Génomique des virus géants, des virophages, et échanges génétiques avec leurs hôtes eucaryotes

Lucie Gallot-Lavallée

AIX-MARSEILLE UNIVERSITÉ
ÉCOLE DOCTORALE SCIENCES DE LA VIE ET DE LA SANTÉ (ED 62)

LABORATOIRE INFORMATION GÉNOMIQUE ET STRUCTURALE

Génomique des virus géants, des virophages, et échanges génétiques avec leurs hôtes eucaryotes

Lucie Gallot-Lavallée

Thèse présentée pour obtenir le grade universitaire de docteur en science de la vie et de la santé

Discipline : Biologie

Spécialité : Génomique et Bioinformatique

Soutenue le 7 novembre 2017 devant le jury :

Elisabeth Herniou	Rapportrice
Marie-Agnès Petit	Rapportrice
Gwenaël Piganeau	Examinatrice
Christophe Robaglia	Examinateur
Guillaume Blanc	Co-directeur de thèse
Jean-Michel Claverie	Co-directeur de thèse

Remerciements

Merci tout d'abord à mes deux directeurs de thèse Guillaume Blanc et Jean-Michel Claverie, sans qui cette thèse n'aurait pas pu exister.

Un grand merci Guillaume pour ces 2 stages de Master suivis de 3 années de thèse passés ensemble. J'aurais pu continuer, mais il est temps que je tente l'envol. Merci pour ta disponibilité, ton implication et ton engagement envers tes thésards et stagiaires, pour ton suivi constant et pour tout ce que j'ai pu apprendre grâce à cette thèse. C'était passionnant !

Merci Seb, Grand Gourou, mon voisin de bureau, tu m'as tout appris niveau code ! Merci pour les conseils BDs, les matins studieux en duo dans l'open space (à l'époque lointaine où je me levais à une heure décente)…

Merci *Adriana Formosa* oups Adrien, cher co-thésard, pour ton soutien, qui a pris des formes multiples : petits biscuits (miam), plaisanteries variées, encouragements… C'est mon tour de te dire : « Allez, courage, dernière ligne droite ! » Merci pour les pâtisseries grandioses, les randos, les baignades, les découvertes du rhyming slang et des essais d'Orwell, les discussions animées (et, on pourrait en douter, passionnantes) sur la concordance des temps. En parlant de ça, merci pour tes commentaires d'ADD (accordeur du dimanche) sur ma thèse, très avisés.

Merci Olivier, les thésards et les stagiaires sont veinards de t'avoir. Toujours prêt à aller crapahuter ! Merci pour ta bonne humeur continuelle, tes récits d'exploits sportifs et d'aventures mémorables, nos sorties footing, bivouac, grimpe.. Merci d'avoir été là pour écouter les petits déboires des journées d'une thésarde.

Merci à la team MIO, Magali, Pascal, Youri, Hassiba, Émilie. Merci Magali d'être si chaleureuse, merci pour l'organisation de belles sorties culturelles ou sportives, les bonnes soirées chez toi. Merci aussi de m'avoir gentiment introduite par mail auprès d'Adam et de Hiro avant mon congrès à Plymouth. Merci Émilie pour ton soutien lors du concours pour décrocher la bourse de thèse, et le partage des bons tuyaux pour mon séjour Beyrouthin. Merci Hassiba de râler plus que moi mais d'être tout de même de si bonne compagnie, pour les cheesecakes délicieux et autres tajines succulents, les compliments gentils, et merci de sublimer mes gribouillis ! Merci Youri pour toutes les discussions stimulantes ; elles m'ont toujours redonnées du cœur à l'ouvrage si jamais il venait à manquer, ou bien une curiosité nouvelle.

Merci Cyril pour le modèle de travailleur persévérant et brillant que tu représentes, pour les belles sorties et les bonnes soirées quand tu étais à Marseille, et pour les visites amicales quand tu rentres de la Réunion.

Merci Miguel d'avoir égayé en papotant mes trajets du bus 47 ou même de la navette maritime.

Merci Élisabeth d'avoir partagé ton expérience d'ex doyenne des thésards maintenant thésée !

Merci à tous les travailleurs de l'open space, Dorothée, Nadège, Adrien, notre mascotte Guillaume Lacroix, Dimitra arrivée tardivement (du point de vue de ma thèse) mais bien vite adoptée, Virgine, Seb, Anna, merci à tous pour la bonne ambiance en rimes, en musique douce (sonnerie de téléphone en provenance de mon voisinage proche), en rires si entraînants, en cafés glacés, en jeux de mots, en post-it élogieux, en énigmes, j'en passe et des meilleurs. Merci aussi à ceux qui passent fréquemment faire une visite à l'open space pour quêter une douceur chocolatée ou le plus souvent en offrir, et aussi pour les bons conseils ciné Lionel !

Merci à tous les membres du labo ; je n'ai pas cité Jean-Marie tellement drôle, Matthieu et ses conseils avisés pour mes présentations, Sandra qui partage mon goût des bons vins et des bons fromages, Chantal et ses coups de pouce pour les figures, Audrey, ses gâteaux de crêpes et apéros gargantuesques, Estelle. Merci à tous pour l'ambiance sympathique lors des célébrations variées : publications, naissances, anniversaires, toute occasion est bonne pour remplir les flûtes et trinquer ! Merci aussi pour les deux congrès auxquels j'ai eu la chance d'assister.

Je remercie également Gwénaël Piganeau, Marie-Agnès Petit, Élisabeth Herniou et Christophe Robaglia qui me font l'honneur d'examiner mon travail de thèse et d'être présents lors de ma soutenance.

Enfin, je remercie ma famille et mes amis pour leur présence chère, leur soutien, et pour tout le reste !

"A quoi servirait-il à la ligne droite d'être la plus courte si la ligne courbe n'était pas la plus agréable ?"
— *Du Baroque*, Eugenio d'Ors y Rovira.

"It is hard to describe the exact route to scientific achievement, but a good scientist doesn't get lost as he travels it."
— Isaac Asimov

Résumé

Les grands virus nucléo-cytoplasmiques à ADN (NCLDV pour *Nucleo Cytoplasmic Large DNA virus* en anglais) forment un groupe très divers de virus à ADN double brin infectant exclusivement des eucaryotes. Les NCLDVs contiennent les-dits virus « géants » visibles au microscope optique, dont le premier spécimen découvert en 2003 a ouvert un champ inexploré de la biodiversité virale. Certaines familles de NCLDVs pourraient être aussi anciennes que l'émergence des domaines cellulaires contemporains. Ces virus habitent une grande variété d'environnements (sols, océans, etc.) dans lesquels ils concourent au contrôle des populations microbiennes eucaryotes (e.g., protistes, phytoplancton). Par ailleurs, certains de ces virus sont parasités par d'autres virus de taille plus modeste, les virophages. Ces derniers utilisent la machinerie de réplication des virus géants pour leur propre réplication dans un hôte cellulaire commun. La compréhension de l'évolution, la biodiversité et des interactions hôte/virus/virophage chez les NCLDVs est aujourd'hui un des grands chantiers de la virologie.

A l'aide d'approches génomiques et bioinformatiques, deux objectifs ont été poursuivis dans ce travail de thèse: (1) Caractériser la biodiversité et l'évolution au sein de la famille taxonomique des *Mimiviridae* (NCLDV) à travers la description et l'analyse comparative du génome du virus « CeV » qui infecte l'algue unicellulaire haptophyte *Haptolina (ex Chrysochromulina) ericina*. Ce volet de ma thèse a donner lieu à la publication d'un article (Gallot-Lavallee et al., 2017) ; (2) Caractériser la co-évolution des NCLDVs et des virophages avec leurs hôtes eucaryotes. Ce deuxième volet a été adressé en étudiant les transferts horizontaux d'ADNs viraux dans les génomes eucaryotes séquencés. Il a abouti à la publication de 2 articles (Blanc et al., 2015; Gallot-Lavallée and Blanc, 2017).

La particule icosahédrique du virus CeV a un diamètre de 160 nm et contient un génome de 474-kb dont l'analyse a confirmé que CeV est apparenté aux virus de la famille des *Mimiviridae*, initialement définie autour des virus géant prototypes Mimivirus et Megavirus infectant des amibes. D'autres membres des Mimiviridae, plus proches de CeV, infectent aussi des espèces d'algues unicellulaires. Ces derniers forment un sous-groupe, que nous proposons d'élever au rang d'une sous-famille, les *Mesomimivirinae*. Les *Mesomimivirinae* possèdent des caractères génomiques exclusifs: une deuxième copie du gène de l'ARN polymérase, la présence d'un gène codant pour seconde protéine de capside, ainsi que 7 autres groupes de gènes orthologues absents chez les autres *Mimiviridae*. En dépit de leurs liens de parenté, chacun de ces virus possède une majorité de gènes qui lui est unique, et dont l'origine est inconnue. D'autre part, plusieurs cas indépendants d'acquisition par transfert horizontal du même gène par CeV et d'autres virus non-apparentés ont été analysés. Cette thèse documente également des évènements génétiques qui se sont produits dans la seule lignée de CeV, comme par exemple des fusions de gènes inédites.

Le criblage bioinformatique des bases de données de séquences a mis en évidence des fragments d'ADNs de tailles variables (jusqu'à ~500Kb) intégrés au sein des génomes de nombreux eucaryotes, ayant originellement appartenus à des NCLDVs ou des virophages. Leur analyse a permis d'élargir le spectre d'hôtes connus des NCLDVs et suggèrent l'existence de nouvelles familles virales qui restent à découvrir. Egalement, ces insertions contiennent des gènes que l'on ne retrouve pas dans les génomes de NCLDVs séquencés, et pourraient coder pour des fonctionnalités inédites du virus donneur. Des copies de génomes de virophages intégrées au génome de l'algue *Bigelowiella natans* suggèrent une nouvelle stratégie de co-infection et de

dissémination du virophage, qui pourrait aussi constituer un mécanisme de défense contre les NCLDVs pour l'hôte eucaryote.

Blanc, G., Gallot-Lavallée, L., and Maumus, F. (2015). Provirophages in the Bigelowiella genome bear testimony to past encounters with giant viruses. Proc. Natl. Acad. Sci. U. S. A. *112*, E5318-5326.

Gallot-Lavallée, L., and Blanc, G. (2017). A Glimpse of Nucleo-Cytoplasmic Large DNA Virus Biodiversity through the Eukaryotic Genomics Window. Viruses *9*.

Gallot-Lavallee, L., Blanc, G., and Claverie, J.-M. (2017). Comparative genomics of Chrysochromulina Ericina Virus (CeV) and other microalgae-infecting large DNA viruses highlight their intricate evolutionary relationship with the established Mimiviridae family. J. Virol. JVI.00230-17.

Abstract

Nucleo Cytoplasmic Large DNA Viruses (NCLDV) form a very diverse group of double-stranded DNA viruses exclusively infecting eukaryotes. The NCLDVs contain the so-called "giant" viruses visible under light microscope, the first specimen of which, discovered in 2003, opened an unexplored field of viral biodiversity. Some families of NCLDVs could be as old as the emergence of contemporary cellular domains. These viruses inhabit a wide variety of environments (soils, oceans, etc.) where they contribute to the control of eukaryotic microbial populations (e.g., protists, phytoplankton). Moreover, some of these viruses are parasitized by smaller viruses, the virophages, that use the giant virus machinery for their own replication in a common cellular host. Understanding NCLDVs' evolution, biodiversity and host/virus/virophage interactions is now one of the major areas of virology.

Using genomics and bioinformatics approaches, I pursued two objectives in this thesis: (1) Characterize the biodiversity and evolution within the taxonomic family of *Mimiviridae* (NCLDV) through the description and comparative analysis of the genome of the virus "CeV" which infects the unicellular alga haptophyte *Haptolina (ex Chrysochromulina) ericina.* This part of my thesis resulted in the publication of an article (Gallot-Lavallee et al., 2017); (2) Characterize the co-evolution of NCLDVs and virophages with their eukaryotic hosts. This second part was addressed by studying horizontal transfers of viral DNA in the available sequences of eukaryotic genomes. It resulted in the publication the following articles (Blanc et al., 2015; Gallot-Lavallée and Blanc, 2017).

The icosahedral particle of the CeV virus has a diameter of 160 nm and contains a 474-kb genome. The analysis of the latter confirmed that CeV is related to the viruses of the *Mimiviridae* family, originally defined around the giant viruses prototypes Mimivirus and Megavirus which infect the amoebae. Other members of the *Mimiviridae*, closer to CeV, also infect unicellular algae. Together with CeV, the latter form a subgroup that we proposed to cluster into the subfamily *Mesomimivirinae*. *Mesomimivirinae* possess exclusive genomic characters: a second copy of the RNA polymerase gene, the presence of a gene coding for a second capsid protein, and 7 other groups of orthologous genes absent in other *Mimiviridae*. Despite their relatedness, each of these viruses possesses a majority of unique genes, the origin of which is unknown. On the other hand, several cases of parallel acquisitions by horizontal transfer of the same gene by CeV and other unrelated viruses were analyzed. This manuscript also documents genetic events that have occurred only in CeV, such as unique gene fusions.

Bioinformatic screening of the sequence databases revealed DNA fragments of varying sizes (up to ~ 500Kb) integrated into the genomes of many eukaryotes, originally belonging to NCLDVs or virophages. Their analysis allowed to widen the spectrum of known NCLDVs' hosts and suggest the existence of new viral families that remain to be discovered. Also, these inserts contain genes that are not found in the sequenced NCLDVs genomes, and could encode novel functionalities of the donor virus. Copies of virophage genomes integrated into the *Bigelowiella natans* algae genome suggest a new strategy for co-infection and dissemination of virophage, which may also constitute a defense mechanism against NCLDVs for the eukaryotic host.

Blanc, G., Gallot-Lavallée, L., and Maumus, F. (2015). Provirophages in the Bigelowiella genome bear testimony to past encounters with giant viruses. Proc. Natl. Acad. Sci. U. S. A. *112*, E5318-5326.

Gallot-Lavallée, L., and Blanc, G. (2017). A Glimpse of Nucleo-Cytoplasmic Large DNA Virus Biodiversity through the Eukaryotic Genomics Window. Viruses *9*.

Gallot-Lavallee, L., Blanc, G., and Claverie, J.-M. (2017). Comparative genomics of Chrysochromulina Ericina Virus (CeV) and other microalgae-infecting large DNA viruses highlight their intricate evolutionary relationship with the established Mimiviridae family. J. Virol. JVI.00230-17.

Sommaire

Introduction

A - Introduction générale sur les virus

1 - Historique et définition

En 1884, alors que sévit à Paris une épidémie de fièvre typhoïde, le biologiste et physicien Chamberland crée un filtre dans le but d'obtenir de l'eau exempte de microorganismes. Ce « filtre de Chamberland » possède des pores de 0,2 µm. En effet, tous les microbes connus à l'époque sont d'une taille supérieure (1).

En 1892, Ivanovski montre qu'une maladie de plants de tabac (la mosaïque du tabac) est due à un agent infectieux capable de traverser le filtre de Chamberland. Il suppose alors qu'il s'agisse d'une bactérie plus petite, ou bien d'une toxine produite par cette bactérie. Ce n'est qu'en 1898 que Martinus Beijerinck met en avant la nouveauté de cet agent infectieux filtrable, et introduit la notion de virus (2).

Jusque dans les années 1950, le concept de virus restait néanmoins assez flou, et résidait principalement en 2 observations : (i) les virus sont des agents infectieux capables de passer à travers le filtre de Chamberland (ii) ils ne sont pas cultivables dans les milieux de cultures utilisés pour multiplier les microorganismes.

Grâce aux avancées en microscopie électronique, la nature des virions (forme extracellulaire du virus) a pu être déterminée: le virion est composé d'une capside qui véhicule et protège le génome du virus. Cette capside peut être ou non couverte d'une enveloppe virale qui est dérivée de portions de la membrane cellulaire de l'hôte additionnée de glycoprotéines. Cependant, le mode de réplication et transmission de l'information génétique chez les virus restait à élucider, le concept de virus a donc continué à évoluer, notamment en parallèle des progrès effectués dans la compréhension de la nature du matériel génétique (3). L'idée que les virus ne sont pas limités à leur forme extracellulaire avait déjà été proposée par Claudio Bândea en 1983 (4), et a été récemment clairement énoncée par le prof. Jean-Michel Claverie (5). Par opposition à l'état de virion qui serait simplement comparable à une « spore », l'usine virale, formée par certains virus dans le cytoplasme de leur hôte et ceinte de membranes, serait le véritable «organisme viral» (5). C'est effectivement pendant ce stade cellulaire que le virus subit mutations, réarrangements et pressions de sélection menant à des adaptations. Sur cette lancée de propositions conceptuelles, la notion de « virocell » (6) a été introduite par le prof. Patrick Forterre. Elle rend compte de la redirection partielle ou totale du métabolisme des cellules vers la réplication du virus qui est observée lors d'une infection virale. Le concept de virus est donc encore actuellement en mutation.

Plutôt que de donner une définition *stricto sensu* des virus, un consensus réside dans une liste de différences entre les organismes cellulaires et les virus. Il est intéressant de voir que ce consensus sur le concept de virus repose sur une absence de propriétés cellulaires : (i) ils ne se divisent pas par fission binaire, (ii) le matériel génétique des virus ne code pas pour des constituants du ribosome ni pour une machinerie de traduction complète, (iii) ils ne possèdent pas de système enzymatique permettant de convertir les nutriments en énergie nécessaire à la synthèse biochimique. Ces deux premières caractéristiques impliquent que les virus sont absolument dépendants d'une cellule hôte pour se répliquer et qu'ils sont considérés comme métaboliquement inertes en dehors de cette cellule hôte. Notons que par contraste, tous les parasites intracellulaires et les endosymbiontes (symbionte vivant dans le cytoplasme de son partenaire) eucaryotes, bactériens et archées (7), bien qu'ils soient également dépendants de leur cellule hôte, possèdent ces deux machineries enzymatiques. Ils les ont effectivement conservés malgré le processus de réduction génomique qui a accompagné l'adaptation à la vie intracytoplasmique et les a ammené à une perte d'autonomie et à un état intracellulaire obligatoire. Ce processus de réduction génomique est inhérent au parasitisme : profitant des machineries de leur hôte, les pressions de sélection sur certains gènes du parasite diminuent. Ces gènes sont alors progressivement « pseudogénisé » puis perdus.

Ces trois propriétés négatives permettent de discriminer les virus des cellules. Etablir une définition au moyen de propriétés positives est une mission complexe, car les virus sont extrêmement diverses (ceci sera explicité en partie 3). Notons également que bien que le concept ait évolué constamment depuis l'isolement du premier virus, le critère « pratique » de filtrabilité a continué à être utilisé jusqu'en 2002 pour isoler la fraction virale de la fraction cellulaire. La découverte des virus géants capables d'être filtrés sur pore de 0.2 um a remis en question l'utilisation de la filtration pour explorer la diversité du monde viral.

2 - Cycle lytique/lysogénique

Nous avons vu que les virus sont absolument dépendants des cellules pour se répliquer. Il existe deux voies principales par lesquelles cette réplication se produit : le cycle lytique et le cycle lysogénique.

Cycle lytique. A la suite de l'entrée du virus (ou de son génome uniquement) dans une cellule compétente, le génome viral est répliqué et ses protéines structurales synthétisées (voir figure 1). Les virions sont ensuite assemblés, puis s'échappent de la cellule hôte de façon à pouvoir infecter d'autres cellules. Au terme de la réplication, la cellule hôte peut être lysée, ce qui a donné son nom au cycle, ou bien les particules peuvent également être relarguées dans le milieu extracellulaire par bourgeonnement à la surface de la membrane plasmique sans que cela entraine la mort de la cellule.

Cycle lysogénique. Cette voie a été initialement mise en évidence chez certains bactériophages (virus infectant des bactéries, aussi appelés phages) qui sont alors qualifiés de « tempérés ». La lysogénie est caractérisée par l'intégration du génome du phage dans celui de la bactérie. Sous cette forme de génome intégré, le virus est appelé prophage. Les gènes du prophage ne sont pas transcrits, et le prophage se réplique passivement lorsque la cellule hôte bactérienne se divise pour former deux cellules filles. Le phage tempéré peut subir de tels cycles lysogéniques pour plusieurs générations avant de commuter vers un cycle lytique par un processus appelé induction. Cette induction survient généralement en réponse à des altérations de l'ADN causées par des facteurs environnementaux comme par exemple les rayons UV. L'intégration et l'excision sont réalisées par des enzymes du phage, les intégrases et les excisionnases respectivement. Récemment, des travaux ont montré que cette modalité de réplication n'est pas l'exclusivité des bactériophages. Effectivement, d'autres virus, également ADN double brin (dsDNA pour *double stranded DNA* en anglais) mais infectant des eucaryotes (plus précisément des algues multicellulaires de la classe *Phaeophyceae*), sont également capables de lysogénie. Les mécanismes moléculaires impliqués dans le cycle lysogénique de ces virus ne sont pas bien compris (8).

3 - Une virosphère extrêmement variable

J'ai mentionné plus haut que les virus présentent une diversité remarquable. La figure 1A expose la variété des virions (la forme extracellulaire du virus) que nous connaissons actuellement. Les différentes morphologies des particules y sont représentées. Ces particules contiennent le génome du virus qui peut être de différentes natures moléculaires: ADN double brin (dsDNA); ADN simple brin (ssDNA) ; ARN double brin (dsRNA) ; ARN simple brin sens + (ssRNA+) ou sens - (ssRNA-). De plus, l'information génétique peut être distribuée sur un ou plusieurs réplicons. Les stratégies employées par les virus pour répliquer ce matériel génomique sont complexes et variables. La figure 1 panels B,C,D et E présente dans les grandes lignes ces modes de réplication se produisant lors d'un cycle lytique. Cette myriade de formes et de mécanismes chez les virus, nourrissent, comme nous allons le voir dans la partie suivante, les controverses concernant l'origine des virus.

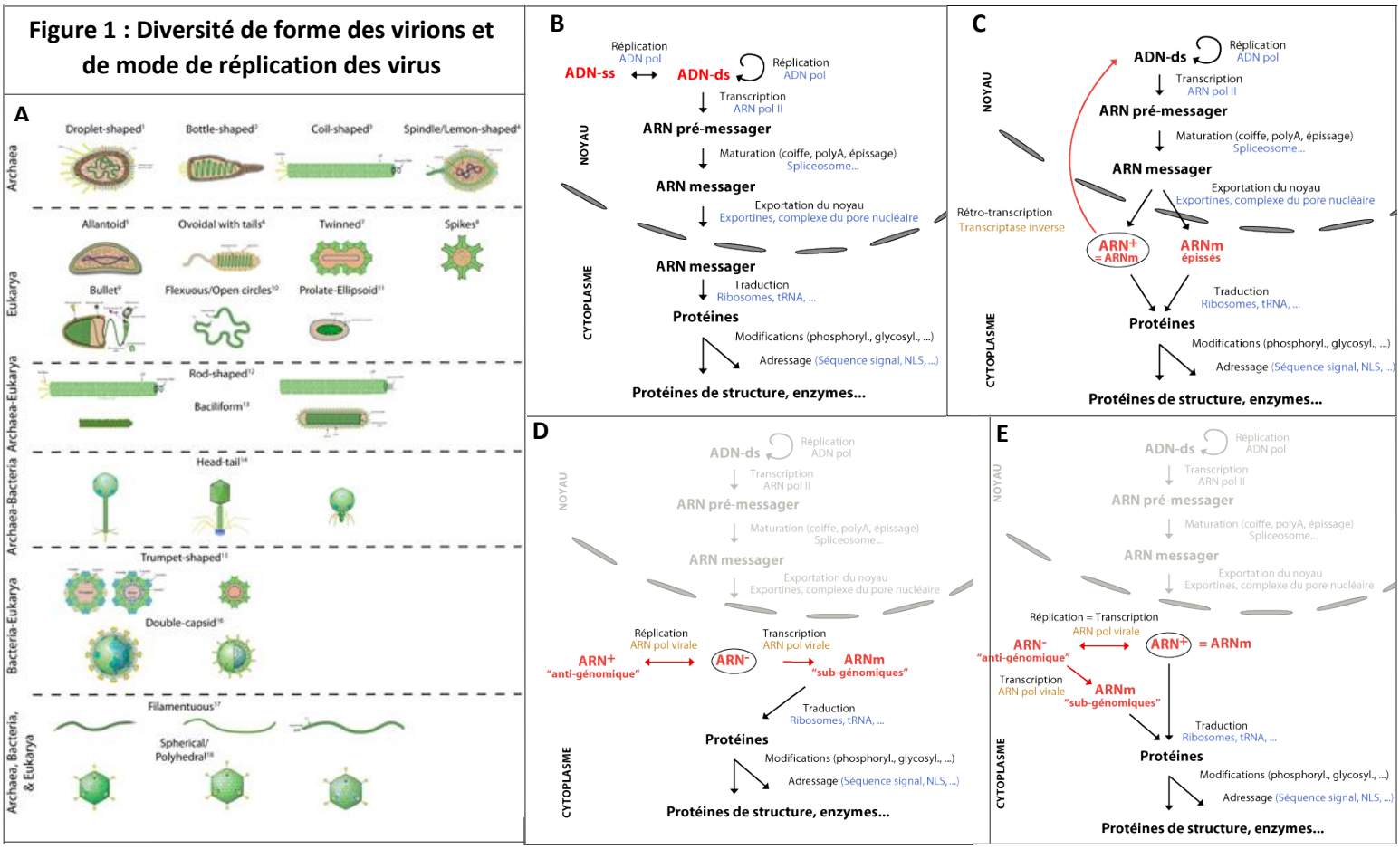


Figure 1: diversité des virus. A/ diversité des virions ; **BCDE** Schémas de différents cycles réplicatifs lytiques de virus ; **B/** réplication des virus à ADN qui ont un cycle nucléocytoplasmique (certains dsDNA virus ont un cycle exclusivement cytoplasmique). Le génome des virus ssDNA est préalablement converti en dsDNA par la polymérase cellulaire. **C/** réplication des rétrovirus : La transcriptase inverse virale copie l'ARN génomique en ADN double brin qui migre dans le noyau et est intégré dans le génome de la cellule. L'ARN polymérase cellulaire transcrit cet ADN en ARNms, également traduits par la cellule, et ARN génomique **D/** réplication des virus à ARN- : l'ARN polymérase virale transcrit l'ARN- en ARNm et assure aussi la réplication du génome : ARN- > ARN+ > ARN- ; **E/** réplication des virus à ARN+ (avec ARNm subgénomique) : Les protéines virales sont traduites en partie à partir de l'ARN génomique directement et en partie à partir d'ARN antigénomique préalablement copié à partir de l'ARN génomique. Ces derniers sont appelés ARNms subgénomiques. La réplication du génome viral est assurée par la polymérase virale qui recopie l'ARN génomique en ARN antigénomique (-) qui est ensuite recopié en ARN génomique (+) ; **BCDE** Après la réplication du génome viral et la synthèse des protéines structurales, les virions sont assemblés et libérés dans le milieu extracellulaire. Tirée de (8) et de ©UCLouvain.

4 - Origine des virus

Cette thématique a ici été scindée en deux sous-questions : (a) Une origine unique ou des origines multiples des virus ? et (b) à quel point les virus sont-ils anciens ? En partie (c) seront ensuite présentées les différentes hypothèses s'attachant à expliquer la naissance d'un parasitisme de type viral.

a - Une origine unique ou multiple des virus ?

J'ai décrit plus haut l'extrême diversité des virus. Dans une perspective évolutive, ceci peut être interprété de deux façons : soit tous les virus ont une origine unique, qui a été suivie d'une divergence extrême conduisant à la diversité virale actuelle, soit l'émergence de pathogènes viraux a eu lieu à plusieurs reprises au cours de l'histoire de la vie. En faveur de cette seconde hypothèse plaide le fait que les virus n'ont apparemment pas de gènes universellement conservés. Même les protéines de capsides, qui forment la structure protégeant le matériel génétique des virions, seraient apparues à plusieurs reprises au cours de l'évolution à partir de différentes protéines cellulaires (10). D'un autre côté, cette diversité peut être le résultat d'une divergence des caractères ancestraux (e.g., gènes, protéines, structures) telle que les liens de parenté entre virus ne sont plus décelables par les approches analytiques actuelles (i.e., similarité de séquences et de structures). Ceci pourrait être dû par exemple à un fort taux de mutation et/ou de transferts de gènes que connaissent la plupart des virus (11). Ce scenario est soutenu par la reconnaissance de séquences homologues entre virus éloignés (mais non universellement conservées) qui constitueraient des indices encore apparent d'une origine commune de ces virus. C'est le cas par exemple pour la protéine portail des *Herpesviridae* et de bactériophages, qui infectent respectivement des eucaryotes et des bactéries (7, 12). Le module commun (double barrel trimer) présent dans la protéine de capside des adenovirus humain, du virus de bactérie PRD1 et du virus d'archée STV1 (7), en est un autre exemple. Pour conclure, cette question n'est encore pas résolue. Je penche personnellement pour le scenario d'une multiplicité des origines car les arguments invoqués pour soutenir cette hypothèse me semblent solides. Ceci n'exclut pas le fait que certains virus éloignés puissent effectivement avoir une origine commune.

b - Les virus sont-ils très anciens ?

Quelle est l'ancienneté du parasitisme de type viral ? Quelle est l'ancienneté des virus actuels ? Ces questions sont bien entendu liées à la précédente dans le sens où elles se réduisent à une seule et même question si nous admettons une origine unique des virus, mais représentent bien deux questions distinctes si nous privilégions le scénario de la multiplicité des origines : certaines familles virales pourraient être très anciennes, d'autres, plus récentes.

Les cas précédemment invoqués d'homologies entre virus infectant des hôtes extrêmement éloignés (e.g., bactéries contre eucaryotes), si nous admettons qu'elles représentent des indices

d'une origine commune de ces virus, sont également révélateurs de l'ancienneté de ces familles de virus concernées. Effectivement, les virus actuels connus qui infectent les bactéries n'infectent pas d'eucaryotes. L'hypothèse est donc que les lignées virales aient co-évolué avec les lignées d'hôtes cellulaires. Ceci est effectivement une conséquence connue (à échelle de temps plus réduite) de leur nature de parasite car ils dépendent de leurs hôtes pour leur réplication. En d'autres termes, l'ancêtre de ces virus était donc présent en même temps que l'ancêtre commun à leurs hôtes actuels, c'est à dire LUCA, et la divergence de ces virus serait contemporaine à celle de leurs hôtes (13, 14). Voici par ailleurs un argument en faveur d'une origine ancienne des virus : la quasi-totalité des organismes que nous connaissons, que ce soit des bactéries, des archées ou des eucaryotes, sont infectés par des virus (9, 15, 16). Ceci suggère que le parasitisme de type viral est intrinsèquement lié à la vie cellulaire. Il apparait alors contre-intuitif que cette caractéristique soit exclusive aux cellules actuelles; il en serait plutôt ainsi depuis les débuts de la vie cellulaire (du moins depuis que les cellules eurent atteint une forme comparable à celles d'aujourd'hui). Autrement dit, les virus existeraient depuis LUCA (17–20). Un autre argument arguant pour une origine ancienne des virus provient d'analyses comparatives des séquences d'enzymes clé de la réplication des génomes, nommément la DNA polymérase et la DNA topoisomérase. Ces analyses suggèrent que les protéines virales ont divergées des homologues cellulaires avant LUCA (21). Enfin, des études phylogénomiques basées sur l'analyse des familles de repliements de protéines (FSF, Fold SuperFamilies), suggèrent que la plupart des plus anciens FSF sont partagés par les cellules et les virus, ce qui signifierait une coexistence précoce des virus et des cellules (15). Cette étude pourrait cependant souffrir de certains biais liés à la petite taille des génomes viraux (22), qui invalideraient les conclusions tirées (23). J'ai présenté ici plusieurs arguments qui sont en faveur de l'existence de virus contemporains à LUCA. Remarquons cependant que des explications alternatives à une ancienneté des virus actuels peuvent parfois également être invoquées pour satisfaire aux observations décrites, comme par exemple, des évènements de transferts latéraux de matériel génétique ainsi que la rapidité d'évolution des séquences virales.

c - Hypothèses quant à l'émergence de parasitisme de type viral

Nous avons présenté l'état actuel des connaissances quant à l'ancienneté et l'origine, unique ou non, des virus. Mais comment un parasite de type viral peut-il émerger ? A partir de quoi ? Quels sont les mécanismes possibles ? Il y a plusieurs possibilités: (1) avant d'être des parasites obligatoires, les virus auraient été des entités autonomes qui auraient subi une réduction génomique et seraient devenu dépendants d'une cellule hôte. Cette théorie est appelée « the *reduction hypothesis* » (5, 7, 14); (2) les virus pourraient provenir de l'évasion d'une partie du matériel génétique d'une cellule sous la forme d'un élément mobile « égoïste » infectieux. Ce dernier pourrait grossir en acquérant petit à petit des gènes cellulaires. Cette théorie est appelée « the *escape theory* » (14, 24). Chacune de ces deux premières hypothèses peut être elle-même divisée en deux selon l'ancienneté de l'entité autonome (hypothèse 1) ou la cellule donneuse (hypothèse 2). Nous avons donc 4 alternatives, qui sont résumées dans la figure 2. (3) Une dernière théorie existante est appelée «*virus first hypothesis* » (5, 7, 25). Elle propose que les virus fussent déjà

présents dans un monde pré-biotique, où ils eussent effectué leur cycle réplicatif en utilisant les fonctionnalités trouvées dans la « soupe primitive ».

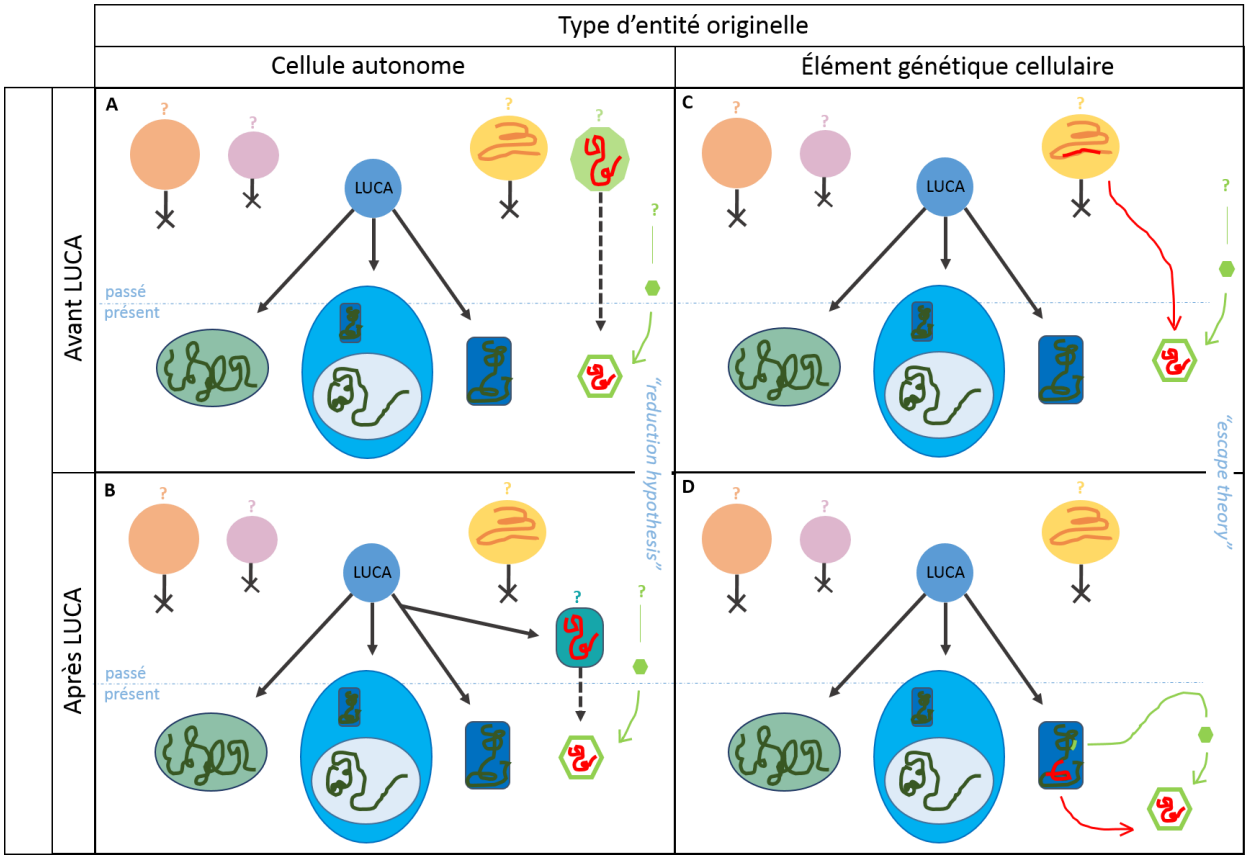


Figure 2 : Émergence des virus : les différentes hypothèses. **A** et **B** : « reduction hypothesis ». **C** et **D**: « escape theory ». Dans chacun des panels, le virus est représenté par un hexagone vert vide contenant de l'ADN rouge. Les archées, eucaryotes et bactéries sont représentés respectivement de gauche à droite en différents ton de bleu sous la ligne en pointillé qui délimite le présent du passé. Les autres formes pleines, situées au dessus de la ligne en pointillé, représentent d'hypothétiques lignées cellulaires aujourd'hui disparues. L'évènement d'acquisition d'une protéine de capside est représenté par une flèche reliant un petit hexagone vert plein. La dernière hypothèse appelée « virus first hypothesis » qui postule que les virus existaient avant les cellules et se repliquaient en utilisant les fonctionnalités trouvé dans la « soupe primitive » n'a pas été représentée car l'aspect de la dite soupe m'a paru hautement sujet à interprétations.

