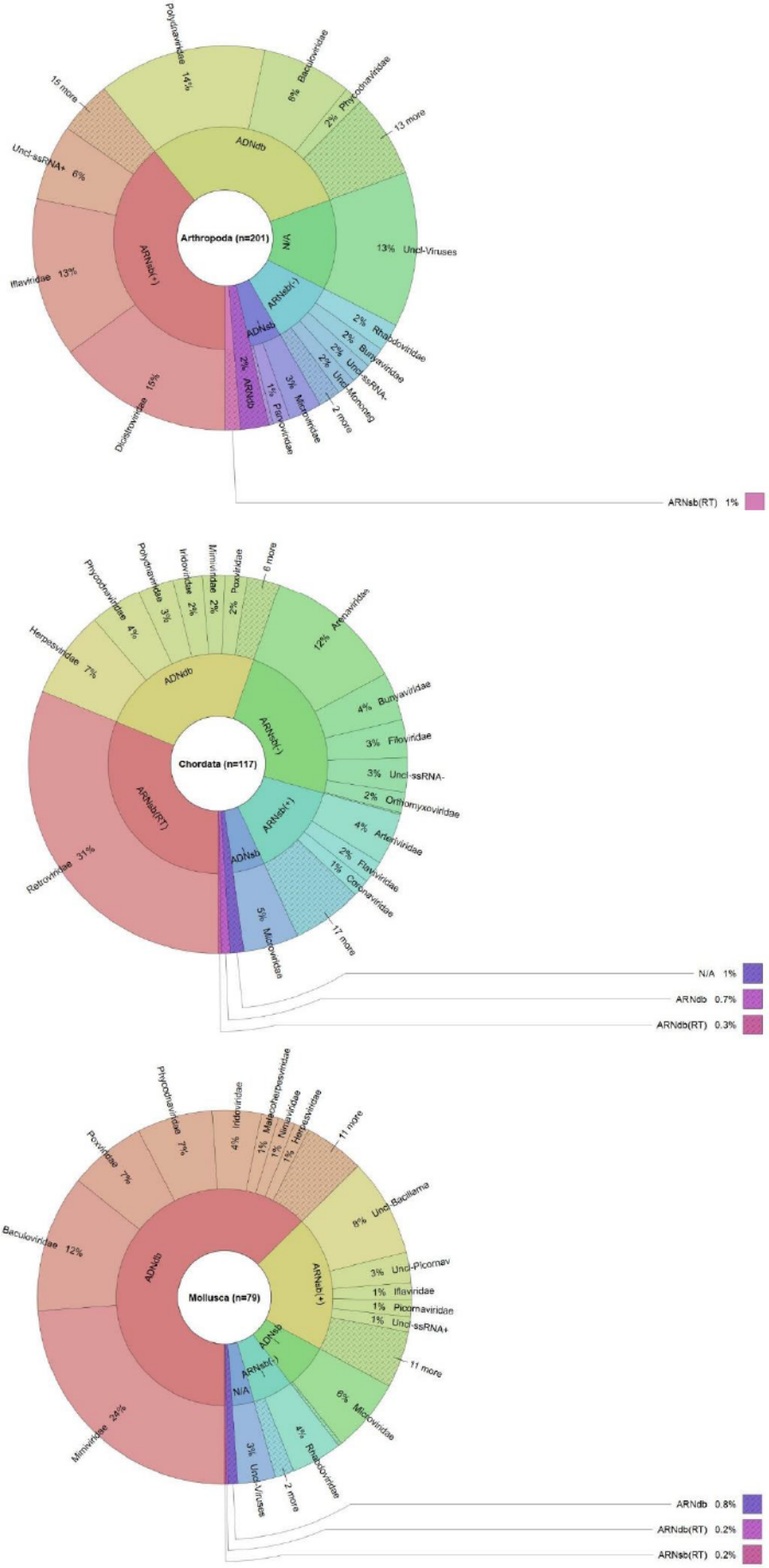
ANNEXES

Annexe 1: Liste des 135 espèces étudiées.

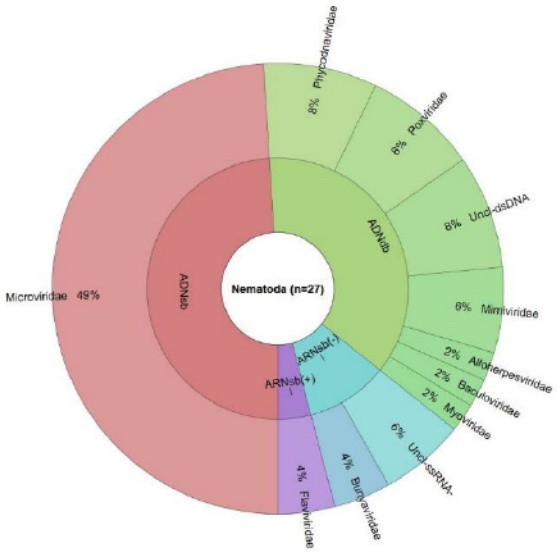
Nom de l'espèce	Nom Vernaculaire
Abatus agassizii	Sea urchin
Abatus cordatus	Sea urchin
Allolobophora chlorotica L1	Green worm
Allolobophora chlorotica L2	Green worm
Allolobophora chlorotica L4	Green worm
Aphaenogaster subterranea	Ant
Aporrectodea icterica	Mottled Worm
Aptenodytes forsteri	Emperor penguin
Aptenodytes patagonicus	King penguin
Artemia franciscana	Brine shrimp
Artemia salina	Brine shrimp
Artemia sinica	Brine shrimp
Artemia tibetiana	Brine shrimp
Artemia urmiana	Brine shrimp
Asbestopluma hypogea	Carnivorous sponge
Basiliscus plumifrons	Green basilisk
Boa constrictor	Boa constrictor
Bostrycapulus aculeatus	Spiny slippersnail
Caenorhabditis brenneri	Roundworm
Caenorhabditis sp.10 JR-2014	Roundworm
Caenorhabditis sp.16 NG-2017	Roundworm
Camponotus aethiops	Ant
Camponotus ligniperdus	Ant
Cephalothrix hongkongiensis	Ribbon worms
Cerebratulus marginatus	Ribbon worms
Chelonoidis carbonarius	Red-footed tortoise
Chelonoidis nigra	Galapagos Giant Tortoise
Ciona cf. intestinalis A AZ-2012	Vase tunicate
Ciona cf. intestinalis B AZ-2012	Vase tunicate
Clavelina lepadiformis	Light bulb sea squirt
Crepidula fornicata	Slipper Shell
Crepidula plana	Eastern White Slippersnail
Crocodylus niloticus	Nile crocodile
Culex hortensis	Mosquito
Culex pipiens	Common house mosquito
Culex torrentium	Mosquito
Cystodytes dellechiajei	Colonial seasquirt
Echinocardium cordatum	Common heart urchin
Echinocardium mediterraneum	Mediterranean urchin
Emiliania huxleyi	Coccolithophore
Emys orbicularis	European Pond Turtle
Escarpia southwardae	Marine worm
Eudyptes chrysolophus	Macaroni penguin
Eudyptes filholi	Southern rockhopper penguin
Eudyptes chrysocome moseleyi	Northern Rockhopper Penguin
Eunicella cavolinii	Yellow sea fan
Eunicella verrucosa	Pink sea fan
Formica cunicularia	Ant
Formica decipiens	Ant
Formica fusca	Ant
Formica gagates	Ant
Formica lusatica	Ant
Formica paralugubris	Ant
Formica polyctena	Ant
Formica pratensis	Ant
Formica sanguinea	Ant
Formica selysi	Ant
Galba truncatula	Lesser pond snail
Gephyrocapsa oceanica	Coccolithophore
Halictus scabiosae	Sweat Bee
Halictus sexcinctus	Bee
Halictus simplex	Bee
Hippocampus guttulatus	Long snouted seahorse
Hippocampus kuda	Common seahorse

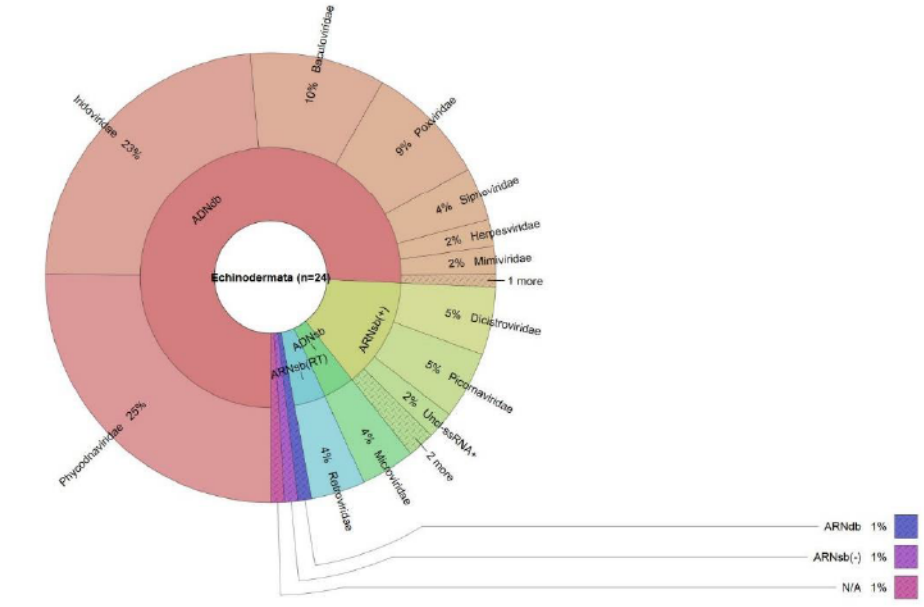
Iguana iguana	Green iguana
Lamellibrachia sp.	Deep-sea tubeworm
Lamellibrachia sp. 1 DAC-2013	Deep-sea tubeworm
Leptogorgia sarmentosa	Corals
Lepus americanus	Snowshoe hare
Lepus granatensis	Granada hare
Lepus timidus	Mountain hare
Lineus lacteus	Ribbon worms
Lineus longissimus	Bootlace worm
Lineus pseudolacteus	Ribbon worms
Lineus ruber	Ribbon worms
Lineus sanguineus	Ribbon worms
Liocarcinus depurator	Harbour crab
Lumbricus terrestris	Common earthworm
Lymnaea fuscus	Marsh pond snail
Lymneae cubensis	Pond snail
Lymneae sp.	Pond snail
Mauremys leprosa	Mediterranean Turtle
Melitaea cinxia	Glanville fritillary
Melitaea didyma	Red-band Fritillary
Mellicta athalia	Heath fritillary
Mellicta parthenoides	Meadow Fritillar
Messor barbarus	Harvester ant
Messor bouvieri	Harvester ant
Messor capitatus	Harvester ant
Messor concolor	Harvester ant
Messor structor	Harvester ant
Metasepia pfefferi	Flamboyant Cuttlefish
Microtus agrestis	Short-tailed vole
Microtus arvalis	Common vole
Microtus glareolus	Bank vole
Myrmica rubra	European fire ant
Mytilus californianus	California mussel
Mytilus edulis	Blue mussel
Mytilus galloprovincialis	Mediterranean mussel
Mytilus trossulus	Pacific blue mussel
Necora puber	Velvet crab
Ophioderma longicauda	Brittle stars
Oscheius sp.	Common soil nematode
Oscheius tipulae	Common soil nematode
Ostrea chilensis	Chilean Oyster
Ostrea edulis	European flat oyster
Ostreola stentina	Dwarf Oyster
Parus caeruleus	Blue tit
Parus major	Great tit
Pectinaria auricoma	Segmented worm
Pectinaria koreni	Trumpet Worm
Periparus ater	Coal Tit
Pheidole pallidula	Ant
Physa acuta	European physa
Physa carolinaea	Physa
Physa gyrina	Tadpole Physa
Physa hendersoni	Bayou Physa
Pleurodeles waltl	Sharp-ribbed Newt
Polyergus rufescens	European Amazon ant
Pontia daplidice	Bath white
Porcellio dilatatus petiti	Woodlouse
Pygoscelis adelia	Adelie Penguin
Pygoscelis papua	Gentoo Penguin
Reticulitermes flavipes	Eastern subterranean termite
Reticulitermes grassei	Termite
Reticulitermes lucifugus	Mediterranean termite
Riftia pachyptila	Giant tube worm
Salamandra salamandra	Fire salamander
Sepia officinalis	Common cuttlefish
Sepiella japonica	Japanese spineless cuttlefish
Thymelicus lineola	Essex skipper
Thymelicus sylvestris	Small Skipper
Trachemys scripta	Common Slider
Tripylus abatoides	Sea urchin
Tupinambis teguixin	Common tegu

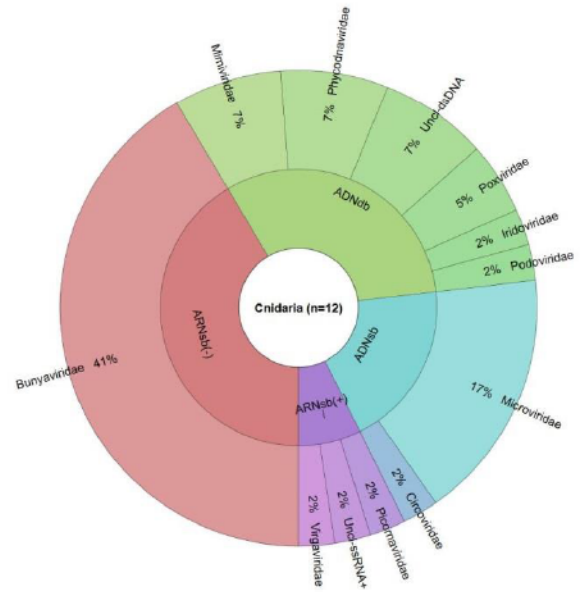
Annexe 2 : Répartition des hits viraux définis par BLAST réciproques dans les différents phyla animaux.

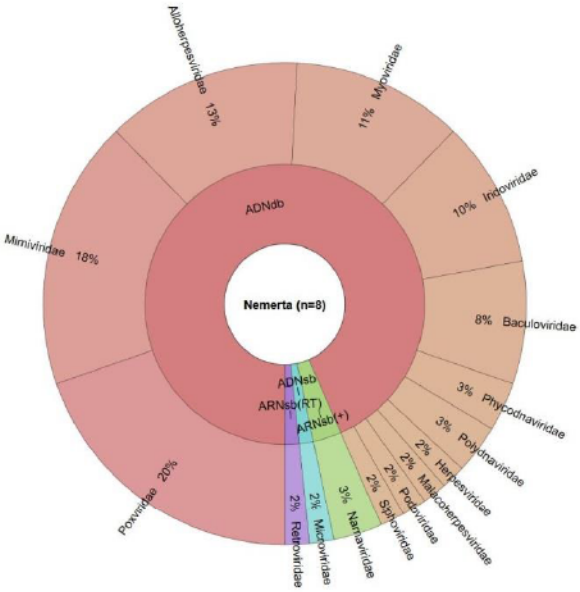


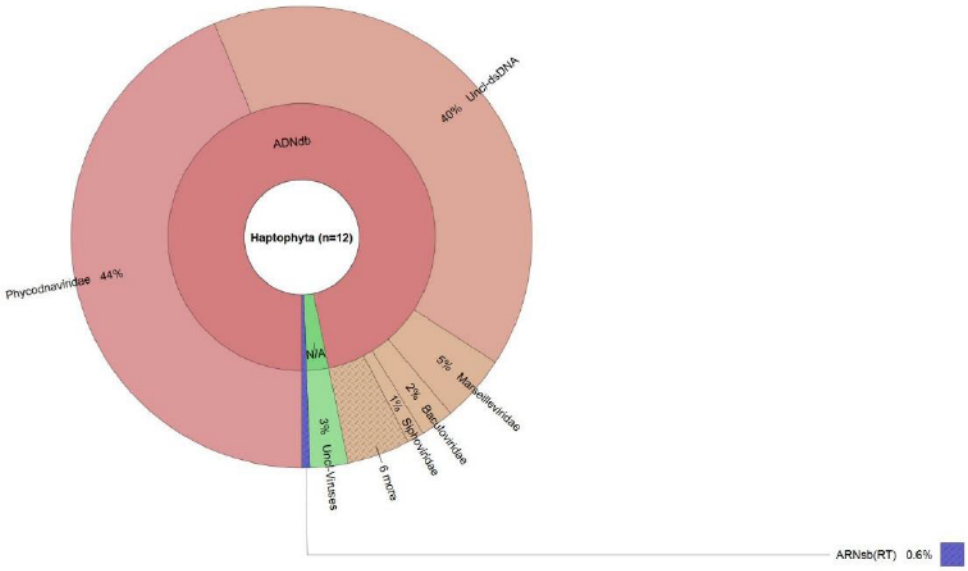

Annelida (n=43)
ADNdb
ARNsb(+)
ARNsb(-)
ADNsb
ARNsb(RT)
N/A
12% Baltiviviridae
4% Und-Picornav
2% Picornaviridae
2% Coronaviridae
4 more
9% Bunyaviridae
6% Partitiviridae
2% Microviridae
4% Und-Viruses
2% Retroviridae
11% Phycodnaviridae
9% Iridoviridae
6% Baculoviridae
5% Poxviridae
4% Mimiviridae
4% Polydnaviridae
3% Myoviridae
3% Siphoviridae
2% Nimaviridae
2% Adenoviridae
11 more
ARNdb(RT) 1%


Nematoda (n=27)
ADNdb
ADNsb
ARNsb(-)
ARNsb(+)
8% Phycodnaviridae
6% Poxviridae
6% Und-dsDNA
6% Mimiviridae
2% Alloherpesviridae
2% Baculoviridae
2% Myoviridae
6% Und-ssRNA-
4% Bunyaviridae
4% Flaviviridae
49% Microviridae


Echinodermata (n=24)
ADNdb
ADNsb
ARNsb(RT)
ARNsb(+)
10% Baculoviridae
9% Poxviridae
4% Siphoviridae
2% Herpesviridae
2% Mimiviridae
1 more
5% Dicistroviridae
5% Picornaviridae
2% Und-ssRNA+
2 more
4% Microviridae
4% Retroviridae
25% Phycodnaviridae
23% Iridoviridae
ARNdb 1%
ARNsb(-) 1%
N/A 1%

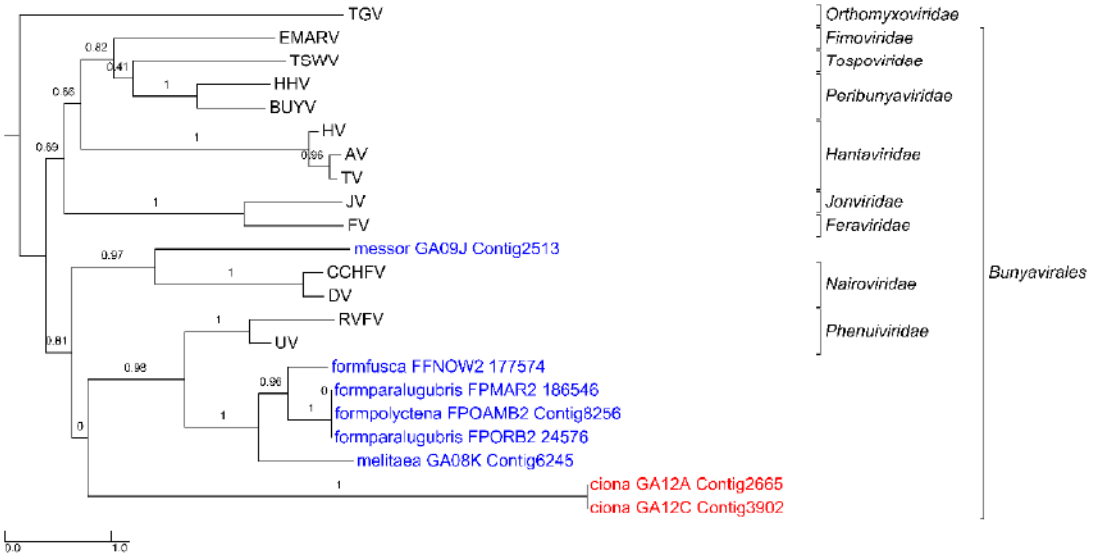

Cnidaria (n=12)
ADNdb
ADNsb
ARNsb(-)
ARNsb(+)
Mimiviridae 7%
Phycodnaviridae 7%
Unc-dsDNA 7%
Poxviridae 5%
Iridoviridae 2%
Podoviridae 2%
Microviridae 17%
Orthoherpesviridae 2%
Picornaviridae 2%
Unc-ssRNA+ 2%
Virgaviridae 2%
Bunyaviridae 41%