(1) - Dans l'hypothèse (1) de réduction, les entités autonomes pourraient être soit (i) postérieures à LUCA (figure 2, panel B), soit (ii) antérieures à LUCA (figure 2, panel A). Ces lignées cellulaires auraient perdu leur autonomie par le même processus de réduction génomique (décrit précédemment) que celui qui opère chez les parasites intracellulaires eucaryotes, bactériens ou

archées. L'hypothèse 1 (i) (panel B) souffre du fait que nous ne connaissons pas d'intermédiaire entre les cellules aux modes de vie obligatoirement parasitaires (ou symbiotiques) et les virus (21). Nous ne connaissons pas par exemple, d'entité biologique codant pour des ribosomes et ayant un mode de propagation dépendant d'un stade virion, sans fission ; Dans l'hypothèse 1 (ii) (panel A), les virus seraient des reliques de formes de vie dont plus aucune forme autonome ne subsisterait de nos jours. Cette hypothèse a été proposée afin d'expliquer les nombreux ORFans (gènes codants pour des protéines, et par extension protéines, qui n'ont de similarité statistiquement significative avec aucun(e) autre gène (protéine) connu(e)) présent dans les protéomes des virus. Cependant, le fait que ces formes de vies originelles (possiblement des lignées cellulaires) dont seraient issus les virus soient aujourd'hui éteintes, complique fortement la possibilité de faire la démonstration de la véracité de cette hypothèse, faute de spécimens d'étude et comparaison.

(2) - L'hypothèse (2), dite «théorie de l'évasion», est supportée par l'observation que les virus d'aujourd'hui peuvent intégrer des gènes cellulaires dans leur propre génome. De la même façon que concernant l'hypothèse (1), l'élément mobile originel aurait pu s'évader d'une cellule descendante de LUCA (hypothèse 2 i) (panel D), ou bien d'une lignée sœur, aujourd'hui éteinte (hypothèse 2 ii) (panel C). La présence de nombreux gènes viraux n'ayant pas d'homologues cellulaires est une faiblesse de la théorie 2 i, dont ne souffre pas la version selon laquelle le matériel génétique évadé proviendrait de lignées éteintes.

(3) - L'hypothèse (3) place la naissance des premiers virus avant les premières cellules, dans un monde à ARN constitué de molécules libres. Le professeur Koonin entre autre, suppose que ce monde pré-cellulaire était effectivement capable de supporter la réplication d'entités de type virus très variées (14). Cependant, cela n'implique pas que ces anciens virus soient les ancêtres des virus actuels. Effectivement, les virus contemporains codent pour des protéines et dépendent donc de la présence de ribosomes. Or, le professeur Forterre remarque que la mise en place des ribosomes peut difficilement être imaginée dans un monde sans sélection darwinienne, mais plutôt requiert que des entités individuelles bien définies (des « proto-cellules », au moins) soient en compétition. D'après lui, cela suggère que cette hypothèse peut être rejetée pour les virus actuels (7). Plus généralement, ceci implique qu'il y a peu de chances que cette hypothèse puisse être prochainement prouvée scientifiquement (5).

L'origine de la protéine de capside est centrale dans l'évolution de la structure des virus, et des scénarios évolutifs se déclinent selon des variations propres à chaque hypothèse. Nous avons déjà mentionné un article récent qui suggère que des protéines de capsides ont émergé de protéines homologues cellulaires ancestrales en de multiples occasions. Les éléments génétiques évadés (hypothèses 2) pourraient avoir recruté l'une d'entre elles. Les éléments autonomes (hypothèse 1), ou les virus du monde pré-cellulaire (hypothèse 3) pourraient de leur côté avoir déjà eu dans leur protéome une *bona fide* capside, ou bien une protéine qui aurait été recrutée pour cela par la suite.

Rappelons que les connaissances actuelles ne permettent pas de trancher quant à l'unicité ou la multiplicité de l'origine des virus. Si les virus sont apparus en de multiples occasions, alors plusieurs de ces possibilités pourraient effectivement avoir eu lieu.

Au-delà de ces questions fondamentales et très théoriques, des faits confondants concernant les virus sont connus et font l'objet des prochains paragraphes.

5 - Impact global des virus

Les virus sont ubiquitaires et sont les entités les plus nombreuses sur Terre. Dans les océans particulièrement, ils surpassent en nombre les procaryotes d'un facteur >10 (26, 27). La majorité de ces virus infectent des microbes, comme des bactéries ou des micro-eucaryotes. Or, ces hôtes microbiens, et particulièrement le phytoplancton (les cyanobactéries et les algues unicellulaires), jouent un rôle prépondérant dans les cycles biogéochimiques à l'échelle planétaire (28), ils sont notamment responsable de 50 % de la production primaire mondiale ; l'impact des virus sur leurs hôtes peut donc avoir des répercussions globales, que voici rapidement résumées :

(i) Impact écologique. Les virus modifient la composition des communautés microbiennes par la lyse de certaines des espèces constituant cette communauté. Le modèle sous-jacent à la théorie dite « killing the winner » avance que les virus représentent un facteur d'équilibre grâce auquel des espèces bactériennes avec des taux de croissance différents peuvent toutes coexister dans un même environnement, malgré des ressources limitantes (29, 30). Effectivement, la théorie propose que la meilleure stratégie pour un virus est de tuer l'espèce majoritaire. Il est assumé que sans ce mécanisme, les espèces non dominantes n'auraient pas accès aux ressources et la biodiversité au sein des écosystèmes serait beaucoup plus étroite qu'observé. Par exemple, plusieurs virus infectant des micro-algues sont responsables de la régulation des populations de leurs hôtes. *Emiliania huxleyi* Virus (EhV) a un rôle prépondérant dans la terminaison des immenses efflorescences (qui peuvent couvrir jusqu'à 10 000 km2) formées par son hôte la micro-algue haptophyte *Emiliania huxleyi* (31). C'est le cas également pour le virus HaV qui régule les populations de son hôte *Heterosigma hakashiwo* (32), une algue raphidophyte qui a des effets délétères sur les cultures notamment d'huîtres à perles (33) et de saumon (34) au large du japon.

(ii) Impact biogéochimique. Le « viral shunt » est le terme définissant la remise à disposition des matières organiques contenues dans la cellule lorsqu'elle est lysée par un virus (35). Au lieu d'être transféré vers les niveaux trophiques supérieurs par ingestion des prédateurs, ou de sédimenter sur le sol océanique après la mort de la cellule, ce matériel peut donc être recyclé par d'autres organismes. Les organismes planctoniques hétérotrophes (36) profitent de ce « viral shunt », et notamment des chaines carbonées rendues disponibles ; mais ce ne sont pas les seuls : la majeure partie des besoins du plancton photosynthétique en Nitrogène (37) et en Fer (38) serait fournie par

la lyse virale. Ceci favoriserait donc indirectement la production primaire. Concernant le fer, le rôle des phages à queue (i.e., *tailed phages*) est parfois même absolument direct, puisque ces queues pourraient contenir jusqu'à 70 % du fer organique suspendu à la surface des océans (39). Selon l'hypothèse nommée « ferrojan horse hypothesis » (39) (le mot ferrojan est un mixe des deux mots fer et trojan signifiant « de Troie », « ferrojan horse » est donc une allusion à la stratégie du cheval de troie), ces queues ferreuses serviraient d'appât pour leurs hôtes bactériens. Effectivement, elles seraient captées par les sidérophores bactériens (des structures spécialisées dans l'assimilation du fer), à la suite de quoi le virus pourrait perforer la membrane afin d'injecter son ADN dans la cellule. Les stocks de fer cellulaire pourrait même être recyclés en étant incorporés dans les queues des nouveaux phages issus de l'infection.

6 - Impact métabolique sur les cellules

Le cycle infectieux d'un virus, lytique (40), ou lysogénique (41), altère, de façons diverses, le métabolisme de l'hôte. Ces modifications du métabolisme microbien peuvent être extrêmement ciblées ou au contraire très globales. Bien que de façon moins directe que concernant la lyse des cellules, ceci peut également perturber les écosystèmes localement ou globalement, particulièrement lorsque cela touche des producteurs primaires. Les mécanismes sous-jacents à ces modifications ne sont pas toujours connus.

Voici quelques exemples d'altérations globales qui ont été décrites: premièrement, l'infection de *Sulfitobacter* et *Pseudomonas aeruginosa* par leur bacteriophage génère une augmentation globale des niveaux de métabolites variés (42, 43). Deuxièmement, l'observation que l'infection par des virus remodèle le métabolisme des lipides de leur hôte a été faite à plusieurs reprises, concernant des virus de taxa diverses (44–47). Dans ces deux cas, il est admis que cela favorise la réplication du virus, mais les mécanismes sous-jacents sont peu connus. Un niveau additionnel de compréhension a été atteint dans l'exemple suivant : des analyses quantitatives du carbone et de l'azote séquestrés dans la biomasse particulaire ont indiqué que l'infection de *Sulfitobacter* par son phage redirige ~75% des nutriments vers les virions (43).

Nous avons également parfois plus d'indices quant aux mécanismes en jeu, notamment lorsque les altérations du métabolisme sont plus ciblées. Ils peuvent être distingués en (i) « indirects » ou (ii) « directs ».

(i) <u>mécanismes indirects</u> : Une infection de plants de tabac par le virus de la mosaïque altère les fonctions photosynthétiques de la plante. Ceci semble être dû à l'accumulation de protéines de capside dans les chloroplastes (48).

(ii) modifications directes : Le virus peut également exercer un contrôle sur le métabolisme de l'hôte de façon plus directe, ciblée. Effectivement, un paradigme émergeant est que les phages possèdent des gènes métaboliques, appelés AMG (pour Auxiliary Metabolic Genes), que l'on supposait précédemment restreints aux cellules. Ces AMGs (revues : (49–51)), provenant probablement de transferts horizontaux depuis les cellules, sont utilisés pour suppléer le métabolisme de l'hôte en enzymes ou nutriments limitants, voire, parfois pour le « reprogrammer ». C'est alors un métabolisme qualifié de « viro-cellulaire » qui se met en place (52). Les AMGs décrits sont impliqués notamment dans la photosynthèse (composants de photosystèmes (53, 54), transporteurs d'électrons (55), ou encore biosynthèse des pigments (56)), dans la voie des pentoses phosphates (57), dans l'acquisition du phosphate (58, 59), du soufre (60). A côté de ces AMGs impliqués dans la production d'énergie ou l'acquisition de nutriments, d'autres prennent part spécifiquement à la synthèse des nucléotides (61, 62), et/ou peuvent toucher des systèmes plus globaux comme la réponse au stress (63). Parfois, ces AMGs ne se contentent pas de compléter le métabolisme de l'hôte mais le « reprogramment » de façon drastique. L'exemple le mieux connu est la redirection de l'énergie de la photosynthèse vers la voie des pentoses phosphates au détriment du cycle de calvin lors de l'infection des cyanobactéries *Prochlorococcus* et *Synechococcus* par leur cyanophages (64). Les AMGs les plus étudiés sont les composants des photosystèmes codés par les cyanophages, pour les raisons suivantes (i) la majorité de ces phages en possèdent ; ceci implique donc qu'une part de la production primaire est codées par des virus ; (ii) parmi ces AMGs, les proteines *hli* qui « récoltent » l'énergie lumineuse, sont modifiées dans la lignée virale par accumulation de mutations. Cette diversité génétique produite par les virus participe à l'évolution des photosystèmes, particulièrement lorsque les gènes viraux sont transférés « en retour » chez les hôtes bactériens (65, 66). Cette expansion de certaines familles de gènes produite chez les hôtes par leur virus pourrait même accélérer la différenciation des niches des hôtes, en leur permettant de s'adapter à une gamme d'environnements plus vaste (67). Parfois, ces AMGs peuvent évoluer jusqu'à accomplir des fonctions modifiées. Par exemple, le produit du gène viral PebS, xenologue (homologue issu de transfert horizontal) du gène bactérien PebA, est capable d'effectuer non seulement la réaction catalysée par PebA, mais également la réaction catalysée chez les bactéries par PebB (56), deux réactions successives de la voie de biosynthèse de pigments du phycobilisome. Ce dernier exemple met en évidence que les virus peuvent être des réservoirs de diversité génétique pour les cellules, et donc influencent leur évolution.

7 - Les virus : des acteurs de l'évolution des cellules.

Ci-après sont exposés les différents aspects de cette mise à contribution des virus dans l'histoire évolutive des cellules. Ceci va de contraintes sélectives exercées localement, jusqu'aux innovations majeures de l'histoire de la vie.

a - Les théories évolutives de la course aux armes (l'hypothèse de la reine rouge) et la stratégie du chat de Cheshire

La plupart des virus (en particulier les virus à ARN) sont connus pour évoluer rapidement (68, 69). Les mécanismes de défenses des cellules pourraient donc rapidement être obsolètes. Cependant, une sélection continue s'opère chez les hôtes pour s'adapter, ce qui donne lieu à une évolution simultanée et également rapide des gènes cellulaires spécialisés dans la défense contre les virus. C'est l'hypothèse de la reine rouge, qui tire son nom d'un épisode fameux du livre de Lewis Caroll : De l'autre côté du miroir (70) (deuxième volet d'Alice au pays des merveilles) au cours duquel le personnage principal et la Reine Rouge se lancent dans une course effrénée. Alice demande alors : « *Mais, Reine Rouge, c'est étrange, nous courons vite et le paysage autour de nous ne change pas ?* » Et la reine répondit : « *Nous courons pour rester à la même place.* ». Cette hypothèse de la reine rouge (71) est une métaphore qui symbolise la course aux innovations sans fin pour des mécanismes de défenses et mécanismes de contournement de ces défenses, qui a lieu entre des espèces en compétition, ici les virus et leurs hôtes cellulaires.

-Un des micro-eucaryotes les plus abondants dans les océans, *Emiliania huxleyi*, a trouvé le moyen de s'affranchir de cette course aux armements infinie de façon élégante : pendant son cycle de vie, *E.huxleyi* alterne les phases diploïdes (phase végétative) et haploïdes (phase de reproduction sexuée). Sous sa forme diploïde, il est sensible aux infections du virus *E.h* virus. Sous sa forme haploïde, qui serait induite par les attaques virales, *E.huxleyi* est par contraste résistant au virus. Ceci a été nommé la stratégie du « chat de Cheshire », en référence au personnage tiré cette fois du 1er tome d' Alice au pays des merveilles (72) de Lewis Caroll. Dans ce roman, le chat de Cheshire échappe à la décapitation en rendant son corps transparent (L'analogie tient en ce que *E. huxleyi* est « invisible » aux virus sous sa forme haploïde). Ceci a mené des spécialistes du plancton marin des laboratoires de Plymouth en Angleterre, Bigelow laboratory aux Etats-Unis et Roscoff en France à proposer que cette différence de susceptibilité entre les formes haploïdes et diploïdes pourrait être une raison de l'apparition et de la maintenance de la reproduction sexuée chez les eucaryotes : la méiose, dissociée temporellement de la fusion des gamètes, permet que la transmission des gènes à la génération suivante se déroule dans un environnement dénué de virus (73).

-Une autre micro-algue, *Ostreococcus tauri* de la lignée verte, aurait quant à elle évolué un chromosome entièrement dédié à l'immunité contre ses virus (74). Effectivement, une partie importante des gènes différenciellement exprimés entre les lignées d'algues résistantes aux virus et les lignées sensibles sont localisés sur un chromosome en particulier, le 19. Ce chromosome possède par ailleurs des propriétés atypiques comparativement aux autres chromosomes : il est plus petit, son taux de GC est plus bas, et il est particulièrement plastique (74, 75). Il est également enrichi en glycosyltransférases, et ces enzymes sont majoritairement exprimées chez les souches résistantes. Par ailleurs, les souches résistantes ont subi des réarrangements variés au niveau de ce chromosome 19. Ceci a mené l'équipe de l'observatoire océanographique de Banyuls à proposer

que cette forte propension aux réarrangements pourrait être un mécanisme adaptatif pour générer rapidement des variants génétiques capables de résister aux virus. Cette résistance serait acquise par des altérations de l'état de glycosilation du protéome et particulièrement des glycanes de surface (74). L'apparition d'un chromosome spécialisé dans la défense contre les virus est unique, mais peut être comparée aux ilots génomiques hypervariables regroupant des gènes impliqués dans la résistance à leur phage chez les cyanobactéries *Prochlorococcus* (74, 76, 77).

Nous verrons par la suite que les gènes responsables de l'immunité anti-virale eux-mêmes peuvent être hérités des virus.

b - Le cas des virus mutualistes

De plus en plus de cas de mutualisme entre une cellule et un virus sont documentés (78, 79). Dans de telles relations, la présence de virus est (ou a été) requise pour l'adaptation d'un organisme cellulaire à un environnement particulier. Ceci peut subséquemment générer une accélération de l'évolution génétique de l'organisme en question. Voici deux exemples qui m'ont semblés particulièrement parlants :

-Une étude publiée sous le titre « A virus in a fungus in a plant : three-way symbiosis required for thermal tolerance » (80) montre que la thermotolérance de la plante *Dichanthelium lanuginosum* est fournie par l'expression de gènes de son symbionte fongique *C. protuberata.* Important pour nous, cette expression a lieu uniquement lorsque le champignon est infecté par son virus CThTV. Les mécanismes sous-jacents n'ont pas été élucidés.

-Un cas que l'on peut rapprocher d'une symbiogènese (fusion de deux symbiontes, habituellement cellulaires, créant une nouvelle entité), est celui des guêpes parasitoïdes et de leur polydnavirus. Ces guêpes sont elles-mêmes des parasites de chenilles, dans lesquelles elles pondent leurs œufs. Ceci est permis grâce à leur virus mutualiste (figure 3): des gènes de la guêpe suppresseurs des défenses immunitaires de la chenille hôte sont encapsidés dans des particules des virus, qui sont ainsi injectées dans la chenille. Les œufs des guêpes peuvent après quoi se développer jusqu'à maturité (81, 82).

c - Les virus, agents de transferts horizontaux de gènes vers les cellules.

Le séquençage à haut débit et l'analyse des génomes microbiens ont permis de révéler le rôle prépondérant des transferts horizontaux de gènes (HGT) dans le façonnage des génomes procaryotes. Les mécanismes sous-jacents sont bien connus (83–86), et l'un d'entre eux qui implique des bactériophages sera décrit dans le paragraphe suivant. Des HGTs ont également été mis en évidence chez les eucaryotes (87, 88), bien que ces évènements semblent cependant plus anecdotiques (89). Au contraire des HGTs bactériens, les mécanismes sont moins compris. Certains cas impliqueraient des virus, qui sont en effet en contact rapproché avec le génome de leur hôte

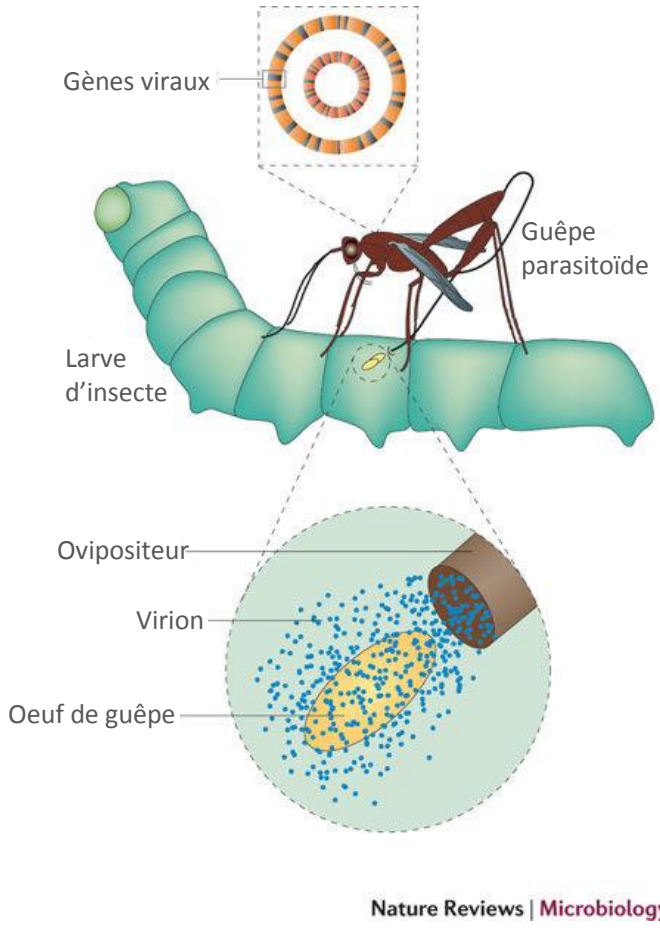


Figure 3 (adaptée de (78)) relations entre polydnavirus, guêpes et chenilles. Quand une guêpe injecte ses œufs, grâce à son ovipositeur, dans une larve de lepidoptère, elle injecte également des virions de son polydnavirus symbiogénique. Ces virions contiennent des gènes de la guêpe qui seront exprimés dans la larve et contreront les défenses immunitaires de l'insecte larvaire.

lors de la réplication. Que l'organisme receveur soit procaryote ou eucaryote, ces transferts de gène intermédiés par les virus peuvent être divisés en deux classes : (i) les virus peuvent véhiculer des gènes cellulaires d'une cellule à une autre. Ils peuvent alors être comparés à des navettes de gènes entre organismes cellulaires. (ii) Alternativement, des gènes viraux (apparus ou modifiés dans une lignée virale) peuvent s'intégrer dans les génomes d'organismes cellulaires.

(i) Les virus, navettes de gènes entre cellules

<u>Bactérie vers bactérie :</u>

-Voici un exemple révélateur de l'importance du phénomène chez les bactéries. 30 % du génome de bactéries pathogéniques ou symbiotiques correspond à des « ilots de pathogénicité » ou des « ilots de symbiose » (83, 85). Or, ces ilots « voyagent » entre bactéries, ce qui constitue autant de transferts. Ceci peut être le fait de bactériophages par un mécanisme appelé transduction (86) : l'ADN de l'hôte bactérien peut être embarqué dans une particule en même temps que le génome du phage. Ceci se produit soit lorsque le prophage a été mal excisé, et donc contient à ses extrémité

de l'ADN de l'hôte, soit lorsque des fragments d'ADN bactériens libres sont encapsidés par erreur. Cet ADN bactérien est ensuite libéré à l'intérieur d'une nouvelle cellule bactérienne lors de l'infection suivante, puis incorporé à son génome (90). Un second exemple est le transport d'une souche bactérienne à l'autre, par leur bactériophages, de plasmides codant pour des gènes de résistance (91, 92).

Eucaryote vers eucaryote :

Plusieurs centaines de transferts latéraux d'élements transposables eukaryote vers eukaryote sont documentés (93). L'un des mécanismes sous-jacents ces transferts, qui demeuraient hypothétiques, a été révélé par un étude publiée en 2014 (94): des virus auraient été utilisés comme vecteurs. En effet, cette étude suggère que 3 espèces de papillons appartenant à un même sous-ordre, et ayant un mode de vie sympatrique, ont échangé des transposons par l'intermédiaire d'un baculovirus.

L'implication de virus dans le transfert de gènes algue vers animal dans le cas de kleptoplastie (une cellule animale récuppère les chloroplastes d'algues dont elle se nourrit) impliquant la limace de mer *Elysia chlorotica* a été proposé (95, 96), mais cette hypothèse semble dorénavant obsolète. Historiquement, la nécéssité de la présence de gènes d'algues chez la limace a été invoquée pour expliquer que certains chloroplastes sont retenus par les limaces pendant la durée exceptionnellement longue de 9 mois, alors qu'une grande partie des gènes nécéssaire à cette maintenance n'est plus codée par les génomes chloroplastiques (ceux-ci ont en effet été transférés vers le génome cellulaire de l'exosymbionte eucaryote au cours du processus d'endosymbiose primaire de la bactérie photosynthétique). L'hypothèse était donc que ces gènes aient été dans le passé transférés depuis une algue vers la lignée germinale de la limace, transferts pour lesquels le concours de virus a été proposé (95, 96). Cependant, le séquençage du génome de la limace (97) a récemment révélé que nul gène d'origine algale n'est présent dans le génome de l'animal. *A fortiori* donc, aucuns de ces gènes n'a été transféré par un virus. La question du maintien des chloroplastes chez l'animal est donc encore ouverte, mais pourrait simplement révéler une capacité de maintien des chloroplastes d'algues supérieure à celle des chloroplastes de plantes supérieures, plus largement étudiées (98). En particulier, la protéine fstH, une protéase impliquée dans la réparation constante de la protéine D1 du photosystème II, est codée par le noyau chez les plantes supérieures, mais est toujours codée par le génome chloroplastiques chez les algues (99), et pourrait être suffisante a la longévité des kleptoplastes.

Bactérie vers eucaryote :

-L'analyse de la distribution du gène de réparation de l'ADN MutS7, ainsi que la reconstruction de sa phylogénie (Figure 4) a mené H.Ogata et J.M Claverie à proposer qu'un virus ait pu servir de navette entre les bactéries epsilon et les mitochondries d'octocoralliaires (100–102) (je vous accorde que l'état eucaryote du récipiendaire peut être contesté, car le gène a été transféré au

génome de l'organelle dérivé d'une bactérie que constitue la mitochondrie). De plus, les mitochondries d'octocoralliaires pourraient ici bénéficier d'une capacité particulière de cette version de MutS7 que ne possèdent pas les classiques MSH mitochondriales : celle de corriger les erreurs d'appariement (mismatch), qui est due à un domaine additionnel d'endonucléase (100).

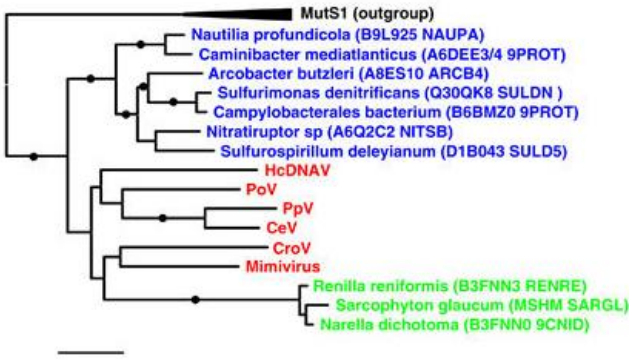


Figure 4. Arbre du maximum de vraissemblance des proteins MutS1 et MutS7. Les espèces en bleue sont des bactéries epsilon, en vert des octocorraliaires. Les virus sont notés en rouge. La barre d'échelle représente 0,5 substitutions par sites. Tirée de (101).

D'autre part, les virus peuvent également agir comme un moteur de diversité génétique pour des gènes cellulaires. Ce processus se déroule en trois étapes. Tout d'abord, un gène cellulaire est capturé par le virus par transfert horizontal, puis modifié dans la lignée virale par accumulation de mutations ou additions de domaines. Le gène viral modifié est alors réintroduit dans une lignée cellulaire lors d'un nouveau transfert horizontal en sens inverse, participant ainsi à l'augmentation de diversité biologique. Parfois, le gène transféré d'origine virale peut remplacer son homologue cellulaire lorsque ce dernier vient à être éliminé par délétion ou mutation délétère. Le cas du composant des photosystèmes hli, évoqué précédemment, est une illustration de ce processus. Celui de la protéine de fusion Mre11-Rad50 en constitue une seconde : (103) le complexe Mre11/Rad50 joue un rôle crucial dans la réparation des cassures double brin de l'ADN. Mre11 et Rad50 sont codés par deux gènes distincts chez la majorité des eucaryotes. La protéine de fusion est majoritairement trouvée chez les virus, mais également chez quelques organismes cellulaires. La distribution taxonomique sporadique du gène fusionné chez ces derniers suggère que ces virus ont joué un rôle dans la formation du gène fusionné, et dans sa propagation subséquente dans différentes lignées cellulaires (103). Des cas ont été décrits impliquant également une fusion : les primases et hélicases d'origine bactérienne responsables de la réplication de la mitochondrie ont été remplacées par une protéine de phage qui possède ces deux domaines fusionnés (104). Enfin l'ARN polymérase mitochondriale est différente de l'enzyme bactérienne ancestrale aux multiples sous-unités qui a été remplacée (sauf chez les jakobea) par une ARN polymérase de phage (105).

(ii) Les transferts de matériel génomique viral vers les cellules

Chez les eucaryotes, les séquences nucléiques dérivées d'éléments mobiles ou de virus endogènes représentent au moins 50% des génomes de mammifères, et ce pourcentage va jusqu'à 90% chez les plantes (84). Ces séquences dérivées d'éléments égoïstes peuvent être recrutées par les cellules de plusieurs façons. Tout d'abord, ceci peut fournir à la cellule du matériel génétique

additionnel d'où peuvent émerger des innovations génétiques, après sa transformation due à des mutations, des délétions, des insertions, ou encore des évènements de recombinaisons, et/ou la sélection naturelle. D'autre part, les éléments mobiles répétés peuvent être utilisés comme éléments de régulation des gènes. Ils peuvent effectivemement fournir une séquence motif qui sera reconnue par un facteur de transcription ce qui induira la coexpression de ces gènes. Enfin, ce sont les gènes en tant que tels présents sur ces séquences qui peuvent être recrutés par la cellule. Quelques cas de domestication de gènes viraux sont effectivement documentés. Cette domestication peut être définie comme suit : l'acquisition d'un gène viral est suivie d'une « fonctionalisation » chez la cellule (l'apparition de séquences promotrices en amont du gène afin de réguler son expression ; ou encore sa fusion à un gène cellulaire, sont des exemples de fonctionnalisations), et enfin le gène est fixé dans la population. Un cas célèbre est celui du gène de la syncitine dérivé d'un rétrovirus, qui est impliqué dans la formation du placenta chez les mammifères (106). Plus récemment décrit, la protéine responsable de la fusion des gamètes lors de la reproduction sexuée serait dérivée d'une protéine de virus médiatrice de la fusion entre la membrane virale et la membrane de la cellule hôte (107). Cette protéine est présente chez plusieurs des clades eucaryotes : les plantes, les animaux, et plusieurs groupes de protozoaires, suggérant que le transfert depuis le virus a eu lieu au temps de LECA (last Eucaryotic Common ancestor). Un dernier exemple suggère que des protéines virales sont responsables de la structure des chromosomes des dinoflagellés. En effet, de façon unique parmi tous les eucaryotes séquencés, les dinoflagellés ne possèdent pas d'histones. Leur ADN est néanmoins compacté (de façon constante d'ailleurs). Ceci serait assuré par une protéine originaire des *Phaeovirus*, dont la fonction chez les virus n'est pas connue (108).

Les virus sont détenteurs d'une part de la diversité génétique, du fait de leur biologie unique. Les nombreux ORFans des génomes viraux, ainsi que les repliements protéiques uniques aux virus, en témoignent (15, 109). Théoriquement, toute cette diversité génétique pourrait être recrutée par les cellules, par des mécanismes analogues à ceux sous-jacents aux exemples présentés précédemment.

d - Les virus médiateurs de transition évolutive.

Un article conceptuel a été publié par Koonin (110) dans lequel il fait l'hypothèse que les virus et les éléments mobiles ont pu être des médiateurs de plusieurs des « transitions évolutives majeures» *sensu* Maynard-Smith et Szathmary (111, 112). Rapidement, la propriété clé de ces transitions est le regroupement de réplicateurs indépendants pour former une entité subissant un niveau supérieur de sélection. Des gènes se réunissent en « proto-cellule », des procaryotes se joignent pour constituer la cellule eucaryote, des protistes se regroupent pour former des organismes multicellulaires, etc. Pour que le succès de la transition soit assuré, l'évolution au niveau inférieur doit être contrainte par le niveau supérieur (111). La transition procaryotes-eucaryotes par exemple, pourrait être le fait d'un virus ; il a effectivement été proposé qu'un virus à ADN double brin soit à l'origine du noyau (112). Cette hypothèse a été appelée l'eucaryogénèse virale. A côté de « l'acquisition » du noyau, le second évènement crucial dans l'émergence des

eucaryotes fut l'acquisition d'une mitochondrie. Plus démonstrativement que dans le cas du noyau, des indices suggèrent que des virus ont été impliqués dans cette étape également. Effectivement, le remplacement, par des enzymes de phages, des enzymes clés des machineries de réplication et de transcription de la mitochondrie primitive, aurait permis la domestication de l'alphaprotéobactérie juste « engloutie ».

Il a été proposé que L'ADN lui-même ait pu être inventé par les virus, pour se prémunir de la dégradation des ARNs par les Rnases cellulaires (dans le contexte d'un monde à ARN) (113). Une hypothèse étayée par le fait que l'ADN de certains phages ne contient pas de thymidine mais de la deoxyuridine (5, 114). L'hypothèse d'une origine virale des protéines de réplication eucaryotes a également été évoquée, et pourrait expliquer leurs phylogénies atypiques (112).

En résumé, les virus semblent jouer et avoir joué un rôle prépondérant dans l'évolution des cellules, des gènes, des écosytèmes. Ils sont moteurs et acteurs d'évolution.

B - Les grand virus nucléo-cytoplasmique à ADN (NCLDVs)

Nous allons maintenant faire un focus sur un groupe de virus en particulier, qui sont étudiés au laboratoire. Ils sont appelés en anglais nucleo-cytoplasmic large DNA viruses (NCLDVs). Pour vous présenter les NCLDVs, je commencerai par un historique de la recherche sur ce groupe viral, ce qui me permettra de décrire succinctement les différentes familles taxonomiques qui le compose. Nous discuterons ensuite des caractéristiques des NCLDVs qui ont récusées certains dogmes historiques de la virologie.

1 - Historique

Le groupe taxonomique des « NCLDVs » a été introduit dans la classification des virus en 2001 (115) pour regrouper 4 familles de virus à ADN double brin infectant des eucaryotes : les *Poxviridae* (date de création : 1971 ; membre prototype : vaccinia virus), les *Iridoviridae* (1975 ; *Tipula* iridescent virus), les *Phycodnaviridae* (1990 ; *Paramecium bursaria Chlorella* virus) et les *Asfarviridae* (1998 ; African swine fever virus). Leur monophylie (caractéristique d'un groupe contenant tous les « descendants » d'un ancêtre) a été inférée sur la base de l'étude phylogénétique de 9 gènes « core » (présents dans toute les familles), et de 22 gènes présents dans au moins trois des familles (115).

Les NCLDVs se répliquent dans des structures appelées « usines à virions », qui sont formées dans le cytoplasme de la cellule hôte lors de l'infection. Le cycle de réplication peut se dérouler entièrement dans le cytoplasme, ou bien être amorcé par une phase nucléaire si le virus n'est pas autonome pour effectuer la transcription de ses gènes.

Suite au groupement initial de ces 4 familles, plusieurs nouvelles familles de virus d'eucaryotes ayant un cycle réplicatif cytoplasmique ou nucléo-cytoplasmique ont été ajoutées aux NCLDVs : les *Ascoviridae,* tout d'abord, dont le membre prototype est *Spodoptera frugiperda* ascovirus, puis se sont succédées en 10 ans, les découvertes de 5 nouvelles familles de virus isolés à partir des amibes.

Comment s'explique ce soudain isolement de nouvelles familles, comment peut-on expliquer que la virologie les avait jusque-là ignorées ? Pour le comprendre, retournons en 1992, soit 11 ans avant la caractérisation de Mimivirus, le tout premier virus de cette série d'isolements. Rappelons-nous qu'alors, le critère principal utilisé pour décider de la nature virale d'une entité était celui de sa « filtrabilité »: sa capacité à traverser les pores de 200 nm du filtre de Chamberland. Les virus étaient par conséquent supposés invisibles au microscope optique. Cette année-là, un micro-organisme provenant d'une tour de refroidissement d'un hôpital à Bradford en Angleterre est isolé et observé au microscope optique : il ressemble à une bactérie Gram positive en forme de coque et a donc été baptisée *Bradfordcoccus*. Cependant, il n'est pas cultivable en milieux bactériens standards, mais seulement en coculture avec des amibes du genre Acanthamoeba. Toutes les tentatives pour amplifier l'ARN ribosomal 16S de *Bradfordcoccus* échouent, malgré l'utilisation d'amorces PCR « bactériennes universelles ». La caractérisation du microorganisme est laissée en suspens jusqu'en 2003. *Bradfordcoccus* est alors de nouveau analysé, au microscope électronique cette fois ; les corps réguliers et icosaédriques observés dans les amibes infectées sont réminiscents des iridovirus. De plus, ils émergent à la surface de structures intracellulaires qui rappellent les usines à virions de ces mêmes iridovirus. Le séquençage de son génome le confirme : il s'agit bien d'un virus, de 450 nm de diamètre (700 nm, si l'on ajoute les fibres entourant la particule). *Bradfordcoccus* est renommé *Acanthamoeba polyphaga* mimivirus. Il est le premier virus observable au microscope optique et qui ne passe pas à travers le filtre de Chamberland. Il est par ailleurs le membre fondateur de la famille des *Mimiviridae*. Débarrassés du « chausse-pied mental » [1], une question : Y en a-t-il d'autres ? La chasse aux virus géants démarrait.

A partir d'échantillons environnementaux de provenances variées (sols, eaux marines et douces), et en utilisant l'amibe comme hôte, sont successivement isolés et décrits plusieurs virus de très grande taille : *Acanthamoeba Polyphaga* Marseillevirus (117) est le membre fondateur des *Marseilleviridae*, famille reconnue par l'ICTV (International Committee on Taxonomy of Viruses). Les deux *Pandoravirus dulcis* et *P. salinus* (118), *Pithovirus* (119) et enfin *Mollivirus*

[1] Cette expression a été employée par Stephen Jay Gould dans le cadre de la reclassification des fossiles de la faune de Burgess, qui avaient été classés _a tort_ dans des clades d'animaux contemporains ou déjà connues (116)

(120) n'ont pas encore de famille officielle respective. Cependant, les familles *Pandoraviridae*, les *Pithoviridae* et *Molliviridae* sont régulièrement citées dans la litterature. En parallèle, le développement de la métagénomique (121–123) et du « single virus sequencing » (i.e., le séquençage de genomes de virus provenant d'échantillons environementaux après l'isolement de chaque particule individuelle) (124, 125) met en évidence une diversité virale immense qui reste aujourd'hui encore largement cachée.

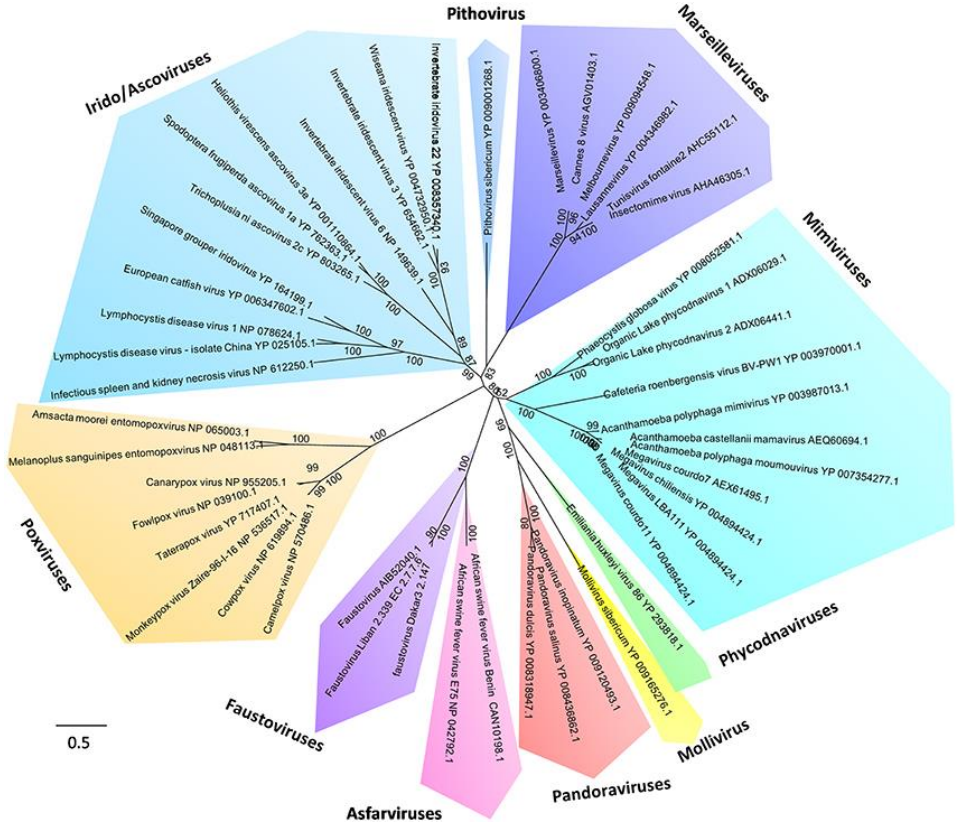


Figure 5: Les dix familles connues de NCLDVs. Cette figure est tirée de (126). La reconstruction phylogénétique est basée sur les ADN polymérases ARN-dépendante.

Étendue à 10 familles (de virus isolés) (voir tableau 1 et figure 5), le nombre de caractères partagés par tous les NCLDVs a chuté. Ils présentent effectivement une diversité importante. Tout d'abord, leur niveau de dépendance à l'hôte varie : les *Mimiviridae* (127), les *Poxviridae* (128) et les *Pithoviridae* (129) codent pour la totalité des machineries de réplication et de transcription. De plus, lors de l'assemblage, la machinerie de transcrition est embarquée dans les particules, ce qui leur permet de démarrer leur cycle sans avoir besoin d'utiliser celle de l'hôte, localisée dans le noyau. Leur cycle est ainsi exclusivement cytoplasmique. Les *Pandoraviridae*, les *Phycodnaviridae* (130), les *Asfarviridae* (131), les *Ascoviridae* et les *Iridoviridae* (132) font par contre un passage obligatoire par le noyau. Certains *Marseilleviridae* enfin, semblent avoir un cycle intermédiaire (133). Ensuite, la forme de leur virion peut être soit ovoïde, soit icosaédrique. De

Tableau 1: Caractéristiques des dix familles connues de NCLDVs

Famille de NCLDV	**Hôte naturel ou expérimental**	**Taille du génome (kb)**	**Forme du virion; taille (nm)**	**réplication**
Asfarviridae (faustovirus)	Porc (Amibe)	170-190 (466)	Icosaédrique; 170-200 (250)	Cytoplasmique, noyau requis
Poxviridae	Vertébrés et arthropodes	130-160	Ovoïde de 300 nm de long	Cytoplasmique
Iridoviridae	Invertébrés et vertébrés à sang froid	140-303	Icosaédrique; membrane interne; 120-350 nm	Phase nucléaire puis phase cytoplasmique
Phycodnaviridae	"microalgues"	160-560	Icosaédrique; 100-220	Phase nucléaire puis phase cytoplasmique
Ascoviridae	Invertébrés	150-186	Ovoïde de 200 nm de long	Phase nucléaire puis phase cytoplasmique
Mimiviridae	Acanthamoeba; Stramenopile; Haptophytes; Chlorophytes	371-1259	Icosaédrique; 140-680	Cytoplasmique
Marseilleviridae	Acanthamoeba	350-390	Icosaédrique; 250	Cytoplasmique, noyau requis
Pandoraviridae	Acanthamoeba	1900-2500	Ovoïde; 1000 x 500	Cytoplasmique, noyau requis
Pithoviridae	Acanthamoeba	610	Ovoïde; 1500 x 500	Cytoplasmique
Molliviridae	Acanthamoeba	650	Sphérique; 500-600	Cytoplasmique, noyau requis

plus, leur mode d'entrée dans la cellule hôte est également diverse : attachement à la paroi cellulaire de l'hôte, comme décrit pour les *Phycodnaviridae* (134) ou internalisation par phagocytose, comme chez les *Marseilleviridae*… La pertinence de cette classification est donc contestée, et il a par exemple été proposé de créer un nouvel ordre, les Megavirales, sur la base de la présence de gènes ancestraux conservés et d'une capside icosaédrique formée de la protéine « β-barrel jelly roll » (135). Cet ordre exclurait les *Poxviridae*, les *Ascoviridae*, ainsi que les *Pandoraviridae*, les *Pithoviridae* et Mollivirus, car ils ne possèdent pas cette protéine de capside. L'ordre des Megavirales n'est à ce jour pas reconnu. Malgré ces variations parmi les NCLDVs, tous construisent une usine virale dans le cytoplasme de l'hôte pour leur réplication. Ceci est le cas également pour certains virus à ARN (136) mais pour aucun autre des virus à ADN connus. Des gènes sont universellement présents chez les NCLDVs, une étude annonçait le chiffre de 5 (la protéine de capside principale appelée MCP pour Major Capsid Protein, la D5 primase-helicase,

DNAP, A32-like packaging ATPase, et un facteur de transcription appelé VLTF3 pour Very Late Transcription factor 3), mais il semblerait que les *Pandoraviridae* ne possèdent pas de MCP, alors que les *Pithoviridae* ne possèdent ni de MCP, ni de packaging ATP-ase. Cependant, la répartition des protéines homologues (clusters of orthologous groups COGs) au sein des NCLDVs semble justifier leur appartenance à un clade viral commun (137), même si des hypothèses alternatives existent et seront mentionnées plus loin.

2 - Pourquoi les NCLDVs, et particulièrement les géants, défraient-ils la chronique ? Un aperçu des dogmes rompus et des points intrigants.

Taille de particule : Cause de leur anonymat passé : la grande taille de leurs particules qui sont filtrées dans les filtres à 200 nm. Elle atteint 1500 nm chez Pithovirus, qui est donc plus grand que certaines bactéries libres.

Autonomie : En parallèle du gigantisme de leur particule, ces virus ont de très grands génomes (le record revient a *Pandoravirus salinus* avec 2,77 Mb), qui codent pour un grand nombre de protéines (jusqu'à 2556, également pour *P. salinus*). Ces protéomes complexes permettent à certains de ces virus d'être autonomes pour la réplication de leur génome et pour la transcription de leurs gènes, qui peut par ailleurs être finement régulée (129, 138).

Enzymes de la traduction : Ils possèdent des enzymes impliquées dans le processus de la traduction, et notamment des aminoacyl-ARNt synthétase (AARS). Ces protéines lient un acide-aminé spécifique à l'ARN de transfert (ARNt) correspondant, avant que le premier soit ajouté à la chaine peptidique naissante. On pensait précédemment que ces protéines étaient exclusives aux cellules. Ceci alimente donc le débat sur une origine cellulaire de ces virus. Cependant, pour la majorité de ces gènes au moins, une acquisition par transfert horizontal, plutôt qu'un héritage vertical, semble être l'explication d'une telle présence (139).

Les ORFans : A côté de cela, les protéomes de ces virus contiennent une large proportion d'ORFans, protéines n'ayant pas d'homologues dans les bases de données. Quels métabolismes ou quelles structures particulières sont codés par ces ORFans ? Et au-delà, quelle est l'origine de ces ORFans, qui n'ont pas d'homologues chez les cellules connues ? Pour certains, ils sont l'indice d'une origine ancienne de ces virus, et distincte des cellules actuelles (the *virus first hypothesis*). Ils mettent au défi « *l'escape theory* », dans le sens où ils ne semblent pas avoir été acquis par transferts de gènes depuis les cellules.

Ancienneté : Cette ancienneté potentielle est soutenue par d'autres indices : Tout d'abord, les NCLDVs, et même certaines familles en particulier comme les *Mimiviridae*, infectent plusieurs des super-groupes eucaryotes. A cela, deux explications ont été proposées (140). Soit l'ancêtre de

ces virus infectait une branche eucaryotes en particulier et il y aurait eu des changements d'hôte d'une branche de la phylogénie eucaryote vers les autres ; ou bien il infectait déjà l'ancêtre de tous les eucaryotes actuels et la diversité contemporaine des NCLDVs serait le fruit d'une coévolution des virus et de leurs hôtes eucaryotes. Cette dernière hypothèse est privilégiée parce-que (i) la possibilité de nombreux « sauts d'hôtes » très éloignés par des virus demeure spéculative, et (ii) les reconstructions phylogénétiques effectuées à partir de protéines communes aux virus et aux cellules, impliquées dans la réplication et la réparation de l'ADN, ainsi que dans la transcription, présentent un clade viral formant une branche basale au côté des branches cellulaires (141). Ces résultats lancèrent la controverse du 4ème domaine de la vie : les NCLDVs formeraient une lignée sœur aux lignées bactériennes, archées et eucaryotes. Pour d'autres cependant, ces arbres sont mal interprétés : ces virus seraient certes ancestraux aux cellules, mais représenteraient plutôt les vestiges de plusieurs lignées cellulaires alors autonomes, et non pas d'un 4ème domaine. Leur regroupement dans les arbres phylogénétiques serait donc un artefact dû à leur antériorité – commune – aux cellules actuelles. Enfin, les gènes en question pourraient avoir été acquis par transfert de gènes (142). L'argument évoqué précédemment dans l'introduction générale provenant de l'analyse des repliements protéiques est également valide pour les NCLDVs.

Leurs particules contiennent et de l'ADN et de l'ARN, alors qu'il était posé que les virions ne possèdent qu'un seul type d'acide nucléique (143).

Ils possèdent des « structures » biochimiques et morphologiques unique : (i) La Cu-Zn superoxide dismutase de Megavirus est capable d'incorporer des facteurs métalliques sans l'aide de chaperonne (144), ce qui lui est exclusif; (ii) Le génome de *Acanthamoeba polyphaga* mimivirus code pour une nucléoside diphosphate kinase (NDK), qui, au contraire des NDKs cellulaires qui fonctionnent pour tous type ribo- and desoxyribonucléotides, exhibe une forte affinité pour les desoxyribonucléotides. La phylogénie serait également en faveur d'une origine ancestrale de l'enzyme (145) ; (iii) La présence de nombreux ORFans, qui pourraient représenter des innovations virales, a été évoquée. La détermination de la structure et de la fonction de tous ces ORFans est un travail de titan, juste entamé. Lorsque ces domaines ORFans sont associés, sous forme de protéines de fusion, à des domaines connus, ceci fournit des indices sur un rôle potentiel, et d'autre part, indique une spécificité de l'enzyme virale. C'est le cas de la sulfhydrile oxydase de Mimivirus, à laquelle est fusionné l'un de ces ORFans. La structure de ce domaine suggère qu'il pourrait être impliqué dans la localisation de l'enzyme au sein de l'usine virale ou du virion, ou bien dans son attachement à sa protéine cible (146) ; (iv) Certains membres de la famille des *Mimiviridae* encodent des voies de biosynthèse de sucre très inhabituels qui par exemple composent les fibres entourant leur particule. Ces voies sont de plus constituées d'enzymes différentes des homologues cellulaires (147–150).

Le « mobilome » des NCLDVs.

Le terme “mobilome” fait référence à l’ensemble des éléments mobiles génétiques. Les NCLDVs possèdent un mobilome varié, regroupant différentes classes d’éléments mobiles. Certaines sont communes aux cellules et aux NCLDVs, comme les intéines ; d’autres sont spécifiques aux NCLDVs, comme les virophages, les transpovirons ou encore MIGE.

(i) Les virophages : Plusieurs NCLDVs de la famille de *Mimiviridae* sont eux-mêmes les cibles d’autres dsDNA virus, les petits (diamètre : 35-60nm) virophages. Ceux-ci co-infectent les cellules hôtes eucaryotes avec leur *Mimiviridae*-hôte associé, et se répliquent dans l’usine virale formée par le ce dernier dans le cytoplasme de la cellule (151). Le spectre d’hôte viral des virophages est variable, mais, pour les cas étudiés, reste cantonné à des *Mimiviridae* d’un même genre (152). La plupart des virophages ont un effet négatif sur la réplication des NCLDVs, qui voient leur taux de multiplication fortement réduit, ou produisent des particules malformées (151, 153, 154). Ces observations ont donné lieu à une petite controverse sur le statut des virophages : représentent-ils ou non une nouvelle entité, comparativement aux virus satellites ? Effectivement, ces derniers ont également besoin d’un « *helper virus* » pour se répliquer et peuvent impacter négativement la réplication de ce dernier. C’est le cas par exemple pour le virus satellite de la nécrose du tabac (155). D’après Mart Krupovic et Virginija Cvirkaite-Krupovic (156), les virophages ne méritent donc pas un statut différent des virus satellites. Matthias Fischer par contre ne se satisfait pas de cette assimilation (157). D’après lui, les deux points suivants suggèrent qu’une distinction mérite d’être faite: (i) au contraire des virus satellites, les virophages ne sont pas des particules défectueuses mais des virus fonctionnels qui encodent leur propre polymérase, (ii) il semblerait que les virophages détournent la machinerie de leur virus-hôte plutôt que celle de cellule-hôte pour effectuer leur cycle réplicatif. Pour clore cette controverse, il a été proposé de créer plusieurs nouveaux genres et familles pour classifier les virus satellites et les virophages (158). Ainsi, les virophages connus, qui notoirement sont les seuls virus qui sont dépendants d’autres virus et dont le génome est une molécule dsDNA, constitueraient la nouvelle famille des *Lavidaviridae*. Au delà de ces questions de classification, des simulations de l’interaction [cellule hôte]-[virus hôte]-virophage suggèrent que du fait de la susceptibilité des Mimiviridae à certains virophages, ces derniers pourraient avoir un rôle protecteur sur les populations de cellules hôtes (159, 160) et donc jouer un rôle dans l’équilibre global des populations de cellules et de NCLDVs. D’autre part, des analyses phylogénétiques de l’intégrase de type rve du virophage Mavirus (161) (isolé lors d’une infection de *Cafeteria roenbergensis* par son virus CroV) avaient montré que cette enzyme est apparentée à des homologues d’éléments mobiles eucaryotes de la famille des Maverick/Polintons (MP) (162, 163). Par ailleurs, il semblerait que les MPs encodent des protéines apparentées à des protéines de capside (164). Ceci avait mené à l’hypothèse que les MPs sont de *bona fide* virus qui auraient émergé d’ancêtre de virophage ayant acquis la capacité de transposition (161). L’hypothèse inverse postulant que les virophages aient pu émerger d’un MP a été proposée par ailleurs (165), et implique également que les virophages aient pu avoir la capacité d’intégrer leur génome dans celui de leur hôte eucaryote.

(ii) Les transpovirons. Les transpovirons forment une classe distincte d'éléments génétiques mobiles associés aux *Mimiviridae* et se répliquant dans les usines à virions. Ils sont linéaires, longs de 6,5 à 7,5 kb et contiennent de 6 à 8 gènes codant pour des protéines. Ils sont flanqués à leurs extrémités de séquences répétées inversées (TIR Terminal Inverted Repeat) (166).

(iii) MIGE (Major Interspersed Genomic Element). Cet élément possède un cadre ouvert de lecture sans homologues connus. Il est présent en plusieurs copies dans les génomes de deux membres de la famille des *Mimiviridae* (167) dans chacun desquels il se serait propagé indépendamment (Gallot-Lavallée et al 2017). Le mécanisme par lequel cet élément se duplique et ou se déplace dans le génome est inconnu.

(iv) Les intéines. Les intéines sont des éléments génétiques que l'on peut qualifier d' « introns protéiques » : ils sont traduits en même temps que leur protéine hôte puis s'en auto-excisent. La protéine hôte est alors fonctionnelle. Les intéines encodent une endonucléase responsable de l'intégration de leurs acides nucléiques dans des sites génomiques spécifiques et par là de la colonisation de nouvelles protéines hôte (168). Plusieurs *Mimiviridae*, Phycodnaviridae et Iridoviridae arborent des intéines (169–171).

Ce mobilome varié associé aux NCLDVs pourrait être un acteur clé de l'évolution des génomes viraux, comme c'est le cas chez les cellules procaryotes et eucaryotes. D'autre part, il pourrait être médiateur de transferts horizontaux de gènes.

Un dernier point concerne spécifiquement la famille des *Mimiviridae*.

Un spectre d'hôte et une diversité hors du commun : Comme précédemment expliqué, la famille des *Mimiviridae* a été historiquement créée lors de l'isolement à partir d'une amibe du genre Acanthamoeba du premier virus géant Mimivirus. Toujours isolés à partir de l'amibe, d'autre *Mimiviridae* (dont certains plus gros encore que Mimivirus) ont ensuite été caractérisés, et forment 2 lignées sœurs de Mimivirus (voir figure 1 de l'article 1). Les membres fondateurs de ces clades sont Megavirus chiliensis (172) et Moumouvirus (173). Une étape importante dans l'étude de cette famille fut l'analyse des données métagénomique du projet Tara Océan, qui suggérait que le spectre d'hôte des *Mimiviridae* ne se limitait pas aux amibes : des séquences apparentées aux *Mimiviridae* étaient trouvées en abondance dans les échantillons marins (121). Cette prédiction se révélait exacte puisque plusieurs virus marins apparentés aux *Mimiviridae* furent successivement isolés. *Cafeteria Roenbergensis* virus (CroV) (174) infecte le stramenopile hétérotrophe *Cafeteria Roenbergensis*, *Aureococcus anophagefferens* virus (AaV) (175) infecte l'algue pelagophyte *Aureococcus anophagefferens* et *Phaeocystis globosa* Virus (PgV) (167) l'algue haptophyte *Phaeocystis globosa*. Avec ces nouveaux virus, le spectre d'hôte connu des *Mimiviridae* est devenu particulièrement étendu, puisque ces protistes ont divergés des amibes au tout début de l'histoire eucaryotes (il y a plus de 1,5 milliard d'années (176) ; la datation de la radiation primitive des eucaryotes est néanmoins sujette à débat) et appartiennent par ailleurs eux-mêmes à plusieurs des

grands groupes eucaryotes. D'autre part, ces hôtes ont des modes de vie très divergents, puisque certains se nourrissent par phagotrophie (les amibes, *Cafeteria Roenbergensis*) quand d'autres sont photosynthétiques (*Phaeocystis globosa*, *Aureococcus anophagefferens*). L'analyse phylogénétique de cette famille a par ailleurs révélé que :

- Les *Mimiviridae* infectant les micro-algues forment un groupe monophylétique d'une part, les virus infectant des organismes hétérotrophes forment un groupe monophylétique d'autre part.

- Les *Mimiviridae* sont de tailles très diverse (AaV : 140 nm de diamètre et 371 Kb ; Megavirus chilensis : 680 nm de diamètre, 1259 Kb), et se répartissent comme suit : les plus gros infectent les amibes, les plus petits infectent les micro-algues.

C - Introduction à mon travail de thèse

Afin d'améliorer nos connaissances concernant la biodiversité, l'écologie et l'évolution des NCLDVs, la stratégie principale adoptée depuis la découverte de Mimivirus a été de se servir de l'amibe pour isoler de nouveaux virus géants et mettre à jour de nouvelles familles. Pendant ma thèse, nous avons utilisé deux autres approches, qui ont constitué les deux volets de mon travail.

(A) Premièrement, nous avons choisi de faire un focus sur une famille de NCLDVs en particulier : les *Mimiviridae*. La diversité et l'évolution au sein de cette famille a été interrogée grâce à l'analyse du génome d'un nouveau membre, *Chrysochromulina ericina* Virus (CeV).

(B) Deuxièmement, nous nous sommes interessés aux échanges génétiques entre les NCLDVs, les virophages et leurs hôtes cellulaires, en nous concentrant sur les transferts orientés virus vers hôtes eucaryotes. Effectivement, des séquences d'origines NCLDVs présentes dans les génomes eucaryotes pourraient nous permettre de révéler de nouveaux couples hôtes/virus ainsi que des capacités inédites de ces derniers. D'autre part, ceci nous a permis d'aborder également la question de l'impact des NCLDVs sur l'évolution des génomes cellulaires.

Résultats et discussion

Cette partie est organisée en deux chapitres correspondants aux deux volets de ma thèse. Chacun a donné lieu à la publication d'articles scientifiques : 1 pour le volet A, 2 pour le volet B. Ils constituent le corps des résultats et sont donc intégrés dans les chapitres correspondants. À chacun des deux articles du volet B correspond une sous-partie de B. Les articles sont suivis de quelques résultats supplémentaires et de discussions.

Chapitre A - CeV et les *Mimiviridae*

Article 1: *Comparative genomics of Chrysochromulina Ericina Virus (CeV) and other microalgae-infecting large DNA viruses highlight their intricate evolutionary relationship with the established Mimiviridae family*

Chapitre B - Séquences de NCLDVs et virophages dans les génomes eucaryotes

Article 2: *Provirophages in the Bigelowiella genome bear testimony to past encounters with giant viruses* (sous partie 2 - *Bigelowiella natans* – un génome analysé en détail)

Article 3: *A Glimpse of Nucleo-Cytoplasmic Large DNA Virus Biodiversity through the Eukaryotic Genomics Window* (sous-partie 3 - Un crible global des eucaryotes)

Chapitre A - CeV et les *Mimiviridae*

1 - Introduction

J'ai introduit précédemment la famille des *Mimiviridae,* qui regroupe un clade de virus « géants » et un clade de virus plus petits dont les membres infectent des micro-algues. L'intérêt du laboratoire est de comprendre la biodiversité et l'évolution au sein de cette famille.

Or, un virus, Chrysochromulina ericina Virus (CeV), a été isolé précédemment par le département de microbiologie de l'université de Bergen (Norvège), et se révèle être apparenté aux *Mimiviridae* infectant les micro-algues, PgV et AaV. Ceci a été mis en évidence par une analyse phylogénétique préalable. L'hôte de CeV est l'haptophyte *Haptolina* (anciennement *Chrysochromulina*) *ericina* (Figure 6). Cette algue a une distribution planétaire et peut occasionnellement former des efflorescences massives (177). CeV se réplique au sein du cytoplasme de son hôte. Son cycle lytique dure de 14 à 19 heures, et abouti à la libération de 1800 particules environ (177). Ces particules sont de forme icosaédrique et leur diamètre mesure 160 nm.

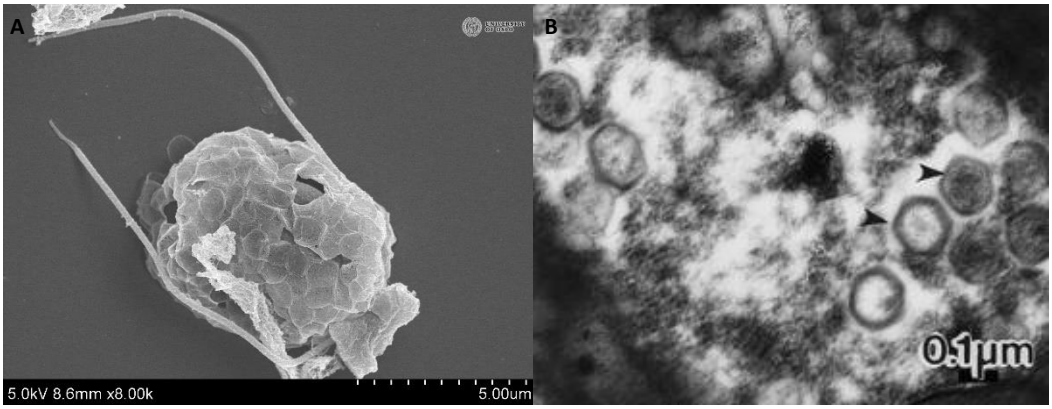


Figure 6. Haptolina ericina et CeV. A/ Photographie de microscopie électronique à balayage de *Haptolina ericina*, l'hôte de CeV. Cette image est tirée du site « BioMarks Data Portal » associé au projet BioMarks. B/ Fine section d' *Haptolina ericina* infectée par CeV. Les particules virales sont marquées de flèches noires. Photographie tirée de (177).

A travers une collaboration internationale, le laboratoire a pris en charge le séquençage et l'analyse du génome de ce nouveau virus. J'ai conduit l'analyse des séquences génomiques. Ces données m'ont permis de caractériser le répertoire génétique et le métabolisme de ce virus. Par ailleurs, j'ai aussi réalisé une étude comparative entre CeV, ses plus proches parents, et les

Mimiviridae « géants », afin d'adresser la question de l'évolution de ces génomes, et de mettre au jour des spécificités qui pourraient être reliées à l'adaptation à leurs hôtes. Les résultats de mes travaux ont fait l'objet de la publication suivante.

2 - Article 1

Comparative genomics of Chrysochromulina Ericina Virus (CeV) and other microalgae-infecting large DNA viruses highlight their intricate evolutionary relationship with the established *Mimiviridae* family. Gallot-Lavallee, L., Blanc, G., and Claverie, J.-M. (2017).

AMERICAN SOCIETY FOR MICROBIOLOGY | Journal of Virology®

GENETIC DIVERSITY AND EVOLUTION

Comparative Genomics of Chrysochromulina Ericina Virus and Other Microalga-Infecting Large DNA Viruses Highlights Their Intricate Evolutionary Relationship with the Established *Mimiviridae* Family

Lucie Gallot-Lavallée,[a] Guillaume Blanc,[a]* Jean-Michel Claverie[a,b]

Information Génomique et Structurale, UMR 7256 (IMM FR 3479) Centre National de la Recherche Scientifique & Aix-Marseille University, Marseille, France[a]; Assistance Publique des Hôpitaux de Marseille, La Timone, Marseille, France[b]

ABSTRACT Chrysochromulina ericina virus CeV-01B (CeV) was isolated from Norwegian coastal waters in 1998. Its icosahedral particle is 160 nm in diameter and encloses a 474-kb double-stranded DNA (dsDNA) genome. This virus, although infecting a microalga (the haptophyceae *Haptolina ericina*, formerly *Chrysochromulina ericina*), is phylogenetically related to members of the *Mimiviridae* family, initially established with the acanthamoeba-infecting mimivirus and megavirus as prototypes. This family was later split into two genera (*Mimivirus* and *Cafeteriavirus*) following the characterization of a virus infecting the heterotrophic stramenopile *Cafeteria roenbergensis* (CroV). CeV, as well as two of its close relatives, which infect the unicellular photosynthetic eukaryotes *Phaeocystis globosa* (Phaeocystis globosa virus [PgV]) and *Aureococcus anophagefferens* (Aureococcus anophagefferens virus [AaV]), are currently unclassified by the International Committee on Viral Taxonomy (ICTV). The detailed comparative analysis of the CeV genome presented here confirms the phylogenetic affinity of this emerging group of microalga-infecting viruses with the *Mimiviridae* but argues in favor of their classification inside a distinct clade within the family. Although CeV, PgV, and AaV share more common features among them than with the larger *Mimiviridae*, they also exhibit a large complement of unique genes, attesting to their complex evolutionary history. We identified several gene fusion events and cases of convergent evolution involving independent lateral gene acquisitions. Finally, CeV possesses an unusual number of inteins, some of which are closely related despite being inserted in nonhomologous genes. This appears to contradict the paradigm of allele-specific inteins and suggests that the *Mimiviridae* are especially efficient in spreading inteins while enlarging their repertoire of homing genes.

IMPORTANCE Although it infects the microalga *Chrysochromulina ericina*, CeV is more closely related to acanthamoeba-infecting viruses of the *Mimiviridae* family than to any member of the *Phycodnaviridae*, the ICTV-approved family historically including all alga-infecting large dsDNA viruses. CeV, as well as its relatives that infect the microalgae *Phaeocystic globosa* (PgV) and *Aureococcus anophagefferens* (AaV), remains officially unclassified and a source of confusion in the literature. Our comparative analysis of the CeV genome in the context of this emerging group of alga-infecting viruses suggests that they belong to a distinct clade within the established *Mimiviridae* family. The presence of a large number of unique genes as well as specific gene fusion events, evolutionary convergences, and inteins integrated at unusual locations document the complex evolutionary history of the CeV lineage.

Received 10 February 2017 **Accepted** 18 April 2017

Accepted manuscript posted online 26 April 2017

Citation Gallot-Lavallée L, Blanc G, Claverie J-M. 2017. Comparative genomics of chrysochromulina ericina virus and other microalga-infecting large DNA viruses highlights their intricate evolutionary relationship with the established *Mimiviridae* family. J Virol 91:e00230-17. https://doi.org/10.1128/JVI.00230-17.

Editor Grant McFadden, The Biodesign Institute, Arizona State University

Address correspondence to Lucie Gallot-Lavallée, Lucie.Gallot-Lavallee@igs.cnrs-mrs.fr, or Jean-Michel Claverie, Jean-Michel.Claverie@univ-amu.fr.

* Present address: Guillaume Blanc, Mediterranean Institute of Oceanography (MIO), Aix Marseille Université, Université de Toulon, CNRS/INSU, IRD, UM 110, Marseille, France.

KEYWORDS Aureococcus anophagefferens virus, Chrysochromulina ericina virus, Haptolina ericina virus, Megamimivirinae, Mesomimivirinae, *Mimiviridae*, nucleocytoplasmic virus, Phaeocystis globosa virus

Several new viral families have been recently created (or proposed) following the discovery of highly diverse double-stranded DNA (dsDNA) giant viruses (initially defined as those with particles visible under a light microscope), all of which exhibit large genomes (>300 kb) and infect unicellular eukaryotes (reviewed in references 1–3). Among those new families, the most populated is the *Mimiviridae*, and it is the only one officially recognized by the International Committee on Viral Taxonomy (ICTV). The *Mimiviridae* comprises two registered genera: the *Mimivirus* and the *Cafeteriavirus*. The *Mimivirus* genus includes several dozen members distributed among three clades (A, B, and C), all of which infect *Acanthamoeba* and have pseudoicosahedral particles approximately 700 nm in diameter and genomes of about one megabase in length (4). The genus *Cafeteriavirus* contains a single member, *Cafeteria roenbergensis virus* (CroV). This virus is markedly different from the mimiviruses, having smaller particles (300 nm in diameter) and a smaller, 730-kb genome, with its host being the heterotrophic stramenopile *C. roenbergensis* (5). Prior to defining these two genera, metagenomic studies had already hinted at the presence of mimivirus relatives in marine environments (6). The successful isolation and characterization of several of these viruses showed that they correspond to smaller icosahedral particles (140 to 180 nm in diameter) packing smaller dsDNA genomes (370 to 475 kb in length) (7–9). In core gene phylogenies, these viruses clearly cluster with the *Mimiviridae*, although they appear to constitute a distinct clade (Fig. 1) (1, 2, 7, 8). This clade comprises only viruses infecting photosynthetic hosts (i.e., microalgae) from different taxa: haptophyceae for Phaeocystis globosa virus (PgV) and Haptolina ericina virus and stramenopile for Aureococcus anophagefferens virus (AaV). This emerging subgroup within the *Mimiviridae* appears to include other lesser characterized members, such as Phaeocystis pouchetii virus (PpV), Prymnesium kappa virus (PkV) (10), and Pyramimonas orientalis virus (PoV). It also includes a number of other nonisolated candidates predicted solely from the assembly of metagenomics sequence data (11), such as the Organic Lake phycodnaviruses (OLPV1 and -2) (12). Since these viruses infect algae, a paraphyletic group of organisms, they were originally classified as members of the *Phycodnaviridae*, although this historical family increasingly encompasses viruses with little phylogenetic relationship (13, 14). One of the goals of the present study is to clarify this issue.