Nemerta (n=8)
ADNdb
ADNsb
ARNsb(RT)
ARNsb(+)
Alphaherpesviridae 13%
Mimiviridae 11%
Iridoviridae 10%
Baculoviridae 8%
Phycodnaviridae 3%
Polydnaviridae 3%
Herpesviridae 2%
Malacoherpesviridae 2%
Podoviridae 2%
Siphoviridae 2%
Nantanaviridae 3%
Microviridae 2%
Retroviridae 2%
Poxviridae 20%
Mimiviridae 18%

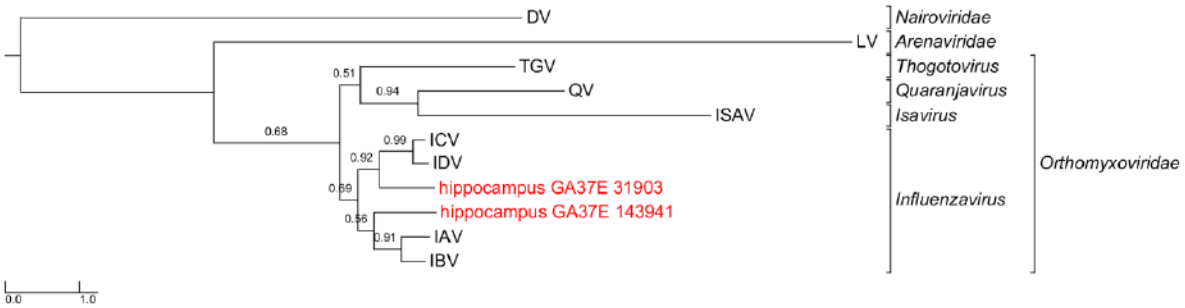

Haptophyta (n=12)
ADNdb
N/A
Unc-Viruses 3%
6 more 1%
Siphoviridae 2%
Baculoviridae 5%
Marseilleviridae 40%
Unc-dsDNA
Phycodnaviridae 44%
ARNsb(RT) 0.6%

Annexe 3 : Phylogénie par Maximum de Vraisemblance des réplicases virales.
En couleur sont indiquées les nouvelles séquences détectées dans cette étude (Bleu: Arthropoda, Rouge: Chordata, Vert: Mollusca, Orange: Annelida). La robustesse de nœuds est calculée par la statistiques aLRT. La barre d'échelle représente le taux de substitutions par sites. Les numéros d'accession et les noms complets des virus sont en Annexe 4.