Aside from their gene-based phylogenetic clustering within the *Mimiviridae*, the members of this emerging clade exhibit additional features, such as AT-rich genomes encoding full DNA transcription and replication machinery (needed for their intracytoplasmic replication) (2) and a special version of the MutS mismatch DNA repair protein strangely related to octocorals (15). Most of them also encode an asparagine synthase (AsnS) (11). The *Mimiviridae* family also harbors the only viruses known to allow the replication of virophages (7, 12, 16–18).

Chrysochromulina ericina virus (CeV) was isolated from Norwegian coastal waters in 1998 but was only recently fully sequenced (9). It replicates within the cytoplasm of *H. ericina* (formerly *Chrysochromulina ericina*) with a lytic cycle lasting 14 to 19 h, resulting in the release of thousands of icosahedral particles 160 nm in diameter (19). Its host is distributed worldwide. Here, we performed a detailed comparative analysis of the genome of CeV and of its closest fully sequenced relatives (PgV and AaV) to provide support for their classification within a new clade in the *Mimiviridae*, as well as to investigate their evolutionary relationship with the rest of the family.

RESULTS

CeV genome global analysis. In line with its virion size, the CeV genome (473,558 bp) is larger than those of PgV and AaV (Table 1) and is the largest among those of alga-infecting viruses from any other family. We identified 512 putative protein-coding

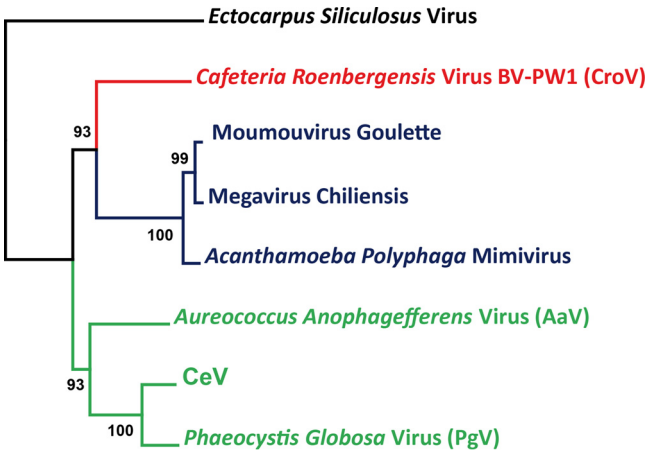


FIG 1 Phylogeny of the concatenation of MCP, DNAPol B, and DNA packaging ATPase. The phylogenetic tree was built using PhyML (58) based on multiple alignments generated using Expresso (60). The tree was rooted with *Ectocarpus siliculosus virus* (*Phaeovirus*). Representatives of the genera *Mimivirus* (blue) and *Cafeteriavirus* (red) have been included, as well as the three fully sequenced alga-infecting *Mimiviridae* relatives (green). The tree was drawn using MEGA7 (61).

genes with an average length of 280 codons (ranging from 41 to 2,317 codons) and 12 tRNA genes (two $tRNA^{Leu}$, two $tRNA^{Ser}$, one $tRNA^{Ala}$, one $tRNA^{Ile}$, two $tRNA^{Lys}$, one $tRNA^{Gln}$, one $tRNA^{Asn}$, one $tRNA^{Arg}$, and one $tRNA^{Gly}$). Intergenic sequences are very short (82.5 nucleotides [nt] on average) and exhibit higher A+T contents than average (85% versus 73.7%), suggesting that the loss of C+G nucleotides is an ongoing evolutionary process only slowed down by the negative selection pressure applied on protein-coding regions (Table 1). Compared to the NCBI nonredundant sequence database NR (including *Mimiviridae*), 43% (218) of the predicted proteins did not exhibit a significant match (E value of $<10^{-5}$ by BLASTP). This proportion of ORFans (i.e., without recognizable homologs within the whole NR database) is similar to that of other alga-infecting *Mimiviridae* (AaV, 45%; OLPV1, 44%; PgV, 43%). Among the 293 predicted proteins with a database homolog, 221 (75.4%) had their best match in eukaryote-infecting large dsDNA viruses, most of which (214/221, i.e., 96.8%) were members of the *Mimiviridae* family, mainly PgV (with 144 best matches) and the two OLPVs (with 30 best matches). The 72 nonviral best matches were distributed between bacteria (30) and eukaryotes (43), including 7 open reading frames (ORFs) in haptophytes (i.e., the taxon of the CeV host), pointing out potential horizontal gene transfers (HGT). A comparison (dot plot) of the orthologous gene positions in the genomes of CeV's closest relatives indicates numerous rearrangements (data not shown).

Phylogenetic analyses. To reconstruct the relationship between the viruses composing the *Mimiviridae* alga-infecting clade, we used a concatenation of the DNA polymerases, ATP DNA-packaging enzymes, and closest orthologs of the major capsid protein (MCP1). Consistent with the distribution of CeV best hits in the NR database, PgV appears to be its closest relative among fully sequenced viruses (Fig. 1). However,

TABLE 1 Genomic features of the *Mimiviridae*

Virus	Genome size (kbp)	No. of ORFs	Avg ORF size (bp)	Genome GC%	Coding density	GC%	
						ORF	Inter-ORF
Megavirus	1,259	1,120	1,015	25.3	0.90	26.5	14.1
Mimivirus	1,181	979	1,080	28.0	0.90	29.1	18.3
CroV	730	544	1,025	23.3	0.76	24.3	20.1
AaV	371	377	870	28.7	0.88	29.9	19.8
OLPV1[a]			697	29.5	0.86	30.2	25.0
PgV	460	434	960	32.0	0.91	33.7	15.3
CeV	474	512	840	25.4	0.91	26.3	15.7

[a]OLPV1 statistics are based on a 344,723-bp-long contig (12).

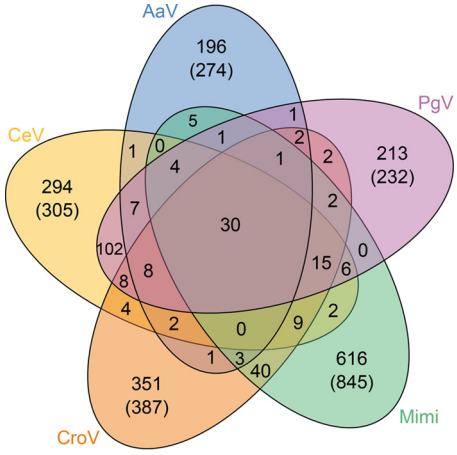


FIG 2 Venn diagram indicating the global proximity in gene content of CeV, its two closest relatives, PgV and AaV, and one member of each genus of the family *Mimiviridae* (*Cafeteriavirus* genus, CroV; *Mimivirus* genus, Mimi). The numbers in parentheses correspond to the raw number of encoded proteins without a homolog in the four other viruses. The numbers without parentheses indicate how many distinct clusters they constitute. The analysis was driven using OrthoMCL software (20), with a 10^{-5} E-value threshold and 1.5-mcl inflation parameter.

when we included OLPV1 in the analysis, no consistent pattern was found for the branching order of CeV, PgV, and OLPV after they diverged from their common ancestor with AaV. Moreover, applying the Shimodaira and Hasegawa (SH) test (see Materials and Methods) to each of the 39 groups of orthologous proteins shared by the 4 viruses (CeV, PgV, OLPV, and AaV) did not produce a conclusive answer. Such ambiguous results could be due to divergence times too close to be resolved, orthologous gene exchanges among these viruses, and/or compositional constraint similarities blurring the phylogenetic signal.

Gene content. Using OrthoMCL (20), we analyzed the groups of orthologues shared by mimivirus, CroV, and the three fully sequenced alga-infecting viruses, CeV, PgV, and AaV (Fig. 2). A striking result is that among the hundreds of proteins encoded by these clearly related viruses, only 30 are shared by all of them (a number dropping down to 19 when including Moumouvirus and Megavirus). Thus, including the new clade of alga-infecting viruses in the *Mimiviridae* causes the extended family to rest on an amazingly small proportion of common core genes. On the other hand, each viral genome exhibits a high number of unique genes (i.e., without recognizable homologs in the other *Mimiviridae*) (305 for CeV, 274 for AaV, 232 for PgV, 387 for CroV, and 845 for mimivirus) (Fig. 2). As 68% of CeV-unique genes correspond to ORFans, postulating a high frequency of horizontal exchanges with known viruses or cellular organisms is clearly not sufficient to explain their origin. Finally, some paralogs are sporadically present in the various viruses. Altogether, these large differences in the gene contents of these various viruses, which nevertheless share a strong phylogenetic signal of common ancestry, are rather puzzling and at least suggest a complex evolutionary history of the family.

Unique features common to the alga-infecting *Mimiviridae*. We further investigated if the alga-infecting viruses possessed common genes/functions not shared with the other *Mimiviridae*. As PgV, CeV, and AaV infect photosynthetic protozoans (at variance with the other *Mimiviridae*), we postulated that genomic features exclusive to them could be linked to this specific lifestyle. Only 7 of such shared genes were identified, among which a single one corresponds to a predicted function: an ERCC4-type DNA repair nuclease (YP_009173624.1, or CeV_369). Such enzymes are usually part of the cellular response against UV-induced DNA damage (21, 22), also known to occur in eukaryotic viruses (23). While infecting photosynthetic hosts, alga-infecting *Mimiviridae* might be exposed to sunlight-induced genome damages, making these nucleases

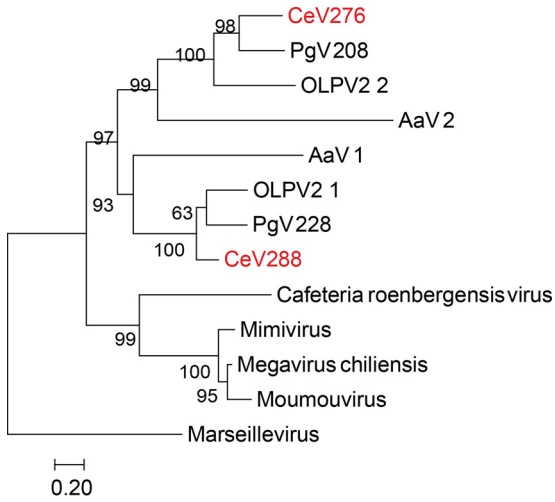


FIG 3 Duplication of the second largest subunit of the DNA-directed RNA polymerase II (RPB2) in CeV, PgV, and AaV. A maximum likelihood phylogenetic tree (58) of RPB2, aligned using Tcoffee (62) (Mcoffee mode), was constructed. Statistical branch supports (in percentages) for SH-like local support tests are given beside nodes. This phylogeny strongly supports a separate lineage leading to CeV, AaV, and PgV.

useful for repair. Such an ERCC4-dependent process extends the known diversity of viral responses to light-induced DNA damage. Other marine viruses possess a variety of mechanisms to repair DNA damage, which are either host dependent or host independent (23–25). These might be essential for the maintenance of viral communities (24), perhaps indirectly by protecting their host, as suggested by the reduced sensitivity to UV-B stress of microalgae cocultured with viruses compared to those in virus-free cultures (23).

Unrelated to the above-described function, another common feature of the alga-infecting *Mimiviridae* is the presence of two paralogs of the second-largest subunit of the DNA-directed RNA polymerase II (RPB2) (CeV_276 and CeV_288). As none of these copies appears defective and their sequences are quite divergent (only 34.2% identical), two distinct forms of transcriptional complexes might be formed following their interaction with the single RPB1 subunit encoded by these viruses. Combinations of different subunits leading to different versions of RNA polymerases IV and V (pol IV and pol V, respectively) are known in plants (26). These complexes evolved specialized roles (e.g., nonredundant gene silencing) (27, 28). Similarly, alternative RNA pol II complexes formed in alga-infecting *Mimiviridae* could play different roles. The duplicated paralogs present longer branches (Fig. 3), consistent with an accelerated rate of evolution. The presence of a C-terminal extension on the sequences of this group of paralogs would also be consistent with a functional modification: such additional residues could cause a change in substrate specificity. Phylogenetic reconstruction (Fig. 3) suggests that the RPB2 paralogs originated from a single duplication that occurred after the divergence from the rest of the *Mimiviridae*. In addition, AaV exhibits two copies of RNA polymerase II large subunits (RPB1) (AaV_242 and AaV_320) (8).

The alga-infecting *Mimiviridae* also share two distinct versions of major capsid proteins. The MCP1 paralogs exhibit a phylogeny consistent with the "species" tree and correspond to the least-divergent version of the major capsid protein common to all *Mimiviridae* (Fig. 4, top). The *Mimivirus* and *Cafeteriavirus* genera possess a second copy of MCP1, the duplication of which clearly predates their divergence (Fig. 4). In contrast, another type of MCP2 paralog (Fig. 4, bottom) is uniquely found in CeV (CeV_86, CeV_87, and CeV_88), PgV, and AaV. It probably results from an exchange with other alga-infecting large DNA viruses.

Altogether, these phyletic patterns specific to AaV, PgV, and CeV further support grouping them in a clade distinct from the other *Mimiviridae*.

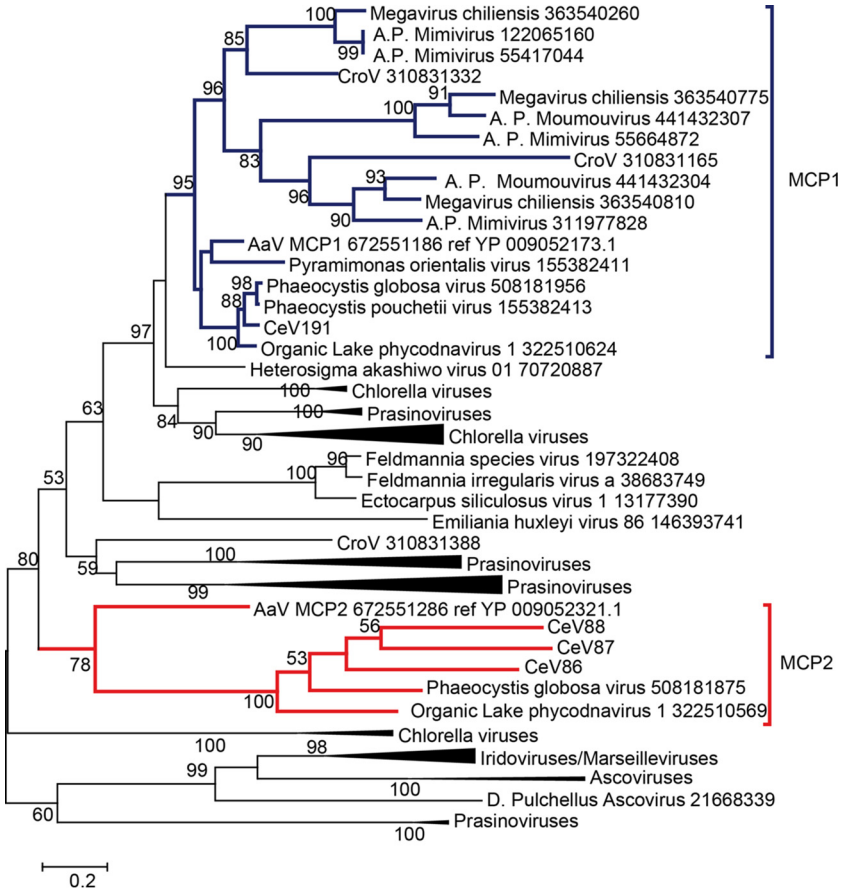


FIG 4 Relationship of the two major capsid protein homologs (MCP) found in CeV, PgV, and AaV. A maximum likelihood phylogenetic tree (58) of MCP, aligned using Tcoffee (62) (Expresso mode), was constructed. Statistical branch supports (in percentages) are given beside nodes.

Specific features shared only by CeV and PgV. CeV, PgV, and OLPV share homologs to the cold shock protein, which is known to act as an RNA chaperone (29). This protein, not found in other eukaryotic large dsDNA viruses, either was acquired by the alga-infecting lineage and later lost by AaV or was acquired after its divergence from AaV. We noticed that some dsDNA phages also encode similar cold shock proteins, suggesting that proper RNA folding is a recurrent constraint among unrelated viruses.

The analysis of PgV's genomic repeats led to the identification of an ORFan present in 12 copies, designated PgV_MIGE (major interspersed genomic element) (7). CeV possesses six copies of MIGE homologs. Interestingly, phylogenetic reconstruction grouped the PgV's MIGE and CeV's MIGE in separate clusters (Fig. 5). This suggests that MIGE was initially a single-copy gene in each virus before undergoing multiple duplications after their divergence. Analysis of CeV-MIGE did not hint at the mechanism by which this genetic element is duplicated and/or moved around. At variance with PgV, MIGE homologs in CeV lack an associated noncoding highly conserved region (7).

CeV and PgV also share a fusion event between a DNA polymerase X (DNAPolX) and a DNA ligase that is discussed in a later section.

Predicted functions unique to CeV. Among the 305 proteins unique to CeV, only 106 (35%) have a database homolog, among which only 52 are associated with functional attributes (listed in Table 2). Except for one light-harvesting complex protein (CeV_128) that will be discussed in a later section, none of these predicted functions have previously been shown to be specifically linked to the viral infection of algae.

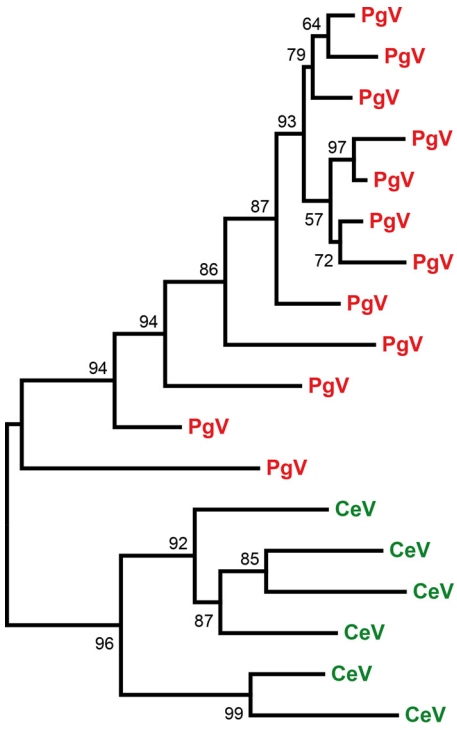


FIG 5 Independent MIGE spreading in CeV and PgV. A maximum likelihood phylogenetic tree (58) of MIGE protein sequences, aligned using Muscle (57), was constructed. Statistical branch supports (in percentages) for SH-like local support tests are given beside nodes.

CeV, a privileged host for spreading inteins? Inteins are mobile genetic elements encoding protein introns that remove themselves from host proteins through an autocatalytic excision and protein-splicing process (30). Fully functional inteins encode a homing endonuclease that mediates their integration at specific genomic sites. Different classes of inteins are found associated with the same, usually highly conserved, proteins (and thus genomic sequences), most of which are essential enzymes involved in DNA processing, replication, or synthesis. Nonallelic inteins (i.e., inteins not associated with the same insertion sites and/or protein-coding genes) do not exhibit significant sequence similarity, although they all presumably share an ancestor (30).

CeV encodes 8 inteins, to date the largest number among viruses. These inteins are inserted in 7 different ORFs: two in the ribonucleotide reductase large subunit (CeV_219), one in a DEAD-like RNA helicase (CeV_416), one in the DNA polymerase (CeV_365), one in an ATP-dependent Lon peptidase (CeV_043), one in an RPB2 paralog (CeV_288), one in a GDP-mannose 4,6-dehydratase (CeV_113), and one at the C terminus of a protein (CeV_451), concatenating an N-terminal U-box domain and a von Willebrand factor (vWA) domain. In general, intein insertions are strongly biased toward DNA-processing enzymes (30, 31). Among other explanations (30, 31), this bias might result from the fact that viruses are the main vectors of inteins within eukaryotes (32) and that viral metabolisms are mostly limited to DNA processing. In that context, the inteins hosted in a CeV's GDP-mannose 4,6-dehydratase and Lon peptidase are noticeable exceptions, to our knowledge the first such cases reported in eukaryotic viruses. With their extended metabolisms, viruses with larger genomes, such as members of the *Mimiviridae*, might thus expand the range of homing genes/enzymes for inteins. According to the current paradigm, inteins can be maintained only within essential genes. This is likely true for 5 of the 7 genes cited above (that belong to the *Mimiviridae* core genes) (3), the exception being the GDP-mannose 4,6-dehydratase that has no homolog in other *Mimiviridae*. Such a function might nevertheless be

TABLE 2 Protein-coding genes unique to CeV and associated with functional attributes

ORF no.	Predicted function or detected domain
CeV_002	Multiple glycosyltransferase domain
CeV_003	Methyltransferase
CeV_008	Ubox domain
CeV_009	Alpha-1,2-fucosyltransferase
CeV_023	Alkylated (methylated) DNA repair protein
CeV_033	Papain-like cysteine peptidase
CeV_053	RING-finger domain
CeV_096	Putative patatin-like phospholipase
CeV_113	Intein containing GDP-mannose 4,6-dehydratase
CeV_128	Light-harvesting complex protein
CeV_137	Superoxide dismutase Cu-Zn
CeV_139	Arginase
CeV_146	Trans-2-enoyl-coenzyme A reductase (TER) and 2,4-dienoyl-coenzyme A reductase (DECR)
CeV_149	Class V aminotransferase
CeV_151	Putative phosphotransferase
CeV_152	Quaternary ammonium transporter
CeV_154	Fe-S cluster assembly scaffold protein
CeV_155	Zinc finger, C_3HC_4 type domain-containing protein, RING superfamily
CeV_161	Putative prenyltransferase
CeV_171	Phospholipase/carboxylesterase
CeV_176	Collagen and repeat-containing protein
CeV_179	Multiple type acyltransferase domains
CeV_180	Glucosyltransferase TPR
CeV_183	Glycosyltransferase family 2
CeV_184	Collagen triple helix
CeV_194	PAN/APPLE-like domain
CeV_195	Repeat containing Hsp70-like protein
CeV_196	Ubox/RING superfamily domain
CeV_201	RING-finger domain
CeV_213	Protein disulfide isomerase (PDIa)
CeV_218	Glycosyl hydrolase family 16
CeV_233	Putative AHH-like nuclease
CeV_252	Proline-rich repeats
CeV_256	Toll-like receptor
CeV_265	Partial perforin domain-like
CeV_267	Putative AAA^+ family ATPase
CeV_311	Galactose binding lectin domain
CeV_323	Putative permease
CeV_324	N-acyltransferase/N-myristoyltransferase
CeV_327	Putative AAA^+ family ATPase
CeV_359	Link (hyaluronan-binding) domain
CeV_361	Ring finger domain
CeV_366	Protein disulfide-isomerase domain
CeV_372	Hsp70-like protein
CeV_373	Acetylpolyamide aminohydrolase (histone deacetylase)
CeV_404	ATP-dependent Clp protease, proteolytic subunit
CeV_415	Putative syntaxin, SNARE domain
CeV_433	Class II HMG-box domain
CeV_440	HMG box domain
CeV_463	RING domain
CeV_464	YABBY domain
CeV_467	Catalytic core of Asn/Asp-ARNt synthetase

important, as it was independently acquired by a number of other large dsDNA viruses that infect various algae (*Chlorella* and *Prasinophyceae*) (33). The CeV homolog might thus be necessary in an alga-infecting context. The closest homologues to CeV GDP-mannose 4,6-dehydratase are intein-free bacterial enzymes (although intein-containing bacterial enzymes exist that may not have been sequenced yet).

Inteins normally insert in highly conserved regions of essential proteins. As the paradigm goes, the strong conservative constraints exerted on these regions ensures the correct excision (from the protein) or homing (into the DNA) processes. Surprisingly, we found that two similar inteins of the same prototype (standard class 1 with a

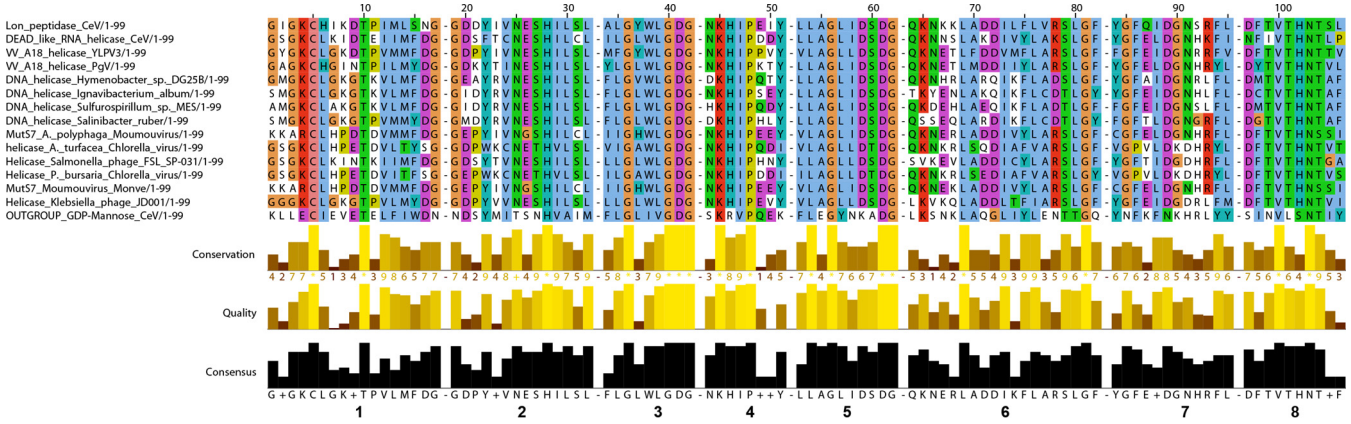


FIG 6 Helicase-type intein inserted in the CeV Lon peptidase. Multiple alignment, computed with Muscle (57), of the 8 blocks (numbered below the alignment) characteristic of inteins (Table 3). The intein hosted by the CeV GDP-mannose 4,6-dehydratase is included for comparison.

homing endonuclease of the LAGLIDADG type) have been inserted in two unrelated CeV genes/proteins: a Lon protease and a DEAD box helicase (Fig. 6 and Table 3). Moreover, a similar intein helicase-targeted allele is inserted in different enzymes in other large DNA viruses: MutS7 in Moumouvirus, the VVA18 helicase in PgV and Yellowstone Lake phycodnavirus (YLP) (percent identity with CeV Lon protease intein ranging from 42.5% to 54%), and, besides the *Mimiviridae*, DNA helicase B of 2 chloroviruses (identity with CeV Lon protease intein, 40% and 38%). Other related inteins are hosted by DNA helicases from phages or bacteria (identity with CeV Lon protease intein ranging from 39% to 43%). The 8 blocks specific to this intein prototype (34) are well conserved (Fig. 6). Although unrelated, these enzymes share a P-loop NTPase domain. All inteins but those in MutS7 are inserted in the ATP/GTP binding site (i.e., the Walker A or P-loop domain), precisely at the GK/T site. This might be sufficient for this prototype of intein to properly excise. More puzzling is the way by which this intein might have jumped from one enzyme to another. Indeed, after being cut by the intein-encoded endonuclease, the free intein gene should proceed with homing by homologous base pairing with the intein-containing allele, which serves as the template for the polymerase. This scenario appears unlikely given the limited similarity of the extein DNA sequences (Fig. 7). Another intriguing fact is that the intein is not inserted in the P-loop domain of MutS7 but is inserted at an AR/S site, 30 amino acids

TABLE 3 High similarity of the CeV Lon peptidase intein to those usually found in various DNA helicases

Intein (no.)	% Similarity to intein no.[a]:													
	1	2	3	4	5	6	7	8	9	10	11	12	13	14
Lon peptidase, CeV (1)	100	44	44	54	42	40	41	39	43	40	42	38	42	43
DEAD-like RNA helicase, CeV (2)	43	100	40	44	36	35	37	33	42	36	40	36	41	37
Putative VV A18 helicase, YLPV3 (3)	44	40	100	49	42	40	44	41	42	37	44	37	42	47
VV A18-like helicase, PgV (4)	54	43	50	100	43	44	46	41	42	42	46	42	42	45
DNA helicase, *Hymenobacter* sp. (5)	42	35	42	43	100	59	62	58	36	33	41	33	35	43
DNA helicase, *Ignavibacterium album* (6)	40	36	40	43	59	100	60	62	37	32	41	32	36	45
DNA helicase, *Sulfurospirillum* sp. (7)	41	36	44	46	62	60	100	55	34	32	38	32	34	45
DNA helicase, *Salinibacter ruber* (8)	39	33	41	41	58	62	55	100	34	33	37	33	34	42
MutS7, A. P. Moumouvirus (9)	43	42	42	42	36	36	34	34	100	37	37	36	94	40
Helicase, ATCV NTS-1 (10)	40	35	37	42	33	31	32	33	37	100	35	75	37	51
Helicase, *Salmonella* phage (11)	42	40	44	46	41	40	38	37	37	35	100	36	38	36
Hypothetical protein B508R, PBCV (12)	38	35	37	42	33	32	32	33	36	75	36	100	35	41
MutS7, Moumouvirus Monve (13)	42	41	42	42	35	36	34	34	94	37	38	35	100	25
Helicase, *Klebsiella* phage (14)	43	37	47	45	43	44	45	42	40	37	51	36	41	100
Outgroup GDP-mannose CeV					24	26	22	20				25		20

[a]The pairwise percentages of identity between these homologous inteins is given. The values for the nonallelic intein hosted by CeV GDP-mannose 4,6 dehydratase is included for comparison when pairwise alignment was possible. YLPV3, Yellowstone Lake phycodnavirus 3; ATCV, Acanthocystis turfacea chlorella virus; PBCV, Paramecium bursaria chlorella virus.

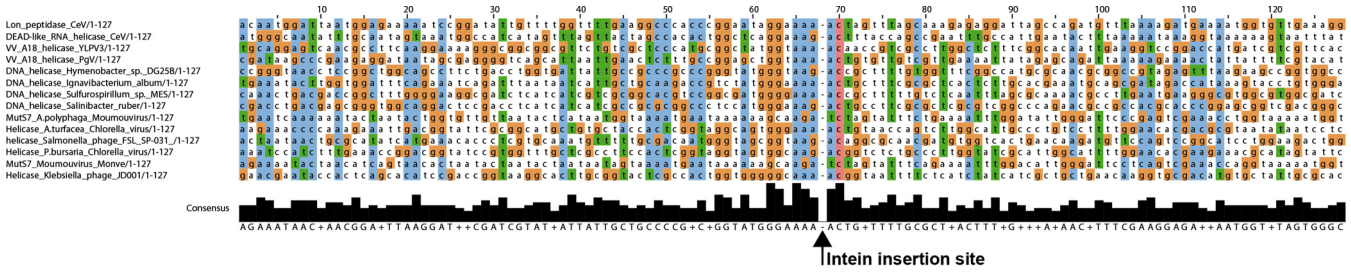


FIG 7 Extein DNA sequences in different genes hosting similar intein alleles (Fig. 6) are not similar.

upstream. Furthermore, these amino acids are not conserved in the intein-free MutS7 found in other *Mimiviridae*. Altogether, these cases constitute violations of the current allele-specific intein paradigm. They suggest a mechanism specific to the *Mimiviridae* spreading inteins while widening their homing range.

Convergent acquisition of host genes. Detailed attention was paid to the CeV-encoded light-harvesting complex protein (LHC) (CeV_128) of the LIL (light-harvesting light) family. The presence of such a gene coding for a component of the photosynthesis apparatus was rather unexpected. Further analyses led to the discovery that other alga-infecting viruses encode proteins of the LIL family (Fig. 8). Their phylogenies clearly suggest that they were acquired from their hosts through three independent events (one for CeV and two for the prasinoviruses) (Fig. 8). Besides the expected chlorophyll binding (CB) motifs and 3 predicted transmembrane helices (35), all viral LIL proteins contain an N-terminal transit peptide targeting them to their respective host chloroplast type (36) (secondary red for *H. ericina* and primary green for the prasinophytes). This suggests that these proteins are functional. Rather than collecting incom-

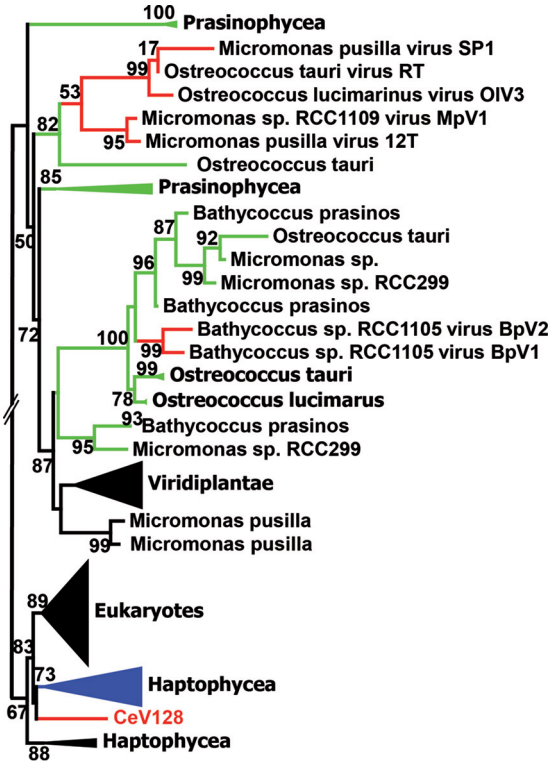


FIG 8 Convergent acquisitions of host LIL proteins. A maximum likelihood phylogenetic tree (58) of LIL proteins, aligned using Muscle (57), was constructed. Statistical branch supports (in percentages) for the S-H-like local support tests are given beside nodes. Branches corresponding to viruses are colored red, with blue for haptophyceae and green for prasinophyceae. Sequences used for the analysis were selected using BLAST Explorer as implemented in Phylogeny.fr (63).

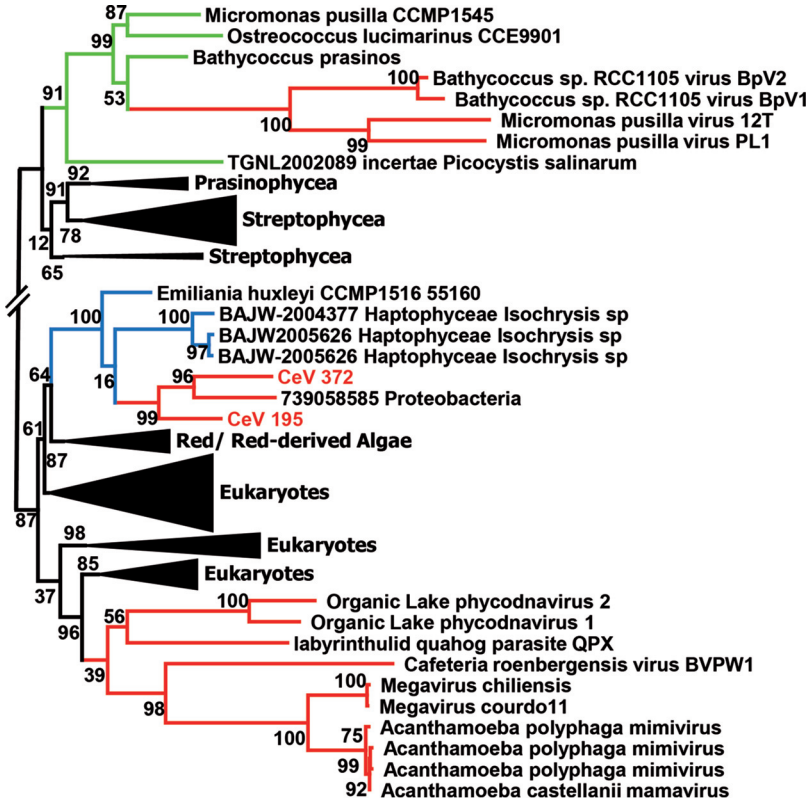


FIG 9 Convergent acquisitions of host HSP70 proteins. A maximum likelihood phylogenetic tree (58) of HSP70, aligned using Muscle (57), was constructed. Statistical branch supports (in percentages) for S-H-like local support test are given beside nodes. Viral branches are colored red, with blue for haptophyceae and green for prasinophyceae.

ing light for photosynthesis, some nucleus-encoded eukaryotic LIL proteins are involved in photoprotection, such as nonphotochemical quenching (NPQ) (35, 37). This might be the role of the viral LIL homologs.