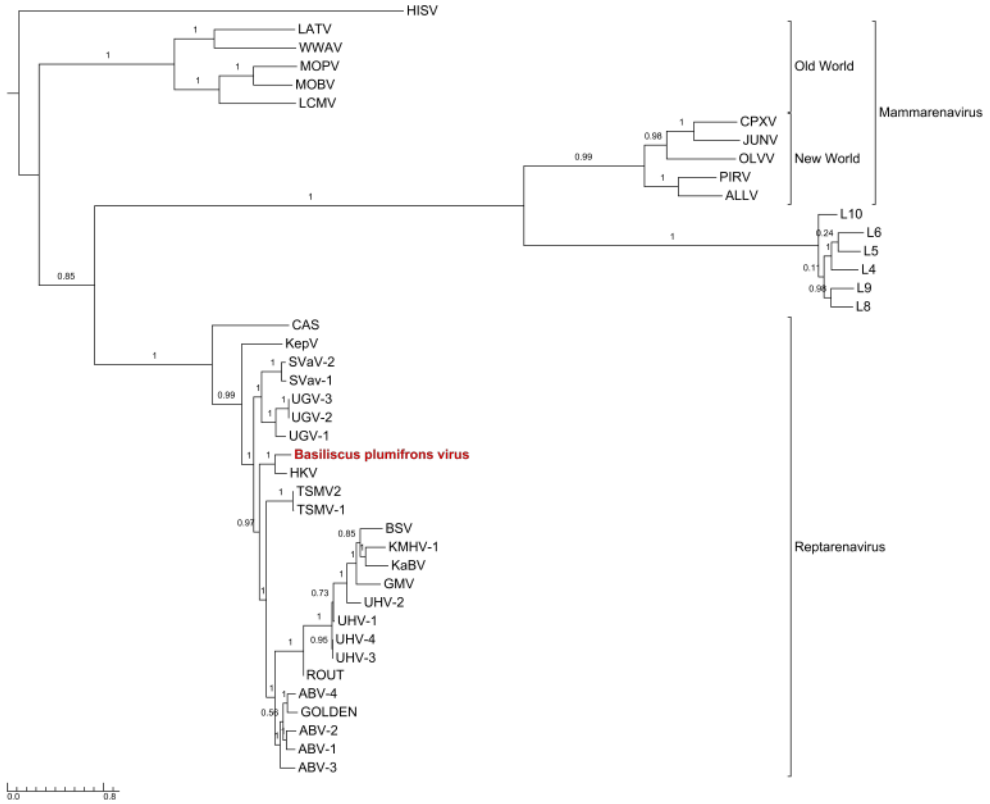
A-*Bunyavirales*



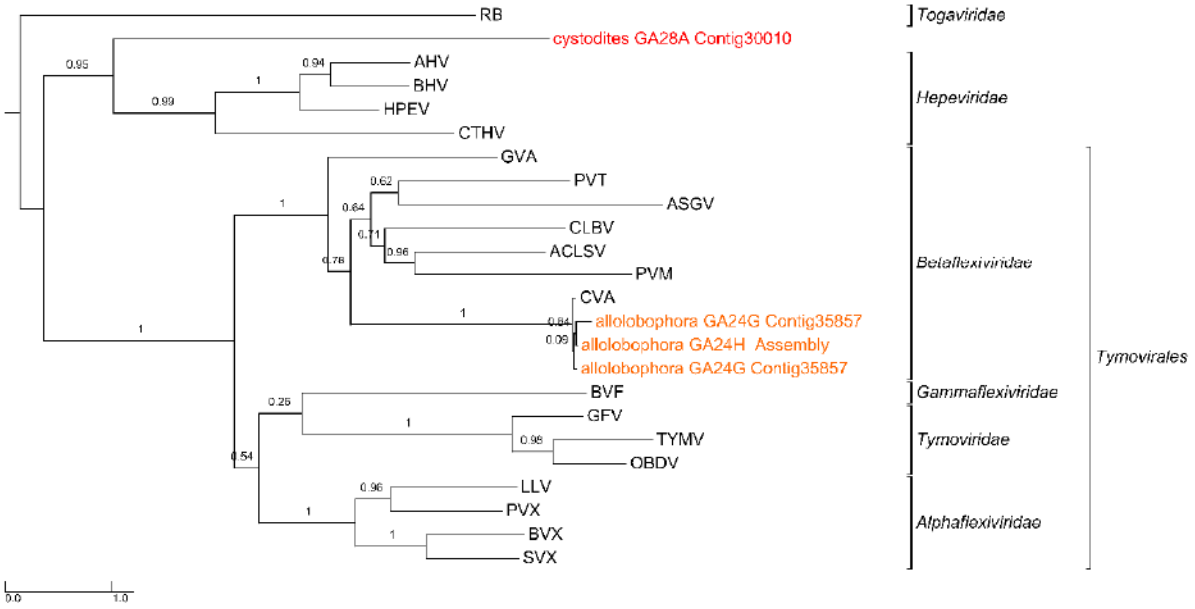
B-*Orthomyxoviridae*



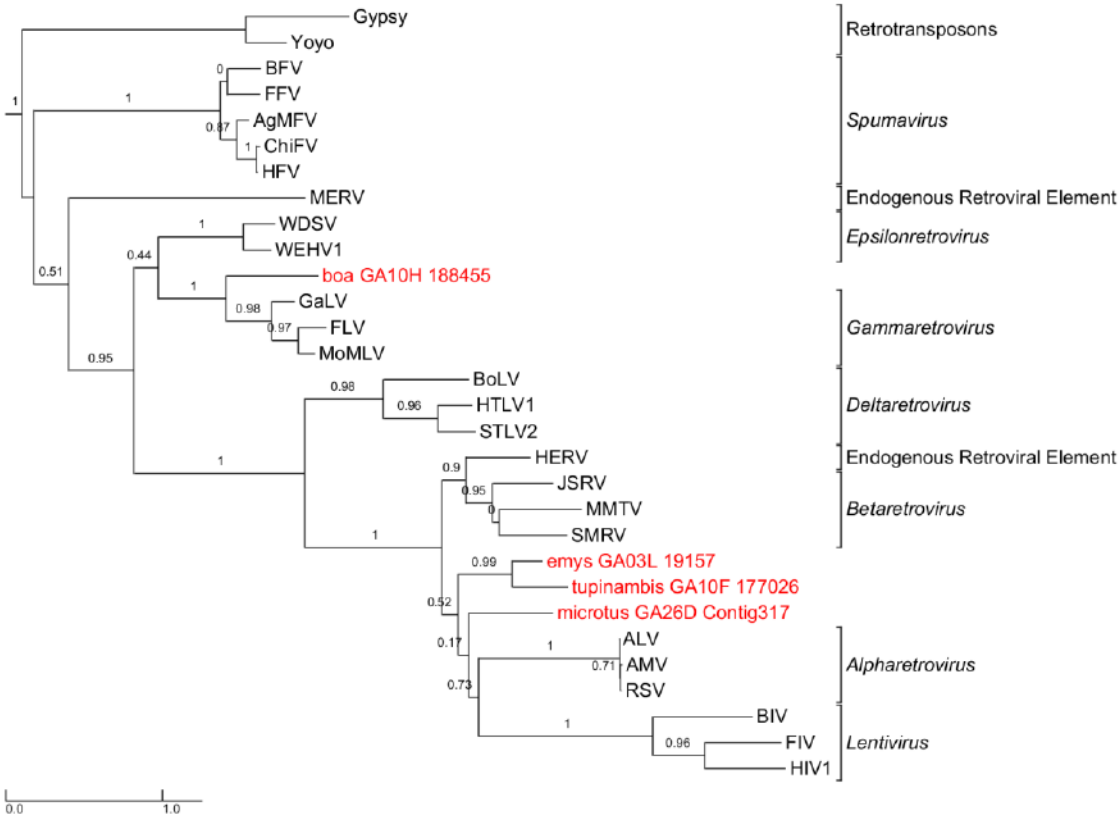
C-*Arenaviridae*



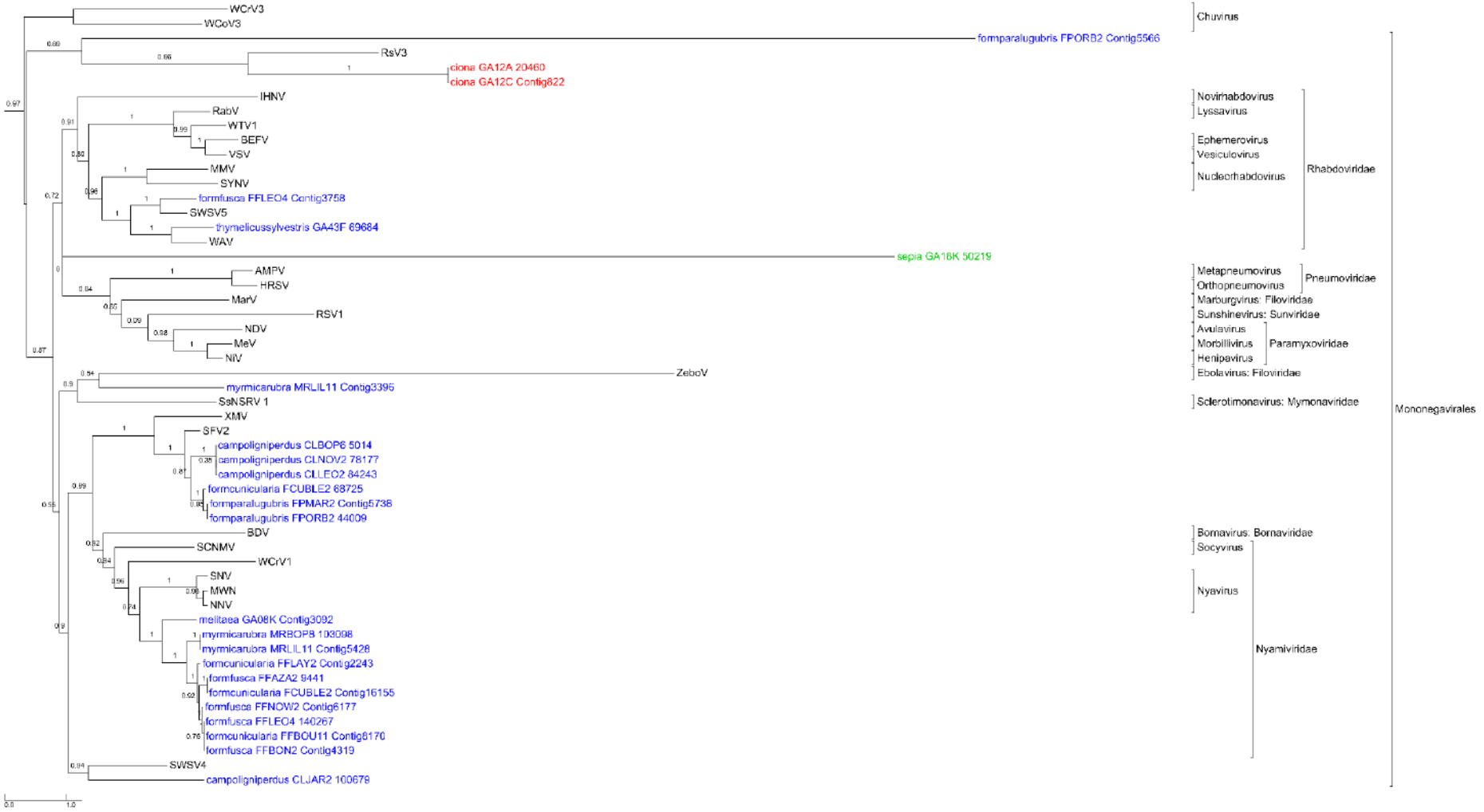
D-*Hepeviridae-Tymovirales*



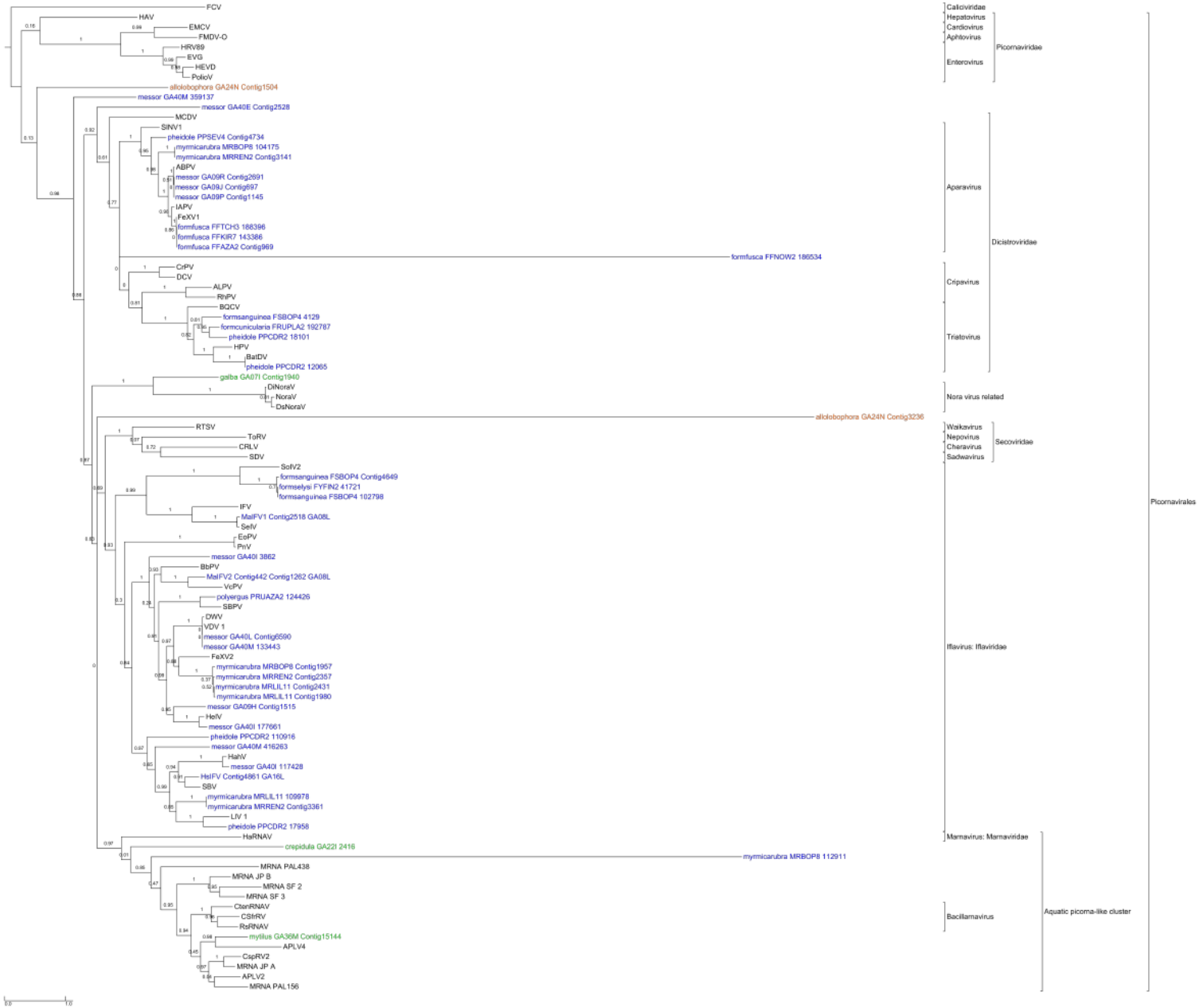
E-*Retroviridae*



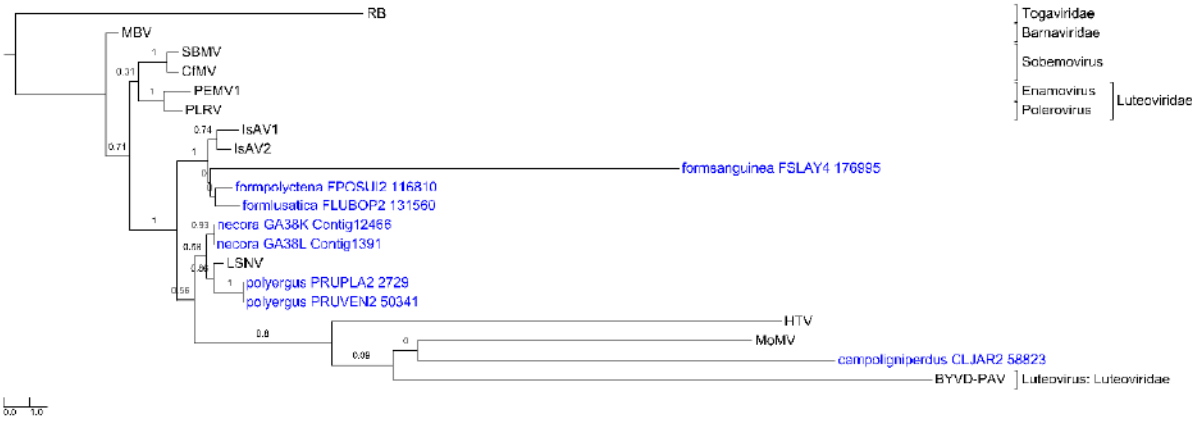
F-*Mononegavirales*



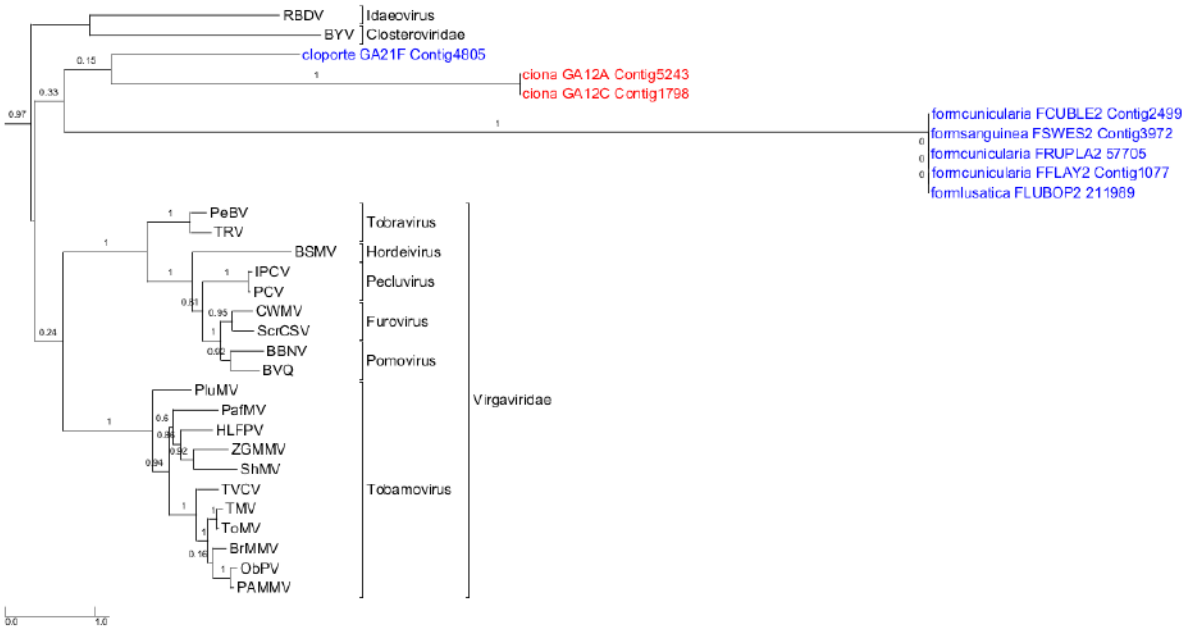
G-Picornavirales



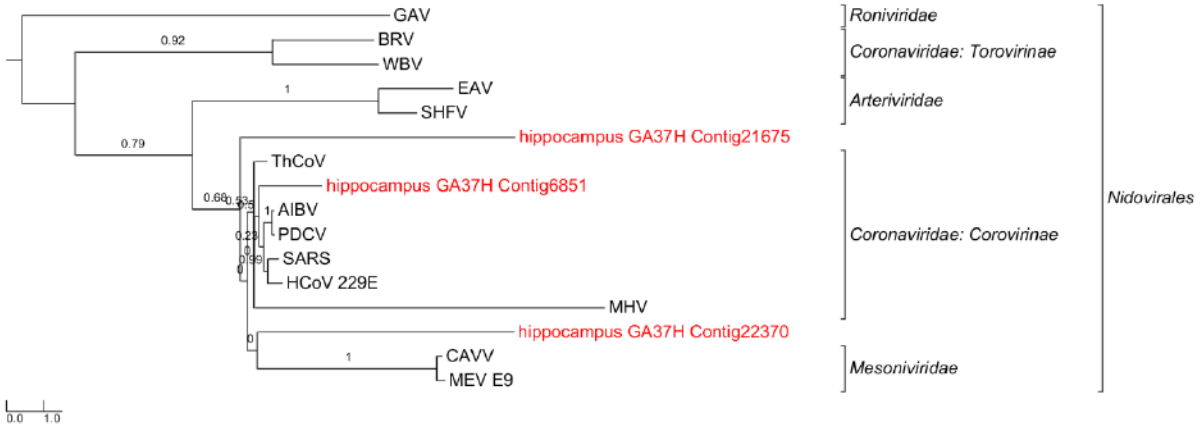
H-Luteoviridae-*Sobemovirus*



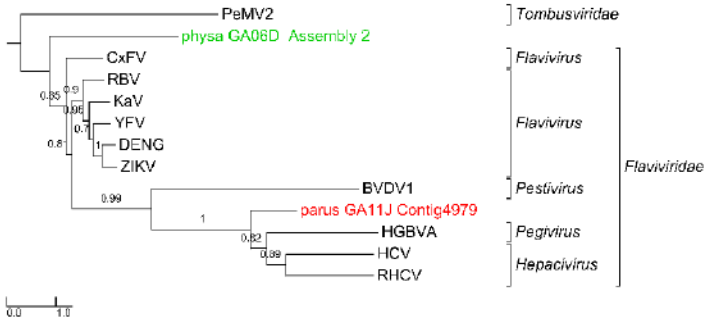
I-*Virgaviridae*



J-*Nidovirales*

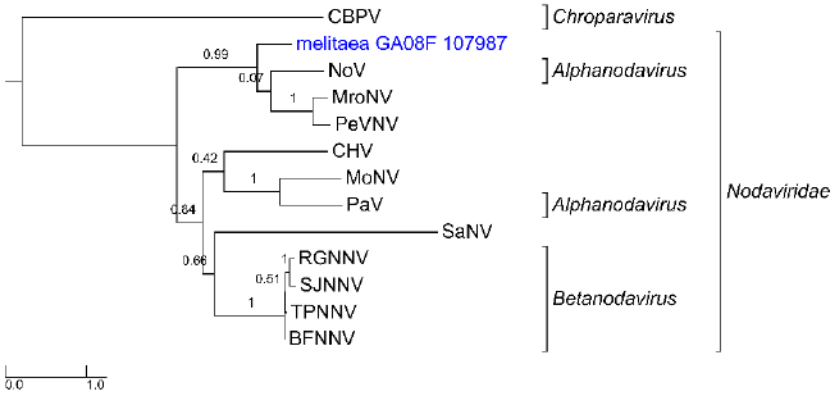


K-*Flaviviridae*

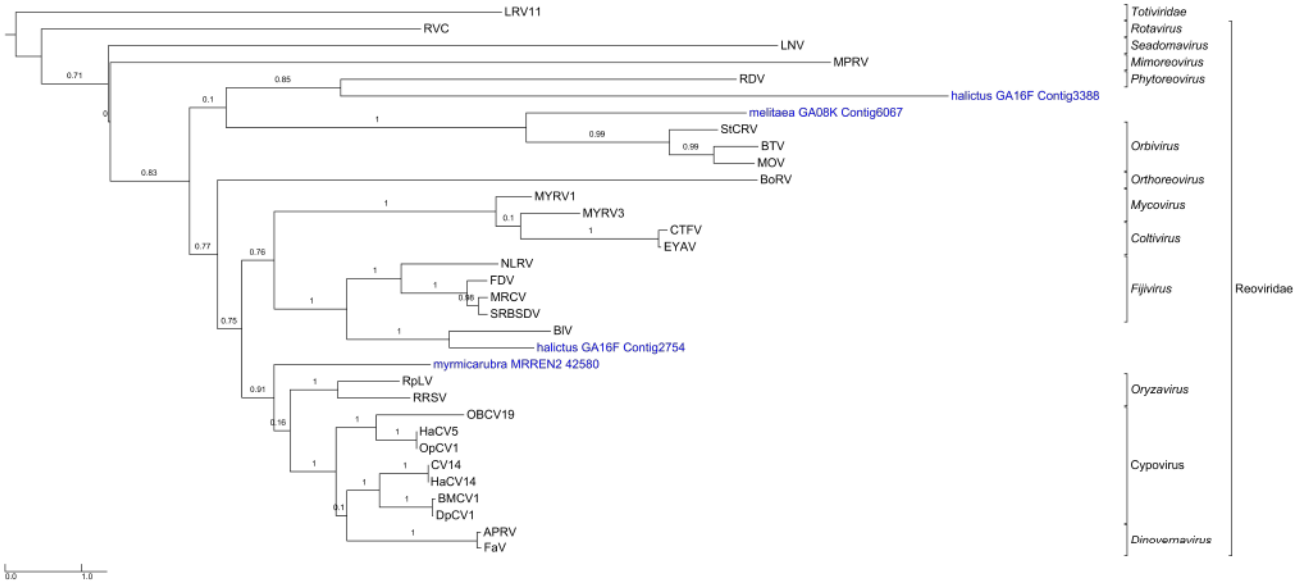


L-*Tombusviridae*

M-*Nodaviridae*



N-*Reoviridae*.



Annexe 4 : Liste des virus utilisés dans les phylogénies réplicases de la Figure 17 et Annexe 3.

INITIALS	VIRUS NAME	ORDER	FAMILY	GENUS	GENOME ACCESSION	REPLICASE PROTEIN ACCESSION
ABPV	*Acute Bee Paralysis virus*	Picornavirales	Dicistroviridae	Aparavirus	NC_002548.1	NP_066241
ABV-1	*Aurora borealis viruses*	-	Arenaviridae	Unclassified		KR870021
ACLSV	*Apple chlorotic leaf spot virus*	Tymovirales	Betaflexiviridae	Trichovirus	NC_001409.1	NP_040551.1
AgMFV	*African green monkey simian foamy virus*	-	Retroviridae	Spumavirus	NC_010820.1	YP_001956722
AHV	*Avian hepatitis E virus*	-	Hepeviridae	Orthohepevirus	NC_023425.1	YP_009001465
AIBV	*Avian Infectious bronchitis virus*	Nidovirales	Coronaviridae	Gammacorona-virus	NC_001451.1	NP_066134.1
ALLV	*Allpahuayo mammarenavirus*	-	Arenaviridae	Mammarenavirus		NC 010249
ALPV	*Aphil lethal paralysis virus*	Picornavirales	Dicistroviridae	Cripavirus	NC_004365.1	NP_733845
ALV	*Avian leukemia virus*	-	Retroviridae	Alpharetrovirus	NC_015116.1	YP_004222728
AMPV	*Avian metapneumovirus*	Mononegavirales	Pneumoviridae	Metapneumovirus	NC_007652.1	YP_443845.1
AMV	*Avian myeloblastosis-associated virus type 1*	-	Retroviridae	Alpharetrovirus	L10922.1	AAA46304
APLV2	*Antarctic picorna-like virus 2*	Picornavirales	Unclassified	env sample	NC_030233.1	YP_009255229.1
APLV4	*Antarctic picorna-like virus 4*	Picornavirales	Unclassified	env sample	NC_030235.1	YP_009255232.1
APRV	*Aedes pseudoscutellaris reovirus*	-	Reoviridae	Dinovernavirus	NC_007667.1	YP_443936
ASGV	*Apple stem grooving virus*	Tymovirales	Betaflexiviridae	Capillovirus	NC_001749.2	NP_044335.1
AV	*Andes virus*	Bunyavirales	Hantaviridae	Orthohantavirus	NC_003468.2	NP_604473.1
BatDV	*Bat dicistrovirus*	Picornavirales	Dicistroviridae	Unclassified	KF170223.1	AGN73377.1
BBNV	*Broad bean necrosis virus*	-	Virgaviridae	Pomovirus	D86636	BAA34692.2
BbPV	*Brevicoryne brassicae picorna-like virus*	Picornavirales	Iflaviridae	Iflavirus	NC_009530.1	YP_001285409
BDV	*Borna disease virus 1*	Mononegavirales	Bornaviridae	Bornavirus	NC_001607.1	NP_042024.3
BEFV	*Bovine ephemeral fever virus*	Mononegavirales	Rhabdoviridae	Ephemerovirus	NC_002526	NP_065409.1
BFNNV	*Barfin flounder nervous necrosis virus*	-	Nodaviridae	Betanodavirus	NC_013458.1	YP_003288756
BFV	*Bovine foamy virus*	-	Retroviridae	Spumavirus	NC_001831.1	NP_044929
BHV	*Bat hepevirus*	-	Hepeviridae	Orthohepevirus	NC_018382.1	YP_006576507
BIV	*Bovine immunodeficiency virus*	-	Retroviridae	Lentivirus	NC_001413.1	NP_040563
BlV	*Bloomfield virus*	-	Unclassified	Unclassified	KP714090	AKH40310.1
BMCV1	*Bombyx mori cypovirus 1*	-	Reoviridae	Cypovirus	AF323782.1	AAK20302
BoLV	*Bovine leukemia virus*	-	Retroviridae	Deltaretrovirus	K02120.1	AAA42785
BoRV	*Baboon orthoreovirus*	-	Reoviridae	Orthoreovirus	NC_015878.1	YP_004769548
BQCV	*Black Queen Cell Virus*	Picornavirales	Dicistroviridae	Triatovirus	NC_003784.1	NP_620564
BrMMV	*Brugmansia mild mottle virus*	-	Virgaviridae	Tobamovirus	NC_010944.1	YP_001974323
BRV	*Breda virus*	Nidovirales	Coronaviridae	Torovirus	NC_007447	YP_337905.2
BSMV	*Barley stripe mosaic virus*	-	Virgaviridae	Hordeivirus	MBSRNAGT	AAA66600.1
BTV	*Bluetongue virus*	-	Reoviridae	Orbivirus	NC_006023.1	YP_052968
BUYV	*Bunyamwera virus*	Bunyavirales	Peribunyaviridae	Orthobunyavirus	NC_001925.1	NP_047211.1
BVDV1	*Bovine viral diarrhea virus 1*	-	Flaviviridae	Pestivirus	NC_001461.1	NP_040937.1
BVF	*Botrytis virus F*	Tymovirales	Gammaflexiviridae	Mycoflexivirus	NC_002604.1	NP_068549.1
BVQ	*Beet virus Q*	-	Virgaviridae	Pomovirus	AJ223596	CAA11457.1
BVX	*Botrytis virus X*	Tymovirales	Alphaflexiviridae	Botrexvirus	NC_005132.1	NP_932306.1
BYDV-PAV	*Barley yellow dwarf virus - PAV*	-	Luteoviridae	Luteovirus	NC_004750.1	NP_840014.2
BYV	*Beet yellows virus*	-	Closteroviridae	Closterovirus	NC_001598	NP_041870.2
CarMV	*Carnation mottle virus*	-	Tombusviridae	Carmovirus	NC_001265.2	YP_009032645.1
CAS	*CAS Virus*	-	Arenaviridae	Reptarenavirus		NC 018484
CAVV	*Cavally virus*	Nidovirales	Mesoniviridae	Alphamesonivirus	NC_015668.1	YP_004598981.2
CBPV	*Chronic Bee Paralysis Virus*	-	Unclassified	Chroparavirus	KY937971.1	ARM39059.1
CCHFV	*Crimean-Congo hemorragic fever virus*	Bunyavirales	Nairoviridae	Orthonairovirus	NC_005301.3	YP_325663.1
CfMV	*Cocksfoot mottle virus*	-	Unclassified	Sobemovirus	NC_002618.2	NP_941957.2
ChiFV	*Chimpanzee simian foamy virus*	-	Retroviridae	Spumavirus	NC_001364.1	Q87040
CHV	*Craigies Hill virus*	-	Nodaviridae	Unclassified	KP714084.1	AKH40302
CLBV	*Citrus leaf blotch virus*	Tymovirales	Betaflexiviridae	Citrivirus	NC_003877.1	NP_624333.1
CMV	*Carrot mottle virus*	-	Tombusviridae	Umbravirus	NC_011515.1	YP_002302259.1
CPXV	*Cupixi mammarenavirus*	-	Arenaviridae	Mammarenavirus		NC010252
CRLV	*Cherry rasp leaf virus*	Picornavirales	Secoviridae	Cheravirus	NC_006271.1	YP_081444
CrPV	*Cricket paralysis virus*	Picornavirales	Dicistroviridae	Cripavirus	NC_003924.1	NP_647481