We then systematically looked for additional cases of convergent acquisition involving other genes. We identified 2 more, concerning the DNAK (HSP70) chaperone (CeV_372 and CeV_195) (Fig. 9) and a U-box domain at the C terminus of the CeV_008-encoded protein (data not shown). CeV and two prasinoviruses have independently acquired HSP70, likely from their hosts (Fig. 9). Interestingly, other *Mimiviridae* also encode homologs of this chaperone. However, these proteins appear to derive from an ancestral version of HSP70 already present at the root of the family tree. Thus, the HSP70 proteins found today in CeV might result from a nonhomologous replacement of the ancestral *Mimiviridae* version of the protein. The U-box domain mediates the ubiquitin conjugation of protein targeted for degradation in the proteasome (38). CeV (CeV_008) and EhV (EMVG_00184) appear to have independently acquired host-derived U-box domains.

Gene fusions in CeV. Gene fusions are genomic rearrangements that are thought to facilitate the coexpression and/or assembly of proteins initially encoded separately. The involved proteins could physically interact or be functionally associated (39, 40).

We systematically screened for potential gene fusions in the CeV genome by comparing the topology of homologous genes in other *Mimiviridae*. Two clear cases were identified: one (CeV_489) is a fusion between the DNAPolX and the NAD-dependent DNA ligase (also found in PgV), and the other (CeV_007) is a fusion between the uridyltransferase and the UDP-glucose 4,6-dehydratase (UDG). We did not detect such fusions in other viruses or cellular organisms.

The DNAPolX/NAD-dependent DNA ligase fusion makes functional sense, as these enzymes normally work in succession when participating in DNA repair: DNAPolX

fills single-nucleotide gaps before their ends are sealed by the ligase (41, 42). CroV homologs of these two proteins are encapsidated, suggesting their role in prereplicative DNA repair (43). We noticed that the CeV and PgV fusion proteins exhibit very different linkers that could have resulted from 2 independent events. However, the analysis of similar fusions in the environmental database showed that such linkers are not conserved. The CeV and PgV proteins thus likely resulted from a unique fusion event prior to their divergence and that of their linkers. Interestingly, some bacteria also exhibit a fusion between an ATP-dependent DNA ligase and a DNA polymerase domain (44). This might constitute a case of functional evolutionary convergence between bacteria and eukaryotic viruses resulting in more efficient DNA repair enzymes.

The unique fusion between a uridyltransferase domain and UDG identified in CeV (CeV_007) might optimize the synthesis of L-rhamnose, known to be involved in the glycosylation of structural proteins in mimivirus and certain chloroviruses (45). This fusion occurred after the duplication of the uridyltransferase gene (CeV_479), originally involved in the three-component UDP-*N*-acetylglucosamine biosynthetic pathway (46) found in large *Mimiviridae*. This duplication might have opened the way to the fusion event (47).

DISCUSSION

Convergent evolution in CeV and other large dsDNA viruses. This study identified several cases where homologs of the same gene were independently acquired by CeV and other viruses, most likely from their hosts. The best example of such convergent acquisition is that of the light-harvesting complex protein (LHC) that is found in CeV and seven species of prasinoviruses (Fig. 8). Such gene transfers are reminiscent of the acquisition of core photosystem components (including a remote member of the LIL family, HLIP) by cyanophages from their bacterial hosts (35, 37, 48). Thus, the manipulation of host photosynthesis by viruses might be a recurrent theme in evolution, most likely to meet the increased energy burden of the production of virions by the infected cell.

DNA repair is another general function that has been the object of convergent innovations, this time through gene fusions. The DNAPolX-NAD-dependent DNA ligase fusion identified in CeV (and shared with PgV) is echoed by that of a DNAPolX-AP-endonuclease found in poxviruses (49) and that of an Mre11 and Rad50 domain in mimivirus (50). These proteins are likely parts of an optimized dsDNA strand break repair machinery.

The unique fusion between a uridyltransferase domain and UDG identified in CeV (CeV_479), the presence of several enzymes for the synthesis of rhamnose in mimivirus (45), and the fused sequential enzymes found in OtV5, OmV1, and OlV1 (YP_001648294.1, YP_009172960.1, YP_004061822.1) are other examples of convergent innovations targeting the same biosynthetic pathway, probably central to the glycosylation of viral structural proteins (45).

CeV also acquired copies of the HSP70 chaperone (CeV_195 and CeV_372) that clearly originated from a different source than the one encoded by its close relative, OLPV, as well as CroV and acanthamoeba-infecting *Mimiviridae*. It is also different from the one acquired by several prasinoviruses from their hosts (Fig. 9).

Finally, CeV and EhV have independently acquired U-box domain-encoding sequences, suggesting the active involvement of the host proteasome in the infectious process, perhaps as a source of recycled amino acids (51).

The puzzling origin of CeV's many unique genes. We previously pointed out the paradox of CeV exhibiting so many unique protein-coding genes (294/512, or 57.4%), while the rest of them overwhelmingly had their closest relative in PgV or other *Mimiviridae*. This paradox is amplified by the small number of core genes, i.e., genes common to all of these obviously related viruses (Fig. 2). This paradox would be partially explained if most of these so-called unique genes were not real but were false-positive calls from an overestimating (albeit standard) (9) bioinformatic annotation procedure. Such ambiguities are best solved by validating gene predictions using

transcriptomic data that unfortunately is not available for CeV. However, we can estimate the actual proportion of real genes among the unique ones using CroV, for which such data exist. According to the original publication (5), CroV has 544 protein-coding genes, of which 438 genes were tested for transcription and 274 (63%) found positive. We then examined how this proportion changed when separately considering unique versus shared CroV genes. Out of the 438 tested genes, 299 are unique, of which 177 (59.2%) fall in the expressed category. On the other hand, 97 of the 139 shared genes were found to be expressed (69.7%). Even though these numbers denote a significant difference in transcript detection in favor of shared genes (P value of <0.05 for the table [177, 122; 97, 42] by Fisher's exact test), they nevertheless suggest that a large proportion (59.2/69.7, or 84.9%) of CroV unique genes are real (postulating that their transcripts should be detected in the same proportion as those of shared genes). Extending the same reasoning to CeV and using the same ratio would reduce the number of unique true genes from 294 to 250, i.e., still half of its total gene complement.

The evolutionary scenario susceptible to leading to the presence of so many CeV-unique genes while preserving the strong phylogenetic affinity globally exhibited by the genes shared among the *Mimiviridae* remains to be elucidated. As most of the unique CeV genes are ORFans, a scenario involving a huge number of gene acquisitions from cellular organisms (or known viruses) is unlikely, short of postulating an equally huge mutation rate erasing their phylogenetic origins. There is no evidence of such genomic instability in these large dsDNA viruses. On the contrary, they are well equipped with high-fidelity DNA replication and repair machineries (our results here and reference 52). The converse model, postulating a reductive evolution from a common *Mimiviridae* ancestor (2) with a genome large enough to accommodate the number of unique genes indicated in Fig. 2, appears increasingly unlikely (although viruses with thousands of genes are known to exist [53]). An alternative to the opposite and equally unlikely accretion and reduction evolutionary scenarios would be to postulate that these DNA viruses *de novo* generate new protein-coding genes (and functions) by a totally unknown mechanism. Demonstrating such a capacity would definitely put viruses on the center stage of biological evolution.

Proposed taxonomy of an extended *Mimiviridae* family. The present analysis of the CeV genome adds to the mounting consensus that despite their large differences in gene content, particle size, host range, and ecology, a group of alga-infecting large dsDNA viruses (CeV, PgV, OLPV, and AaV) and the acanthamoeba-infecting *Mimiviridae* (genus, *Mimivirus*) belong to the same family and share an ancestor (3, 7, 8, 13). This family also includes CroV (5), a virus that infects the heterotrophic stramenopile *Cafeteria roenbergensis*, as the sole member of the genus *Cafeteriavirus*. In addition to the sharing of unique genes (such as the mismatch DNA repair protein MutS7 and the puzzling asparagine synthase), large A+T rich genomes, and a full DNA transcription and replication apparatus, divergent members of this virus group (mimivirus, CroV, Pgv, and OLPV) exhibit a unique association with virophages, small dsDNA viruses replicating as parasites of their intracytoplasmic virion factories. Since CeV, PgV, and their relatives (such as Phaeocystis pouchettii virus or the Organic Lake phycodnaviruses) infect unicellular algae, they are referred to as unclassified new members of the family *Phycodnaviridae* in the literature as well as in sequence databases. As presently recognized by the ICTV, the family *Phycodnaviridae* includes six genera: *Raphidovirus*, *Coccolithovirus*, *Phaeovirus*, *Chlorovirus*, *Prasinovirus*, and *Prymnesiovirus*. As more alga-infecting viruses are characterized, it is clear that an increasing number of them do not fit within this established family, the name of which ("phyco" means "algae") has become a source of confusion. This highlights the danger in classifying viruses within clades named after their hosts, as there is increasing evidence that the same host can be infected by phylogenetically distinct viruses (such as *Acanthamoeba* being infected by five different types of giant dsDNA viruses: mimivirus, marseillevirus, pandoravirus, pithovirus, and mollivirus) (2). Conversely, the *Mimiviridae* family (described here) (but also the *Asfarviridae* family [54]) shows that viruses with

strong phylogenetic relationships can infect hosts belonging to branches that diverged at the earliest time of eukaryote history.

To help clarify the classification of alga-infecting viruses and acknowledge the phylogenetic affinity of CeV, PgV, and AaV with CroV and the mimivirus group (Fig. 1), we propose to divide the family *Mimiviridae* into two subfamilies. One, tentatively named the "Megamimivirinae" (i.e., the largest *Mimiviridae*), should include the 3 clades (A, B, and C) of the existing *Mimivirus* genus (4) and CroV as the prototype of the existing *Cafeteriavirus* genus. A new subfamily, named the "Mesomimivirinae" (i.e., the still large but smaller *Mimiviridae*), should include CeV and PgV (as well as the partially sequenced OLPV1 and OLPV2) as a new genus and the outlier AaV as the prototype of yet another distinct genus. Redefined as proposed, the new family *Mimiviridae* would clearly separate the above-described large alga-infecting viruses from the *Phycodnaviridae* family while acknowledging their relationship with the acanthamoeba-infecting mimiviruses. At the same time, the range of sequence divergence exhibited by the core proteins (such as DNA pol B, MutS7, or the packaging ATPase) within the *Mimiviridae* will remain comparable to that observed within other large dsDNA virus families, such as the *Poxviridae* (also divided into two subfamilies, *Chordopoxvirinae* and *Entomopoxvirinae*). As this paper was in review, findings from metagenomic studies suggested that yet another lineage of large *Mimiviridae* remains to be characterized (55).

MATERIALS AND METHODS

Genome sequencing, assembly, and annotation. The procedures for genome sequencing, assembly, and annotation were described previously in reference 9.

Identification of gene fusions. We mapped the CeV predicted proteins onto other *Mimiviridae* genomes using TBLASTN. When different segments of a CeV protein were found to best match at two distant locations in the target genome, the corresponding ORFs were submitted to further phylogenetic analyses to confirm their orthologous relationship with the different parts of the candidate CeV fusion protein. BLAST best-scoring sequences belonging to bacteria, archaea, eukaryotes, and viruses were included in the analysis to establish that the fusion occurred within the *Mimiviridae* lineage.

S-H test. To discriminate between the three possibilities, (i) OLPV emerged first after AaV, (ii) PgV emerged first, or (iii) CeV emerged first, we performed multiple Shimodaira and Hasegawa (S-H) tests (56) as follows. First, we identified the 39 clusters of orthologs present in the proteomes of AaV, OLPV, PgV, and CeV using OrthoMCL. Second, 4 proteins of each of the 39 clusters were aligned using MUSCLE (57), the resulting 39 multiple alignments were visually validated, their gapped positions removed, and the corresponding likelihood matrices for the three tree topology were computed with PhyML (58).

Third, the CONSEL procedure (59) (www.sigmath.es.osaka-u.ac.jp/shimo-lab/prog/consel/) was applied to perform the S-H test itself (i.e., a computation of the *P* value for each of the possible topologies by comparison of the matrix of the log likelihoods).

The test was not conclusive. Among the 117 (3×39) trees, we could reject only three of them at a *P* value of <0.05.

ACKNOWLEDGMENTS

We thank Matthieu Legendre, Sebastien Santini, Adrien Villain, and Olivier Poirot for their helpful discussions and for help with the sequence analysis software used for this work. We also thank Chantal Abergel for her help with the final versions of the figures.

The IGS laboratory is supported by the Centre National de la Recherche Scientifique and Aix-Marseille University. We acknowledge the use of the PACA-Bioinfo Platform, supported by France-Génomique (ANR-10-INBS-0009) and Institut Français de Bioinformatique (ANR-11-INSB-0013). L. Gallot-Lavallée is supported by a PhD award from Aix-Marseille University.

REFERENCES

1. Fischer MG. 2016. Giant viruses come of age. Curr Opin Microbiol 31:50–57. https://doi.org/10.1016/j.mib.2016.03.001.
2. Abergel C, Legendre M, Claverie J-M. 2015. The rapidly expanding universe of giant viruses: mimivirus, pandoravirus, pithovirus and mollivirus. FEMS Microbiol Rev 39:779–796. https://doi.org/10.1093/femsre/fuv037.
3. Yutin N, Colson P, Raoult D, Koonin EV. 2013. *Mimiviridae*: clusters of orthologous genes, reconstruction of gene repertoire evolution and proposed expansion of the giant virus family. Virol J 10:106. https://doi.org/10.1186/1743-422X-10-106.
4. Yoosuf N, Yutin N, Colson P, Shabalina SA, Pagnier I, Robert C, Azza S, Klose T, Wong J, Rossmann MG, La Scola B, Raoult D, Koonin EV. 2012. Related giant viruses in distant locations and different habitats: *Acan-*

thamoeba polyphaga moumouvirus represents a third lineage of the *Mimiviridae* that is close to the megavirus lineage. Genome Biol Evol 4:1324–1330. https://doi.org/10.1093/gbe/evs109.
5. Fischer MG, Allen MJ, Wilson WH, Suttle CA. 2010. Giant virus with a remarkable complement of genes infects marine zooplankton. Proc Natl Acad Sci U S A 107:19508–19513. https://doi.org/10.1073/pnas.1007615107.
6. Monier A, Larsen JB, Sandaa R-A, Bratbak G, Claverie J-M, Ogata H. 2008. Marine mimivirus relatives are probably large algal viruses. Virol J 5:12. https://doi.org/10.1186/1743-422X-5-12.
7. Santini S, Jeudy S, Bartoli J, Poirot O, Lescot M, Abergel C, Barbe V, Wommack KE, Noordeloos AAM, Brussaard CPD, Claverie J-M. 2013. Genome of *Phaeocystis globosa* virus PgV-16T highlights the common ancestry of the largest known DNA viruses infecting eukaryotes. Proc Natl Acad Sci U S A 110:10800–10805. https://doi.org/10.1073/pnas.1303251110.
8. Moniruzzaman M, LeCleir GR, Brown CM, Gobler CJ, Bidle KD, Wilson WH, Wilhelm SW. 2014. Genome of brown tide virus (AaV), the little giant of the *Megaviridae*, elucidates NCLDV genome expansion and host–virus coevolution. Virology 466-467:60–70.
9. Gallot-Lavallée L, Pagarete A, Legendre M, Santini S, Sandaa R-A, Himmelbauer H, Ogata H, Bratbak G, Claverie J-M. 2015. The 474-kilobase-pair complete genome sequence of CeV-01B, a virus infecting *Haptolina* (Chrysochromulina) *ericina* (Prymnesiophyceae). Genome Announc 3:e01413-15.
10. Johannessen TV, Bratbak G, Larsen A, Ogata H, Egge ES, Edvardsen B, Eikrem W, Sandaa R-A. 2015. Characterisation of three novel giant viruses reveals huge diversity among viruses infecting Prymnesiales (Haptophyta). Virology 476:180–188. https://doi.org/10.1016/j.virol.2014.12.014.
11. Mozar M, Claverie J-M. 2014. Expanding the *Mimiviridae* family using asparagine synthase as a sequence bait. Virology 466-467:112–122.
12. Yau S, Lauro FM, DeMaere MZ, Brown MV, Thomas T, Raftery MJ, Andrews-Pfannkoch C, Lewis M, Hoffman JM, Gibson JA, Cavicchioli R. 2011. Virophage control of Antarctic algal host-virus dynamics. Proc Natl Acad Sci U S A 108:6163–6168. https://doi.org/10.1073/pnas.1018221108.
13. Wilson WH, Van Etten JL, Allen MJ. 2009. The *Phycodnaviridae*: the story of how tiny giants rule the world. Curr Top Microbiol Immunol 328:1–42.
14. Maruyama F, Ueki S. 2016. Evolution and phylogeny of large DNA viruses, Mimiviridae and Phycodnaviridae including newly characterized *Heterosigma akashiwo* virus. Front Microbiol 7:1942.
15. Ogata H, Ray J, Toyoda K, Sandaa R-A, Nagasaki K, Bratbak G, Claverie J-M. 2011. Two new subfamilies of DNA mismatch repair proteins (MutS) specifically abundant in the marine environment. ISME J 5:1143–1151. https://doi.org/10.1038/ismej.2010.210.
16. Villain A, Gallot-Lavallée L, Blanc G, Maumus F. 2016. Giant viruses at the core of microscopic wars with global impacts. Curr Opin Virol 17:130–137. https://doi.org/10.1016/j.coviro.2016.03.007.
17. Desnues C, La Scola B, Yutin N, Fournous G, Robert C, Azza S, Jardot P, Monteil S, Campocasso A, Koonin EV, Raoult D. 2012. Provirophages and transpovirons as the diverse mobilome of giant viruses. Proc Natl Acad Sci U S A 109:18078–18083. https://doi.org/10.1073/pnas.1208835109.
18. Fischer MG, Hackl T. 2016. Host genome integration and giant virus-induced reactivation of the virophage mavirus. Nature 540:288–291. https://doi.org/10.1038/nature20593.
19. Sandaa R-A, Heldal M, Castberg T, Thyrhaug R, Bratbak G. 2001. Isolation and characterization of two viruses with large genome size infecting *Chrysochromulina ericina* (Prymnesiophyceae) and *Pyramimonas orientalis* (Prasinophyceae). Virology 290:272–280. https://doi.org/10.1006/viro.2001.1161.
20. Li L, Stoeckert CJ, Roos DS. 2003. OrthoMCL: identification of ortholog groups for eukaryotic genomes. Genome Res 13:2178–2189. https://doi.org/10.1101/gr.1224503.
21. Westerveld A, Hoeijmakers JHJ, van Duin M, de Wit J, Odijk H, Pastink A, Wood RD, Bootsma D. 1984. Molecular cloning of a human DNA repair gene. Nature 310:425–429. https://doi.org/10.1038/310425a0.
22. Busch D, Greiner C, Lewis K, Ford R, Adair G, Thompson L. 1989. Summary of complementation groups of UV-sensitive CHO cell mutants isolated by large-scale screening. Mutagenesis 4:349–354. https://doi.org/10.1093/mutage/4.5.349.
23. Jacquet S, Bratbak G. 2003. Effects of ultraviolet radiation on marine virus-phytoplankton interactions. FEMS Microbiol Ecol 44:279–289. https://doi.org/10.1016/S0168-6496(03)00075-8.
24. Weinbauer MG, Wilhelm SW, Suttle CA, Garza DR. 1997. Photoreactivation compensates for UV damage and restores infectivity to natural marine virus communities. Appl Environ Microbiol 63:2200–2205.
25. Furuta M, Schrader JO, Schrader HS, Kokjohn TA, Nyaga S, McCullough AK, Lloyd RS, Burbank DE, Landstein D, Lane L, Van Etten JL. 1997. Chlorella virus PBCV-1 encodes a homolog of the bacteriophage T4 UV damage repair gene denV. Appl Environ Microbiol 63:1551–1556.
26. Tucker SL, Reece J, Ream TS, Pikaard CS. 2010. Evolutionary history of plant multisubunit RNA polymerases IV and V: subunit origins via genome-wide and segmental gene duplications, retrotransposition, and lineage-specific subfunctionalization. Cold Spring Harbor Symp Quant Biol 75:285–297. https://doi.org/10.1101/sqb.2010.75.037.
27. Ream TS, Haag JR, Wierzbicki AT, Nicora CD, Norbeck AD, Zhu J-K, Hagen G, Guilfoyle TJ, Pasa-Tolić L, Pikaard CS. 2009. Subunit compositions of the RNA-silencing enzymes Pol IV and Pol V reveal their origins as specialized forms of RNA polymerase II. Mol Cell 33:192–203. https://doi.org/10.1016/j.molcel.2008.12.015.
28. Haag JR, Pikaard CS. 2011. Multisubunit RNA polymerases IV and V: purveyors of non-coding RNA for plant gene silencing. Nat Rev Mol Cell Biol 12:483–492. https://doi.org/10.1038/nrm3152.
29. Rudan M, Schneider D, Warnecke T, Krisko A. 2015. RNA chaperones buffer deleterious mutations in E. coli. eLife 4:e04745. https://doi.org/10.7554/eLife.04745.
30. Gogarten JP, Senejani AG, Zhaxybayeva O, Olendzenski L, Hilario E. 2002. Inteins: structure, function, and evolution. Annu Rev Microbiol 56:263–287. https://doi.org/10.1146/annurev.micro.56.012302.160741.
31. Novikova O, Jayachandran P, Kelley DS, Morton Z, Merwin S, Topilina NI, Belfort M. 2016. Intein clustering suggests functional importance in different domains of life. Mol Biol Evol 33:783–799. https://doi.org/10.1093/molbev/msv271.
32. Novikova O, Topilina N, Belfort M. 2014. Enigmatic distribution, evolution, and function of inteins. J Biol Chem 289:14490–14497. https://doi.org/10.1074/jbc.R114.548255.
33. Piacente F, Gaglianone M, Laugieri ME, Tonetti MG. 2015. The autonomous glycosylation of large DNA viruses. Int J Mol Sci 16:29315–29328. https://doi.org/10.3390/ijms161226169.
34. Perler FB, Olsen GJ, Adam E. 1997. Compilation and analysis of intein sequences. Nucleic Acids Res 25:1087–1093. https://doi.org/10.1093/nar/25.6.1087.
35. Engelken J, Brinkmann H, Adamska I. 2010. Taxonomic distribution and origins of the extended LHC (light-harvesting complex) antenna protein superfamily. BMC Evol Biol 10:233.
36. Patron NJ, Waller RF. 2007. Transit peptide diversity and divergence: a global analysis of plastid targeting signals. Bioessays News Rev Mol Cell Dev Biol 29:1048–1058. https://doi.org/10.1002/bies.20638.
37. Büchel C. 2015. Evolution and function of light harvesting proteins. J Plant Physiol 172:62–75. https://doi.org/10.1016/j.jplph.2014.04.018.
38. Hatakeyama S, Yada M, Matsumoto M, Ishida N, Nakayama KI. 2001. U box proteins as a new family of ubiquitin-protein ligases. J Biol Chem 276:33111–33120. https://doi.org/10.1074/jbc.M102755200.
39. Marcotte EM, Pellegrini M, Ng HL, Rice DW, Yeates TO, Eisenberg D. 1999. Detecting protein function and protein-protein interactions from genome sequences. Science 285:751–753. https://doi.org/10.1126/science.285.5428.751.
40. Yanai I, Wolf YI, Koonin EV. 2002. Evolution of gene fusions: horizontal transfer versus independent events. Genome Biol 3:research0024.1-research0024.13.
41. Prasad R, Singhal RK, Srivastava DK, Molina JT, Tomkinson AE, Wilson SH. 1996. Specific interaction of DNA polymerase beta and DNA ligase I in a multiprotein base excision repair complex from bovine testis. J Biol Chem 271:16000–16007. https://doi.org/10.1074/jbc.271.27.16000.
42. Yamtich J, Sweasy JB. 2010. DNA polymerase family X: function, structure, and cellular roles. Biochim Biophys Acta 1804:1136–1150. https://doi.org/10.1016/j.bbapap.2009.07.008.
43. Fischer MG, Kelly I, Foster LJ, Suttle CA. 2014. The virion of *Cafeteria roenbergensis* virus (CroV) contains a complex suite of proteins for transcription and DNA repair. Virology 466-467:82–94.
44. Della M, Palmbos PL, Tseng H-M, Tonkin LM, Daley JM, Topper LM, Pitcher RS, Tomkinson AE, Wilson TE, Doherty AJ. 2004. Mycobacterial Ku and ligase proteins constitute a two-component NHEJ repair machine. Science 306:683–685. https://doi.org/10.1126/science.1099824.
45. Parakkottil Chothi M, Duncan GA, Armirotti A, Abergel C, Gurnon JR, Van Etten JL, Bernardi C, Damonte G, Tonetti M. 2010. Identification of an

L-rhamnose synthetic pathway in two nucleocytoplasmic large DNA viruses. J Virol 84:8829–8838. https://doi.org/10.1128/JVI.00770-10.
46. Piacente F, Bernardi C, Marin M, Blanc G, Abergel C, Tonetti MG. 2014. Characterization of a UDP-N-acetylglucosamine biosynthetic pathway encoded by the giant DNA virus Mimivirus. Glycobiology 24:51–61. https://doi.org/10.1093/glycob/cwt089.
47. Ohno S. 1970. Evolution by gene duplication. Springer-Verlag, New York, NY.
48. Lindell D, Jaffe JD, Johnson ZI, Church GM, Chisholm SW. 2005. Photosynthesis genes in marine viruses yield proteins during host infection. Nature 438:86–89. https://doi.org/10.1038/nature04111.
49. Afonso CL, Tulman ER, Lu Z, Oma E, Kutish GF, Rock DL. 1999. The genome of *Melanoplus sanguinipes* entomopoxvirus. J Virol 73:533–552.
50. Yoshida T, Claverie J-M, Ogata H. 2011. Mimivirus reveals Mre11/Rad50 fusion proteins with a sporadic distribution in eukaryotes, bacteria, viruses and plasmids. Virol J 8:427. https://doi.org/10.1186/1743-422X-8-427.
51. Price CT, Al-Quadan T, Santic M, Rosenshine I, Abu Kwaik Y. 2011. Host proteasomal degradation generates amino acids essential for intracellular bacterial growth. Science 334:1553–1557. https://doi.org/10.1126/science.1212868.
52. Doutre G, Philippe N, Abergel C, Claverie J-M. 2014. Genome analysis of the first *Marseilleviridae* representative from Australia indicates that most of its genes contribute to virus fitness. J Virol 88:14340–14349. https://doi.org/10.1128/JVI.02414-14.
53. Philippe N, Legendre M, Doutre G, Couté Y, Poirot O, Lescot M, Arslan D, Seltzer V, Bertaux L, Bruley C, Garin J, Claverie J-M, Abergel C. 2013. *Pandoraviruses*: amoeba viruses with genomes up to 2.5 Mb reaching that of parasitic eukaryotes. Science 341:281–286. https://doi.org/10.1126/science.1239181.
54. Reteno DG, Benamar S, Khalil JB, Andreani J, Armstrong N, Klose T, Rossmann M, Colson P, Raoult D, La Scola B. 2015. *Faustovirus*, an asfarvirus-related new lineage of giant viruses infecting amoebae. J Virol 89:6585–6594. https://doi.org/10.1128/JVI.00115-15.
55. Schulz F, Yutin N, Ivanova NN, Ortega DR, Lee TK, Vierheilig J, Daims H, Horn M, Wagner M, Jensen GJ, Kyrpides NC, Koonin EV, Woyke T. 2017. Giant viruses with an expanded complement of translation system components. Science 356:82–85. https://doi.org/10.1126/science.aal4657.
56. Shimodaira H, Hasegawa M. 1999. Multiple comparisons of log-likelihoods with applications to phylogenetic inference. Mol Biol Evol 16:1114–1116. https://doi.org/10.1093/oxfordjournals.molbev.a026201.
57. Edgar RC. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res 32:1792–1797. https://doi.org/10.1093/nar/gkh340.
58. Guindon S, Gascuel O. 2003. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. Syst Biol 52:696–704. https://doi.org/10.1080/10635150390235520.
59. Shimodaira H, Hasegawa M. 2001. CONSEL: for assessing the confidence of phylogenetic tree selection. Bioinformatics 17:1246–1247. https://doi.org/10.1093/bioinformatics/17.12.1246.
60. Armougom F, Moretti S, Poirot O, Audic S, Dumas P, Schaeli B, Keduas V, Notredame C. 2006. Expresso: automatic incorporation of structural information in multiple sequence alignments using 3D-Coffee. Nucleic Acids Res 34:W604–W608. https://doi.org/10.1093/nar/gkl092.
61. Kumar S, Stecher G, Tamura K. 2016. MEGA7: Molecular Evolutionary Genetics Analysis version 7.0 for bigger datasets. Mol Biol Evol 33:1870–1874. https://doi.org/10.1093/molbev/msw054.
62. Notredame C, Higgins DG, Heringa J. 2000. T-Coffee: a novel method for fast and accurate multiple sequence alignment. J Mol Biol 302:205–217. https://doi.org/10.1006/jmbi.2000.4042.
63. Dereeper A, Guignon V, Blanc G, Audic S, Buffet S, Chevenet F, Dufayard J-F, Guindon S, Lefort V, Lescot M, Claverie J-M, Gascuel O. 2008. Phylogeny.fr: robust phylogenetic analysis for the non-specialist. Nucleic Acids Res 36:W465–W469. https://doi.org/10.1093/nar/gkn180.

3 - Analyses complémentaires et discussion

a - Évolution de la taille des génomes des *Mimiviridae*.

En plus de ces travaux publiés, nous avons déterminé, par une méthode fondée sur le critère du maximun de parcimonie, le nombre minimal de gènes présents à chaque nœud (i.e., ancêtre) de l'arbre des *Mimiviridae.* Pour cela, j'ai reconstruit les familles de protéines de 8 *Mimiviridae* appartenant aux différents clades connus (voir figure 7) en utilisant le logiciel OrthoMCL (178). A partir de cette classification des protéines homologues, le schéma suivant était appliqué : si une protéine était présente chez au moins un « *Mesomimivirinae* » (i.e., nom de sous-famille proposé dans l'article pour le clade réunissant CeV, PgV et AaV), et au moins un « *Megavirinae* » (i.e., groupe comprenant CroV, et les *Mimiviridae* qui infectent les amibes), alors ce gène était considéré comme présent dans le génome de l'ancêtre commun des *Mimiviridae*. Ce principe était appliqué à chaque nœud de l'arbre phylogénétique des *Mimiviridae* (i.e., chaque nœud représente l'ancêtre commun des virus descendants). Les différences dans les répertoires de gènes entre chaque nœud permettaient d'inférer le nombre de gains et de pertes de gènes pour chaque branche de l'arbre (i.e., chaque branche représente l'évolution de la lignée entre deux nœuds). Ceci a été réalisé dans l'optique de détecter de potentielles tendances dans l'évolution des génomes de *Mimiviridae :* gagnent-ils, ou perdent-ils majoritairement des gènes ? Les résultats de cette analyse sont résumés dans la figure 7.

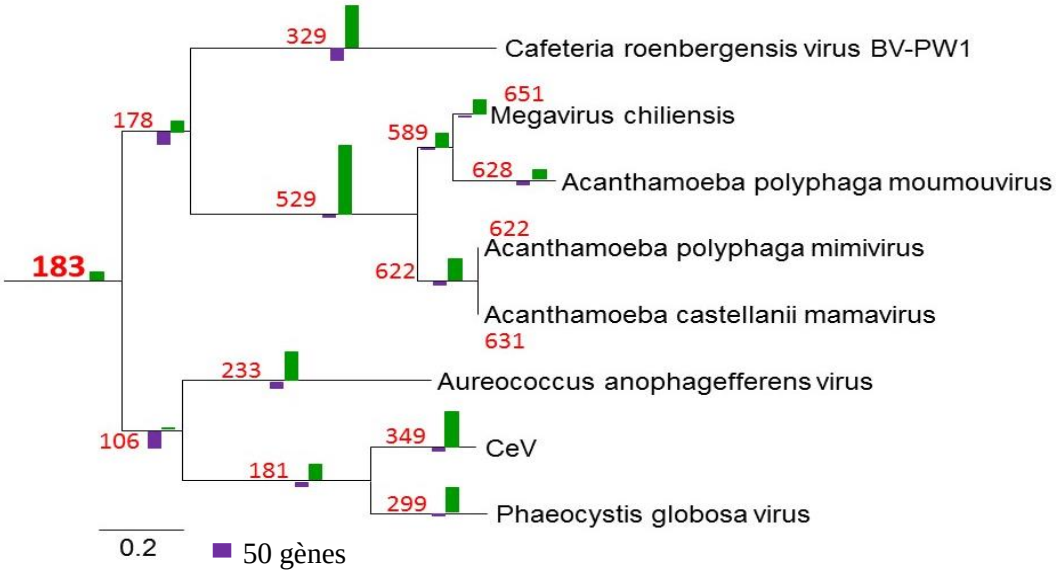


Figure 7 : Reconstruction du nombre de gains et de pertes de gènes par une méthode de parcimonie. Les nombres en rouge sont le nombre de familles de protéines inférées à chaque nœud, les barres vertes et violettes représentent respectivement le nombre de gains et de pertes inféré à chaque branche.

Selon notre analyse, l'ancêtre des *Mimiviridae* possédait donc au moins 183 gènes, que l'on retrouve dans tout, ou une partie, des descendants. Ce nombre de gènes ancestraux est plus faible que le nombre de gènes trouvés chez les virus séquencés (i.e., de 384 gènes pour AaV à 1120 gènes

pour Megavirus). Ce résultat est compatible avec l'hypothèse que le génome de l'ancêtre des *Mimiviridae* était moins complexe que ceux des virus actuels. Cependant, l'inférence par l'approche du maximum de parcimonie souffre de certaines limitations (voir ci-dessous) qui peuvent conduire à une estimation erronée du nombre de gènes ancestraux. Un second message de cette analyse est qu'il n'y a pas d'unidirectionnalité dans l'évolution des génomes de *Mimiviridae* : que ce soit pour les plus petits ou les plus gros génomes, il semblerait que tous ont à la fois perdu certains gènes ancestraux, et gagné de nouveaux gènes ; certes à des rythmes différents ce qui explique les différences de taille des génomes actuels.

La méthode du maximum de parcimonie comporte certains biais qui méritent d'être étudiés en plus amples détails car ils pourraient affecter les estimations des nombres de gènes ancestraux :

-Une surestimation du nombre de gains, qui par effet de vases communiquants conduit à une sous-estimation du nombre de gènes chez l'ancêtre, peut avoir lieu lorsque des gènes ont été perdus tôt dans l'histoire des *Mimiviridae*, ou bien encore dans plusieurs lignées distinctes de la famille, de telle façon que les virus possédant ces gènes n'appartiennent plus qu'à un sous ensemble de l'arbre. La méthode du maximum de parcimonie conclura alors de manière erronée à un gain dans la branche de l'ancêtre des virus qui ont conservé le gène. Ce biais pourra être en partie corrigé lorsque de nouveaux génomes de *Mimiviridae* seront séquencés et révèleront la nature ancestrale de certains des gènes rangés dans la catégorie « gain ».

-Le phénomène d'acquisition parallèle de gènes homologues peut mener à une sous-estimation des gains de gènes, si ces acquisitions se produisent indépendamment dans plusieurs lignées émergeant de l'ancêtre. J'ai par exemple décrit, dans l'article ci-dessus, 3 cas évidents d'acquisitions parallèles de gènes homologues impliquant CeV et d'autres NCLDVs infectant des algues. D'autres cas ont dernièrement fait l'objet d'une publication (139) : de nouveaux *Mimiviridae* tout récemment isolés et formant un 4ème clade au sein des « Megavirinae » possèdent un vaste répertoire de gènes codant pour des composants de la traduction. L'analyse de l'origine des enzymes de ce nouveau clade de *Mimiviridae* révèle que la plupart des gènes correspondant ont été acquis indépendamment pour chaque virus, de diverses sources cellulaires. Concernant ce dernier point, il faut cependant faire la remarque qu'il pourrait également s'agir d'un remplacement fonctionnel d'un gène à l'origine présent dans le génome viral par un gène d'origine cellulaire, et non d'une « acquisition *de novo* » de cette fonctionnalité. Ceci a par exemple été proposé pour la chaperonne HSP70 de CeV, ou encore l'ADN ligase des NCLDVs (179).

Peu après que nous ayons conduit cette analyse en interne, le laboratoire de Shoko Ueki au Japon a publié une étude similaire (en y incluant les virus de la famille des *Phycodnaviridae* (180)) qui a été réalisée à l'occasion de l'analyse du génome du virus *Heterosigma akashiwo* virus (HaV) infectant l'algue raphydophyte *Heterosigma akashiwo*. Ceci nous a permis de comparer nos résultats et nos interprétations avec les leurs. Leur étude a été menée grâce au logiciel COUNT, qui utilise non pas le principe de parcimonie, mais une méthode basée sur le maximum de vraissemblance. A partir d'un arbre groupant un ancêtre et tous ses descendants et d'une matrice

de présence/absence de gènes orthologues chez les différents virus actuels, sont inférés la taille des génomes ancestraux ainsi que les évènements de gains et de pertes de gènes le long des branches de l'arbre. Cette méthode est supposée moins sensible aux biais énoncés plus haut. Cependant, l'utilisateur de COUNT doit arbitrairement choisir la valeur des pénalités de gains et de pertes de gènes qui seront appliquées lors de l'analyse. L'équipe japonaise a voulu tester l'impact de la valeur des paramètres choisis sur le résultat, et a donc réalisé deux séries d'analyse : la première avec une pénalité de gain élevée et la seconde avec une pénalité de gain faible, tout en gardant constante la pénalité de perte. De façon saisissante, changer la valeur de ce paramètre change également l'interprétation générale de l'évolution des *Mimiviridae* : lorsque la pénalité de gain est forte, alors l'hypothèse d'une réduction génomique est privilégiée, tandis que lorsqu'elle est faible, c'est l'hypothèse inverse, d'accrétion de gènes à partir de petits génomes, qui est favorisée. Nos résultats sont intermédiaires, mais sensiblement plus proche des résultats obtenus avec une pénalité de gain forte.

L'étude de Shoko Ueki montre que les conclusions tirées dépendent en partie d'*a priori* apposés sur les données. Il est effectivement très difficile d'estimer la valeur juste pour ces pénalités. Ci-après sont présentées quelques pistes de réflexion concernant la pertinence des scénarii de gains et des scenarii de pertes.

Si nous faisons une analogie avec les parasites cellulaires et les parasites viraux que sont les *Mimiviridae*, ceci va à l'encontre d'un scénario de gains. Effectivement, les génomes des premiers ne grossissent pas après qu'ils deviennent obligatoirement intracellulaires. Au contraire, c'est un phénomène de réduction génomique qui est observé (181, 182) : la baisse des pressions de sélection sur certains gènes, due au fait que l'endosymbionte profite des machineries de son hôte, entraine une « pseudogénisation » et une perte progressive de ces gènes. Il pourrait en être de même pour les *Mimiviridae*.

Cependant, il existe des exceptions à la trajectoire réductioniste des génomes d'organismes intracellulaires : une étude suggère que certains parasites intracellulaires d'amibes ont des génomes plus complexes que leurs proches parents libres (183). Or, les plus gros *Mimiviridae* sont des parasites d'amibes. Ceci va plutôt dans le sens d'un scénario de gain de gènes.

De façon importante, les gènes rangés dans la catégorie « gains » sont pour la majorité des ORFans, c'est-à-dire qu'ils n'ont pas d'homologues chez les organismes cellulaires séquencés. Donc à moins d'admettre que nous ne connaissons qu'une faible proportion des protéines existantes dans le monde cellulaire, ce qui apparait de plus en plus improbable avec les progrès de la métagénomique notamment, la probabilité que tous ces « gains » soient le résultat de transfert latéraux de gènes « volés » aux organismes cellulaires est faibles. Ainsi, l'origine de ces ORFans reste à déterminer.

D'autre part, présupposer une faible fréquence de pertes de gènes implique qu'une majorité des gènes a un fort impact sur la valeur adaptative globale (i.e., « fitness ») des virus à leur environnement. Concernant ce point, des informations contradictoires nous viennent de deux études concernant les grands virus à ADN d'amibe : D'une part, une analyse des pressions de sélections qui sont à l'œuvre sur les gènes de plusieurs *Marseilleviridae* (184) a montré que la majorité des gènes de ces virus a évolué sous un régime de sélection négative, ce qui suggère que ces gènes contribuent à la valeur adaptative globale des virus concernés. Au contraire, les résultats d'une étude sur Mimivirus mis en « culture » avec des amibes exemptes de germes indiquent que le virus a subi une délétion représentant 1/6 de son génome (185). Ceci semble montrer que les gènes contenus dans le fragment d'ADN perdu sont dispensables en condition de laboratoire. Une autre observation est compatible avec cette dernière conclusion : les pertes parallèles de gènes. Ce phénomène est mis en évidence lors de la construction de diagramme de Venn représentant la répartition des familles de protéines au sein des différents membres de la famille des *Mimiviridae*, comme représenté dans la figure 2 de Gallot-Lavallée et al 2017 par exemple. Certaines familles de protéines sont retrouvées sporadiquement au sein des *Mimiviridae*. Ceci a également mis en évidence dans d'autres études concernant les *Mimiviridae* ou les NCLDVs en général (167, 175, 186, 187). Ce profil peut résulter d'acquisitions convergentes comme déjà mentionné. Cependant, les analyses phylogénétiques montrent majoritairement que les protéines forment un groupe monophylétique. Il s'agirait donc plus probablement de pertes différentielles le long des branches. Ce phénomène est imaginable dans le cas de gènes « dispensables ». Il peut par ailleurs brouiller le signal phylogénétique lorsque celui-ci est inféré à partir de patrons de présence/absence de gènes.

Pour conclure, aucune réponse claire quant au gigantisme n'a pu être apportée ni avec l'une, ni avec l'autre de ces études. Elles révèlent cependant un mode d'évolution complexe des *Mimiviridae*, avec d'une part les disparités dans le répertoire de gènes entre virus mêmes proches (nombreux ORFans uniques à chaque virus), et d'autre part les cas de convergences évolutives, qui font se ressembler des virus éloignés. Concernant les gènes ORFans, une idée proposée dans certaines études (188, 189) ainsi que dans notre papier sur CeV est qu'ils pourraient provenir en partie de la « création » *de novo* de gènes. La création *de novo* de gènes à partir de séquences non codantes a été décrite chez le nématode *Caenorhabditis elegans* et la drosophile (190, 191) et pourrait également se produire de façon particulièrement importante chez les NCLDVs.

b - Acquisitions convergentes de gènes : les AMGs ne sont pas l'apanage des phages

Trois cas d'acquisitions indépendantes du même gène chez leur hôte respectif, de CeV d'une part, et d'autres NCLDVs non apparentés mais infectant également des algues ont été décrit dans l'article. L'un d'entre eux, codant pour le composant du photosystème II LHC (Light Harvesting Protein), a également été acquis par des cyanophages à partir de leurs hôtes bactériens photosynthétiques. Des virus éloignés infectant des organismes photosynthétiques semblent donc adopter une stratégie commune pour s'adapter à leur hôte, qui vise probablement à modifier les caractéristiques des photosystèmes. Ce cas spécifique pourrait illustrer un phénomène fréquent

d'adaptation convergente des virus à un environnement donné. Après l'écriture de l'article, j'ai identifié des cas additionnels d'acquisitions convergentes impliquant différentes branches de NCLDVs et/ou des phages, et répertorié d'autres cas décrits dans la littérature :

b.1 - Protéine Fe-S de type A (ATC pour « A type carrier »). Ces protéines prennent part à l'assemblage des centres Fe/S puisqu'elles lient des centres $[2Fe\_2S]^{2+}$ ou $[4Fe\_4S]^{2+}$ après qu'ils aient été synthétisés puis les transfèrent aux apoprotéines. Des gènes codant pour des ATC ont été acquis par CeV d'une part, le prasinovirus *Ostreococcus lucimarus* Virus 7 (OlV7) d'autre part et des cyanophages enfin, de sources cellulaires différentes (figure 8).

La majorité des bactéries possèdent un seul type d'ATC, tandis que les alpha, beta et gamma protéobactéries en possèdent plusieurs types, qui sont classés en ATCI et ATCII. Ainsi, lors de l'endosymbiose primaire de l'alphaprotéobacterie, évènement qui a engendré les mitochondries, les eucaryotes ont hérité de deux ATCs (nommées Isa1 (ATCII) et Isa2 (ATCI) chez les eucaryotes). Ensuite, lors de l'endosymbiose secondaire de la cyanobactérie qui résulta en l'établissement du chloroplaste, certains eucaryotes ont acquis une ATC supplémentaire.

J'ai réalisé une reconstruction phylogénétique des ATCs, en utilisant d'une part des protéines qui présentaient un bon score d'alignement blast contre les protéines virales, et d'autre part les protéines utilisées dans une étude sur la répartition et la phylogénie des ATCs (192) réalisée par le laboratoire LCB à Marseille. L'arbre phylogénétique que j'ai obtenu (figure 8) concorde remarquablement (à l'exception des branches virales) avec les résultats de l'étude susnommée.

La protéine du prasinovirus Olv7 fait partie des ATCII. Elle émerge au sein des protéines Isa2 des prasinophytes (i.e., la classe d'algue de son hôte) (figure 8), ce qui suggère qu'elle est issue d'un transfert de gène depuis une algue prasinophyte. De plus, la protéine virale présente un fort taux d'identité (80%) avec la protéine de son hôte *Ostreococcus lucimarus*, ce qui est compatible avec la possibilité que la protéine virale soit bien adaptée pour supplémenter le métabolisme de son hôte. Ni la protéine de l'algue, ni la protéine du virus, ne possèdent de peptide signal (TP pour target peptid en anglais), portion peptidique permettant l'adressage à la mitochondrie de protéine codée par le noyau, dans le cas où l'organelle est le compartiment d'action de la protéine. Ceci suggère que ces deux protéines sont responsables du transport des centres Fe/S cytoplasmiques.

Par contraste, la protéine ATC de CeV n'émerge pas au sein d'une branche eucaryote (figure 8). En fait, elle semble n'appartenir à aucun des clades d'ATCs caractérisés par l'étude de l'équipe Marseillaise, qui pourtant s'est basée sur les ATCs détectées dans la totalité des génomes procaryotes disponible en 2008 (date de leur étude), ainsi que dans la totalité des protéines eucaryotes disponible à la même date. Autre point intriguant : des protéines ATCs de phages appartiennent elles aussi à ce clade restreint (figure 8). La présence au sein de ce clade, également,

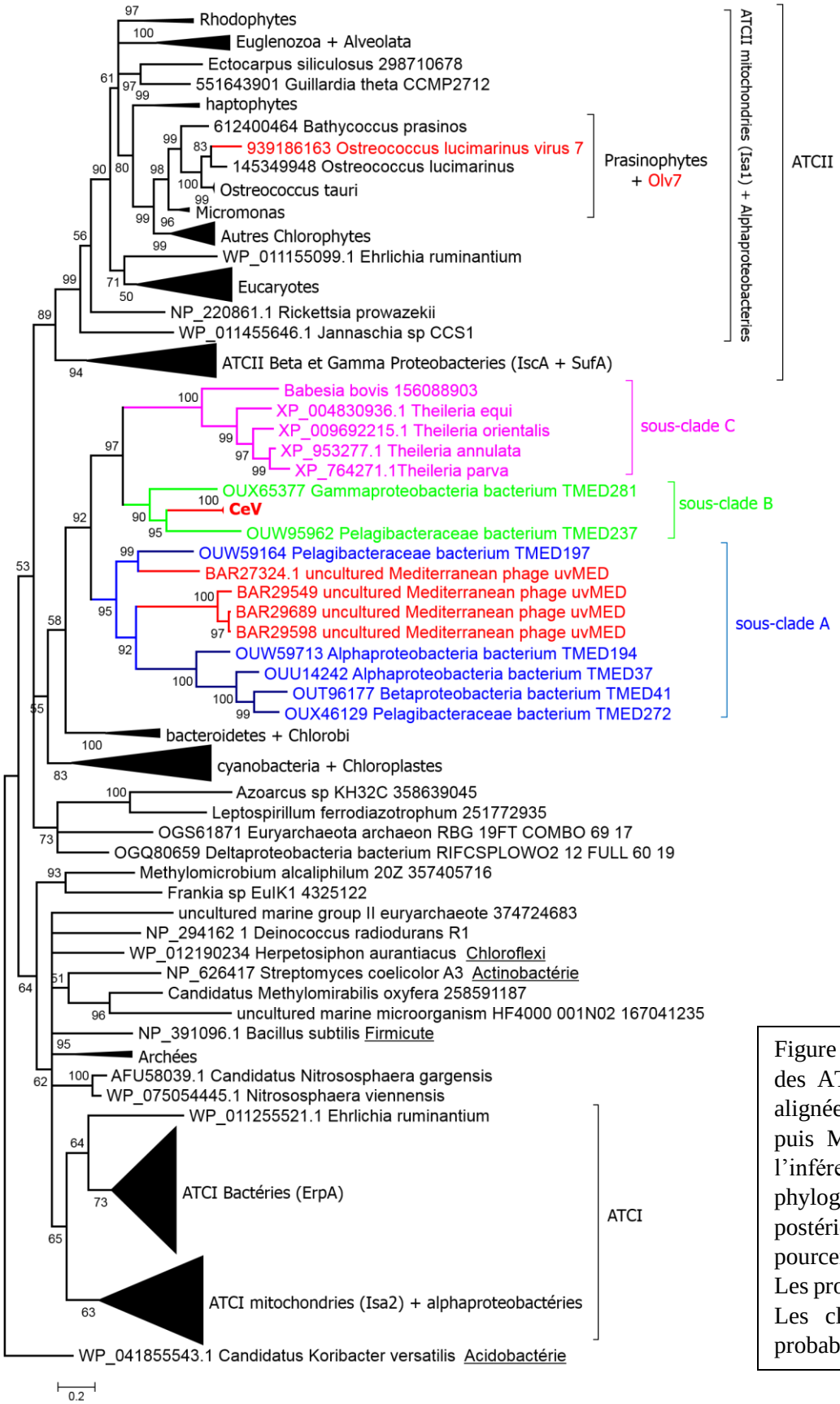


Figure 8: Arbre phylogénétique des ATCs. Les protéines ont été alignées grâce au logiciel Muscle puis MrBayes a été utilisé pour l'inférence des relations phylogénétiques. Les probabilités postérieures sont indiquées en pourcentages le long des branches. Les protéines virales sont en rouge. Les clades A et B constituent probablement des clades viraux.

de quelques protéines de taxas bactériens variés m'a paru suspecte. De plus, les contigs contenant les gènes codant pour les ATCs de chacune de ces bactéries sont assez courts (8835-53735 bp). J'ai donc analysé l'origine taxonomique des protéines de la base de donnée nr obtenant les meilleurs scores d'alignement blast contre chacun des gènes présents sur les contigs, dans le but de savoir s'il s'agit effectivement de contigs bactériens, ou bien s'il pourrait s'agir de contamination. Concernant les contigs contenant les ATCs du sous-clade A (en bleu sur la figure 8), les résultats semblent confirmer une contamination par des contigs viraux. En effet, 51% des protéines codées par ces contigs ayant un match dans nr ont leur meilleur match chez des phages, contre 46% chez des bactéries. En conséquence, le sous-clade A constituerait en fait un clade viral. La même analyse a été réalisée sur les contigs du sous-clade B. L'un présente 66,7% de meilleurs matchs NCLDVs et donc constitue très probablement un contigs NCLDVs. C'est moins clair pour le second contig du sous-clade B : sur la base des meilleurs scores blast de chacune des 6 protéines du contig, deux seraient bactériennes, l'une NCLDV, tandis que les 3 autres seraient eucaryotes. Parmi ces 3 dernières, 2 présentent des scores d'alignement peu significatifs (2e-8 et 1e-8). La dernière possède des matchs significatifs (5e-18 – 2e-64) mais au nombre très restreint de 5. Or, ceux-ci appartiennent à des eucaryotes phylogénétiquement éloignés. De plus, 3 d'entre eux, dont les deux les plus significatifs, appartiennent à des organismes que nous avons détectés dans notre étude des inserts NCLDVs dans les assemblages eucaryotes: *Shaeroforma artica* (2 hits) et *Gonapodya prolifera*. Tous ces indices suggèrent qu'il s'agit plutôt d'un contigs d'origine NCLDV. Le sous-clade B serait donc également viral. Cette branche est constituée d'un dernier sous-clade, noté sous-clade C sur la figure 8. Il est constitué de protéines appartenant à des piroplasmida, une sous-classe du phylum des apicomplexes. Nous avons donc un clade constitué d'un premier sous-clade phages, d'un second sous-clade NCLDVs et d'un dernier sous-clade eucaryote. Le phénomène d'attraction des longues branches pourrait expliquer ce regroupement. Effectivement, les apicomplexes sont des parasites et pourraient donc avoir un taux de mutation élevé. Si la phylogénie est juste, plusieurs évènements de transferts de gènes doivent être invoqués pour expliquer cette distribution : le fait que les apicomplexes soient les seuls eucaryotes de cette branche fait pencher la balance en faveur d'un transfert virus vers apicomplexes. Les apicomplexes encodent une Isa2 (ATCI) mais n'ont pas de protéine Isa1 (ATCII) typique des eucaryotes. La protéine d'origine virale pourrait avoir remplacé la protéine eucaryote, comme décrit pour certains composants de l'appareil de réplication et de transcription des mitochondries (voir introduction). La branche du clade dans sa totalité émerge de façon très basale, ce qui permet d'envisager les scenarii suivants : (i) soit un transfert bactérie vers virus a eu lieu très tôt après la divergence des Chlorobi, la séquence aurait ensuite évolué dans la lignée virale, (ii) soit l'enzyme a émergé dans une branche bactérienne disparue (ou n'ayant plus d'ATC) avant d'être transférée aux virus ; (iii) soit enfin, les ATCs sont apparus chez les virus. Les apicomplexes sont, comme les phages et les NCLDVs, des parasites obligatoires. On peut alors imaginer que cette protéine a la capacité de s'adapter à une variété d'hôtes eucaryotes, et que c'est la raison pour laquelle, après transfert, elle aurait été retenue par ces 3 entités parasitaires. Ceci demeure hautement spéculatif.

b.2 - PhytanoylCoA-dioxygenase. Ce gène est présent d'une part, dans plusieurs génomes de cyanophages (193). D'autre part, une protéine de CeV et une protéine de PgV possèdent un domaine PFAM (la base de données de domaines protéiques de l'EMBL-EBI) PhytanoylCoA-dioxygenase. Ces virus semblent donc également encoder cette protéine, mais ceci mériterait cependant une démonstration expérimentale ; effectivement, les homologues cellulaires les plus proches (des fungi) présentent 30 % d'identité avec les protéines de CeV et PgV. Les protéines de phages d'un côté et de NCLDVs de l'autre sont trop divergentes entre elles pour les intégrer dans un alignement multiple commun. Je n'ai pas donc pas effectué la reconstruction phylogénétique qui aurait permis de vérifier qu'il s'agit bien d'acquisitions indépendantes. Cependant, parce que (1) les séquences de phages et de NCLDVs sont très divergentes, et (2) leurs homologues cellulaires les plus proches appartiennent à deux 2 domaines différents (i.e., bactéries pour les phages et eucaryotes pour PgV et CeV), ceci semble être l'hypothèse la plus vraisemblable. Chez les cyanophages, il a été proposé que cette enzyme convertisse le 2-oxuglutarate (accumulé chez l'hôte bactérien, en réponse à l'infection par le phage) en succinate, un donneur d'électron majeur de la chaine respiratoire chez les bactéries. Ce dernier prendrait part à la production de l'énergie requise pour le processus infectieux.

b.3 - Il a été montré autre part que le transporteur de phosphate PHO4 a été indépendamment acquis par EhV, certains prasinovirus et certains cyanophages (194); tandis qu'une ATPase spécifiquement induite lors de carence en phosphate (phoH) est présente chez plusieurs prasinovirus (194) ainsi que chez plusieurs cyanophages (193). Conjointement, ceci suggère que manipuler le métabolisme du phosphate de l'hôte est une adaptation importante des virus marins. D'autres observations sont compatibles avec cette hypothèse : (i) la distribution des prasinovirus est affectée par la disponibilité du phosphate (195), (ii) le fait que des Chlorella virus, des prasinovirus et *Ectocarpus Siliculosus* virus encodent également des canaux à ions potassium. Leur phylogénie et l'analyse de leur séquence suggérant qu'il s'agit cette fois d'un caractère ancestral de ces virus (voire même que ces canaux pourraient avoir été inventés chez ces virus avant d'être transférés au monde cellulaire) (196, 197).

b.4 - Enfin, le cycle du soufre est lui aussi la cible de différents virus. Ceci a déjà été mis en lumière par la présence d'enzymes responsables de l'oxydation du soufre chez plusieurs phages (198). La présence de l'exporteur de composés soufrés TauE, indépendamment acquis de sources bactériennes par CeV d'une part, et PgV d'autre part, semble en être un autre exemple (Figure 9). Chez les bactéries, TauE est impliqué dans l'exportation de sulfite ou sulfoacetate résultant spécifiquement de la consommation de taurine (199, 200). Le rôle de TauE pendant une infection par CeV ou PgV reste cependant à déterminer puisque, à l'exception de certaines algues vertes (201, 202), seules les bactéries sont connues pour cataboliser la taurine. De façon intéressante (pour les producteurs et/ou amateurs de vin au moins), un autre virus encode TauE. Il s'agit d'*Oenococcus oeni* phage, qui infecte *Oenococcus oeni* la bactérie responsable de la fermentation malolactique. Ce processus, qui suit la fermentation alcoolique, participe à une désacidification, ainsi que, indirectement, à une modification des propriétés organoleptiques

(arômes, goût, couleur) des vins (203, 204). L'enzyme du phage est à 99% identique à celle de plasmides de son hôte bactérien, ce qui suggère qu'elle est héritée de l'hôte. La présence de sulfites, soit résultant de la fermentation alcoolique réalisée par les levures, soit ajoutée dans les cuves par les viticulteurs, influence la fermentation malolactique (205). La possession de plasmides codant pour TauE a donc été présentée comme une adaptation au vin de ces bactéries (205). L'acquisition secondaire de cette enzyme par le phage (qui, via l'infection de son hôte bactérien, pourrait indirectement influencer la fermentation malolactique (203, 206)) en serait une seconde, du phage cette fois.

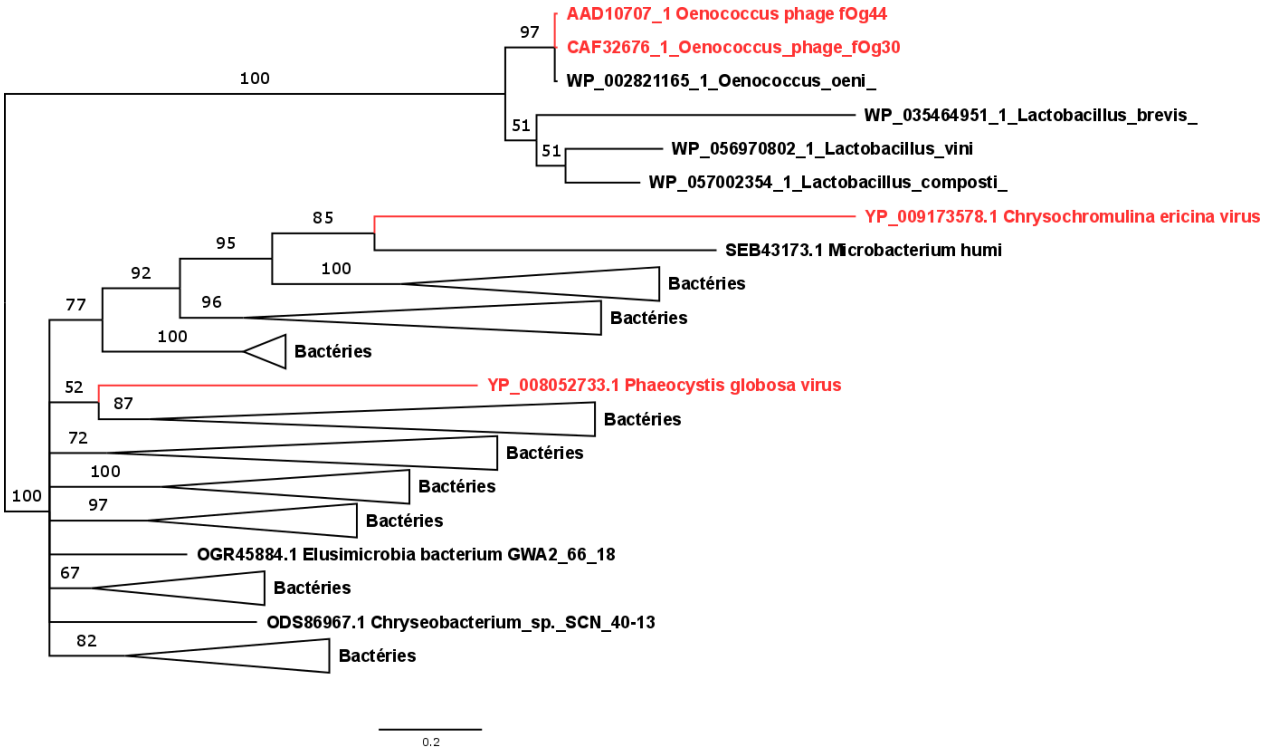


Figure 9. Phylogénie des protéines TauE. Les protéines ont été alignées grâce au logiciel Muscle puis MrBayes (modèle mixed) a été utilisé pour l'inférence des relations phylogénétiques. Les probabilités postérieures sont indiquées en pourcentages le long des branches. Les branches virales sont en rouge.

Conclusion : ces quelques exemples d'acquisitions indépendantes du même gène (mais provenant de sources différentes) de plusieurs branches de NCLDVs infectant des algues et de cyanophages (à l'exception du dernier exemple), mettent en relief des convergences évolutives de virus dont les hôtes sont éloignés phylogénétiquement, mais partageant certaines caractéristiques environnementales et/ou certains métabolismes. Par ailleurs, la possession de certains de ces AMGs par les NCLDVs pourrait avoir un impact sur les cycles des nutriments et du carbone, comme le font ceux décrits chez les phages.

c - Les Asparaginyl-tRNA synthétase (AsnRS) de NCLDVs

Les aminoacyl-tRNA synthétase (AARSs) codées par Mimivirus ont été les premières AARS virales reportées (207). Or, ces enzymes ont un rôle central dans le processus de traduction des protéines, et ce processus était supposé exclusivement cellulaire. Ces AARS alimentent le débat sur une potentielle origine cellulaire de ces virus (208, 209), une attention particulière leur est donc portée. Ceci m'a mené à m'intéresser à 2 asparaginyl tRNA synthétase (AsnRS) de NCLDVs. La première est codée par CeV et a été détectée lors de l'annotation du génome. La seconde est codée par *Heterosigma akashiwo* Virus, un NCLDV qui infecte une micro-algue raphydophyte (stramenopile) et dont le génome a été publié pendant ma thèse (180). Megavirus et Moumouvirus encodent également une AsnRS. Une question que l'on peut se poser est la suivante : Est-ce que la possession d'une AsnRS est un caractère ancestral (i.e., provenant d'un virus ancestral possédant un appareillage autonome pour la traduction), ou bien est-ce un caractère acquis ? Un article récent (139) suggère qu'il s'agit d'un caractère acquis mais n'avais pas encore été publié quand nous étudiions CeV au laboratoire. Effectivement ni l'AsnRS de CeV, ni celle d'HaV ne forment un groupe monophylétique avec celles de Megavirus et Moumouvirus (ces deux dernières par contre sont bien regroupées dans la phylogénie) (Figure 10). Cette configuration polyphylétique des AsnRSs virales soutient l'hypothèse qu'il s'agit plutôt d'un caractère acquis récemment par des NCLDVs. Notons cependant que des cas de replacements fonctionnels sont connus (179, 186) et pourraient également constituer une explication alternative. Les AsnRS de CeV et d'HaV sont polyphylétiques. Fait intriguant, elles ont tout de même un caractère en commun : elles ne possèdent qu'un seul des deux domaines nécessaires à l'activité AARS. Effectivement, il leur manque le domaine de reconnaissance de l'anticodon sur l'ARNt, et ne possèdent que le domaine « catalytique ». Celui-ci lie l'acide aminé et effectue habituellement la liaison entre l'acide aminé et l'ARNt lors de la traduction. Ces AsnRS-like virales pourraient donc avoir une fonction différente des AsnRSs classiques. Plusieurs études décrivent de tel cas (210–214) pour des AsnRS ou d'autre AARS. Il semblerait effectivement que d'une manière générale, les paralogues des AARS soient de façon réccurente l'objet de cooption pour une autre fonction que celle de lier un ARNt et son acide aminé associé lors de la traduction.

Quelle pourrait être celle des Asn-like de CeV et HaV? Je présente ci-dessous quelques éléments de réponse fondés sur l'analyse phylogénétique de ces protéines.

c.1 - HaV

Bien que les enzymes apparentés à celle de HaV forment un groupe sœur aux AsnRS canoniques (bactérienne, archée, eucaryotes) (Figure 10), aucunes ne possèdent le domaine de reconnaissance de l'anticodon l'AsnRS.

L'histoire des enzymes de cette branche a été bien décrite (210, 215, 216): cette protéine résulterait d'une duplication de l'AsnRS chez l'ancêtre cellulaire universel LUCA, aurait perdu

son domaine de reconnaissance de l'anticodon, puis évolué vers une activité de synthèse de l'asparagine à partir d'aspartate et d'ammonium. Elle est appelée AsnA (pour asparagine synthase de type A). Ce cas d'école montre que les voies empruntées par l'évolution sont parfois sinueuses: cette AsnRS-like a été recrutée pour participer au métabolisme de son acide aminé associé. HaV n'a donc pas d'AARS, mais bien une asparagine synthase.

Or, les *Mimiviridae* ainsi que les prasinovirus, ont également un gène codant pour la synthèse de l'asparagine, mais de type S (AsnS) (le NH3 n'est plus tiré l'ammonium mais résulte d'une hydrolyse de la glutamine). Il s'agit donc ici d'un autre type de convergence entre HaV d'un côté et les *Mimiviridae* et prasinovirus de l'autre côté, puisque cette fois un même produit est obtenu mais grâce à deux enzymes non homologues. Le rôle de l'asparagine chez ces virus, qui semble primordial aux vues de ces recrutements à répétition, reste à déterminer.

c.2 - AsnRS de CeV

De même que celle d'HaV, les enzymes de la branche de CeV sont composées uniquement du domaine catalytique sans domaine liant l'anticodon. Elles forment elles aussi une branche sœur de toutes les AsnRS (bactérienne, archée, eucaryotes), et donc résulteraient également d'une duplication ancestrale de l'AsnRS. Au contraire de celle d'HaV, cette branche n'a jamais été décrite, et ce alors que l'histoire des AsnRS a été bien étudiée (210, 215, 217). De façon étonnante, les homologues les plus proches de la protéine de CeV sont des protéines de cyanophages infectant les cyanobactéries *Prochlorococcus* et *Synechococcus*. Moins proche, mais toujours dans cette branche sont présentes des protéines aux taxonomies aussi variées que des bactéries, des « poissons » téléostéens, un fungi, un brachiopode, un cnidaire et un alveolata. A l'exception des virus, tous ces organismes encodent bien les *bona fide* AspRS et AsnRS. Certaines de ces protéines (celles de téléostei et de Hydra) présentent des domaines additionnels en N-term. Ceci pourrait nous donner des indications sur la fonction de la Asn-like de CeV. En effet, lorsque deux protéines distinctes A et B dans un organisme 1 sont fusionnées dans un organisme 2, ceci est révélateur d'une relation fonctionelle putative entre les domaines A et B qui la compose. Dans de nombreux cas, les deux protéines interagissent physiquement ou participent à des étapes consécutives d'une même voie métabolique (218). La recherche de domaines fusionnés a été utilisée pour prédire la fonction de protéines de fonction inconnue (219). Cette technique basée sur le principe de « coupable par association » est appelée « méthode de la pierre de rosette » (220) parce que la protéine AB contient de l'information sur A et sur B.

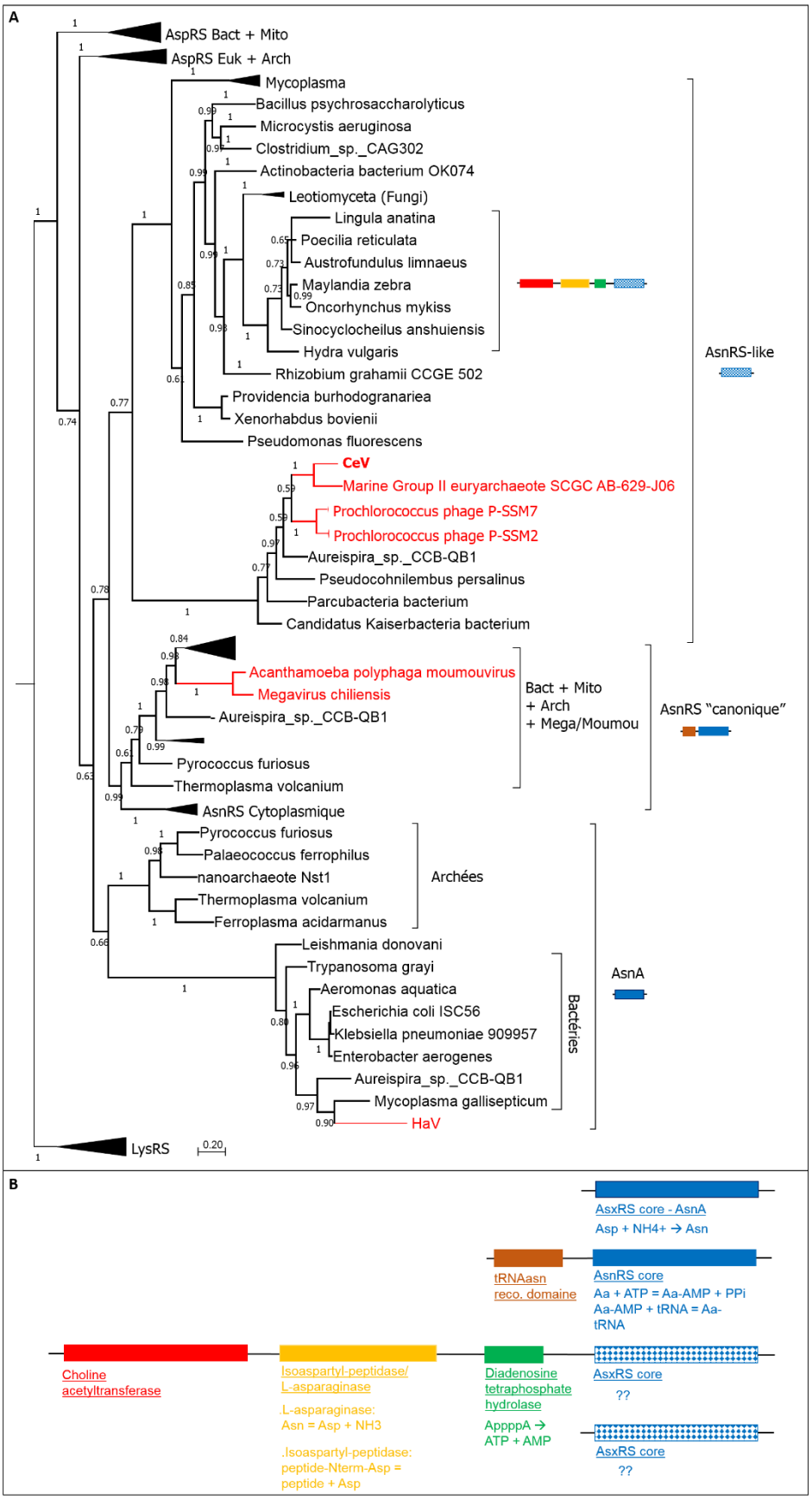


Figure 10: **A -** Phylogénie générale des AsnRS réalisée avec MrBayes à partir d'un alignement exécuté avec PSICOFFEE (voir la partie matériel et méthodes). Les probabilités postérieures sont notées le long des branches. L'arbre a été enraciné avec les LysRS, dont la duplication engendra l'ancêtre aux AspRS et AsnRS. Les branches et feuilles rouges sont virales. « Marine groupe euryarchaeote SCGC AB 629 J06 » est en fait un virus proche de CeV qui a été mal annoté. La composition en domaines des AsnRS est schématisée à côté des branches correspondantes. **B -** Compositions en domaines des différentes protéines contenant un domaine AsnRS-core. Les noms des domaines sont soulignés et les réactions

Les 3 domaines additionnels prédits des protéines de poisson et de hydra sont représentés sur la figure 10, ainsi que les réactions catalysées par ces domaines. En postulant que l'AsnRS-like participe aux mêmes voies métaboliques que ces domaines, voici les quelques pistes de réflexion auxquelles ont abouti la méthode de la pierre de rosette, additionnée de recherches dans la littérature :

Un point à prendre en compte est que nous ne pouvons pas savoir (sur la base de la phylogénie) si le domaine AsnRS-like lie l'aspartate ou l'asparagine. Effectivement, la branche de cette AsnRS-like est imbriquée entre l'AsnA (qui lie l'aspartate (210), comme nous avons vu au paragraphe précédent) et l'AsnRS canonique qui lie l'asparagine.