CSfrRV	*Chaetoceros socialis f radians RNA virus 1*	Picornavirales	Unclassified	Bacillarnavirus	NC_012212.1	YP_002647032.1
CspRV2	*Chaetoceros sp. RNA virus 2*	Picornavirales	Unclassified	env sample	AB639040.1	BAK40203.1
CtenRNAV	*Chaetoceros tenuissimus RNA virus 1*	Picornavirales	Unclassified	Bacillarnavirus	AB375474.1	BAG30951.1
CTFV	*Colorado tick fever virus*	-	Reoviridae	Coltivirus	NC_004181.1	NP_690891
CTHV	*Cutthroat trout virus*	-	Hepeviridae	Piscihepevirus	NC_015521.1	YP_004464917
CV14	*Cypovirus 14*	-	Reoviridae	Cypovirus	NC_003006.1	NP_149135
CVA	*Cherry virus A isolate WY*	Tymovirales	Betaflexiviridae	Capillovirus	KY445749.1	ASL72081.1
CWMV	*Chinese wheat mosaic virus*	-	Virgaviridae	Furovirus	AJ012005	CAB41769.1
CxFV	*Culex flavivirus*	-	Flaviviridae	Flavivirus	NC_008604.2	YP_899469.2
DCV	*Drosophila C virus*	Picornavirales	Dicistroviridae	Cripavirus	NC_001834.1	NP_044945
DENG	*Dengue virus 1*	-	Flaviviridae	Flavivirus	NC_001477.1	NP_059433.1
DiNoraV	*Drosophila immigrans Nora virus*	-	Unclassified	Unclassified	NC_024488.1	YP_009047190
DpCV1	*Dendrolimus punctatus cypovirus 1*	-	Reoviridae	Cypovirus	AY147187.1	AAN46860
DsNoraV	*Drosophila subobscura Nora virus*	-	Unclassified	Unclassified	NC_024487.1	YP_009047186
DV	*Dugbe virus*	Bunyavirales	Nairoviridae	Orthonairovirus	NC_004159.1	NP_690576.1
DWV	*Deformed Wing Virus*	Picornavirales	Iflaviridae	Iflavirus	NC_004830.2	NP_853560
EAV	*Equine arteritis virus*	Nidovirales	Arteriviridae	Simartevirus	NC_002532	NP_127506.1
EMARV	*European mountain ash ringspot-associated virus*	Bunyavirales	Fimoviridae	Emaravirus	AY563040.2	AAS73287.2
EMCV	*Encephalomyocarditis virus*	Picornavirales	Picornaviridae	Cardiovirus	NC_001479.1	NP_056777.1
EoPV	*Ectropis obliqua picorna-like virus*	Picornavirales	Iflaviridae	Iflavirus	NC_005092	NP_919029.1
EVG	*Enterovirus G*	Picornavirales	Picornaviridae	Enterovirus	KY498017.1	ARC95293.1
EYAV	*Eyach virus*	-	Reoviridae	Coltivirus	NC_003696.1	NP_620280
FaV	*Fako virus*	-	Reoviridae	Dinovernavirus	NC_025486.1	YP_009104379
FCV	*Feline calicivirus*	-	Caliciviridae	Vesivirus	M86379.1	AAA79326.1
FDV	*Fiji disease virus*	-	Reoviridae	Fijivirus	NC_007159.1	YP_249762
FeXV1	*Formica exsecta virus 1*	Picornavirales	Dicistroviridae	Unclassified	NC_023021.1	YP_008888535
FeXV2	*Formica exsecta virus 2*	Picornavirales	Iflaviridae	Iflavirus	NC_023022.1	YP_008888537
FFV	*Feline foamy virus*	-	Retroviridae	Spumavirus	KC292054.1	AGC11908
FIV	*Feline immunodeficiency virus*	-	Retroviridae	Lentivirus	NC_001482.1	NP_040973
FLV	*Feline leukemia virus*	-	Retroviridae	Gammaretrovirus	NC_001940.1	NP_955577
FMDV-O	*Foot-and-mouth disease virus - type O*	Picornavirales	Picornaviridae	Aphthovirus	NC_004004.1	NP_658990.1
FV	*Ferak virus*	Bunyavirales	Feraviridae	Orthoferavirus	KP710246.1	AKN56888.1
GaLV	*Gibbon ape leukemia virus*	-	Retroviridae	Gammaretrovirus	NC_001885.2	NP_056790
GAV	*Gill-associated virus*	Nidovirales	Roniviridae	Okavirus	NC_010306	YP_001661452.1
GFV	*Grapevine fleck virus*	Tymovirales	Tymoviridae	Maculavirus	NC_003347.1	NP_542612.1
GOLDEN	*Golden Gate Virus*	-	Arenaviridae	Reptarenavirus		NC_018482
GRV	*Groundnut rosette virus*	-	Tombusviridae	Umbravirus	NC_003603.1	YP_009162058.1
GVA	*Grapevine virus A*	Tymovirales	Betaflexiviridae	Vitivirus	NC_003604.2	NP_619662.1
Gypsy	*Gypsy*	-	-	Retrotransposons	EF591043.1	ABQ84946
HaCV14	*Heliothis armigera cypovirus 14*	-	Reoviridae	Cypovirus	DQ242048.1	ABB51571
HaCV5	*Heliothis armigera cypovirus 5*	-	Reoviridae	Cypovirus	NC_010668.1	YP_001883321
HahV	*Halyomorpha halys virus*	Picornavirales	Iflaviridae	Unclassified	NC_022611.1	YP_008719809
HaRNAV	*Heterosigma akashiwo RNA virus*	Picornavirales	Marnaviridae	Marnavirus	NC_005281.1	NP_944776.1
HAV	*Hepatitis A virus*	Picornavirales	Picornaviridae	Hepatovirus	NC_001489.1	NP_041007.1
HCoV-229E	*Human coronavirus 229E*	Nidovirales	Coronaviridae	Alphacoronavirus	NC_002645.1	NP_073549.1
HCV	*Hepatitis C virus genotype 1*	-	Flaviviridae	Hepacivirus	NC_004102.1	NP_671491.1
HeIV	*Heliconius erato iflavirus*	Picornavirales	Iflaviridae	Iflavirus	NC_024016.1	YP_009026409
HERV	*Human endogenous retrovirus-like element*	-	Retroviridae	Endogenous Retroviral Element	AF164614.1	AAD51797.1
HEVD	*Human enterovirus D*	Picornavirales	Picornaviridae	Enterovirus	NC_001430.1	NP_040760.1
HFV	*Human foamy virus*	-	Retroviridae	Spumavirus	Y07723.1	CAA68997
HGBVA	*Hepatitis GB virus A*	-	Flaviviridae	Pegivirus	NC_001837.1	NP_045010.1
HHV	*Herbert herbevirus*	Bunyavirales	Peribunyaviridae	Herbevirus	JQ659256.1	AFR34023.1
HISV	*Haartman Institute snake virus*	-	Arenaviridae	Unclassified		KR870031
HIV1	*Human immunodeficiency virus 1*	-	Retroviridae	Lentivirus	NC_001802.1	NP_789740
HKV	*Kompis virus*	-	Arenaviridae	Unclassified		KR870028
HLFPV	*Hibiscus latent Fort Pierce virus*	-	Virgaviridae	Tobamovirus	FJ196834	ACI24009.1
HPEV	*Hepatitis E virus*	-	Hepeviridae	Orthohepevirus	NC_001434.1	NP_056779
HPV	*Himetobi P virus*	Picornavirales	Dicistroviridae	Triatovirus	NC_003782.1	NP_620560
HRSV	*Human respiratory syncytial virus*	Mononegavirales	Pneumoviridae	Orthopneumovirus	NC_001781.1	NP_056866.1
HRV89	*Human rhinovirus 89*	Picornavirales	Picornaviridae	Enterovirus	NC_001617.1	NP_042288.1
HTLV1	*Human T-lymphotropic virus 1*	-	Retroviridae	Deltaretrovirus	NC_001436.1	NP_057860.1

HTV	*Humaita-Tubiacanga virus*	-	Unclassified	Unclassified	KR003801	AKP18618.1
HV	*Hantaan virus*	Bunyavirales	Hantaviridae	Orthohantavirus	NC_005222.1	NP_941982.1
IAPV	*Israeli Acute Paralysis virus*	Picornavirales	Dicistroviridae	Aparavirus	KY243933.1	APZ86807.1
IAV	*Influenza A virus (A/Puerto Rico/8/1934(H1N1))*	-	Orthomyxoviridae	Influenzavirus A	NC_002021.1	NP_040985.1
IBV	*Influenza B virus (B/Lee/1940)*	-	Orthomyxoviridae	Influenzavirus B	M14880.1	AAA43767.1
ICV	*Influenza C virus (C/Ann Arbor/1/50)*	-	Orthomyxoviridae	Influenzavirus C	NC_006308.1	YP_089653.1
IDV	*Influenza D virus (D/bovine/Minnesota/628/2013)*	-	Orthomyxoviridae	Influenzavirus D	KF425653.1	AGS48810
IFV	*Infectious flacherie virus*	Picornavirales	Iflaviridae	Iflavirus	AB000906	BAA25371.1
IHNV	*Infectious hematopoietic necrosis virus*	Mononegavirales	Rhabdoviridae	Novirhabdovirus	NC_001652.1	NP_042681.1
IPCV	*Indian peanut clump virus*	-	Virgaviridae	Pecluvirus	NC_004729.1	NP_835282.1
ISAV	*Infectious salmon anemia virus*	-	Orthomyxoviridae	Isavirus	NC_006503.1	YP_145804.1
IsAV1	*Ixodes scapularis associated virus 1*	-	Unclassified	Unclassified	KM048318.1	AII01797
IsAV2	*Ixodes scapularis associated virus 2*	-	Unclassified	Unclassified	KM048319.1	AII01812
JSRV	*Jaagsiekte sheep retrovirus*	-	Retroviridae	Betaretrovirus	NC_001494.1	NP_041186
JUNV	*Junin mammarenavirus*	-	Arenaviridae	Mammarenavirus		AY216507
JV	*Jonchet virus*	Bunyavirales	Jonviridae	Orthojonvirus	KP710232.1	AKN56871.1
KaV	*Karshi virus*	-	Flaviviridae	Flavivirus	NC_006947.1	YP_224133.1
LATV	*Latino mammarenavirus*	-	Arenaviridae	Mammarenavirus		EU627612
LCMV	*Lymphocytic choriomeningitis*	-	Arenaviridae	Mammarenavirus		AY847351
LLV	*Lolium latent virus*	Tymovirales	Alphaflexiviridae	Lolavirus	NC_010434.1	YP_001718499.1
LIV-1	*Lygus lineolaris virus 1*	Picornavirales	Iflaviridae	Iflavirus	JF720348.1	AEL30247.1
LNV	*Liao ning virus*	-	Reoviridae	Seadomavirus	NC_007736.1	YP_460026
LRV11	*Leishmania RNA virus 1-1*	-	Totiviridae	Leishmaniavirus	NC_002063.1	NP_041191.1
LSNV	*Laem Singh virus*	-	Unclassified	Unclassified	DQ127905	AAZ95951.1
LV	*Lassa virus*	-	Arenaviridae	Mammarenavirus	NC_004297.1	NP_694872.1
MarV	*Marburg marburgvirus*	Mononegavirales	Filoviridae	Marburgvirus	M92834	AAA46562.1
MBV	*Mushroom bacilliform virus*	-	Barnaviridae	Barnavirus	NC_001633.1	NP_042510
MCDV	*Mud crab dicistrovirus*	Picornavirales	Dicistroviridae	Unclassified	NC_014793.1	YP_004063985
MERV	*Murine endogenous retrovirus-like element*	-	Retroviridae	Endogenous Retroviral Element	Y12713.1	CAA73251
MeV	*Measles virus*	Mononegavirales	Paramyxoviridae	Morbillivirus	NC_001498.1	NP_056924.1
MEV-E9	*Meno virus strain E9/CI/2004*	Nidovirales	Mesoniviridae	Alphamesonivirus	NC_020900	YP_007697636.1
MHV	*Murine hepatitis virus*	Nidovirales	Coronaviridae	Betacoronavirus	NC_001846.1	NP_045298.1
MMTV	*Mouse mammary tumor virus*	-	Retroviridae	Betaretrovirus	NP_056880.1	NP_955564
MMV	*Maize mosaic virus*	Mononegavirales	Rhabdoviridae	Nucleorhabdovirus	NC_005975.1	YP_052855.1
MOBV	*Mobola mamarenavirus*	-	Arenaviridae	Mammarenavirus		DQ328876
MoMLV	*Moloney murine leukemia virus*	-	Retroviridae	Gammaretrovirus	NC_001501.1	NP_057933
MoMV	*Motts Mill virus*	-	Unclassified	Unclassified	KP714076	AKH40290.1
MoNV	*Mosinovirus*	-	Nodaviridae	Unclassified	KJ632942.1	AIO11151
MOPV	*Mopeia mammarenavirus*	-	Arenaviridae	Mammarenavirus		DQ328875
MOV	*Mobuck virus*	-	Reoviridae	Orbivirus	NC_022626.1	YP_008719912
MPRV	*Micromonas pusilla reovirus*	-	Reoviridae	Mimoreovirus	NC_008172.1	YP_654545
MRCV	*Mal de Rio Cuarto virus*	-	Reoviridae	Fijivirus	NC_008733.1	YP_956848
MRNA-JP-A	*Marine RNA virus JP-A*	Picornavirales	Unclassified	env sample	NC_009757.1	YP_001429581.1
MRNA-JP-B	*Marine RNA virus JP-B*	Picornavirales	Unclassified	env sample	NC_009758.1	YP_001429583.1
MRNA-PAL156	*Marine RNA virus PAL156*	Picornavirales	Unclassified	env sample	NC_029307.1	YP_009230120.1
MRNA-PAL438	*Marine RNA virus PAL438*	Picornavirales	Unclassified	env sample	NC_029308.1	YP_009230122.1
MRNA-SF-2	*Marine RNA virus SF-2*	Picornavirales	Unclassified	env sample	KF412901.2	AGZ83339.2
MRNA-SF-3	*Marine RNA virus SF-3*	Picornavirales	Unclassified	env sample	KF478836.2	AHA44480.1
MroNV	*Macrobrachium rosenbergii nodavirus*	-	Nodaviridae	Unclassified	NC_005094.1	NP_919036
MWN	*Midway nyavirus*	Mononegavirales	Nyamiviridae	Nyavirus	NC_012702.1	YP_002905331.1
MYRV1	*Mycoreovirus 1*	-	Reoviridae	Mycovirus	NC_010743.1	YP_001936004
MYRV3	*Mycoreovirus 3*	-	Reoviridae	Mycovirus	NC_007535.1	YP_392478
NDV	*Newcastle disease virus B1*	Mononegavirales	Paramyxoviridae	Avulavirus	NC_002617.1	NP_071471.1

NiV	*Nipah henipavirus*	Mononegavirales	Paramyxoviridae	Henipavirus	NC_002728.1	NP_112028.1
NLRV	*Nilaparvata lugens reovirus*	-	Reoviridae	Fijivirus	NC_003654.1	NP_619776
NNV	*Nyamanini nyavirus*	Mononegavirales	Nyamiviridae	Nyavirus	NC_012703.1	YP_002905337.1
NoraV	*Nora virus*	-	Unclassified	Unclassified	NC_007919.3	YP_004849308
NoV	*Nodamura virus*	-	Nodaviridae	Alphanodavirus	NC_002690.1	NP_077730.1
OBCV19	*Operophtera brumata cypovirus 19*	-	Reoviridae	Cypovirus	DQ192251.1	ABB17221
OBDV	*Oat blue dwarf virus*	Tymovirales	Tymoviridae	Marafivirus	NC_001793.1	NP_044447.1
ObPV	*Obuda pepper virus*	-	Virgaviridae	Tobamovirus	MTVGRNA	BAA02700.1
OCSV	*Oat chlorotic stunt virus*	-	Tombusviridae	Avenavirus	NC_003633.1	NP_619751.1
OLVV	*Oliveros mammarenavirus*	-	Arenaviridae	Mammarenavirus		NC 010250
OpCV1	*Orgyia pseudotsugata cypovirus 5*	-	Reoviridae	Cypovirus	KC588358.1	AHJ14783
PafMV	*Passion fruit mosaic virus*	-	Virgaviridae	Tobamovirus	NC_015552.1	YP_004465358
PAMMV	*Paprika mild mottle virus*	-	Virgaviridae	Tobamovirus	AB089381	BAC07266.1
PaV	*Pariacoto virus*	-	Nodaviridae	Alphanodavirus	NC_003691.1	NP_620109
PCV	*Peanup clump virus*	-	Virgaviridae	Pecluvirus	X78602	CAA55335.1
PDCV	*Pigeon-dominant Coronavirus isolate PdCoV/PG/Guangdong/1418/2014*	Nidovirales	Coronaviridae	Unclassified	KT254279.1	AKQ98478.1
PeBV	*Pea early browning virus*	-	Virgaviridae	Tobravirus	X14006	CAB37343.1
PEMV1	*Pea enation mosaic virus-1*	-	Luteoviridae	Enamovirus	NC_003629.1	NP_620026.2
PeMV2	*Pea enation mosaic virus-2*	-	Tombusviridae	Umbravirus	NC_003853.1	NP_620846.3
PeVNV	*Penaeus vannamei nodavirus*	-	Nodaviridae	Unclassified	NC_014978.1	YP_004207810
PIRV	*Pirital mammarenavirus*	-	Arenaviridae	Mammarenavirus		AY216505
PLRV	*Potato leafroll virus*	-	Luteoviridae	Polerovirus	NC_001747.1	NP_056748.3
PluMV	*Plumeria mosaic virus*	-	Virgaviridae	Tobamovirus	NC_026816.1	YP_009130653.1
PMV	*Panicum mosaic virus*	-	Tombusviridae	Panicovirus	NC_002598.1	NP_068342.1
PnV	*Perina nuda picorna-like virus*	Picornavirales	Iflaviridae	Iflavirus	AF323747	AAL06289.1
PolioV	*Poliovirus*	Picornavirales	Picornaviridae	Enterovirus	NC_002058.3	NP_041277.1
PVM	*Potato virus M*	Tymovirales	Betaflexiviridae	Carlavirus	NC_001361.2	NP_056767.1
PVT	*Potato virus T*	Tymovirales	Betaflexiviridae	Tepovirus	NC_011062.1	YP_002019748.1
PVX	*Potato virus X*	Tymovirales	Alphaflexiviridae	Potexvirus	M63141.1	AAA47172.1
QV	*Quaranfil virus*	-	Orthomyxoviridae	Quaranjavirus	FJ861695.1	ACY56282.1
RabV	*Rabies virus*	Mononegavirales	Rhabdoviridae	Lyssavirus	NC_001542.1	NP_056797.1
RB	*Rubella virus*	-	Togaviridae	Rubellavirus	NC_001545.2	NP_062883
RBDV	*Raspberry bushy dwarf virus*	-	Unclassified	Idaeovirus	NC_003739	NP_620465.1
RBV	*Rio Bravo virus*	-	Flaviviridae	Flavivirus	NC_003675.1	NP_620044.1
RDV	*Rice dwarf virus*	-	Reoviridae	Phytoreovirus	NC_003773.1	NP_620544
RGNNV	*Redspotted grouper nervous necrosis virus*	-	Nodaviridae	Betanodavirus	NC_008040.1	YP_611155
RHCV	*Rodent hepacivirus isolate RHV-339*	-	Flaviviridae	Hepacivirus	NC_021153.1	YP_007905733.1
RhPV	*Rhopalosiphum padi virus*	Picornavirales	Dicistroviridae	Cripavirus	NC_001874.1	NP_046155
ROUT	*ROUT Virus*	-	Arenaviridae	Reptarenavirus		NC 023762
RpLV	*Raspberry latent virus*	-	Reoviridae	Oryzavirus	NC_014600.1	YP_003934919
RRSV	*Rice ragged stunt virus*	-	Reoviridae	Oryzavirus	NC_003771.1	NP_620541
RsRNAV	*Rhizosolenia setigera RNA virus 01*	Picornavirales	Unclassified	Bacillarnavirus	NC_018613.1	YP_006732323.1
RSV	*Rous sarcoma virus Schmidt-Ruppin B*	-	Retroviridae	Alpharetrovirus	AF052428.1	AAC08988
RSV1	*Reptile sunshine virus 1*	Mononegavirales	Sunviridae	Sunshinevirus	NC_025345.1	YP_009094051.1
RsV3	*Rhizoctonia solani negative-strand virus 3*	Mononegavirales	Unclassified	Unclassified	KP900903.1	ALD89111.1
RTSV	*Rice tungro spherical virus*	Picornavirales	Secoviridae	Waikavirus	NC_001632.1	NP_042507
RVC	*Rotavirus C*	-	Reoviridae	Rotavirus	NC_007547.1	YP_392464
RVFV	*Rift Valley Fever virus*	Bunyavirales	Phenuiviridae	Phlebovirus	NC_014397.1	YP_003848704.1
SaNV	*Santeuil nodavirus*	-	Nodaviridae	Unclassified	NC_015069.1	YP_004221742
SARS	*Human SARS coronavirus*	Nidovirales	Coronaviridae	Betacoronavirus	NC_004718.3	NP_828849.2
SBMV	*Southern bean mosaic virus*	-	Unclassified	Sobemovirus	NC_004060.2	YP_007438858.2
SBPV	*Slow bee paralysis virus*	Picornavirales	Iflaviridae	Iflavirus	NC_014137.1	YP_003622540
SBV	*Sacbrood virus*	Picornavirales	Iflaviridae	Iflavirus	NC_002066.1	NP_049374
SCNMV	*Soybean cyst nematode midway virus*	Mononegavirales	Nyamiviridae	Socyvirus	NC_024702.1	YP_009052467.1
ScrCSV	*Sorghum chlorotic spot virus*	-	Virgaviridae	Furovirus	AB033691	BAA94802.1
SDV	*Satsuma dwarf virus*	Picornavirales	Secoviridae	Sadwavirus	NC_003785.2	NP_620566
SeIV	*Spodoptera exigua iflavirus*	Picornavirales	Iflaviridae	Iflavirus	JN091707	AET36829.1

SFV2	*Shuangao Fly virus 2*	Mononegavirales	Unclassified	Unclassified	KM817638	AJG39135.1
SHFV	*Simian hemorrhagic fever virus*	Nidovirales	Arteriviridae	Simartevirus	NC_003092	YP_009109556.3
ShMV	*Sunn-hemp mosaic virus*	-	Virgaviridae	Tobamovirus	U47034.1	AAB38492.1
SINV1	*Solenopsis invicta virus 1*	Picornavirales	Dicistroviridae	Aparavirus	NC_006559.1	YP_164440.1
SJNNV	*Striped Jack nervous necrosis virus*	-	Nodaviridae	Betanodavirus	NC_003448.1	NP_599247.1
SMRV	*Squirrel monkey retrovirus*	-	Retroviridae	Betaretrovirus	NC_001514.1	NP_041261
SNV	*Sierra Nevada virus*	Mononegavirales	Nyamiviridae	Nyavirus	NC_024376.1	YP_009044201.1
SoIV2	*Solenopsis invicta virus 2*	-	Unclassified	Unclassified	NC_009544.1	YP_001285729.1
SRBSDV	*Southern rice black-streacked dwarf virus*	-	Reoviridae	Fijivirus	NC_014714.1	YP_004021936
SsNSRV-1	*Sclerotinia sclerotiorum negative-stranded RNA virus 1*	Mononegavirales	Mymonaviridae	Sclerotimonavirus	NC_025383.1	YP_009094317.1
StCRV	*St Croix River virus*	-	Reoviridae	Orbivirus	NC_005997.1	YP_052942
STLV2	*Simian T-lymphotropic virus 2*	-	Retroviridae	Deltaretrovirus	Y14570.1	CAA74901
SVaV	*Suri Vanera virus*	-	Arenaviridae	Unclassified		KR870024
SVX	*Shallot virus X*	Tymovirales	Alphaflexiviridae	Allexivirus	NC_003795.1	NP_620648.1
SWSV4	*Sanxia Water Strider virus 4*	Mononegavirales	Unclassified	Unclassified	KM817633	AJG39115.1
SWSV5	*Sanxia Water Strider virus 5*	Mononegavirales	Unclassified	Unclassified	KM817634	AJG39120.1
SYNV	*Sonchus yellow net virus*	Mononegavirales	Rhabdoviridae	Nucleorhabdovirus	NC_001615.2	NP_042286.1
TBSV	*Tomato bushy stunt virus*	-	Tombusviridae	Tombusvirus	NC_001554.1	NP_062897.1
TBTV	*Tobacco bushy top virus*	-	Tombusviridae	Umbravirus	NC_004366.1	NP_733848.2
TGV	*Thogoto virus*	-	Orthomyxoviridae	Thogotovirus	NC_006508.1/NC_006495.1	YP_145810.1/YP_145794.1
ThCoV	*Thrush coronavirus HKU12-600*	Nidovirales	Coronaviridae	Deltacoronavirus	NC_011549.1	YP_002308496.1
TMV	*Tobacco mosaic virus*	-	Virgaviridae	Tobamovirus	V01408	CAA24688.1
TNVD	*Tobacco necrosis virus D*	-	Tombusviridae	Betanecrovirus	NC_003487.1	NP_608311.1
ToMV	*Tomato mosaic virus*	-	Virgaviridae	Tobamovirus	X02144	CAA26085.1
ToRV	*Tomato ringspot virus*	Picornavirales	Secoviridae	Nepovirus	NC_003840.1	NP_620765
TPNNV	*Tiger puffer nervous necrosis virus*	-	Nodaviridae	Betanodavirus	NC_013460.1	YP_003288759
TRV	*Tobacco rattle virus*	-	Virgaviridae	Tobravirus	AF166084	AAD48027.2
TSMV	*Tavallinen suomalainen mies virus*	-	Arenaviridae	Unclassified		KR870026
TSWV	*Tomato spotted wild virus*	Bunyavirales	Tospoviridae	Orthotospovirus	NC_002052.1	NP_049362.1
TV	*Tula virus*	Bunyavirales	Hantaviridae	Orthohantavirus	NC_005226.1	NP_942124.1
TVCV	*Turnip vein-clearinf virus*	-	Virgaviridae	Tobamovirus	BRU03387	AAC02783.1
TYMV	*Turnip yellow mosaic virus*	Tymovirales	Tymoviridae	Tymovirus	NC_004063.1	NP_663297.1
UGV-1	*University of Giessen viruses*	-	Arenaviridae	Unclassified		KR870022
UHV	*University of helsinki virus*	-	Arenaviridae	Reptarenavirus		NC_023765
UV	*Uukuniemi virus*	Bunyavirales	Phenuiviridae	Phlebovirus	NC_005214.1	NP_941973.1
VcPV	*Venturia canescens picorna-like virus*	Picornavirales	Picornaviridae	Unclassified	AY534885	AAS37668.1
VDV-1	*Varroa destructor virus 1*	Picornavirales	Iflaviridae	Iflavirus	NC_006494	YP_145791.1
VSV	*Vesicular stomatitis Indiana virus*	Mononegavirales	Rhabdoviridae	Vesiculovirus	NC_001560.1	NP_041716.1
WAV	*Wuhan Ant virus*	Mononegavirales	Unclassified	Unclassified	KM817645	AJG39155.1
WBV	*White bream virus*	Nidovirales	Coronaviridae	Bafinivirus	NC_008516	YP_803213.1
WCoV3	*Wuchang Cockroack virus 3*	-	-	Chuvirus	KM817604	AJG39067.1
WCrV1	*Wenzhou crab virus 1*	Mononegavirales	Unclassified	Unclassified	KM817644	AJG39154.1
WCrV3	*Wenzhou crab virus 3*	-	-	Chuvirus	KM817603	AJG39066.1
WDSV	*Walleye dermal sarcoma virus*	-	Retroviridae	Epsilonretrovirus	EF428979.1	ABO25842
WEHV1	*Walleye epidermal hyperplasia virus 1*	-	Retroviridae	Epsilonretrovirus	AF133051.1	AAD30048
WTV1	*Wuhan Tick virus 1*	Mononegavirales	Unclassified	Unclassified	KM817660	AJG39223.1
WWAV	*Whitewater Arroyo mamma-renavirus*	-	Arenaviridae	Mammarenavirus		EU646186
XMV	*Xincheng Mosquito virus*	Mononegavirales	Unclassified	Unclassified	KM817661	AJG39227.1
YFV	*Yellow Fever virus*	-	Flaviviridae	Flavivirus	NC_002031.1	NP_041726.1
Yoyo	*Yoyo*	-	-	Retrotransposons	U60529.1	AAC28743
ZeboV	*Zaire ebolavirus*	Mononegavirales	Filoviridae	Ebolavirus	NC_002549	NP_066244.1
ZGMMV	*Zucchini green mottle mosaic virus*	-	Virgaviridae	Tobamovirus	AJ295949	CAC82482.1
ZIKV	*Zika virus*	-	Flaviviridae	Flavivirus	NC_012532.1	YP_002790881.1

Préambule à l'Article 6

J'ai participé durant ma thèse à l'analyse bioinformatique (assemblage des transcriptomes, détection des ORFs, assignation taxonomique, annotation fonctionnelle avec CAZymes Analysis Toolkit) dans le cadre d'une collaboration avec Franck Dedeine, sur l'annalyse transcriptomique et fonctionnelle des gènes du métabolisme de la cellulose chez trois espèces de termites *Reticulitermes*.

Ces résultats ont étés valorisé dans une publication parue en 2015 dans PlosONE (doi : 10.1371/journal.pone.0145596).

Annexe 5 : Analyse comparative de transcriptomes de reproducteurs secondaires de trois espèces de termites *Reticulitermes* (**Article 6**).

PLOS | ONE

RESEARCH ARTICLE

Comparative Analysis of Transcriptomes from Secondary Reproductives of Three *Reticulitermes* Termite Species

Franck Dedeine[1], Lucy A. Weinert[2¤], Diane Bigot[1], Thibaut Josse[1], Marion Ballenghien[2], Vincent Cahais[2], Nicolas Galtier[2], Philippe Gayral[1]*

1 Institut de Recherche sur la Biologie de l'Insecte, UMR 7261, CNRS—Université François Rabelais, 37200, Tours, France, **2** Institut des Sciences de l'Evolution, UMR 5554, Université de Montpellier—CNRS—IRD—EPHE, Montpellier, France

¤ Current address: University of Cambridge, Department of Veterinary Medicine, Madingley Road, Cambridge, CB3 0ES, United Kingdom
* philippe.gayral@univ-tours.fr

OPEN ACCESS

Citation: Dedeine F, Weinert LA, Bigot D, Josse T, Ballenghien M, Cahais V, et al. (2015) Comparative Analysis of Transcriptomes from Secondary Reproductives of Three *Reticulitermes* Termite Species. PLoS ONE 10(12): e0145596. doi:10.1371/journal.pone.0145596

Editor: Erjun Ling, Institute of Plant Physiology and Ecology, CHINA

Received: February 27, 2015

Accepted: December 7, 2015

Published: December 23, 2015

Data Availability Statement: Reads used for this study originated from a previous work [31] on R. flavipes (Sequence Read Archive accession no. SRX565295 and SRX565296) and R. grassei (SRA accession no. SRX565297 to SRX565305), and from the present study for R. lucifugus (SRA accession no. SRX565306 and SRX565307).

Funding: Termite collection was supported by a a French ANR grant to FD (EvoSymTer 08-JCJC-007601). Transcriptome sequencing was supported by a European Research Council (ERC) grant to NG (ERC PopPhyl 232971).

Abstract

Termites are eusocial insects related to cockroaches that feed on lignocellulose. These insects are key species in ecosystems since they recycle a large amount of nutrients but also are pests, exerting major economic impacts. Knowledge on the molecular pathways underlying reproduction, caste differentiation or lignocellulose digestion would largely benefit from additional transcriptomic data. This study focused on transcriptomes of secondary reproductive females (nymphoid neotenics). Thirteen transcriptomes were used: 10 of *Reticulitermes flavipes* and *R. grassei* sequenced from a previous study, and two transcriptomes of *R. lucifugus* sequenced for the present study. After transcriptome assembly and read mapping, we examined interspecific variations of genes expressed by termites or gut microorganisms. A total of 18,323 orthologous gene clusters were detected. Functional annotation and taxonomic assignment were performed on a total of 41,287 predicted contigs in the three termite species. Between the termite species studied, functional categories of genes were comparable. Gene ontology (GO) terms analysis allowed the discovery of 9 cellulases and a total of 79 contigs potentially involved in 11 enzymatic activities used in wood metabolism. Altogether, results of this study illustrate the strong potential for the use of comparative interspecific transcriptomes, representing a complete resource for future studies including differentially expressed genes between castes or SNP analysis for population genetics.

Introduction

Termites (Blattodea, Termitoidae) constitute an ecologically and evolutionary diversified group of social insects (>2600 species) that share a common ancestry with cockroaches [1]. The ecological success of termites is often attributed to the combination of their sophisticated social organization with their unique ability to feed on recalcitrant plant matters such as wood

Competing Interests: The authors are also grateful to the Genotoul bioinformatics 405 platform Toulouse Midi-Pyrenees for providing computing and storage resources. This does not alter the authors' adherence to PLOS ONE policies on sharing data and Materials.

(lignocellulose) [2]. Lignocellulose digestion relies on a complex enzymatic system which is synthesized by termites and a diverse intestinal microbial community composed of numerous prokaryotes and, in some termites, unicellular eukaryotes (flagellated protists) [3]. Termites are major decomposers in many tropical and subtropical ecosystems and therefore, are crucial for recycling organic matter [2]. Conversely, some termites are pests, causing serious damage to human-built structures and woody plant crops.

Aside from a specialized nutritional regime, another characteristic of termites allowing them to successfully diversify worldwide is their sophisticated social organization. As in other social insects, such as social Hymenoptera (ants, some bees and wasps), termites live in complex societies where individuals are morphologically, physiologically and behaviorally specialized into distinct castes. The castes work together to accomplish specific and complementary tasks within a colony. Division of labor among castes is the key to efficient colony development, survival and reproduction. The social organization of termites also represents a primary reason why termite infestations can be difficult to control and eradicate. Therefore, a detailed understanding of the expressed genes of termites is not only interesting for academic research but is also essential in the development of new termite-specific insecticides [4].

The acquisition of genetic data in termites and their gut microbial community has been of recent interest to the scientific community. This is mostly due to the development and accessibility of new sequencing technologies such as 454 pyrosequencing and Illumina sequencing [4]. To date, (meta-) genomic and (meta-) transcriptomic studies in termites have been principally aimed at identifying host and/or symbiont genes underlying lignocellulose digestion [5–12], caste differentiation [13–17], reproduction [18] or defense [19]. Large-scale EST libraries have also been constructed in a few distantly related termite species belonging to different families for comparative purpose [13]. Despite these efforts and the recent publication of the first two termite genomes [20,21], genetic data are only available for a limited number of termite species and comparative studies remain scarce. The diversity and identity of genes expressed have rarely been compared between closely related species.

Reticulitermes (Rhinotermitidae) represent an important genus of termites with multiple pest species, particularly in temperate regions [22]. They have cryptic nesting habits and form complex colonies with diffuse nests and multiple feeding sites connected by underground tunnels [23]. Termite colonies are typically founded by a single pair of winged reproductives (i.e., the "queen" and the "king") following a nuptial flight. However, reproduction is not always reserved only to the primary couple within colonies. Another type of reproductives can indeed differentiate among the offspring of the primary couple. Such secondary reproductives are called 'nymphoid neotenics' when they differentiate from nymphs and 'ergatoid neotenics' when they differentiate from workers. Although these two types of neotenics are morphologically different, both are wingless and have no pigmentation; they stay in their native colonies where they may replace or supplement the primary couple's reproduction. The presence of productive neotenics within colonies has tremendous genetic and dynamic impacts on colonies [23], and some authors have argued that the acquisition of this caste has played a major role in the evolution of social life [24,25]. In *Reticulitermes* termites, the number of neotenics is extremely variable among and within species, and several studies have argued that a high number of neotenics could improve the capacity of colonies to develop and disperse in urban areas [23,26,27]. Despite their importance, the conditions under which neotenic reproductives differentiate within colonies as well as the molecular mechanisms underlying such a differentiation remain unknown. In comparison with other castes (workers, soldiers, primary reproductives), only a few studies focus on determining the genes expressed in neotenic reproductives of termites [15,28,29].

The present study compares the transcriptome content obtained from 13 nymphoid neotenic females of three *Reticulitermes* species: two West European species, *R. grassei* and *R. lucifugus* [30], and one North American species, *R. flavipes*, that has been introduced to France [26]. Eleven out of the thirteen transcriptomes analyzed in this study (i.e., those obtained in *R. flavipes* and *R. grassei*) were obtained in a previous study [31]. This dataset was used for SNP detection and population genomics inference only [32,33] and no gene content and functional analysis was performed. Using the same protocol, two additional transcriptomes were generated from nymphoid female neotenics collected in two distinct colonies of *R. lucifugus*. BLAST analysis allowed functional and taxonomic assignation of the predicted contigs, assembled from the whole species from pools of individuals and compared between termite species. Functional analysis of contigs associated with wood degradation was performed using Gene Ontology (GO) term analysis and contributed to the characterization of *Reticulitermes* transcriptomes.

Materials and Methods

Termite samples

We sampled 13 termite colonies representative of 3 *Reticulitermes* species: 9 colonies of *R. grassei*, 2 colonies of *R. flavipes* and 2 colonies of *R. lucifugus*. All samples were collected in 2010 in France from 11 locations (Table 1). None of these termite species are endangered or protected, and no specific permission was required for collecting them since they were taken from unprotected areas. Five hundred to 1000 individuals per colony were brought to the laboratory where they were maintained for 90 days under controlled conditions in their original wood piece at 25°C and 80% humidity. For each colony, the transcriptome of a single fecund nymphoid neotenic female was obtained using the protocol described below. The transcriptomes obtained from *R. grassei* and *R. flavipes* females were obtained and assembled in a previous study [31]. In this study, we used a single nymphoid neotenic female isolated from 2 different colonies of *R. lucifugus* for further RNA isolation, cDNA library construction and transcriptome sequencing (see below).

RNA isolation

Total RNA was isolated independently from the whole body of a single neotenic using an adapted protocol using Guanidinium Thiocyanate-Phenol solution supplemented with

Table 1. Information and sequencing results of nymphoid neotenic female termites used in this study.

Identification name	*Reticulitermes* sp.	Locality	Latitude	Longitude	No of illumina reads (x10^6)
GA15A	*R. grassei*	Pissos	44.307069	-0.778946	14.95
GA15B	*R. grassei*	Poitiers	46.580224	0.340375	44.77
GA15C	*R. grassei*	Gatseau	45.826772	-1.208771	27.21
GA15D	*R. grassei*	Gatseau	45.826772	-1.208771	37.83
GA15E	*R. grassei*	Ustaritz	43.395936	-1.45459	22.20
GA15F	*R. grassei*	Eguilles	43.5686589	5.354234	28.57
GA15G	*R. grassei*	Eguilles	43.5686589	5.354234	27.60
GA15H	*R. grassei*	Cugnaux	43.537141	1.344995	24.52
GA15I	*R. grassei*	Petit Bôo	44.353499	-1.042781	23.30
GA15K	*R. flavipes*	Olonne	46.536402	-1.772826	58.86
GA15L	*R. flavipes*	Oléron	45.9159335	-1.2716743	46.98
GA15M	*R. lucifugus*	Allauch	43.336148	5.4825739	43.65
GA15N	*R. lucifugus*	Le Rove	43.369912	5.252998	44.80

doi:10.1371/journal.pone.0145596.t001

glycogen [34]. Quality and quantity of total RNA were determined using agarose gel electrophoresis, NanoDrop spectrophotometry and analysis on Agilent bioanalyzer 2100 system using the Eukaryote Total RNA Nano assay. RNA isolation for the two other species *R. grassei* and *R. flavipes* used in this study was obtained with the same protocol [31].

Transcriptome sequencing

For each sample, 5 μg of total RNA of *R. lucifugus* were used to build 3'-primed, non-normalized cDNA libraries. Although prokaryotic RNA might be present in the libraries since the gut was not removed before extraction, these transcripts were not specifically targeted. Oligo(dT)-primed first-strand synthesis and cap-primed second-strand synthesis were performed using the SMART cDNA library construction kit (Clontech, Mountain View, CA, USA). Libraries were sequenced using Genome Analyzer II (Illumina) with 5 tagged libraries pooled per lane. Fifty bp single-end reads were produced. The cDNA library construction and sequencing were performed by GATC biotech company (Constanz, Germany). After tag-removing, low quality bases, adaptors and primers were removed with SeqClean software (http://compbio.dfci.harvard.edu/tgi/) with default parameters. Reads used for this study originated from a previous work [31] on *R. flavipes* (Sequence Read Archive accession no. SRX565295 and SRX565296) and *R. grassei* (SRA accession no. SRX565297 to SRX565305), and from the present study for *R. lucifugus* (SRA accession no. SRX565306 and SRX565307).

Transcriptome assembly

Transcriptomes were assembled by pooling the reads obtained from individuals belonging to the same species (N = 9 for *R. grassei* and N = 2 for both *R. flavipes* and *R. lucifugus*). Assemblies were performed using ABYSS [35] with Kmer set at 40, followed by two consecutive runs of CAP3 [36] as described in [37]. Contigs shorter than 100 bases were discarded.

ORF detection

Complete and 5'- or 3'-truncated open reading frames (ORF) were detected using Prodigal software for metagenomic data [38] using standard genetic code. ORF with a stretch of N (undetermined nucleotides) inside the sequence were not discarded. When several ORF were detected on the same contig, only the longest was kept since it would more likely correspond to a true protein. The software Cd-hit [39] was used to remove ORF redundancy from our dataset by detecting sequences showing 100% identity in homologous regions.

Orthology prediction

BLAST-based pairwise orthologs relationships between *R. grassei-R. flavipes*, *R. grassei-R. lucifugus*, and *R. flavipes-R. lucifugus* species pairs were first assessed by InParanoid 4.1 program using default parameters [40]. Non-redundant translated contigs of the 3 species displaying length > 100 bases were used for that purpose. MultiParanoid [41] was then used to analyze the orthology relationship of gene clusters between the 3 species.