1/ Domaine asparaginase. L'asparaginase hydrolyse l'asparagine en aspartate et amonium. Un lien entre ce domaine et l'AsnRS-like semble évident : l'asparagine (ou l'aspartate). Cet indice est maigre, mais est légèrement en faveur d'une intervention du domaine AsnRS-like dans le métabolisme des acides-aminés plutôt que dans la traduction.

2/ Domaine Diadenosine tetraphosphate hydrolase HIT: Ce domaine hydrolyse les Diadenosine tetraphosphate (Ap4A) en ATP + AMP (221). La fonction précise de Ap4A n'est pas connue, mais la molécule pourrait être impliquée dans la régulation de la réplication et/ou intervenir dans la réponse au stress (221). Un lien possible avec l'AsnRS a été détecté par une recherche dans la littérature. Effectivement, il apparait que les AARS sont recrutées pour la synthèse de Ap4A. Cela se déroule de la façon suivante : la réaction de ligation de l'aa à l'ARNt nécessite un intermédiaire aa-AMP. Une molécule d'ATP peut ensuite réagir avec cet intermédiaire à la place de l'ARNt ce qui induit la synthèse de Ap4A (222). Ainsi cette molécule d'Ap4A est commune aux réactions catalysée par les domaines HIT et AARS. La co-présence de ces deux domaines dans une protéine de fusion est compatible avec un domaine AsnRS-like dédié à la synthèse de Ap4A.

3/ Aucun lien entre le domaine « choline acetyltransférase » et AsnRS n'a pu être mis en évidence.

En conclusion, en utilisant la méthode de la pierre de rosette, nous pouvons émettre l'hypothèse que l'AsnRS-like de CeV pourrait être impliquée dans le métabolisme des acides aminés aspartate ou asparagine, et/ou dans dans le métabolisme de Ap4A. Tout ceci reste hautement spéculatif et demande à être étudié par des expériences d'activités.

Par ailleurs, j'ai mentionné que l'enzyme de CeV et celle de cyanophages (phages infectant des cyanobactéries) forment un groupe monophylétique. Il pourrait donc s'agir d'un transfert entre virus infectant des royaumes cellulaires différents. Certaines algues, dont des haptophytes, possèdent un microbiote bactérien varié dont peuvent faire partie certaines cyanobactéries (223, 224). Dès lors, ces communautés rapprochées pourraient être le lieu de transferts de gènes entre organismes cellulaires (entre eux), organismes cellulaires et virus, ou virus entre eux.

d - Spectre d'hôte et date de divergence des *Mimiviridae*

L'isolement de ces nouveaux *Mimiviridae* qui infectent des protistes dont la divergence avec les amibes est un évènement très basal dans l'histoire du domaine eucaryote, suggère pour certains que l'origine des *Mimiviridae* pourrait avoir précédé cette radiation et donc être contemporaine de l'émergence du domaine eucaryote (102). Cette hypothèse ne fait pas l'unanimité puisqu'un autre scénario possible, bien que moins parcimonieux, propose que plusieurs sauts d'hôte soient responsables de la distribution d'hôtes actuelle (140). Un mixe de ces deux scenarii est également envisageable. D'autres éléments sont compatibles avec l'hypothèse de co-radiation. Nous avons déjà évoqué le cas des enzymes impliquées dans la réplication et la réparation de l'ADN ainsi que de celles impliquées dans la transcription qui forment une branche monophylétique émergeant à la racine du domaine eucaryote (141).

Un autre argument nous vient de l'analyse de la distribution du gène MutS7, de la famille des MSH (mutation suppressor homolog), exclusif aux proteobactéries epsilon, aux *Mimiviridae* (tous) et aux octocoralliaires. En effet, la phylogénie suggère que l'explication la plus parcimonieuse expliquant cette distribution serait qu'un *Mimiviridae* ait transféré ce gène depuis les bactéries vers les mitochondries des coraux (102). Toujours d'après l'arbre, ceci ce serait passé après la divergence des « *Megavirinae* » et « *Mesomimivirinae* », et avant la radiation des octocoralliaires (mais après la divergence de ces dernier avec les *Hexacorallia*, car leurs mitochondries n'encodent pas MutS7). Ce scenario s'il est correct, implique que la divergence des *Mimiviridae* a eu lieu il y a au moins 750 millions d'année (date d'estimation récente de la radiation des coraux (225)). Ceci est certes longtemps après la date d'émergence supposée du domaine eucaryote, mais pose déjà une limite au caractère récent de l'évènement.

Que l'hypothèse exacte soit celle d'une divergence des *Mimiviridae* aussi ancienne que le domaine eucaryote ou celle de plusieurs sauts d'hôtes éloignés, nous pouvons faire la prédiction que des virus de la famille des *Mimiviridae* pourraient (ou aurait pu) infecter des hôtes cellulaires appartenant à d'autres super-groupes eucaryotes en plus des 4 hébergeant des hôtes connus (figure 13). L'analyse de traces laissées dans les génomes eucaryotes (ceci me permet de ménager une transition vers la seconde partie de mon travail de thèse), pourraient nous permettre de valider cette prédiction.

4 - Conclusion

L'analyse du génome de CeV a permis de confirmer que les *Mimiviridae* infectant les algues forment leur propre clade au sein de cette famille. A côté de ce trait commun – avoir un hôte photosynthétique – une autre des rares caractéristiques communes à ces virus est leur taille inférieure à celle des autres *Mimiviridae*. L'isolement d'autres virus de ce clade et/ou de cette

famille permettra de confirmer l'universalité de cette tendance, ou au contraire, de l'infirmer. Effectivement, la question de l'état ancestral de la taille du génome des *Mimiviridae* – « géante », ou bien plus petite – n'a pas pu être résolue par les méthodes utilisées. D'autre part, les 3 virus de ce sous-clade de *Mimiviridae* exhibent un vaste répertoire de gènes uniques à chacun, dont la majorité n'est apparemment pas héritée des cellules. Leur origine, ainsi que leur fonction, restent à déterminer. CeV et d'autres virus de taxa variés infectant des hôtes photosynthétiques présentent des convergences évolutives sous la forme d'acquisitions parallèles de gènes homologues, chez leurs hôtes ou bien chez d'autres organismes cellulaire. Les cas répertoriés ici concernent majoritairement l'acquisition de gènes impliqués dans le métabolisme, c'est pourquoi ils ont été qualifiés chez les phages d'AMG pour « Auxiliary metabolic genes ». Ces AMGs pourraient complémenter ou rediriger le métabolisme de leur hôte. Des expériences concernant la fonction, lors de l'infection, de la protéine LHC du virus OtV, ont été initiées par l'observatoire océanographique de Banyuls en collaboration avec un laboratoire du Mans après que j'ai présenté un poster décrivant les acquisitions parallèles des protéines LHC par CeV et OtV lors du congrès « Aquatic virus workshop 8 » à Plymouth en juillet 2016. D'autre part, de la même façon que chez les phages, ces AMGs pourraient avoir un impact sur les cycles globaux des nutriments et du carbone. Leurs acquisitions parallèles par des virus non apparentés suggèrent qu'ils sont hautement bénéfiques pour les virus lors de l'infection, et a confirmé que la possession de gènes homologues par différents virus n'est pas toujours révélatrice d'une origine ancestrale. Ainsi, l'AsnRS-like de CeV n'est pas l'orthologue de celles de Megavirus et Moumouvirus, mais forme un clade avec des enzymes de phage, qui émerge à la base des AsnRS cellulaire. Elle est par ailleurs dépourvue du domaine liant l'ARNt. Sa fonction chez les virus ainsi que chez les quelques organismes cellulaires codant pour ce même prototype d'AsnRS-like reste à déterminer. La famille des *Mimiviridae* présente un spectre d'hôte étendu. Concernant ce point, le second volet de ma thèse, qui a consisté à rechercher et analyser des inserts NCLDVs dans les génomes eucaryotes, pourrait apporter des informations additionnelles.

Chapitre B - Séquences de NCLDVs dans les génomes eucaryotes

1 - Introduction générale

a - Les NCLDVs, des oubliés de la paléovirologie

Nous avons également voulu savoir quelle relation intime les virus géants entretiennent-ils avec leurs hôtes au cours de l'évolution. Il a souvent été avancé que les virus géants ont acquis des gènes cellulaires par transfert horizontal. Cette hypothèse a été validée par des cas précis mais l'ampleur du phénomène est controversée. Certains chercheurs comme Filée et Chandler soutiennent que les virus géants sont des « gene-pickpockets » (226–229), quand d'autres pensent que les évidences ne sont qu'anecdotiques et que le phénomène n'est pas plus fréquent que le HGT bactérien (188, 189, 230–232). Ces derniers insistent sur le fait que, dans les cas où un gène est partagé par les NCLDVs et les cellules, l'orientation du transfert cellules vers virus ou virus vers hôte reste souvent indéterminée (188). Il y a peu d'études qui se sont attachées aux transferts clairement orientés NCLDVs vers hôte cellulaire. En effet, la littérature sur la « paléovirologie », i.e., l'étude des séquences virales anciennes intégrées dans les génomes (233, 234), concerne surtout les autres groupes viraux, notamment les retrovirus pour lesquels le mécanisme sous-jacent est évident. En effet, ceux-ci s'intègrent dans le génome de leur hôte pour se répliquer (dépendant pour cela d'une enzyme appelé intégrase). Ils peuvent alors perdre leur capacité à effectuer un cycle lytique à la suite de mutation par exemple. Le génome du provirus est alors progressivement dégradé. Il arrive cependant que l'origine virale de certaines portions du génome cellulaire puisse être détectée, soit parce que cet épisode est récent, soit parce qu'au lieu d'être dégradé, ils ont été conservés par la cellule. Ceci a permis de révéler le rôle des virus dans l'apparition de phénotypes inédits chez les cellules (voir les cas décrits en introduction), ainsi que des couples hôtes/virus passés.

Les NCLDVs n'ont pas de leur côté de mécanismes actifs permettant une insertion dans le génome de leur hôte, à l'exception des Phaeovirus qui ont un cyle lysogénique. Néanmoins, le génome viral est en contact très intime avec le génome de la cellule (en tout cas ceux qui ont une phase nucléaire) lors du processus de réplication. Ces contacts rapprochés entre les deux génomes peuvent donner lieu à des évènements d'intégration d'ADN viral dans le génome cellulaire par accident. Par exemple, l'utilisation d'ADN viral pour combler des cassures double brin a été décrit lors de la réplication d'un hepadnaviridae dans une cellule tumoral de foie de poulet (234, 235). Ceci pourrait également arriver durant un processus de réparation de l'ADN d'un hôte infecté par un NCLDV. Si la cellule ne meurt pas, soit, par exemple, parce qu'elle possède des mécanismes de défense stoppant la réplication du virus, ou bien parce que le virus n'est pas compétant pour terminer son cycle réplicatif dans l'hôte infecté, alors elle est maintenant dotée de matériel génétique d'origine NCLDV. Dans le cas d'organismes unicellulaires, cet ADN viral intégré sera

transmis à la descendance. Dans le cas d'organismes multicellulaires, ceci n'arrivera que si l'évènement d'intégration a eu lieu dans la lignée germinale. Les génomes de virus géants contiennent des centaines voire des milliers de gènes, ce qui offre d'autant plus d'opportunités dans la nature des gènes échangés ; une grande partie des gènes de virus géants ont des fonctions inconnues et n'ont pas d'homologues reconnaissables chez les êtres cellulaires: il s'agit donc là d'un formidable réservoir de nouvelles fonctions susceptibles d'être domestiquées par les eucaryotes pour leur donner un avantage sélectif original. Quelques études initiales (236–240), décrites au paragraphe suivant, suggèrent qu'effectivement les NCLDVs laissent des traces de leur passage dans les cellules.

Nous avons donc cherché des évidences de HGT virus->eucaryote pour observer ce phénomène.

b - État des lieux avant ma thèse, historique

Voici un bref historique des études qui ont révélé la présence d'insertions de séquences/génomes de NCLDVs dans les génomes eucaryotes.

1998 - Découverte du premier virus ADN double brin d'eucaryote ayant un mode persistent d'infection. (241). *E. siliculosus* virus est détecté dans le génome de l'algue brune (phaeophycea) *E.siliculosus* (236). Il constitue le membre fondateur des virus du genre *Phaeovirus*, seul genre de NCLDV connu dont les membres sont lysogénique.

2010-2012 - Des génomes d'algues de taxa variés sont séquencés. Certaines de ces algues sont par ailleurs des hôtes connus de NCLDVs. L'analyse de leur génome a révélé la présence de segments de génomes NCLDVs endogénisé (*E.huxleyi*, haptophycea) (238) ou bien de gènes partagés avec les virus, qui pourraient donc provenir des virus (*Chlorella variabilis NC64*, chlorophycea) (239). Les NCLDVs qui infectent ces algues ne sont pas lysogéniques, ce qui suggère que d'autres mécanismes peuvent donner lieu à des intégrations de segments géniques viraux.

2014 - Florian Maumus de l'INRA de Versailles et Guillaume Blanc à l'IGS réalisent une recherche grande échelle de gènes de NCLDVs dans les 13 génomes de plantes terrestres disponibles dans les banques de données. Cette étude fondatrice révèle la présence de séquences de NCLDVs endogènes dans les plantes embryophyte *P.patens* et *Selaginella moellendorffii* (240). La phylogénie des séquences virales suggère que le virus donneur appartient à une nouvelle famille de NCLDVs. Par ailleurs, ces plantes ne sont pas des hôtes connus de NCLDVs. Ce type d'étude a donc permis de mettre en évidence des associations eucaryotes/NCLDVs, actuelles ou passées, jusque-là ignorées, et d'augmenter notre connaissance de la biodiversité des NCLDVs.

2014 - Filée a proposé que la DNA primase eucaryote, qui sert à initialiser la réplication de l'ADN, pourrait avoir été remplacée par une D5 primase-helicase de *Mimiviridae* (242). Ainsi, de la même

façon que documenté pour les retrovirus, les gènes NCLDVs intégrés peuvent être domestiqués par les cellules.

c - Et les virophages ..

D'autre part, j'ai précédemment présenté les virophages et mentionné qu'ils possèdent une intégrase. Il a été montré que le virophage Sputnik (151) (isolé avec Mimivirus) peut s'intégrer dans le génome de son hôte virus (166). Nous ne connaissons pas d'exemple d'insertions de virophages dans les génomes eucaryotes, mais ceci apparait comme une possibilité.

d - Problématique

La découverte préliminaire des insertions virales dans les génomes de plantes nous a conduit à nous interroger sur l'importance des transferts horizontaux d'origine NCLDVs à l'échelle de tous les eucaryotes. Dans cette perspective, nous avons réalisé une étude beaucoup plus exhaustive des séquences génomiques eucaryotes disponibles dans les bases de données publiques, dans lesquelles nous nous sommes efforcés d'identifier les vestiges d'anciennes insertions d'ADN apparentés aux génomes de NCLDVs. Nous avons utilisé la même démarche pour rechercher de possibles insertions de génomes de virophages chez les eucaryotes. Les résultats de cette analyse sont ici présentés dans les deux chapitres suivants : le premier révèle l'existence de copies de génomes de virophages et de NCLDVs insérées dans le génome d'une algue unicellulaire, tandis que le second dresse un état des lieux de l'incidence des insertions génomiques de NCLDV dans l'ensemble des eucaryotes séquencés.

2 - *Bigelowiella natans* – un génome analysé en détail

a - Introduction

Florian Maumus, intéressé par les éléments répétés de la famille des Maverick-Polinton, a détecté des séquences apparentées dans le génome de l'algue Chlorarachniophyte (Rhizaria) *Bigelowiella natans*. Sur le principe de sa collaboration précédente avec Guillaume Blanc sur les plantes, il a fait appel à nous pour réaliser une étude approfondie du génome de l'algue. Nous avons finalement pu montrer que les séquences qu'il avait initialement détectées appartenaient en fait à des virophages dont des copies de génome étaient intégrées dans le génome de *B.natans*. Par ailleurs, nous avons aussi découvert des fragments de génomes de NCLDVs insérés dans le génome de cette même algue. Dans cette étude conduite à trois personnes, j'ai pris en charge l'analyse des séquences d'origine NCLDVs présentes dans le génome de *Bigelowiella natans*.

b - Article 2

Provirophages in the Bigelowiella genome bear testimony to past encounters with giant viruses. Blanc, G., Gallot-Lavallée, L., and Maumus, F. (2015).

Provirophages in the Bigelowiella genome bear testimony to past encounters with giant viruses

Guillaume Blanc[a,1,2], Lucie Gallot-Lavallée[a], and Florian Maumus[b,1,2]

[a]Laboratoire Information Génomique et Structurale, UMR7256 (Institut de Microbiologie de la Méditerranée FR3479) CNRS, Aix-Marseille Université, 13288 Marseille cedex 9, France; and [b]INRA, UR1164 Unité de Recherche Génomique-Info, Institut National de la Recherche Agronomique de Versailles-Grignon, 78026 Versailles, France

Edited by Peter Palese, Icahn School of Medicine at Mount Sinai, New York, NY, and approved July 24, 2015 (received for review April 1, 2015)

Virophages are recently discovered double-stranded DNA virus satellites that prey on giant viruses (nucleocytoplasmic large DNA viruses; NCLDVs), which are themselves parasites of unicellular eukaryotes. This coupled parasitism can result in the indirect control of eukaryotic cell mortality by virophages. However, the details of such tripartite relationships remain largely unexplored. We have discovered ∼300 predicted genes of putative virophage origin in the nuclear genome of the unicellular alga *Bigelowiella natans*. Physical clustering of these genes indicates that virophage genomes are integrated into the *B. natans* genome. Virophage inserts show high levels of similarity and synteny between each other, indicating that they are closely related. Virophage genes are transcribed not only in the sequenced *B. natans* strain but also in other *Bigelowiella* isolates, suggesting that transcriptionally active virophage inserts are widespread in *Bigelowiella* populations. Evidence that *B. natans* is also a host to NCLDV members is provided by the identification of NCLDV inserts in its genome. These putative large DNA viruses may be infected by *B. natans* virophages. We also identify four repeated elements sharing structural and genetic similarities with transpovirons—a class of mobile elements first discovered in giant viruses—that were probably independently inserted in the *B. natans* genome. We argue that endogenized provirophages may be beneficial to both the virophage and *B. natans* by (*i*) increasing the chances for the virophage to coinfect the host cell with an NCLDV prey and (*ii*) defending the host cell against fatal NCLDV infections.

virophage | nucleocytoplasmic large DNA virus | microbial community | endogenous virus | Maverick/polinton

Sputnik was first described in 2008 as a new class of small icosahedral viruses with an ∼20-kb circular double-stranded DNA genome (1). Sputnik is a satellite virus, because its replication depends upon proteins produced by the nucleocytoplasmic large DNA virus [NCLDV; also giant virus or proposed order *Megavirales* (2)] *Acanthamoeba polyphaga* Mimivirus (APMV; *Mimiviridae*) and replicates in APMV viral factories. Sputnik was shown to inhibit replication of its helper virus and thus acted as a parasite of that virus. In analogy to the term bacteriophage it was called a virophage, but this designation has been challenged (3). Three additional virophages infecting members of the *Mimiviridae*, e.g., Sputnik 2, Rio Negro, and Zamilon, were subsequently reported (4–6). Virophages that prey on giant viruses that infect heterotrophic nanoflagellates and microalgae have also been discovered, including Organic Lake virophage 1 [OLV1 (7)], Mavirus (8), and a virophage of the *Phaeocystis globosa* virus (PgVV) (9), yet the classification of the latter as a virophage sensu stricto is uncertain. In addition, complete or near-complete virophage genomes have been assembled from environmental DNA: Yellowstone Lake virophages 1–7 (YSLV1–7) and Ace Lake Mavirus (ALM) (10, 11).

Overall, virophage genomes have similar sizes (∼18–28 kb) and low G+C content (∼27–39%) and are related to Sputnik by genetic and structural homologies (12). Among the 20–34 protein-coding sequences predicted in virophage genomes, the putative core gene set comprises six genes encoding the FtsK-HerA family DNA-packaging ATPase (ATPase), primase-superfamily 3 (S3) helicase, cysteine protease (PRO), and zinc-ribbon domain (ZnR) as well as major and minor capsid proteins (MCPs and mCPs, respectively) (12). In addition, genes encoding two different families of integrases have been identified in several virophages: A putative rve integrase was found in Mavirus and ALM (8, 10), whereas Sputnik encodes a putative tyrosine integrase (1). Among virophage genes, only PRO, ATPase, MCP, and mCP support the monophyly of virophages, whereas the remaining gene complement shows complex phylogenies suggestive of gene replacement (12).

Remarkably, phylogenetic analysis of the Mavirus rve integrase indicated that it is mostly related to homologs from eukaryotic mobile elements of the Maverick/polinton (MP) family (8). The polintons are widely distributed in diverse protists and animals and were initially classified as transposable elements (TEs) (13, 14). However, convincing arguments support the hypothesis that polintons encode capsid proteins and might be bona fide viruses (15). Because Mavirus was reported to display further synapomorphy with a putative MP from the slime mold *Polysphondylium pallidum*, it was hypothesized that MPs may have originated from ancient Mavirus relatives that would have acquired the capability of intragenomic transposition (8). However, this hypothesis was recently challenged by Yutin et al. (12). A critical prerequisite for such an evolutionary scenario is the integration of virophage DNA in the genome of a eukaryotic host that would permit vertical transmission and adaptation to an intracellular parasitic lifestyle. However, although Sputnik 2 was shown to integrate into the genome of its Mimivirus host (4), evidence of virophage insertions in eukaryotic genomes is lacking.

Significance

Virophages are viruses that hijack the replication machinery of giant viruses for their own replication. Virophages negatively impact giant virus replication and improve the survival chances of eukaryotic cells infected by giant viruses. In this study, we identified segments of the *Bigelowiella natans* genome that originate from virophages and giant viruses, revealing genomic footprints of battles between these viral entities that occurred in this unicellular alga. Interestingly, genes of virophage origin are transcribed, suggesting that they are functional. We hypothesize that virophage integration may be beneficial to both the virophage and *B. natans* by increasing the chances for the virophage to coinfect the cell with a giant virus prey and by defending the host cell against fatal giant virus infections.

Author contributions: G.B. and F.M. designed research; G.B., L.G.-L., and F.M. performed research; G.B., L.G.-L., and F.M. analyzed data; and G.B. and F.M. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

See Commentary on page 11750.

[1]G.B. and F.M. contributed equally to this work.

[2]To whom correspondence may be addressed. Email: guillaume.blanc@igs.cnrs-mrs.fr or fmaumus@versailles.inra.fr.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1506469112/-/DCSupplemental.

The coinfection of host cells by NCLDVs and virophages has been shown to limit the production of NCLDV particles, accompanied by greater survival of the eukaryotic host (1, 5, 6, 8, 16). At the community level, these parasitic relationships result in complex interplays between virophages, NCLDVs, and eukaryotic hosts. As a result, virophages indirectly positively regulate the population size of eukaryotic hosts. At the global scale, these interactions may have significant impacts on biogeochemical cycles. For instance, in the marine environment, the co-occurrence of giant viruses and virophages in the context of algal blooms may influence the overall carbon flux as proposed for Antarctic lakes (7). Nevertheless, such tripartite community networks remain poorly explored except on theoretical grounds (17, 18).

The ecological prominence and diversity of virophages are largely unknown and wait for the isolation and sequencing of new specimens. Recently, we demonstrated the value of searching integrated viral DNA in the genomes of potential eukaryotic hosts to identify new members of the NCLDVs (19). Here we analyzed the nuclear genome assemblies of 1,153 fully sequenced eukaryotes and report the identification of integrated virophage elements in the genome of the Chlorarachniophyte *Bigelowiella natans* (supergroup Rhizaria). This discovery led to the prediction that this alga is also the host of viruses that are members of the NCLDVs. In support of this prediction, we also identified inserts of likely NCLDV origin. We investigated the transcriptional activity of the *B. natans* genome using RNA-sequencing (seq) data and show that virophage-like genes are actively transcribed in different *B. natans* strains whereas NCLDV-like genes tend to be silent—albeit with notable exceptions. Finally, we identified repeated genetic elements that have structural and genetic similarities to transpovirons, a distinct class of mobile genetic elements associated with giant viruses that were first discovered in members of the *Mimiviridae* (4). We discuss the biological relevance of integrated, actively transcribed virophages and propose a model for the mode of virophage–NCLDV coinfection. Altogether, our results contribute to the understanding of the genetic interactions occurring within microbial communities between eukaryotes, virophages, and NCLDVs.

Results

Integrated Virophage-Like Elements in the Algal Genome. A previous comparative analysis of fully sequenced virophage genomes revealed six core proteins or protein domains that are universally conserved, including S3 helicase, zinc-ribbon domain, major capsid protein, minor capsid protein, DNA-packaging ATPase, and a cysteine protease (12). The four latter proteins were shown to produce consistent monophyletic clades that contrasted virophages from polintons, a class of repeated elements related to virophages. We therefore used these proteins as markers of DNA inserts of putative virophage origin in eukaryotic genome sequences. In practice, we searched 1,153 predicted proteomes of protists, fungi, and basal metazoans for homologs of the four virophage markers using BLASTP. The proteome of *B. natans* was the only one to exhibit homologs for each of the four virophage core protein families. No homolog for any of the virophage markers was identified in the other proteomes using predefined family-specific score thresholds.

To better delineate the subset of *B. natans* proteins that have a potential virophage origin, we used a score plot approach. BLAST scores obtained between *B. natans* predicted proteins and their best virophage matches were plotted against the respective BLAST scores obtained between the *B. natans* predicted proteins and their best cellular matches (Fig. 1). Blue dots below the diagonal identify *B. natans* proteins that have higher similarity to a virophage protein than to a homolog in a cellular organism. Overall, 103 *B. natans* proteins had a match within a virophage proteome, of which 64 had a higher score with virophages than with cellular organisms. Furthermore, examination of the physical location of the virophage-like protein genes revealed that they tend to cluster in specific loci in the genome assembly, revealing large regions of possible virophage origin.

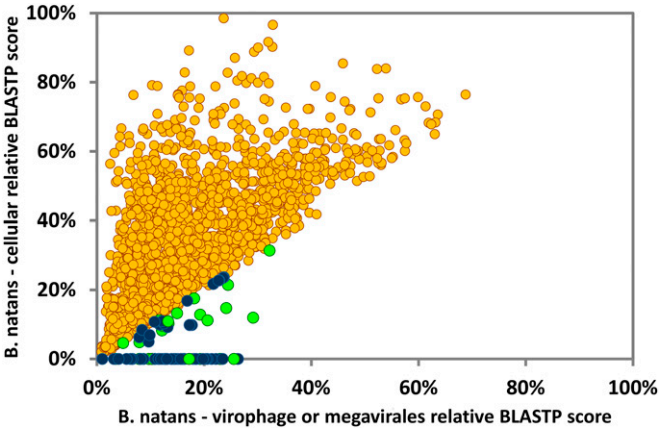


Fig. 1. Similarity plot of *B. natans* proteins against virophage/NCLDV and cellular best hits. Circles represent relative BLASTP scores of *B. natans* proteins aligned against their best cellular hits in the NR database (*y* axis) and their best viral hits among NCLDVs or virophages. When no cellular hit was recorded whereas a viral hit was obtained, the cellular score was set to zero. BLAST scores were normalized by dividing them by the score of the alignment of the query sequence against itself. Circles are colored according to the origin of the best overall scoring hit (yellow, cellular organisms; blue, virophages; green, NCLDVs).

A total of 38 virophage-like elements (VLEs) ranging from 100 base pairs (bp) to 33.3 kb were detected by nucleotide alignments in the *B. natans* genome assembly. Many VLEs correspond to truncated copies of larger elements, suggesting recurrent insertions followed by degradation. However, the number of VLEs may be misestimated because some of them lie at the end of contigs, suggesting that VLEs are difficult to assemble. None of the VLEs were located on the nucleomorph chromosomes, mitochondrial genome, or chloroplast genome. The cumulated size of VLE sequences reaches 327 kb. The VLEs were highly similar between each other—that is, nucleotide identities averaged 91.3%—indicating that they belong to the same family of closely related elements. However, sequence conservation was occasionally interrupted by unique sequences containing one or more genes as shown in Fig. S1, revealing insertion or deletion events that occurred subsequent to their divergence. Some VLEs may be unable to produce viable virophages (i.e., unable to complete a full replication cycle) because important genes may have been lost following integration. Alternatively, the difference in gene content between VLEs may reflect the genetic diversity of virophages before their integration. Six of the identified VLEs contained terminal inverted repeats (TIRs) of 2.0–2.6 kb at their two extremities. TIRs were also described at the extremities of the PgVV virophage-like element [associated with the virus PgV-16T infecting *P. globosa* (9)] and polintons, and are common among poxviruses, chloroviruses, and asfarviruses (7). In *B. natans*, each TIR contained at least two putative ORFs. Other VLEs contained a single TIR copy at one of their extremities, most probably because these elements were truncated.

As shown in Fig. 2*A*, the G+C content of VLEs (36.4% on average) was markedly lower than the background G+C content of the host genome (44.9% on average). Such a difference in G+C content suggests that the VLEs have been acquired horizontally in the relatively recent past.

Virophage-Like Element Genes. Overall, 298 ORFs (>90 codons) were predicted out of the 38 VLEs and organized into 54 gene families (Dataset S1). The largest element of 33.3 kb was identified on scaffold 2 (positions 1,655,224–1,688,550; Fig. 2*B*) and contained 27 predicted ORFs representing 25 distinct gene families listed in Table 1. Functional annotation could be predicted for only 14 of the pan-VLE gene families. Furthermore, 39 gene

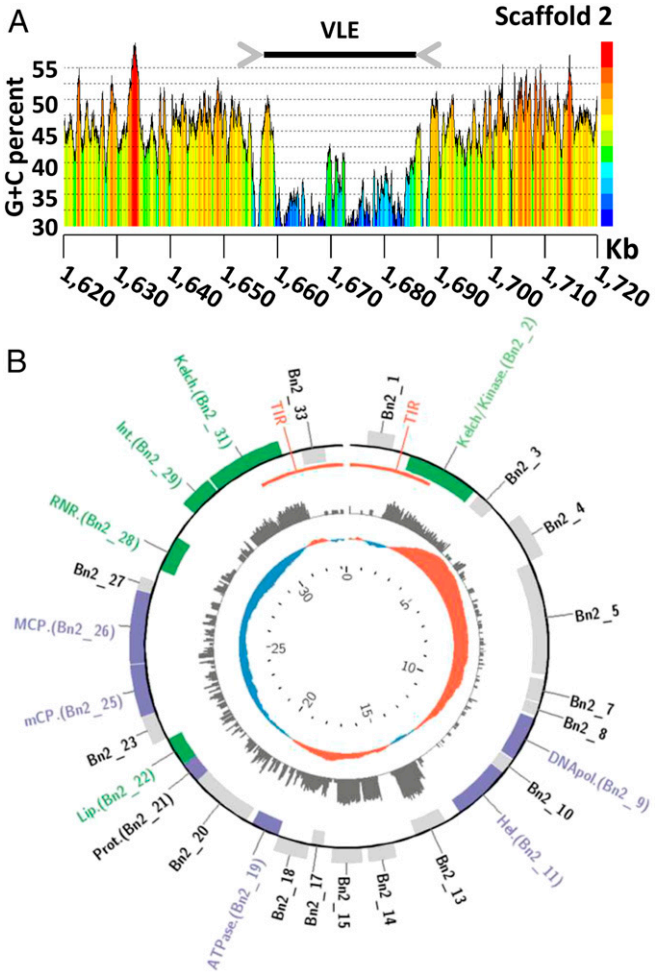


Fig. 2. Schematic representation of the largest virophage-like element of *B. natans*. (*A*) G+C content curve in the region of the virophage-like element on scaffold 2. (*B*) Gene map of the linear virophage-like element. Genes displayed on the outer and inner rims of the outermost track are coded on the direct (Watson) and reverse (Crick) strand, respectively. Purple and green genes correspond to core virophage genes and genes with annotated functions, respectively. The middle track indicates the number of RNA-seq reads mapped to the regions (logarithmic scale). The inner track represents the G+C skew, with positive values in red and negative values in blue. The genome coordinates (in kb) are indicated in the innermost track of the map.

families did not have a significant match in sequence databases (BLASTP E value < 0.01). Such a small fraction of predicted function is a trademark of virophages that generally contain a majority of orphan genes (1, 7, 8). Also, the functions encoded by the *B. natans* VLEs are reminiscent of bona fide virophages. Homologs could be identified for each of the six core genes that are ubiquitously conserved in complete virophage genomes (12). Importantly, these include the major and minor capsid protein genes that are among the viral hallmark proteins distinguishing viruses from other types of mobile elements. In addition, the VLEs encode two key proteins involved in virion maturation, namely the protease and the packaging ATPase. The core virophage zinc-ribbon domain is fused with a GIY-YIG nuclease domain like in Mavirus and OLV; noticeably, the corresponding protein (Bn119_7) is encoded by only one of the VLEs on scaffold 119, suggesting that the other VLEs have lost the gene at some point in their evolution. Furthermore, the VLEs were found to encode an rve family integrase closely related to the one encoded in Mavirus and ALM. This protein might have catalyzed the integration of the original virophage-like DNA into the algal genome. The VLEs also contain five gene families homologous to sequences that have a more patchy distribution among virophages, including genes for lipase, ribonucleotide reductase small subunit, B family DNA polymerase, and two ORFs with unknown function.