Taxonomic assignation

BLAST results from the annotation step were parsed to retrieve sequence identifier (GI) and NCBI taxonomic identifier (TaxID) from the NCBI database (ftp://ftp.ncbi.nih.gov/pub/taxonomy). Each contig was assigned to 5 taxonomic ranks (i.e., superkingdom, kingdom, phylum, genus and species). Contigs were assigned to Bacteria when the indicated superkingdom was 'Bacteria'; they were assigned to Termites when the superkingdom was 'Eukaryota' and the

kingdom was 'Metazoa'; they finally were assigned to protists when the superkingdom was 'Eukaryota' and when the kingdom was not 'Fungi, Viridiplantae and/or metazoa'. For protists, downstream taxonomic ranks (phylum, genus and species) were inspected manually to verify that each of them corresponded to a protist taxon.

Functional annotation

Amino acid homologies of the non-redundant predicted ORFs were analyzed using BLASTp program [42] against the Genbank nr database (March 1, 2011). The first BLAST hits were kept and the minimum E-value was set at 0.001. HSP length cut off was 33 and the lower capacity filer was enabled. GI identifiers from BLAST results were used to retrieve UniProt IDs from the PIR Protein Sequence Database, the latter served to associate GO terms with our predicted ORFs. In addition, known protein signatures were detected using the software InterProScan [43] based on InterPro collection databases. Annotation steps were performed with Blast2go software V.2.2.6 [44] using default parameters. GO terms analysis was performed on the three termite species independently. Based on the KEGG metabolism pathways database V.64.0 [45,46] implemented in Blast2go, GO terms of biological processes were used to retrieve contigs belonging to the starch and sucrose metabolism process (map00500). GO results were displayed at GO level 2. A GO level referred to the hierarchical structure of the Gene Ontology as the number of GO terms between a given term and the Root Term of the ontology. Contigs assigned to a function related to the starch and sucrose metabolism process were further analyzed for the presence of hallmarks of Carbohydrate-Active Enzymes (CAZ) domains using CAZymes Analysis Toolkit (CAT) [47] with the CAZy database updated on 09/20/2013 [48]. In this software, the assignment method is based on both similarity search on proteins sequences and the presence of Pfam conserved domains (option 'Pfam rules based annotation').

Quantification of transcript abundance

For each species, a pool of individual reads were mapped to the species transcriptome using BWA [49] with default parameters and served for FPKM (i.e., Fragment Per Kilobase per Millions fragments mapped) calculation using Cufflinks [50]. To avoid biases due to differences in read numbers across individual data sets, a random subsampling of reads was performed before pooling. A total of 14.95, 49.98 and 43.65 million reads were subsampled for each individual of *R. grassei*, *R. flavipes* and *R. lucifugus*, respectively. This number corresponded to the number of reads of the individual displaying the lower number of reads for the species. Only the 250 most abundant contigs per species (with higher FPKM values) with a Blast2GO hit were kept. As previously described, the functional annotation of these contigs was performed using Blast2GO. Preliminary analyses indicated that GO term levels 2 and 3 were not appropriated due to a limited number of contigs producing too few functional categories. The final analysis was therefore conducted at the GO term level 4.

Results and Discussion

Sequencing, assembly and annotation of transcriptomes

The numbers of illumina reads obtained for all of the 11 termite samples are indicated in Table 1. For each termite species, reads from conspecific individuals were pooled together before assembly (N = 9 for *R. grassei*, N = 2 for *R. flavipes* and *R. lucifugus*). The obtained species-level assemblies were comparable among species and exhibited high N50 values (Table 2). In total, 64,328, 65,814 and 79,404 non-redundant ORF were predicted in *R. flavipes*, *R.*

Table 2. Assemblies of transcriptomes, ORF predictions and functional annotations for interspecific comparisons and Gene Ontology analyses. Contigs < 100 bases were removed from analyses.

Species	*R. flavipes*	*R. grassei*	*R. lucifugus*
No of contigs	93,420	111,549	91,280
N50	998	1,041	1,067
No. of predicted ORF	64,342	79,640	65,855
Non-redundant ORF	64,328	79,404	65,814
No of ORF with BLASTp hit	19,375 (30.1%)	21,671 (27.3%)	20,377 (31%)
No of annotated ORF (GO term)	5,389 (8.4%)	4,922 (6.2%)	5,214 (7.9%)

doi:10.1371/journal.pone.0145596.t002

lucifugus and *R. grassei*, respectively. As expected with non-model organisms lacking complete and well-annotated genome sequences, only a third of ORF showed a significant BLAST hit in the nr protein database. After the functional annotation performed with Blast2Go program, only a small fraction (6–8%) of the initial ORF set could be assigned with one or more GO terms (Table 2). Whatever the number of individual reads sets pooled together (2 or 9), the number and quality of the obtained contigs were similar among species, suggesting that pooling reads from 2 individuals produced satisfactory results.

Orthology relationships

Gene clusters were identified within each species from transcript data. The presence of homologous clusters between species (i.e. orthologues) was assessed to better understand the genetic relationships between the 3 termite species. For each species, transcript clusters were composed of representative unigenes (i.e. alternative transcripts derived from a single locus) or transcripts derived from young duplicated genes (i.e. paralogues). Interspecific comparison showed that a large number (18,323) of orthologous genes clusters (i.e. homologous genes clusters identified in other species) were detected in the 3 species (Fig 1). Orthologous genes were more abundant between *R. grassei* and *R. lucifugus* (8,605) than between the two other pairs of species (6,737 between *R. grassei* and *R. flavipes*, and 6,431 between *R. lucifugus* and *R. grassei*). This result is in accordance with the phylogenetic relationships between these termite species [30]. The North American species, *R. flavipes*, is indeed distantly related to the two European species, *R. grassei* and *R. lucifugus*, which are closely related and probably sister species. Transcripts of gut microbiota may also reinforce this phylogenetic relationship between the taxa, albeit to a lesser extent since they do not contribute much in terms of contig numbers (see next paragraph). This hypothesis is supported by recent studies showing how microbial communities are usually more similar between closely related species than distant species [51–54].

Taxonomic assignment

Table 3 shows the taxonomic distribution of the contigs displaying a significant BLAST hit of their coding sequence against the nr protein database. In total, 41,287 contigs were assigned. Most of them (94.9%) were assigned to the termite genome, whereas the remaining contigs were assigned to diverse lineages of microorganisms (i.e. protists, bacteria, archeae, virus) (3.9%), fungi (0.9%) or plants (0.4%). Contigs assigned to fungi and plants most likely represent environmental contaminations since no endosymbiotic association has been described so far between these organisms and *Reticulitermes* termites. Most contigs assigned to microorganisms were probably expressed by diverse microorganisms living in the hindgut of *Reticulitermes* termites. These well-known microbial communities are composed of protists (two main lineages: Parabasalia and Oxymonadida), Bacteria (the most abundant: Spirochaetes,

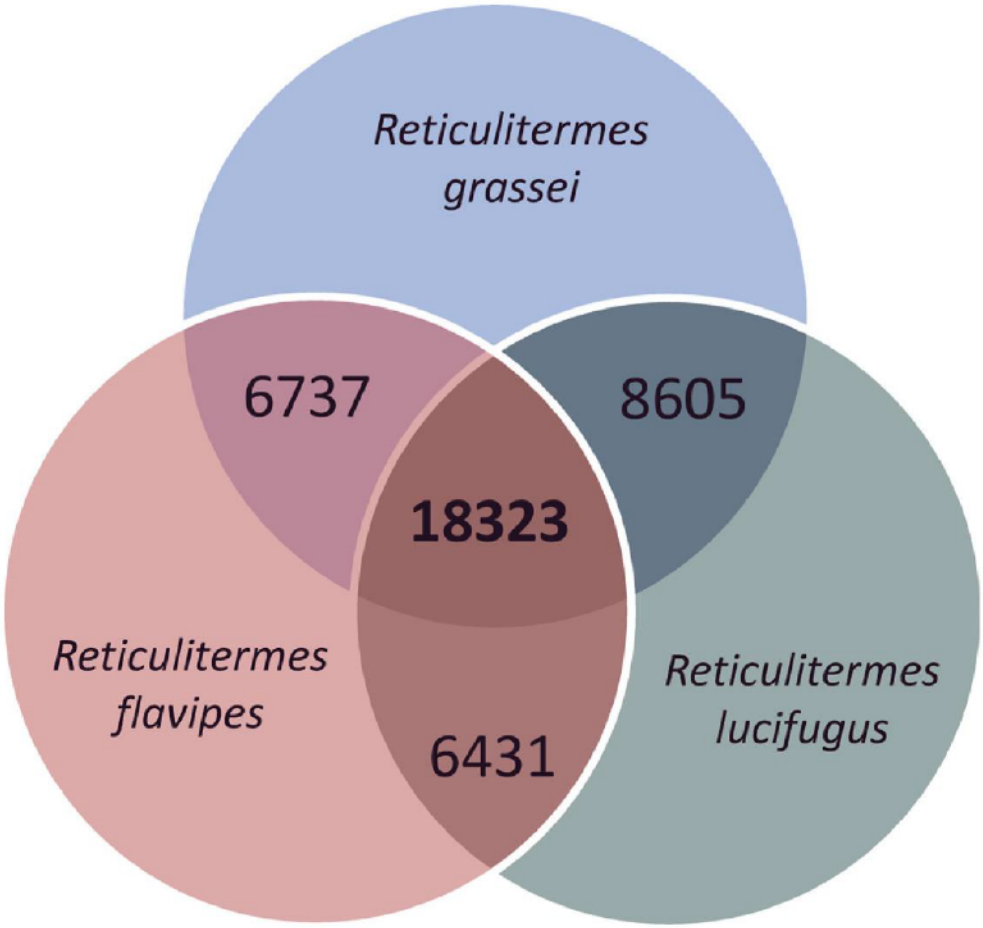


Fig 1. Orthology relationships between *R. grassei*, *R. flavipes* and *R. lucifugus* contigs. The number of orthologous gene clusters is indicated inside Venn diagram.

doi:10.1371/journal.pone.0145596.g001

Bacteroidetes, Firmicutes and Elusimicrobia), methanogenic Archeae (Methanobacteriaceae family) and bacteriophage virus infecting Spirochaetes [3,55,56].

Several reasons could explain the low proportion of microbial contigs in our dataset. First, contigs assigned to prokaryotes were particularly scarce (< 1% for the 3 termite species), probably because the cDNA library protocol underwent an enrichment of mRNA based on the existence of poly-A tails, which are mostly absent in prokaryotic transcripts. Second, as expected with non-model organisms, which lack complete and well-annotated genome sequences, expressed genes in gut microbial communities of termites have not been fully characterized yet. Therefore, the genomic database is likely incomplete and could thus result in an underestimation of microbial genes. Third, all transcriptomes were generated from nymphoid neotenic females. Like other castes or developmental stages in subterranean termites, reproductive castes do not necessarily feed on the wood themselves and instead are fed by workers who provide them with nutrient-rich salivary trophallactic transfers [57]. Since the gut microbiota might be not essential for extracting nutrients from wood in these secondary reproductives,

microorganisms could be less abundant in the hindgut of reproductives compared to that of wood-feeding castes. This hypothesis is supported by previous work in primary reproductives (alates) in 3 species of *Reticulitermes* [58], as well as even earlier work in *R. flavipes* [15] which suggests a reduced microbiota in neotenic reproductives.

Description of gene ontologies of expressed genes

The function of assembled contigs was evaluated by retrieving GO terms according to their termite, protist or bacterial origin. The 3 descriptive ontologies 'cellular components', 'molecular function' and 'biological process' were analyzed. For the cellular components ontology, 10 localizations were found. Transcripts products mainly localized in cell (37–38% depending on termite species), membrane (24–27%), organelle (19–20%) and macromolecular complex (11–13%) (Fig 2). Twenty-one classes of biological processes were found with 2 dominant classes corresponding to metabolic (30–32%) and cellular processes (31–32%) (Fig 3). Finally, 13 GO terms were found in 'molecular function', among which putative catalytic activity (40–42%) and binding (40%) were the most abundant (Fig 4). A very similar distribution of GO terms for the 3 types of ontologies was observed between the 3 termite species, suggesting unbiased transcriptome assembly. The relative taxonomic distribution (termite, protist, bacteria) was consistent among the 3 ontologies and with the total number of contigs (Table 3). GO analysis showed that protists accounted for a significant part of cellular components (mainly cell, organelle and macromolecular complex) of biological processes (metabolic and cellular process) and molecular functions (catalytic activity, binding and structural molecule activity).

The putative biological functions of the 250 most expressed contigs were studied for each of the 3 termite species (Fig 5). These 750 contigs corresponded to highly expressed transcripts, having a FPKM value ranging from 5,940 in *R. grassei* to 3,448,530 in *R. lucifugus*. Eighty-three functional categories were assigned in total among the 3 termite species. The 31 functional categories displaying the most numerous contigs were related to metabolic process whereas the 52 remaining categories corresponded to diverse other biological functions. This result appeared consistent with the global analysis of transcriptomes (Fig 4), and suggests that the most expressed genes have metabolism-related functions.

Contigs putatively associated with wood-degradation enzymes

Transcriptome annotation was used to retrieve contigs associated with functions belonging to the starch and sucrose metabolism. Some of these may be involved in cellulose degradation

Table 3. Taxonomic assignation of contigs obtained in the three termite species.

Species		*R. flavipes*	*R. grassei*	*R. lucifugus*	Total
No. of transcriptomes		2	9	2	13
No. of contigs (%)	All	13,889 (100.00)	14,453 (100.00)	12,945 (100.00)	41,287 (100)
	Termites	12,806 (92.11)	13,620 (94.05)	12,798 (94.05)	39,224 (94.87)
	Microorganisms [a]	1,006 (7.23)	438 (3.02)	95 (0.73)	1,539 (3.92)
	Protists	893 (6.42)	300 (2.07)	39 (0.30)	1,232 (3.14)
	Bacteria	101 (0.72)	120 (0.82)	46 (0.35)	267 (0.68)
	Archeae	6 (0.04)	9 (0.62)	3 (0.02)	18 (0.04)
	Viruses	6 (0.04)	9 (0.62)	7 (0.05)	22 (0.05)
	Fungi	38 (0.28)	302 (2.08)	26 (0.20)	367 (0.93)
	Viridiplantae	38 (0.27)	93 (0.64)	26 (0.20)	157 (0.40)

[a]Contigs of microorganisms include those of Protists, Bacteria, Archeae and Viruses.

doi:10.1371/journal.pone.0145596.t003

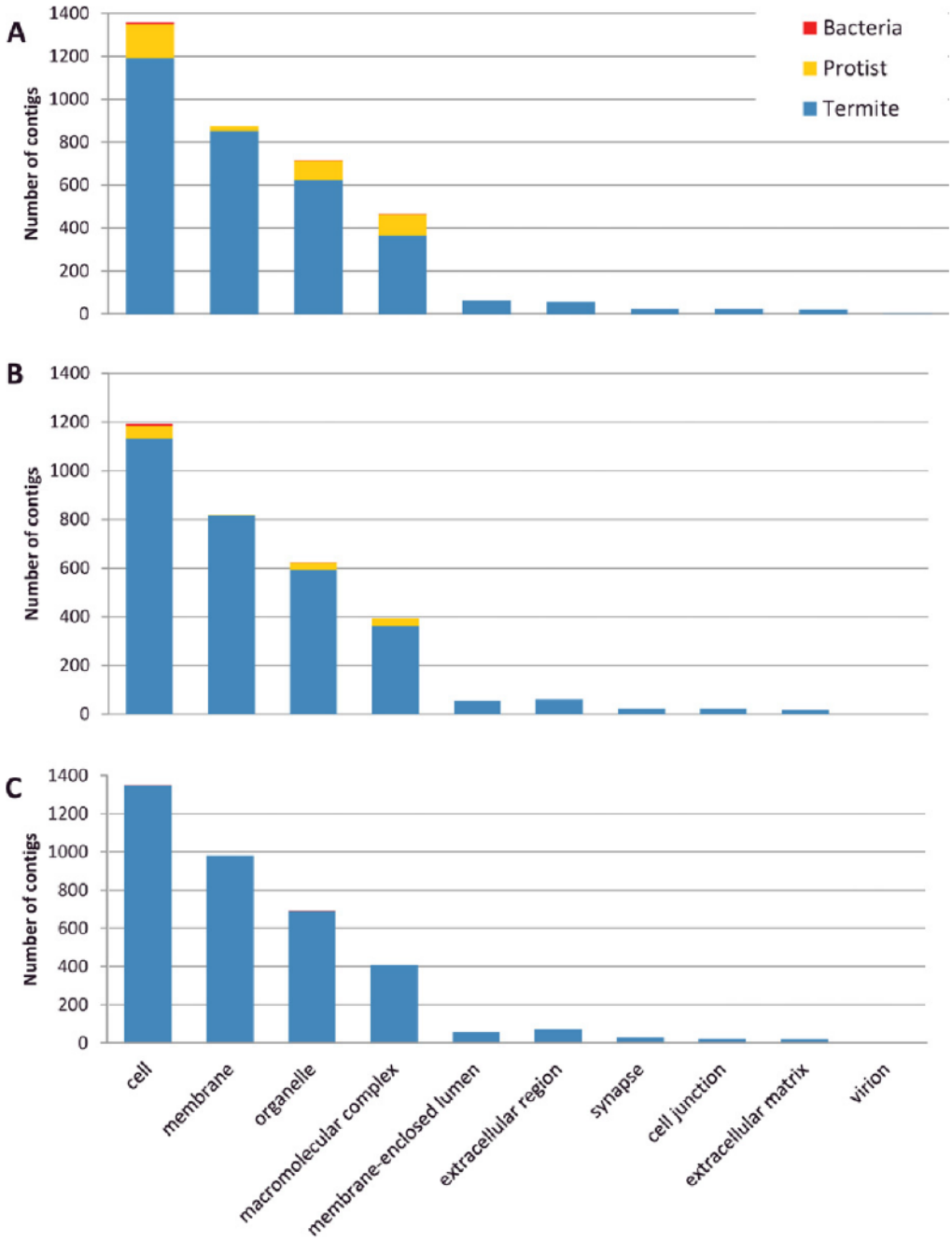


Fig 2. Gene Ontology analysis of Cellular Components (GO level 2). A: *R. flavipes*. B: *R. grassei*, C: *R. lucifugus*.

doi:10.1371/journal.pone.0145596.g002

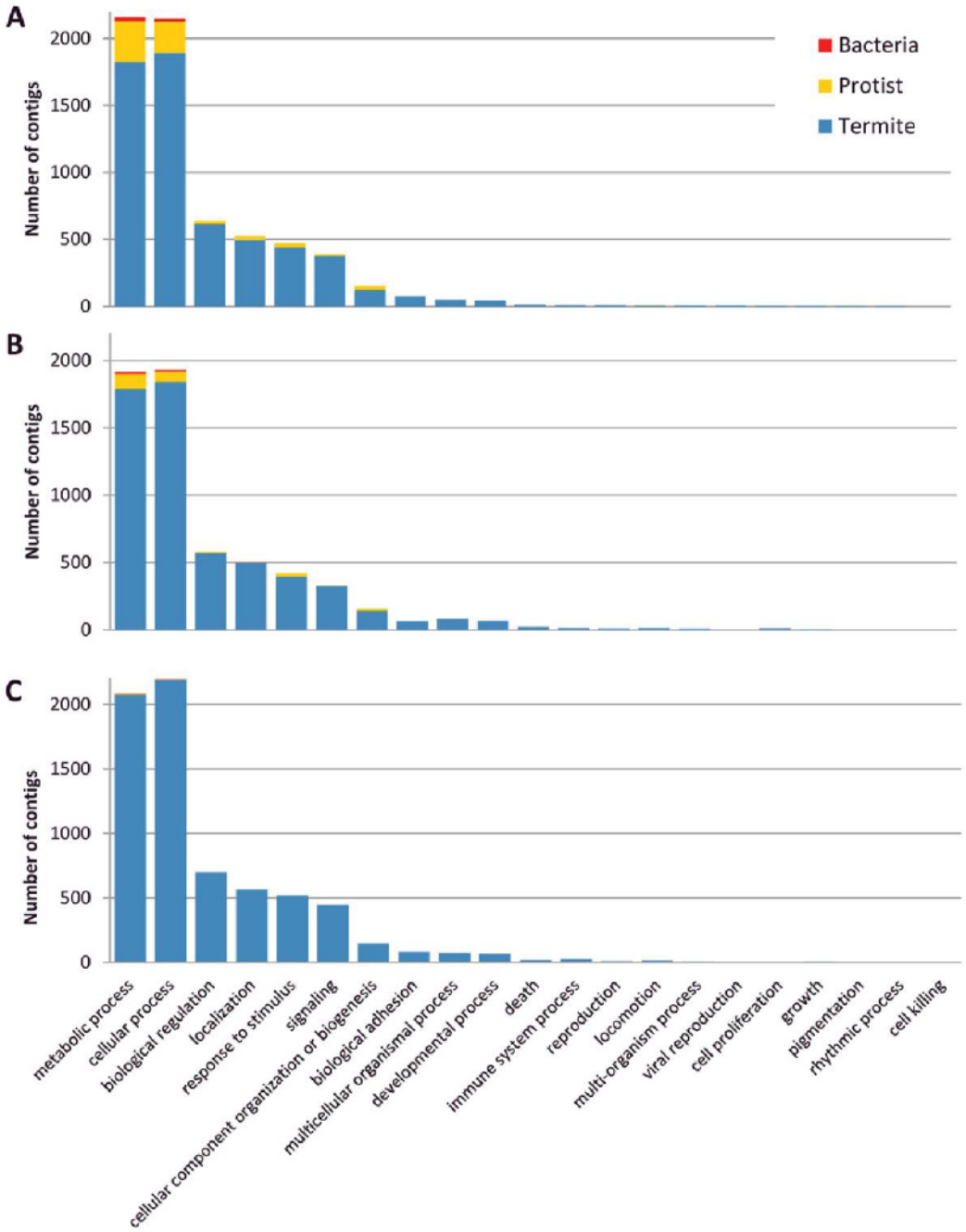


Fig 3. Gene Ontology analysis of Biological Processes (GO level 2). A: *R. flavipes*. B: *R. grassei*, C: *R. lucifugus*.

doi:10.1371/journal.pone.0145596.g003

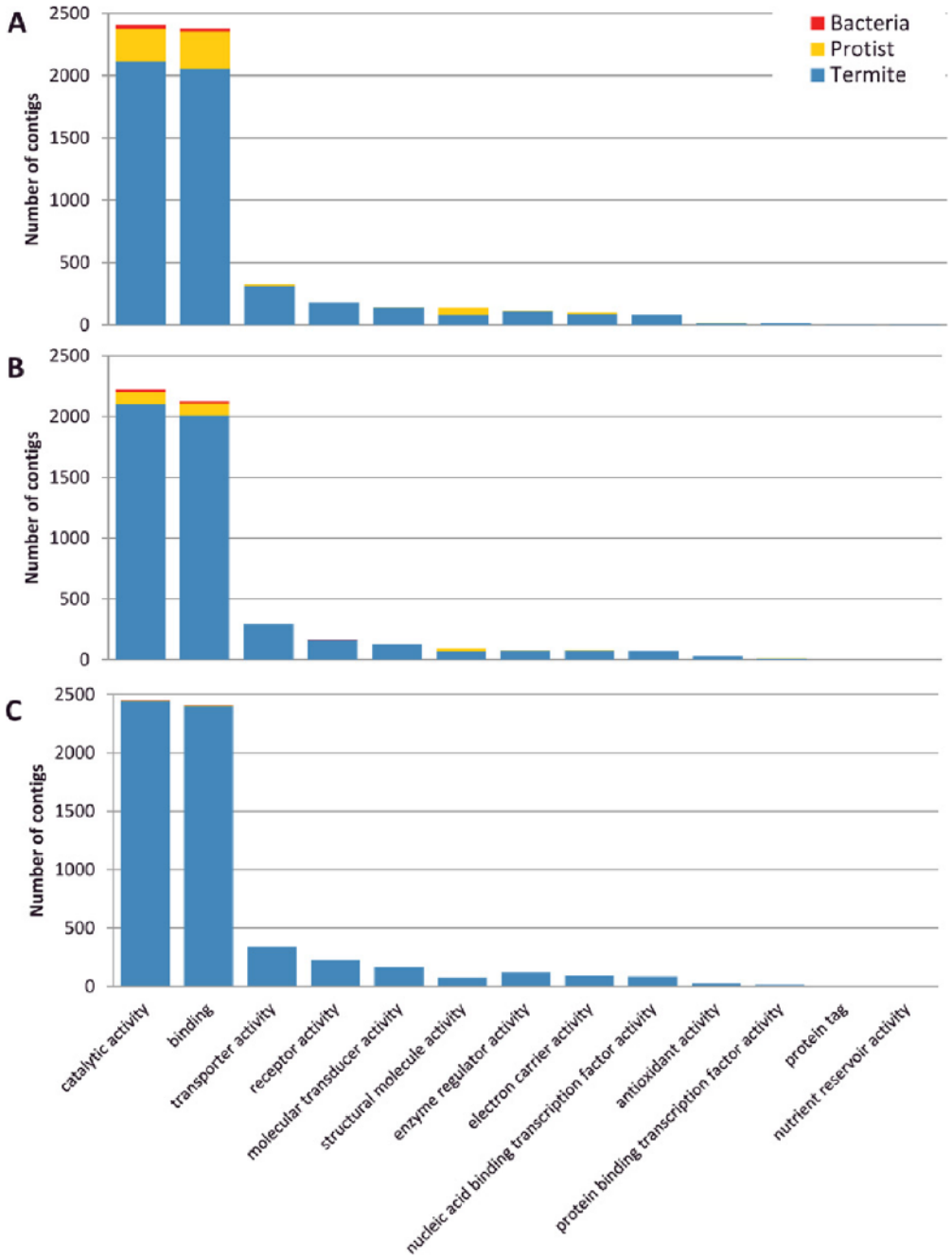


Fig 4. Gene Ontology analysis of Molecular Functions (GO level 2). A: *R. flavipes*. B: *R. grassei*, C: *R. lucifugus*.

doi:10.1371/journal.pone.0145596.g004

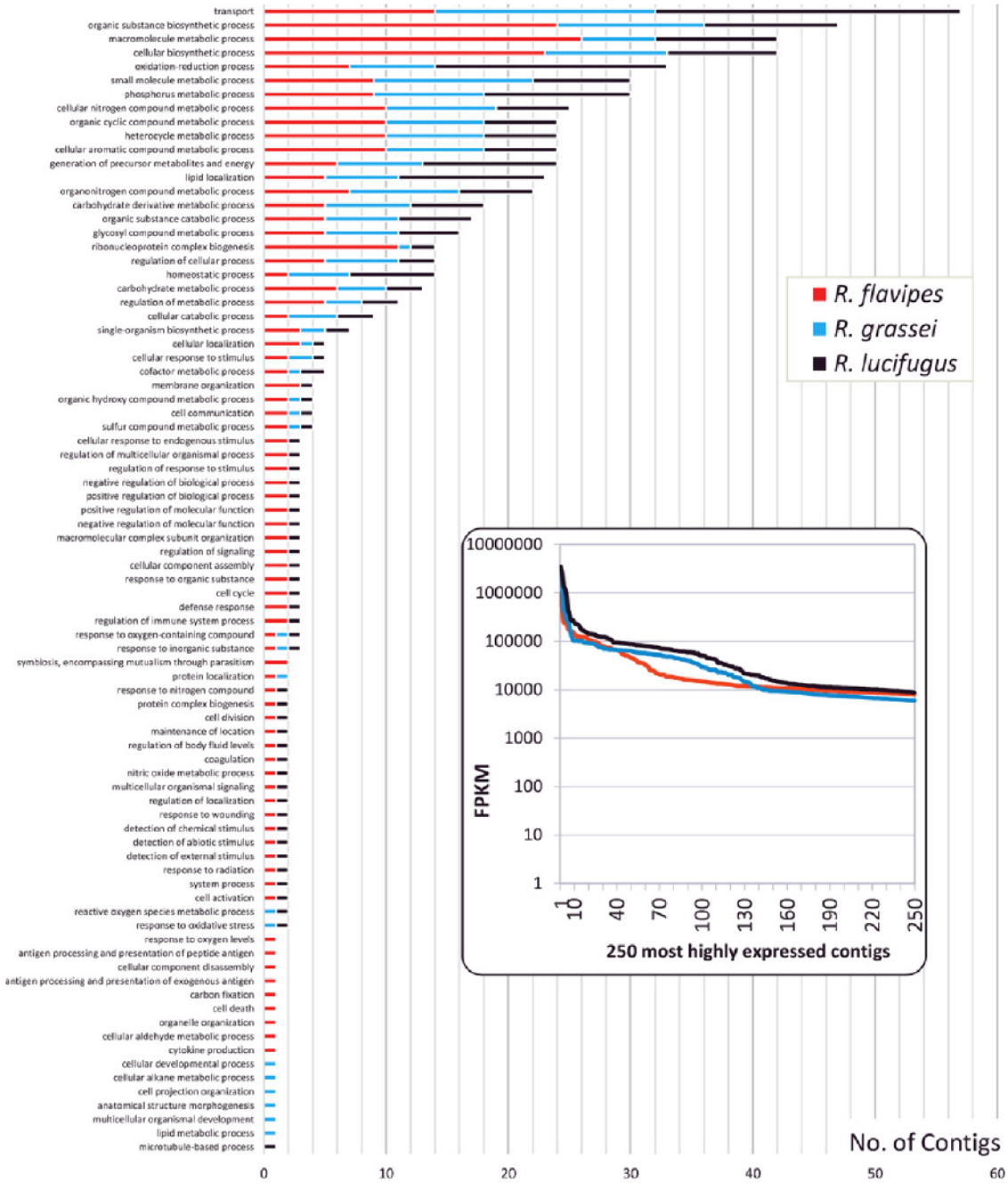


Fig 5. Distribution of the 250 most expressed contigs across their biological functions detected by GO term analysis at the GO level 4. The relative abundance of contigs estimated by FPKM is shown in the box.

doi:10.1371/journal.pone.0145596.g005

[59]. The complete list of contigs displaying functions related to starch and sucrose metabolism and their amino acid sequences are shown in supplementary information (S1 Table and S1 Dataset). Fig 6 shows the enzymatic processes detected in our dataset and plots them on starch and sucrose metabolism map. A total of 11 functions putatively associated with known enzymatic activities were found in the assemblies. Among these functions, we found contigs putatively associated with cellulytic activities. Cellulase is a general term for cellulytic enzymes, of which three main classes are recognized on the basis of the mode of enzymatic actions and substrate specificities: endoglucanases (EGs; EC 3.2.1.4), cellobiohydrolases (CBHs; EC 3.2.1.91) and β-glucosidases (BGs; EC 3.2.1.21). These three categories of enzymes work synergistically to efficiently degrade chains of cellulose. Whereas EGs and BGs are quite common in microorganisms, animals and plants, CBHs are apparently more rare and appear to be restricted to bacteria, fungi and protists [59,60]. In our dataset, 9 contigs were assigned to putative EGs and 11 contigs were assigned to putative BGs (Fig 6). However, we found no evidence for the presence of genes encoding CBHs.

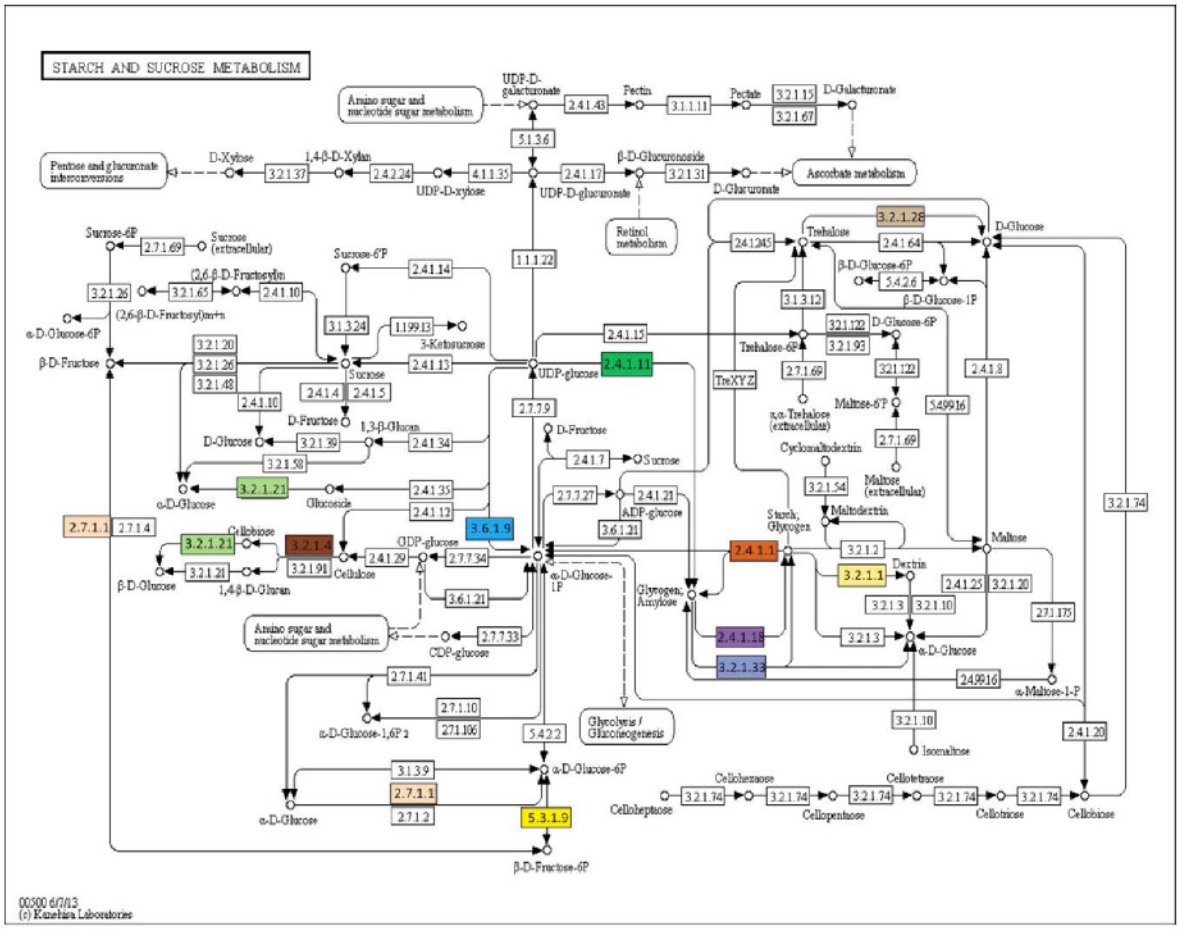
Using CAZymes Analysis Toolkit [47], 9 families of putative genes of Carbohydrate-Active Enzymes (CAZy) were detected in the 3 transcriptomes (Table 4): 5 glycoside hydrolases (GH1, GH9, GH13, GH37, GH45), 2 glycoside transferases (i.e., GT3, GT35) and 2 carbohydrate-binding modules families (i.e., CBM6, CBM48). Genes of a same GH family are usually considered to share not only structural motifs and the catalytic machinery, but also an evolutionary origin [61]. Among the putative GH genes detected in our analyses, GH1 represents a single family of BGs, whereas GH9 and GH45 are two families of EGs. Previous studies show GH1 and GH9 to be mostly expressed by the genome of termites either in their salivary glands and/or in the hindgut [62]. However, genes encoding GH45 can be expressed by both the hosts and their symbiotic protists or prokaryotes living in the hindgut. In addition, CBMs are usually considered to be expressed by microorganisms only (Watanabe & Tokuda 2010). Therefore, the detection of putative genes encoding GH45 and CBM in our dataset suggests that the gut microbial community of neotenics may play a role in the synthesis of the enzymatic system involved in the degradation of cellulose.

Interspecific variation of microbial gene expression patterns

The proportion of microbial contigs varies among the three *Reticulitermes* species (Table 3). This variation is particularly evident in the contigs assigned to protists. Representing only 0.3% of the sequences in *R. lucifugus*, protists contigs were more abundant in *R. grassei* (2.1%) and in *R. flavipes* (6.4%). This general pattern has been found in the functional analysis also since the most important proportion of microbial contigs was assigned in *R. flavipes* transcriptome followed by those of *R. grassei* and *R. lucifugus* (Figs 2–4). We cannot exclude that a part of this observed variation results from methodological fluctuations in mRNA isolation, cDNA library construction and sequencing. However, this pattern might also result from variation of gene expression patterns among individual neotenics of different species, either due to transcriptomic noise or associated to a biological function. In any case, our results suggest that gut microbial communities are not totally absent from *Reticulitermes* neotenics, in spite of their feeding lifestyle, which probably does not directly involve lignocellulose digestion. Abundance, role and regulation mechanisms of gut microbial communities in reproductive termites will require further investigations.

Conclusion

Comparison of a set of assembled transcriptomes of nymphoid neotenic reproductives was performed from 13 colonies belonging to 3 related termite species based on high throughput



		R. flavipes	*R. grassei*	*R. lucifugus*
ec:2.4.1.1	phosphorylase	(14)	(4)	(4)
ec:2.4.1.11	glycogen(starch) synthase	(2)	(1)	(1)
ec:2.4.1.18	1,4-alpha-glucan branching enzyme	(2)	(0)	(1)
ec:2.7.1.1	hexokinase	(4)	(4)	(2)
ec:3.2.1.1	alpha-amylase	(1)	(0)	(0)
ec:3.2.1.21	beta-glucosidase	(7)	(1)	(3)
ec:3.2.1.28	alpha,alpha-trehalase	(2)	(2)	(4)
ec:3.2.1.33	amylo-alpha-1,6-glucosidase	(1)	(0)	(1)
ec:3.2.1.4	cellulase	(4)	(3)	(2)
ec:3.6.1.9	nucleotide diphosphatase	(1)	(1)	(0)
ec:5.3.1.9	glucose-6-phosphate isomerase	(3)	(2)	(2)

Fig 6. Contigs of *R. flavipes*, *R. grassei* and *R. lucifugus* displaying putative enzymatic activities involved in starch and sucrose metabolism. Top panel: termite contigs (colored boxes) mapped on the starch and sucrose metabolism KEGG map (black and white boxes). Bottom panel: number of contigs (brackets) associated to the 11 putative functions of starch and sucrose metabolism.

doi:10.1371/journal.pone.0145596.g006

Table 4. Number of Carbohydrate-Active Enzyme (CAZy) families detected in *Reticulitermes* transcriptomes.

CAZy Families	Pfam domains	Domain descriptions	Putative enzymatic activities	*R. flavipes*	*R. grassei*	*R. lucifugus*
CBM48 /GH13	CBM_48/Alpha-amylase/ Alpha-amylase_C	Carbohydrate-binding module 48 (Isoamylase N-terminal domain)	-	1	0	1
CBM6	Phosphodiest	Type I phosphodiesterase / nucleotide pyrophosphatase	-	1	1	0
GH1	Glyco_hydro_1	Glycosyl hydrolase family 1	3.2.1.21	4	0	1
GH13	Alpha-amylase	Alpha amylase, catalytic domain	3.4.1.183.2.1.1	1	0	0
GH37	Trehalase	Trehalase	3.2.1.28	2	2	3
GH45	Glyco_hydro_45	Glycosyl hydrolase family 45	3.2.1.4	1	1	0
GH9	Glyco_hydro_9	Glycosyl hydrolase family 9	3.2.1.4.	1	1	2
GT3	Glycogen_syn	Glycogen synthase	2.4.1.11	2	1	1
GT35	Phosphorylase	Carbohydrate phosphorylase	2.4.1.1	11	3	3

doi:10.1371/journal.pone.0145596.t004

Illumina sequencing. Intraspecific variation was addressed by pooling two to nine individuals per species. As expected with non-model organisms, a large fraction of contigs had no detectable homologs in the public database. The majority of recovered transcripts had a termite origin, although transcripts from microorganisms provided evidence for the presence of an active gut microbiome in this non-wood feeding life stage. These transcripts were indeed over-represented in starch and sucrose metabolism pathways, and some of them are likely to encode enzymes involved in cellulose degradation.

Supporting Information

S1 Table. Contigs putatively involved in enzymatic activities linked to the starch and sucrose metabolism pathway and detected by GO terms analysis.
(DOCX)

S1 Dataset. Amino acid sequence of contigs putatively involved in enzymatic activities linked to the starch and sucrose metabolism pathway and detected by GO terms analysis in FASTA format.
(DOCX)

Acknowledgments

We would like to thank Simon Dupont and Sylvain Guyot for termite sampling, Vincent Ranwez for help with the NCBI taxonomy database and M. Rivera for English revisions. Analyses largely benefited from the ISEM computing cluster platform with the help of Khalid Belkhir. We are also grateful to the Genotoul bioinformatics platform Toulouse Midi-Pyrenees for providing computing and storage resources. This does not alter our adherence to PLOS ONE policies on sharing data and Materials.

Author Contributions

Conceived and designed the experiments: FD NG PG. Performed the experiments: FD LW MB. Analyzed the data: FD DB VC TJ PG. Wrote the paper: FD PG.

References

1. Inward DJ, Vogler AP, Eggleton P (2007) A comprehensive phylogenetic analysis of termites (Isoptera) illuminates key aspects of their evolutionary biology. Mol Phylogenet Evol 44: 953–967. PMID: 17625919
2. Bignell DE, Roisin Y, Lo N (2011) Biology of termites: a modern synthesis: Springer.
3. Brune A (2014) Symbiotic digestion of lignocellulose in termite guts. Nat Rev Microbiol.
4. Scharf ME (2015) Omic research in termites: an overview and a roadmap. Front Genet 6: 76. doi: 10.3389/fgene.2015.00076 PMID: 25821456
5. He S, Ivanova N, Kirton E, Allgaier M, Bergin C, Scheffrahn RH, et al. (2013) Comparative metagenomic and metatranscriptomic analysis of hindgut paunch microbiota in wood- and dung-feeding higher termites. PLoS ONE 8: e61126. doi: 10.1371/journal.pone.0061126 PMID: 23593407
6. Raychoudhury R, Sen R, Cai Y, Sun Y, Lietze VU, Boucias DG, et al. (2013) Comparative metatranscriptomic signatures of wood and paper feeding in the gut of the termite *Reticulitermes flavipes* (Isoptera: Rhinotermitidae). Insect Mol Biol 22: 155–171. doi: 10.1111/imb.12011 PMID: 23294456
7. Tartar A, Wheeler MM, Zhou X, Coy MR, Boucias DG, Scharf ME. (2009) Parallel metatranscriptome analyses of host and symbiont gene expression in the gut of the termite *Reticulitermes flavipes*. Biotechnol Biofuels 2: 25. doi: 10.1186/1754-6834-2-25 PMID: 19832970
8. Todaka N, Inoue T, Saita K, Ohkuma M, Nalepa CA, Lenz M, et al. (2010) Phylogenetic analysis of cellulolytic enzyme genes from representative lineages of termites and a related cockroach. PLoS ONE 5: e8636. doi: 10.1371/journal.pone.0008636 PMID: 20072608
9. Warnecke F, Luginbuhl P, Ivanova N, Ghassemian M, Richardson TH, Stege JT, et al. (2007) Metagenomic and functional analysis of hindgut microbiota of a wood-feeding higher termite. Nature 450: 560–565. PMID: 18033299
10. Xie L, Zhang L, Zhong Y, Liu N, Long YH, Wang SY, et al. (2012) Profiling the metatranscriptome of the protistan community in *Coptotermes formosanus* with emphasis on the lignocellulolytic system. Genomics 99: 246–255. doi: 10.1016/j.ygeno.2012.01.009 PMID: 22326742
11. Zhang D, Lax AR, Henrissat B, Coutinho P, Katiya N, Nierman WC, et al. (2012) Carbohydrate-active enzymes revealed in *Coptotermes formosanus* (Isoptera: Rhinotermitidae) transcriptome. Insect Mol Biol 21: 235–245. doi: 10.1111/j.1365-2583.2011.01130.x PMID: 22243654
12. Bastien G, Arnal G, Bozonnet S, Laguerre S, Ferreira F, Faure R, et al. (2013) Mining for hemicellulases in the fungus-growing termite *Pseudacanthotermes militaris* using functional metagenomics. Biotechnol Biofuels 6.
13. Hayashi Y, Shigenobu S, Watanabe D, Toga K, Saiki R, Shimada K, et al. (2013) Construction and characterization of normalized cDNA libraries by 454 pyrosequencing and estimation of DNA methylation levels in three distantly related termite species. PLoS ONE 8.
14. Huang Q, Sun P, Zhou X, Lei C (2012) Characterization of head transcriptome and analysis of gene expression involved in caste differentiation and aggression in *Odontotermes formosanus* (Shiraki). PLoS ONE 7: e50383. doi: 10.1371/journal.pone.0050383 PMID: 23209730
15. Scharf ME, Wu-Scharf D, Zhou X, Pittendrigh BR, Bennett GW (2005) Gene expression profiles among immature and adult reproductive castes of the termite *Reticulitermes flavipes*. Insect Mol Biol 14: 31–44. PMID: 15663773
16. Sen R, Raychoudhury R, Cai YP, Sun YJ, Lietze VU, Boucias DG, et al. (2013) Differential impacts of juvenile hormone, soldier head extract and alternate caste phenotypes on host and symbiont transcriptome composition in the gut of the termite *Reticulitermes flavipes*. BMC Genomics 14.
17. Steller MM, Kambhampati S, Caragea D (2010) Comparative analysis of expressed sequence tags from three castes and two life stages of the termite *Reticulitermes flavipes*. BMC Genomics 11: 463. doi: 10.1186/1471-2164-11-463 PMID: 20691076
18. Husseneder C, McGregor C, Lang RP, Collier R, Delatte J (2012) Transcriptome profiling of female alates and egg-laying queens of the Formosan subterranean termite. Comp Biochem Physiol Part D Genomics Proteomics 7: 14–27. doi: 10.1016/j.cbd.2011.10.002 PMID: 22079412
19. Hojo M, Maekawa K, Saitoh S, Shigenobu S, Miura T, Hayashi Y, et al. (2012) Exploration and characterization of genes involved in the synthesis of diterpene defence secretion in *Nasute* termite soldiers. Insect Mol Biol 21: 545–557. doi: 10.1111/j.1365-2583.2012.01162.x PMID: 22984844
20. Terrapon N, Li C, Robertson HM, Ji L, Meng X, Booth W, et al. (2014) Molecular traces of alternative social organization in a termite genome. Nat Commun 5: 3636. doi: 10.1038/ncomms4636 PMID: 24845553

21. Korb J, Poulsen M, Hu H, Li C, Boomsma JJ, Zhang G, et al. (2015) A genomic comparison of two termites with different social complexity. Front Genet 6: 9. doi: 10.3389/fgene.2015.00009 PMID: 25788900

22. Su NY (2002) Novel technologies for subterranean termite control. Sociobiology 40: 95–101.

23. Vargo EL, Husseneder C (2009) Biology of subterranean termites: insights from molecular studies of *Reticulitermes* and *Coptotermes*. Annu Rev Entomol 54: 379–403. doi: 10.1146/annurev.ento.54.110807.090443 PMID: 18793101

24. Howard K, Thorne B (2011) Eusocial evolution in Termites and Hymenoptera. In: Bignell DE, Roisin Y, Lo N, editor. Biology of termites: a modern synthesis: Springer. pp. 97–132.

25. Korb J, Hartfelder K (2008) Life history and development—a framework for understanding developmental plasticity in lower termites. Biol Rev 83: 295–313. PMID: 18979593

26. Perdereau E, Bagneres AG, Bankhead-Dronnet S, Dupont S, Zimmermann M, et al. (2013) Global genetic analysis reveals the putative native source of the invasive termite, *Reticulitermes flavipes*, in France. Mol Ecol 22: 1105–1119. doi: 10.1111/mec.12140 PMID: 23205642

27. Perdereau E, Bagneres AG, Vargo EL, Baudouin G, Xu Y, Labadie P, et al. (2015) Relationship between invasion success and colony breeding structure in a subterranean termite. Mol Ecol 24: 2125–2142. doi: 10.1111/mec.13094 PMID: 25641360

28. Weil T, Korb J, Rehli M (2009) Comparison of queen-specific gene expression in related lower termite species. Mol Biol Evol 26: 1841–1850. doi: 10.1093/molbev/msp095 PMID: 19541881

29. Weil T, Rehli M, Korb J (2007) Molecular basis for the reproductive division of labour in a lower termite. BMC Genomics 8: 198. PMID: 17598892

30. Clement JL, Bagneres AG, Uva P, Wilfert L, Quintana A, Reinhard J, et al. (2001) Biosystematics of *Reticulitermes* termites in Europe: morphological, chemical and molecular data. Insect Soc 48: 202–215.

31. Gayral P, Melo-Ferreira J, Glemin S, Bierne N, Carneiro M, Nabholz B, et al. (2013) Reference-free population genomics from next-generation transcriptome data and the vertebrate-invertebrate gap. PLoS Genet 9: e1003457. doi: 10.1371/journal.pgen.1003457 PMID: 23593039

32. Romiguier J, Gayral P, Ballenghien M, Bernard A, Cahais V, Chenuil A, et al. (2014) Comparative population genomics in animals uncovers the determinants of genetic diversity. Nature 515: 261–263. doi: 10.1038/nature13685 PMID: 25141177

33. Romiguier J, Lourenco J, Gayral P, Faivre N, Weinert LA, Ravel S, et al. (2014) Population genomics of eusocial insects: the costs of a vertebrate-like effective population size. J Evol Biol 27: 593–603. doi: 10.1111/jeb.12331 PMID: 26227898

34. Gayral P, Weinert L, Chiari Y, Tsagkogeorga G, Ballenghien M, Galtier N (2011) Next-generation sequencing of transcriptomes: a guide to RNA isolation in nonmodel animals. Mol Ecol Ressour: 650–661.

35. Simpson JT, Wong K, Jackman SD, Schein JE, Jones SJ, Birol I (2009) ABySS: a parallel assembler for short read sequence data. Genome Res 19: 1117–1123. doi: 10.1101/gr.089532.108 PMID: 19251739

36. Huang X, Madan A (1999) CAP3: A DNA sequence assembly program. Genome Res 9: 868–877. PMID: 10508846

37. Cahais V, Gayral P, Tsagkogeorga G, Melo-Ferreira J, Ballenghien M, Weinert L, et al. (2012) Reference-free transcriptome assembly in non-model animals from next-generation sequencing data. Mol Ecol Ressour 12: 834–845.

38. Hyatt D, LoCascio PF, Hauser LJ, Uberbacher EC (2012) Gene and translation initiation site prediction in metagenomic sequences. Bioinformatics 28: 2223–2230. doi: 10.1093/bioinformatics/bts429 PMID: 22796954

39. Li W, Godzik A (2006) Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics 22: 1658–1659. PMID: 16731699

40. Östlund G, Schmitt T, Forslund K, Köstler T, Messina DN, Roopra S, et al. (2010) InParanoid 7: new algorithms and tools for eukaryotic orthology analysis. Nucleic Acids Res 38: D196–D203. doi: 10.1093/nar/gkp931 PMID: 19892828

41. Alexeyenko A, Tamas I, Liu G, Sonnhammer ELL (2006) Automatic clustering of orthologs and inparalogs shared by multiple proteomes. Bioinformatics 22: e9–e15. PMID: 16873526

42. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. J Mol Biol 215: 403–410. PMID: 2231712

43. Zdobnov EM, Apweiler R (2001) InterProScan–an integration platform for the signature-recognition methods in InterPro. Bioinformatics 17: 847–848. PMID: 11590104

44. Gotz S, Garcia-Gomez JM, Terol J, Williams TD, Nagaraj SH, Nueda MJ, et al. (2008) High-throughput functional annotation and data mining with the Blast2GO suite. Nucleic Acids Res 36: 3420–3435. doi: 10.1093/nar/gkn176 PMID: 18445632

45. Kanehisa M, Goto S (2000) KEGG: kyoto encyclopedia of genes and genomes. Nucleic Acids Res 28: 27–30. PMID: 10592173

46. Kanehisa M, Goto S, Sato Y, Furumichi M, Tanabe M (2012) KEGG for integration and interpretation of large-scale molecular data sets. Nucleic Acids Res 40: D109–114. doi: 10.1093/nar/gkr988 PMID: 22080510

47. Park BH, Karpinets TV, Syed MH, Leuze MR, Uberbacher EC (2010) CAZymes Analysis Toolkit (CAT): web service for searching and analyzing carbohydrate-active enzymes in a newly sequenced organism using CAZy database. Glycobiology 20: 1574–1584. doi: 10.1093/glycob/cwq106 PMID: 20696711

48. Lombard V, Golaconda Ramulu H, Drula E, Coutinho PM, Henrissat B (2014) The carbohydrate-active enzymes database (CAZy) in 2013. Nucleic Acids Res 42: D490–495. doi: 10.1093/nar/gkt1178 PMID: 24270786

49. Li H, Durbin R (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 25: 1754–1760. doi: 10.1093/bioinformatics/btp324 PMID: 19451168

50. Trapnell C, Williams BA, Pertea G, Mortazavi A, Kwan G, van Baren MJ, et al. (2010) Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. Nat Biotechnol 28: 511–515. doi: 10.1038/nbt.1621 PMID: 20436464

51. Colman DR, Toolson EC, Takacs-Vesbach CD (2012) Do diet and taxonomy influence insect gut bacterial communities? Mol Ecol 21: 5124–5137. doi: 10.1111/j.1365-294X.2012.05752.x PMID: 22978555

52. Dietrich C, Köhler T, Brune A (2014) The cockroach origin of the termite gut microbiota: patterns in bacterial community structure reflect major evolutionary events. Appl Environ Microbiol 80: 2261–2269. doi: 10.1128/AEM.04206-13 PMID: 24487532

53. Sabree Z, Moran N (2014) Host-specific assemblages typify gut microbial communities of related insect species. SpringerPlus 3: 138. doi: 10.1186/2193-1801-3-138 PMID: 24741474

54. Tai V, James ER, Nalepa CA, Scheffrahn RH, Perlman SJ, Keeling PJ (2015) The role of host phylogeny varies in shaping microbial diversity in the hindguts of lower termites. Appl Environ Microbiol 81: 1059–1070. doi: 10.1128/AEM.02945-14 PMID: 25452280

55. Ohkuma M, Brune A (2011) Diversity, structure, and evolution of the termite gut microbial community. In: Bignell DE Roisin Y, Lo N, editor. Biology of termites: a modern synthesis: Springer. pp. 413–438.

56. Tadmor AD, Ottesen EA, Leadbetter JR, Phillips R (2011) Probing individual environmental bacteria for viruses by using microfluidic digital PCR. Science 333: 58–62. doi: 10.1126/science.1200758 PMID: 21719670

57. Su NY, La Fage JP (1987) Initiation of worker-soldier trophallaxis by the Formosan subterranean termite (Isoptera: Rhinotermitidae). Insect Soc 34: 229–229.

58. Lewis JL, Forschler BT (2004) Protist communities from four castes and three species of *Reticulitermes* (Isoptera: Rhinotermitidae). Ann Entomol Soc Am 97: 1242–1251.

59. Lo N, Tokuda G, Watanabe D (2011) Evolution and function of endogenous termite cellulases. In: Bignell DE, Roisin Y, Lo N, editor. Biology of termites: a modern synthesis: Springer. pp. 51–67.

60. Watanabe H, Tokuda G (2010) Cellulolytic systems in insects. Annu Rev Entomol. Palo Alto: Annual Reviews. pp. 609–632. doi: 10.1146/annurev-ento-112408-085319 PMID: 19754245

61. Henrissat B, Bairoch A (1993) New families in the classification of glycosyl hydrolases based on amino acid sequence similarities. Biochemical Journal 293: 781–788. PMID: 8352747

62. Ni JF, Tokuda G (2013) Lignocellulose-degrading enzymes from termites and their symbiotic microbiota. Biotechnol Adv 31: 838–850. doi: 10.1016/j.biotechadv.2013.04.005 PMID: 23623853

Diane BIGOT

Biodiversité et évolution des virus présents dans les métagénomes animaux

Résumé

Les virus font partie des entités les plus abondantes sur Terre, mais la diversité des virus est très peu connue puisque biaisée en faveur d'hôtes animaux d'intérêts sociétal, agronomique et économique. L'apport des nouvelles techniques de séquençage permet actuellement d'obtenir des informations qui étaient tout simplement inaccessibles. Le but de mon travail de thèse a été l'étude de la diversité virale présente au sein d'un grand nombre d'animaux non-modèles. Pour répondre à cette problématique il m'a fallu mettre en place une méthodologie analytique innovante de découverte de nouveaux virus par une approche de méta-transcriptomique. Ce travail i) montre que la méthodologie bioinformatique mise en place est pertinente, ii) permet de découvrir de nouveaux virus ayant des caractéristiques génomiques particulières relevant de nouveaux genres ou familles de virus, iii) révèle de nouveaux hôtes pour des virus appartenant à des familles virales très étudiées et iv) montre que la gamme d'hôte de virus connus peut être plus étendue qu'attendu grâce à un focus sur la diversité des virus d'hyménoptères. D'une manière plus globale, mon travail permet de combler quelques lacunes existantes dans les connaissances liées à l'étude de la diversité virale et met en évidence l'importance de l'étude des animaux non-modèles.

Mots-clefs : Découverte de virus, méta-transcriptomique, animaux non-modèles, NGS, évolution et diversité virale, virus d'abeilles.

Abstract

Viruses are among the most abundant entities on Earth, but the viral diversity remains mostly unknown as currently biased in favour of animals of social, agronomic and economic interest. Next Generation Sequencing technologies provide access to so far inaccessible information. The aim of my PhD thesis was the study of the viral diversity within a large range of non-model animals. To address this question I set up an innovative analytical framework to discover new viruses based on a meta-transcriptomic approach. This work i) shows that this bioinformatics method is efficient and powerful, ii) allows the discovery of new viruses with particular genomic organisations suggesting they belong to new virus genera of families, iii) uncovered new viruses from new hosts in well-known viral families and iv) shows wider viral host range than previously expected based on a particular focus on hymenopteran viral diversity. Overall, my work allows to fill some gaps in the knowledge of viral diversity and shows the importance of studying non-model animal species in virology.

Key-words: Virus discovery, meta-transcriptomic, non-model animals, NGS, evolution and diversity of viruses, honey bee viruses.