Each TIR region encompasses two ORFs, resulting in two pairs of duplicated genes in complete VLEs. The outward duplicated ORFs within TIRs lack functional annotation. On the other hand, the inward duplicated ORFs encode Kelch-repeat domains most similar to eukaryotic and poxvirus homologs (Fig. S2*A*) that may be involved in protein–protein interaction. Interestingly, these ORFs terminate outside the TIR regions, so that the N termini differ between the two protein copies. One of the N termini (Bn2_31) encodes a domain with no match in sequence databases, whereas the N-terminal region of the other protein (Bn2_2) encodes a protein kinase (PK) domain that is indirectly related to those encoded in various giant viruses (Fig. S2*B*). Sequenced virophages have been shown to contain genes that are closely related to homologs from their respective giant virus hosts (1, 7, 8, 12). These apparent host-derived genes encode different repetitive proteins (distinct forms of collagen-like repeats in Sputnik and OLV, and FNIP repeats in Mavirus) that could be implicated in the interaction of the virophages with their giant virus hosts (12). The VLEs encode two gene families that are most similar to giant virus homologs (i.e., Bn2_18 and Bn2_28), but not one encodes repetitive proteins. However, the Kelch domain proteins (i.e., Bn2_2 and Bn2_31), ankyrin domain protein (Bn161_7), and adhesion-like protein (Bn2_5) are proteins with repetitive domains likely implicated in the interaction of the virophages with their giant virus hosts; however, they have no significant match in sequenced giant viruses.

Origin of Virophage-Like Elements. We performed a phylogenetic analysis of the virophage-like proteins that had sufficient evolutionary conservation with homologs in sequence databases. For each protein family, the *B. natans* paralogs aggregated in a single clade containing short branches, further confirming the very close relationships between the different VLEs. The packaging ATPase and protease were previously shown to support the monophyly of virophages (12). The corresponding *B. natans* proteins cluster within the virophage clade (Fig. 3 *A* and *B*). In both these trees, the *B. natans* proteins group with Sputnik, OLV, and YSLVs, whereas Mavirus and ALM cluster in a sister clade. The phylogenetic tree of the major capsid proteins is compatible with the scenario of a closer evolutionary proximity of the *B. natans* VLEs to OLV, YSLVs, and Sputnik than to the ALM-Mavirus clade (Fig. 3*C*). The same conclusion can be drawn from the phylogenetic reconstruction of the virophage lipase (Fig. S2*C*). Moreover, the fast-evolving minor capsid protein was conserved between Sputnik, OLV, Zamilon, YSLVs, and the *B. natans* elements, whereas the Mavirus and ALM homologs were comparatively too diverged to allow accurate phylogenetic reconstruction, further pointing to a more distant relationship with the latter (Fig. 3*D*). Altogether, these features suggest that the *B. natans* virophage elements share a common ancestor with Sputnik-OLV-YSLVs. However, the exact timing of speciation events within this subtree could not be established, owing to incongruences between the phylogenetic reconstructions.

The phylogenetic trees for the other *B. natans* virophage-like proteins disclosed more complex evolutionary scenarios. It was suggested that gene replacements and horizontal gene transfers might explain the entangled phylogenetic relationships between virophages, large DNA viruses, and certain classes of mobile elements related to virophages such as polintons (12). An elegant example is the ORFs encoding a superfamily 3 helicase domain that exists in all fully sequenced virophages (core gene) but fails to support monophyly of the latter, revealing multiple origins of the virophage genes (Fig. S2*D*). Nevertheless, the corresponding *B. natans* proteins cluster with the YSLV3 homolog, suggesting that

Table 1. Virophage-like genes

ORF no.	Putative function	RPKM	Percentile expression,* %	Best hit†
Scaffold 2 largest elements				
Bn2_1	Orphan protein	0.2	3.9	
Bn2_2	Kelch and kinase domain protein	16.7	74.5	*Dendroctonus ponderosae* 478256302 (2e-32)
Bn2_3	Orphan protein	6.3	55.1	
Bn2_4	Unknown virophage protein	1.8	27.5	Zamilon 563399747 (2e-04)
Bn2_5	Adhesin-like protein	1.0	19.0	*Escherichia coli* 693111543 (9e-13)
Bn2_7	Orphan protein	0.9	16.7	
Bn2_8	Orphan protein	0.0	0.0	
Bn2_9	DNA polymerase B	0.9	17.2	Mavirus 326439151 (2e-09)
Bn2_10	Orphan protein	0.5	10.4	
Bn2_11	S3 helicase	0.8	15.1	YSL5 701905635 (2e-28)
Bn2_13	Orphan protein	604.1	99.2	
Bn2_14	Orphan protein	125.7	95.6	
Bn2_15	Orphan protein	194.4	97.2	
Bn2_17	Orphan protein	21.8	78.7	
Bn2_18	DnaJ domain protein	24.4	80.4	*Ostreococcus lucimarinus* virus 313843979 (2e-39)
Bn2_19	DNA-packaging ATPase	14.2	71.8	OLV 322510450 (7e-25)
Bn2_20	Orphan protein	9.4	63.8	
Bn2_21	Cysteine protease	1.5	24.5	OLV 322510453 (5e-08)
Bn2_22	Lipase	5.1	50.5	Mavirus 326439161 (6e-04)
Bn2_23	Orphan protein	4.2	46.1	
Bn2_25	Minor capsid protein	4.2	45.8	OLV 322510454 (2e-11)
Bn2_26	Major capsid protein	3.9	44.0	OLV 322510455 (3e-16)
Bn2_27	Unknown protein	1.4	24.0	*Guillardia theta* 551643434 (4e-16)
Bn2_28	Ribonucleotide reductase small subunit	1.6	25.9	*Cafeteria roenbergensis* virus 310831442 (3e-80)
Bn2_29	rve integrase	2.6	35.3	*Dictyostelium fasciculatum* 470248944 (3e-39)
Bn2_31	Kelch domain protein	33.4	84.5	*Strongylocentrotus purpuratus* 390342441 (6e-35)
Bn2_33	Orphan protein	0.2	5.2	
Other remarkable ORFs found in smaller elements				
Bn119_7	ZnR and GIY-YIG domains	3.6	42	*Phytophthora sojae* 695382398 (3e-05)
Bn161_7	Ankyrin domain protein	93.4	94	*Amoebophilus asiaticus* 501449850 (3e-34)
Bn92_2	Unknown phage protein	0.0	1	*Synechococcus* phage 472343273 (1e-05)
Bn92_3	Ankyrin domain protein	0.1	4	*Pseudogymnoascus pannorum* 682412062 (3e-08)
Bn187_9	Unknown phage protein	20.4	78	*Vibrio* phage 510792797 (5e-06)
Bn84_33	Unknown virophage protein	3.0	39	YSLV5 701905611 (2e-07)

*Percentile rank calculated over all *B. natans* genes.
†Species name and GenBank identifier (BLAST *E* value).

they originate from a common virophage ancestor. *B. natans* genes encoding rve family integrase (Fig. S2*E*), an unknown protein family represented by Bn2_27 (Fig. S2*F*), DNA polymerase (Fig. S2*G*), and GIY-YIG nuclease domains (Fig. S2*H*) exhibit preferential phylogenetic affinities, albeit with moderate bootstrap support with eukaryotic homologs encoded by polinton-related elements. These mixed clades are nested within larger clades containing virophages and environmental sequences, suggesting that gene acquisitions or replacements most likely occurred in the polinton-like elements. In contrast, the phylogenetic trees of the unknown protein family represented by Bn2_18 (Fig. S2*I*), the ribonucleotide reductase small subunit (Fig. S2*J*), and the Kelch protein family (i.e., PK and Kelch domains have distinct origins; Fig. S2 *A* and *B*) support scenarios of gene acquisition from different sources (bacteria, eukaryotes, or dsDNA viruses). Thus, the VLE genes reveal a mosaic origin that is typical of bona fide virophages (1, 7, 8). Altogether, the structure, gene content, and phylogenetic affinities of VLEs provide substantial evidence that they represent remains of integrated virophage genomes.

VLE Genes Are Transcribed. To investigate the transcriptional activity of the *B. natans* virophage-like elements, we analyzed a previously published RNA-seq dataset generated from cultivated *B. natans* cells (20). A total of 45.3 million Illumina paired-end reads were aligned onto the *B. natans* genome assembly (Dataset S2), of which 116,671 mapped within one of the virophage-like regions (Dataset S1). Two hundred seventy-eight out of the 302 predicted virophage-like ORFs had at least one read mapped to

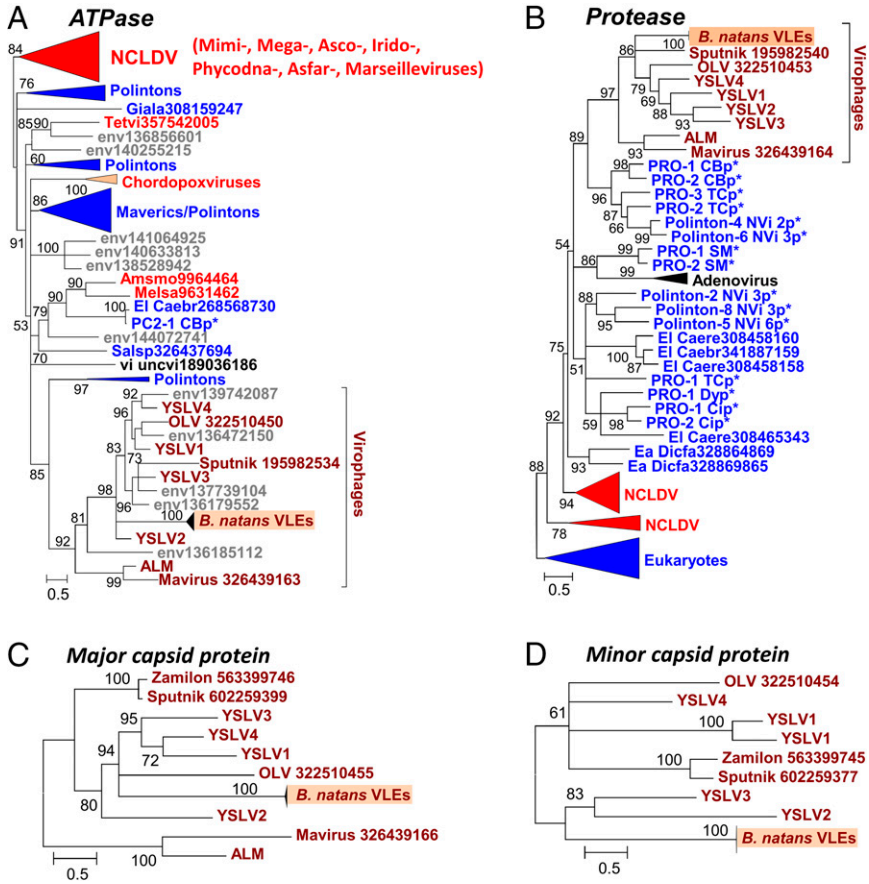


Fig. 3. Maximum-likelihood phylogenetic trees of conserved virophage proteins. (*A*) Packaging ATPase. (*B*) Maturation protease. (*C*) Major capsid protein. (*D*) Minor capsid protein. Branches with bootstrap support (expected-likelihood weights) less than 50% were collapsed. Sequences marked with an asterisk were taken from Repbase (38). For other sequences, the species name abbreviation and GenBank accession number are indicated; env, marine metagenome. Species: Amsmo, *Amsacta moorei* entomopoxvirus "L"; Caebr, *Caenorhabditis brenneri*; Caere, *Caenorhabditis remanei*; Dicfa, *Dictyostelium fasciculatum*; Giala, *Giardia lamblia*; Melsa, *Melanoplus sanguinipes* entomopoxvirus; Salsp, *Salpingoeca rosetta*; Tetvi, *Tetraselmis viridis* virus; uncvi, uncultured virus. Taxa: Ea, Amoebozoa; El, Opisthokonta; u2, Entomopoxvirinae. Dark red, virophages; blue, (predicted) polintons and related elements; light red, NCLDV; gray, unassigned environmental sequences. The numbers of validated amino acid positions in cleaned alignments are 210 (*A*), 166 (*B*), 548 (*C*), and 408 (*D*).

it. The levels of transcription between *B. natans* genes were compared by the mean of the RPKM metric (reads per kilobase per million mapped reads). Ninety-three virophage-like ORFs had RPKM values >5, which ranks them in the top half of the most-transcribed genes in *B. natans*, including 10 genes that figure in the top 10% (i.e., RPKM >50). Interestingly, these highly transcribed virophage-like ORFs encode one major and one minor capsid protein, one DNA-packaging ATPase, one rve integrase, two Kelch domain proteins, and three families of orphan proteins. Other virophage core genes are generally transcribed at low to moderate levels (i.e., the majority of them have RPKM values ranking between the 20th and 36th percentiles), yet 6 of the 10 ATPase gene copies have substantial transcription levels, as indicated by RPKM values ranking between the 50th and 90th percentiles. Interestingly, we identified transcript sequences closely related to the VLEs in assembled RNA-seq datasets generated from various *B. natans* isolates that were sequenced as part of the Marine Microbial Eukaryote Transcriptome Sequencing Project (MMETSP) (21) (Dataset S3). This suggests that the same virophage elements are present and transcribed in other *B. natans* strains, possibly because they were inherited vertically from a common ancestor.

This observation questions the biological significance of the expression of virophage-like genes. One possibility is that the integrations of virophage genomes were accidental events, and that the residual transcriptional activity reflects the fortuitous recognition of regulatory signals of the virophage genes by the cellular host transcription complex. According to this scenario, the expression of the virophage genes has no expected biological effect (neutral) and would disappear as the virophage sequences decay by accumulating random mutations. Alternatively, the observed transcriptional activity might reflect an adaptive strategy that benefits the cellular host population, the virophage, or both. Experimental evidence indicates that virophages have a positive effect on the host-cell population. Mavirus interferes with *Cafeteria roenbergensis* virus propagation and increases the survival of the host-cell population (8). Sputnik causes a 70% decrease in infective Mamavirus particles and a threefold decrease in amoeba cell lysis (1); it also delays or abolishes replication of Marseillevirus (22). A model of population dynamics suggests that the presence of virophages reduces overall mortality of the host algal cell after a bloom (7). Hence, eukaryotes that are susceptible to infection by giant viruses will gain a selective advantage if they can stably associate themselves with virophages (8). An analogous hypothesis was briefly exposed by Katzourakis and Aswad (23) to explain the possible emergence of Maverick/polinton elements from hypothetical integrated virophage genomes in eukaryotes. Under this scenario, the hijacked virophage genes evolve under negative selection in the new eukaryotic genome environment.

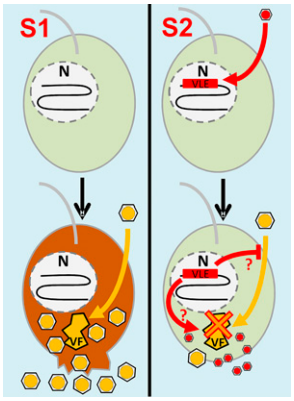


Fig. 4. Scenarios of infection of *B. natans* cells by NCLDVs and virophages. Scenario (S)1 starts with *B. natans* cells devoid of VLE insertions. An NCLDV particle (yellow hexagon) or its DNA enters the cell and establishes viral factories (VFs) that produce new NCLDV particles. The infection causes cell lysis and death accompanied by release of NCLDV particles. Scenario 2 begins with *B. natans* cells carrying VLEs in the form of functional provirophages integrated in the nuclear genome (N). VLEs may have been produced by independent entry of a virophage followed by active DNA integration in the host nuclear genome (delayed-entry mode). Upon NCLDV infection, expressed VLE proteins may inhibit virus penetration or trigger reactivation and excision of the provirophage, which in turn inhibits NCLDV replication and takes advantage of the viral factories for its own replication. As a result, a limited number of NCLDV particles are created compared with S1, leading to increased rates of cell survival. Potentially, new virophages and a limited number of NCLDV particles are released in the environment through exocytosis or another unknown mechanism that does not kill the host cell.

Another possible explanation, which is not mutually exclusive with the former, relates to the adaptive strategy of virophages to increase the frequency of coinfection with a host virus. Different modes of virophage entry into the cell have been evaluated on theoretical grounds (18). However, these scenarios exclusively consider the simultaneous entry of the virophage and the host virus, either independently or in a paired mode [i.e., the virophage adheres to the virus before infection; see Taylor et al. (18)]. Indirect evidence supports the paired-entry mode for Sputnik (1, 22), whereas an independent-entry mode has been observed for Mavirus (8). The transcribed *B. natans* virophage elements bring out a third hypothetical mode that we have dubbed the delayed-entry mode (Fig. 4). According to this scenario, the virophage is expected to enter the cell first by an unknown mechanism; its DNA reaches the cell nucleus, integrates into the cellular genome, and remains latent until superinfection by a virus reactivates and rescues the virophage, allowing its replication in the virus factory. This hypothetical scenario gets support from the observation that the *B. natans* virophage elements encode integrases, which suggests that genome integration is an active process rather than an incidental event. To go further into the delayed-entry scenario, it is possible that the transcription of virophage genes leads to the production of sentinel proteins that are able to detect infection of the host cell by another virus and transduce a signal triggering virophage reactivation. There is an obvious advantage of the delayed-entry mode over the independent-entry mode when the simultaneous independent entry of both the virus and virophage is a rare event due to, for example, high dilution of the virus particles in the environment. Furthermore, the integrated virophage can be passed on to the next generations of host cells, contributing to its spread and multiplication. In the form of an integrated provirophage, the virophage can potentially wait for virus superinfection during long periods of time, whose length depends on the rate at which random mutations inactivate the endogenous virophage element.

Putative NCLDV Insertions. All characterized virophages so far have been shown to infect members of the *Mimiviridae* (NCLDVs). Thus, although no large DNA virus of *B. natans* has been identified so far, the discovery of virophage sequences in *B. natans* suggests that this alga is the prey of giant viruses that are themselves infected by virophages. DNA fragments that are relics of integrated NCLDV genomes have been discovered in various eukaryotic genomes (24–26), revealing footprints of past interaction between viruses and hosts (19). We therefore searched the *B. natans* genome for sequences related to NCLDVs using the BLAST score plot approach. As shown in Fig. 1, a total of 36 *B. natans* predicted proteins have better alignment scores with homologs in NCLDVs than in cellular organisms (Dataset S4). The closer relationship of these proteins with giant virus homologs was confirmed by phylogenetic reconstruction (for examples, see Fig. S3). This phylogenetic affinity may reflect horizontal gene transfer between viruses and eukaryotic hosts, the polarity of which (virus to host or host to virus) cannot always be determined with certainty. However, some proteins are homologous to typical NCLDV core genes, including major capsid protein and packaging ATPase (distantly related to the virophage ones), which have most likely been acquired by the *B. natans* host (Fig. S3 *A* and *B*). Interestingly, some NCLDV-like genes are physically clustered in discrete regions of the *B. natans* genome assembly (Fig. S4). For example, we identified a 165-kb region of scaffold 10 (position ~340–505 kb) that is transcriptionally silent according to our analysis of the RNA-seq dataset (Dataset S4). Such a large untranscribed region is unique in the *B. natans* genome. This region contains 83 predicted genes, of which 9 have obvious NCLDV affinities, including genes for DNA-packaging ATPase, exonuclease, major capsid protein, and transcription factor. Of the remaining genes, only 7 have best matches in eukaryotes (albeit with low similarity) and 4 have best matches with homologs in bacteria and phages. The large majority of predicted genes (63/83 = 76%) are orphans, a characteristic shared with giant virus genomes. Thus, it is likely that this large DNA stretch is the remnant of an integrated NCLDV genome similar to those previously observed in various eukaryotic genomes (19, 24–27). Inserts of likely NCLDV origin identified in the moss genome were also reported as transcriptionally silent (19). These hypothesized viral inserts probably behave like neutrally evolving nonfunctional DNA. The G+C content of the NCLDV-like region (45.5%) is similar to the background G+C content of the *B. natans* genome (44.8%), which prevents precisely identifying the insert boundaries.

Outside of the large NCLDV-like insert, some *B. natans* genes with viral phylogenetic affinity were found to be transcribed (Dataset S4). Six genes have transcription levels among the top 50% of *B. natans* genes. These genes may have been inherited from the *B. natans* ancestor and captured by large DNA viruses from eukaryotic hosts that are closely related to *B. natans*. This scenario can explain the preferential affinity between *B. natans* proteins and NCLDV homologs, whereas the corresponding genes actually have a eukaryotic origin. Alternatively, the corresponding genes may have been recruited in the metabolism of a Bigelowiella ancestor subsequent to their acquisition from viral donors, perhaps now fulfilling new cellular functions. For example, *B. natans* has one copy for each of the four regular B family eukaryotic DNA polymerase catalytic subunits alpha, delta, epsilon, and zeta. However, the alga possesses two extra delta DNA polymerases that group in phylogenetic positions compatible with distinct viral origins (Fig. S3*C*). The two viral-like DNA polymerase catalytic subunits are transcribed (i.e., 43th and 53th percentiles) at higher levels than those of the four regular eukaryotic isoforms (25th–40th percentiles, respectively). Furthermore, they have a large number of introns (i.e., 19 and 30 introns, respectively), whereas viral genes are generally devoid of introns or, in rare cases, only contain a small number of them.

The two extra DNA polymerases have no identifiable orthologs in other sequenced eukaryotes, except in the transcriptomes of other *Bigelowiella* isolates (Dataset S4). We can therefore not exclude a scenario in which the corresponding genes have been captured from viral donors and were progressively "eukaryogenized" through the accumulation of introns.

Putative Transpoviron Insertions. Four of the six NCLDV-like DNA-packaging ATPase genes are carried by four closely related repeated genetic elements (18.8–21.8 kb in length) flanked by TIRs 300–900 bp in length (Fig. 5). These repeated elements contain between 17 and 22 ORFs (Dataset S5), but some of the original genes seem to have accumulated internal stop codons and frameshifts, resulting in truncated translation. Very few proteins encoded by the repeated elements exhibit detectable similarity in public databases. They include a homolog of the *Aureococcus anophagefferens* virus protein AaV202, a homolog of the *P. globosa* virus virophage protein PgvV_00016, a homolog of the Mavirus virophage protein MV06, and a homolog of a hypothetical protein of the fungus *Batrachochytrium dendrobatidis*. Interestingly, two additional proteins are homologous to core transpoviron proteins, including the C-terminal superfamily I helicase domain protein and C2H2 Zn-finger protein. Transpovirons form a distinct class of mobile genetic elements (6.5–7.5 kb) associated with *Mimiviridae* (4). The genes for helicase and C2H2 protein are adjacent in the *B. natans* repeated elements and Mimivirus-associated transpovirons. Furthermore, Mimivirus-associated transpovirons are flanked by TIRs (~530 bp). Thus, the *B. natans* repeated elements and transpovirons share structural and genetic similarities, suggesting that the *B. natans* elements might belong to the transpoviron family. However, the putative integrated *B. natans* transpovirons are substantially bigger in size relative to their Mimivirus counterparts, possibly due to the incorporation of foreign genes of diverse origins, including NCLDVs and virophages. Some of the integrated transpoviron genes show evidence of transcription, including six genes that have transcription levels in the *B. natans* top 50%. In addition, homologous transcripts were identified in the transcriptome data of the other *Bigelowiella* isolates (Dataset S6).

Conclusion

The first discovered virophages have the form of small icosahedral virion particles and have been shown to infect giant viruses. However, integrated virophages have been found more recently in a Mimivirus genome (4). Here we show for the first time, to our knowledge, that virophage genomes can also integrate in a eukaryotic genome. This finding led us to predict that *B. natans* might be the prey for NCLDVs. Additional integrated DNA fragments that most probably originate from NCLDV genomes provide data showing that *B. natans* or its recent ancestor had physical contacts with NCLDV members. Furthermore, we also identified repeated genetic elements that resemble transpovirons associated with *Mimivirus*. Thus, the *B. natans* genome appears to have recorded genetic footprints of molecular "battles" between virophages, transpovirons, and giant viruses. *B. natans* belongs to the Chlorarachniophytes, a group of unicellular marine algae that acquired a plastid by secondary endosymbiosis involving engulfment of a green alga by a eukaryotic heterotroph host (28). They are typically mixotrophic, ingesting bacteria and smaller protists as well as conducting photosynthesis. *B. natans* is the only species of the supergroup Rhizaria for which a complete genome sequence is available (29). During endosymbiosis, hundreds of genes of green origin have been transferred toward the host genome in a process called endosymbiotic gene transfer (29). The acquisition of DNA from giant viruses and transpovirons as well as from virophages through horizontal gene transfer represents an additional component in the melting pot of genes composing the *B. natans* nuclear genome.

One of the most intriguing findings of our analysis is that integrated virophage genes are highly transcribed, suggesting that they are biologically functional. We speculated on three potential adaptive scenarios to explain this observation. Certainly the most interesting of them is the possibility that both the cellular host and virophage take advantage of the integration strategy, by providing the cell with a defense mechanism against giant viruses and providing the virophage with a mechanism to increase the rate of coinfection with a viral prey (i.e., delayed-entry mode S2, schematized in Fig. 4). From the perspective of the cellular host, it is tempting to speculate that integrated virophages can act as

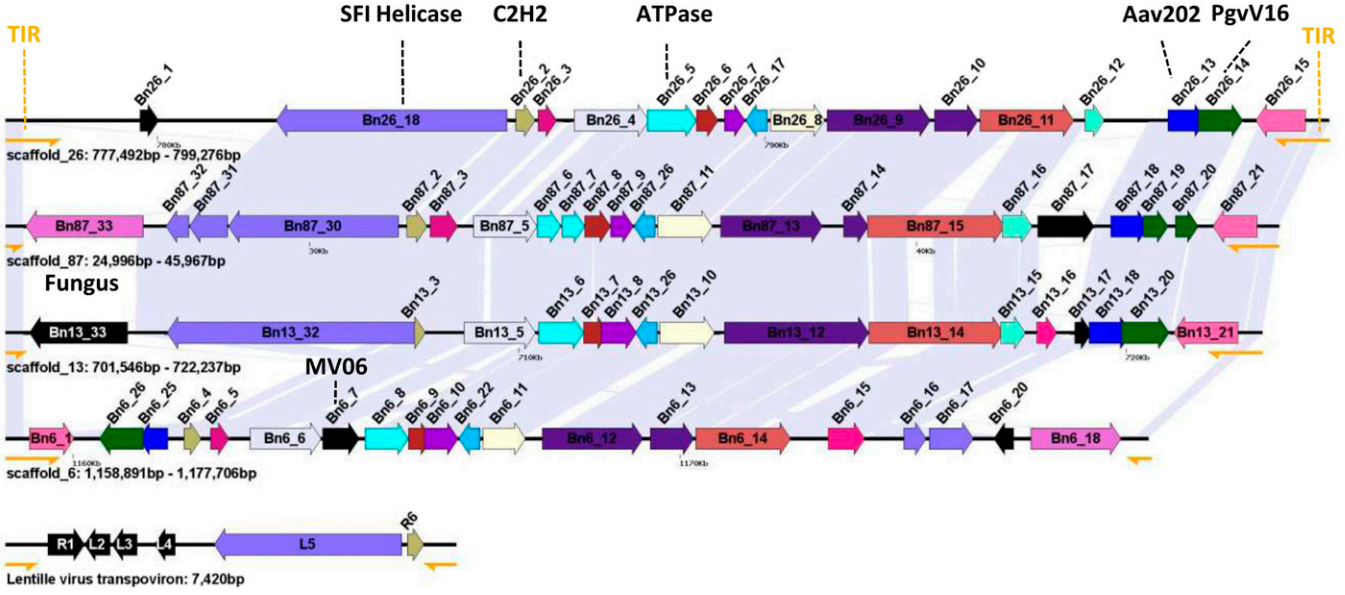


Fig. 5. Organization of putative inserted transpoviron genomes. The position and strand orientation of each ORF (>90 codons) are indicated by an arrow. Homologous ORFs across transpoviron-like elements are indicated with the same color. ORFs with no homologs in the other transpovirons are shown in black. TIRs are represented by yellow semiarrows. Shaded areas between elements indicate regions of high nucleotide similarity.

molecular weapons against viral pathogens, conferring a sort of immunity transmissible from generation to generation. Sputnik and Zamilon strains have been shown to have a broad host spectrum among the *Mimiviridae* (5, 16), and Sputnik even affects the replication of Marseillevirus, which is distantly related to the *Mimiviridae* (22). By extrapolation, we can further hypothesize that the putative defense function triggered by integrated virophages can be efficient over a diversity of viral pathogens.

It is currently difficult to estimate the prevalence and biological significance of virophage insertion in eukaryotes. At first glance, the fact that only *B. natans* showed virophage genes out of 1,153 screened eukaryotic genomes might suggest that genomic integration of virophages is highly unusual. However, there is a historical bias among sequenced eukaryotes, which include a majority of model organisms, crop plants, fungi, animals, and pathogens (30), all of which are apparently not infected by giant viruses and, hence, virophages. As a result, the organisms tested here included relatively few potential hosts of giant viruses and virophages [i.e., known hosts are amoebas and microalgae (1, 5, 8); Dataset S6]. This bias could explain the apparent low prevalence of virophage insertions detected in our study. In contrast, giant viruses and virophages are readily identified in environmental metagenomes (7, 10, 11, 31), suggesting that these viral entities are common in natural ecosystems. Thus, the question of the prevalence and biological significance of virophage insertions can only be reasonably addressed when more genomes of putative hosts are sequenced.

Methods

Sequence Analysis. Annotations and sequences of 1,153 eukaryotic genome assemblies representing various protists, fungi, and basal metazoans were downloaded from GenBank. Dataset S6 lists the investigated species together with their GenBank accession numbers, including members of the Alveolate (103), Amoebozoa (35), Apusozoa (1), Chlorophyta (12), Choanoflagellida (2), Cryptophyta (1), Euglenozoa (49), Fonticula (1), Fornicata (6), Fungi (858), Haptophyta (1), Heterolobosea (2), Parabassalia (1), Metazoa (8), Opisthokonta (2), Rhizaria (3), Rhodophyta (4), Stramenopiles (63), and Streptophyta (2). We also downloaded the genome annotations of nine sequenced virophages, including Mavirus (GenBank accession no. GCF_000890715), Sputnik (GenBank accession no. EU606015), Zamilon (GenBank accession no. NC_022990), ALM (GenBank accession no. KC556923), OLV (GenBank accession no. HQ704801), YSLV1 (GenBank accession no. KC556924), YSLV2 (GenBank accession no. KC556925), YSLV3 (GenBank accession no. KC556926), and YSLV4 (GenBank accession no. KC556922).

For each virophage, the major capsid protein, minor capsid protein, DNA-packaging ATPase, and cysteine protease were aligned against the predicted eukaryotic proteomes using BLASTP. Experience showed that computational annotation of eukaryotic genomes can be inefficient in predicting genes of viral origin, because they have become pseudogenes, often resulting in truncation or in-frame stop codons and/or because they can have very distinct GC content relative to the host genome and no introns. Therefore, we also aligned the virophage markers against the translated products of ORFs (>90 codons) lying between predicted genes. Based on prior analysis of the distribution of BLASTP scores against the nonredundant (NR) database, we applied family-specific score thresholds to avoid false detection of remote homologs that are of cellular origin or nonhomologs with similar low-complexity sequences. The score threshold was set as the minimal BLASTP score between any two virophage proteins in the family. Scores obtained between any virophage proteins and cellular homologs were always lower. These score thresholds were 44.7 for proteases, 46.6 for ATPases, 41.2 for mCPs, and 45.1 for MCPs.

The genome assembly of *B. natans* CCMP2755 exhibited homologs for each of the four marker proteins with BLASTP *E* value >1E-5, except proteases, for which we used an *E*-value threshold of 1E-3, because this protein is less conserved among virophages. To identify additional candidate viral-like protein genes in *B. natans*, we performed a BLAST-score plot analysis previously described in Maumus et al. (19). Briefly, the full complement of *B. natans* predicted proteins together with intergenic ORFs was used to probe the National Center for Biotechnology Information database using BLASTP (*E* value < 1E-5). For each protein query, the alignment scores with the best cellular hit and the best virophage or NCLDV hit were recorded. When no cellular hit was recorded whereas a viral hit was obtained, the cellular score was set to zero. BLAST scores were then normalized by the score of the alignment of the query sequence against itself (i.e., self-score), resulting in relative scores expressed in percent of self-score. Nonviral hit scores are plotted against viral hit scores in Fig. 1.

Identification and Delineation of Individual VLEs. The physical location of the virophage-like protein genes identified by BLASTP revealed that they tend to cluster in specific loci in the genome assembly, unveiling large regions of possible virophage origin. Six of these regions were bordered by long inverted repeats on each side, which coincide with sharp changes in GC content (e.g., Fig. 2). We made the assumption that the long inverted repeats mark the beginning and the end of VLEs. We extracted the nucleotide sequences of these putative complete VLEs from the genome assembly and used BLASTN to align the VLEs back to the genome assembly to identify additional truncated VLEs. Adjacent BLASTN matches that had an *E* value <1E-25 and a minimal length of 100 pb were assembled to identify a total of 38 VLEs up to 33.3 kb in length. Every candidate VLE was checked and validated manually.

Phylogenetic Analysis. Construction of adequate homologous protein sets for phylogenetic analysis was performed using the BLAST-EXPLORER website (32). Homologous proteins were aligned using MUSCLE (33), and amino acid positions in multiple alignments containing >30% gaps were removed. We used this criterion for alignment cleaning to keep coherence with the pioneering study of Yutin et al. (12), which produced a comprehensive phylogenetic study of virophage genes. Maximum-likelihood (ML) phylogenetic reconstruction was performed using the PhyML program (34). Before phylogenetic reconstruction, the best-fitting substitution model for each sequence dataset was determined using the ProtTest program (35). Sequences, alignments, ProtTest outputs, and phylogenetic trees are available in Dataset S7.

RNA-Seq Analysis. To analyze the transcriptional activity of the *B. natans* genome, we downloaded an RNA-seq dataset (ID MMETSP0045) from the CAMERA database (36). This dataset contained 61.6 million Illumina paired reads generated from polyadenylated RNA extracted from *B. natans* CCMP2755 cells grown in f/2-Si media for a month under a 12-h:12-h light:dark cycle at room temperature (20). Reads were aligned onto the reference genome sequence using Bowtie 2 (37) with default parameters. Due to the high nucleotide similarity between VLEs, 63% of the reads that mapped onto a VLE also had valid alignments in at least another VLE. For these cases the origin of the read is ambiguous, and we only picked one of the alignments at random for read-count purposes. In addition, we downloaded assembled RNA-seq datasets (contigs) of other *Bigelowiella* isolates publicly available in the CAMERA database: MMETSP1052 (*B. natans* CCMP623), MMETSP1054 (*B. natans* CCMP1259), MMETSP1055 (*B. natans* CCMP1258.1), MMETSP1358 (*B. natans* CCMP1242), and MMETSP1359 (*Bigelowiella longifila* CCMP242). These datasets were generated as part of the Marine Microbial Eukaryote Transcriptome Sequencing Project (21). Homologs of virophage-like encoded proteins were searched in the transcribed sequences using TBLASTN.

ACKNOWLEDGMENTS. We thank Yongjie Wang for providing the YSLV and ALM sequences ahead of publication. We thank Jean Michel Claverie and Deborah Byrne for critical reading of the manuscript. The Information Génomique et Structurale laboratory is partially supported by the CNRS and Aix-Marseille University. We acknowledge the use of the Paca-Bioinfo platform, supported by Infrastructures en Biologie Santé et Agronomie and France-Génomique (ANR-10-INBS-0009).

1. La Scola B, et al. (2008) The virophage as a unique parasite of the giant mimivirus. *Nature* 455(7209):100–104.
2. Colson P, et al. (2013) "Megavirales", a proposed new order for eukaryotic nucleocytoplasmic large DNA viruses. *Arch Virol* 158(12):2517–2521.
3. Krupovic M, Cvirkaite-Krupovic V (2011) Virophages or satellite viruses? *Nat Rev Microbiol* 9(11):762–763.
4. Desnues C, et al. (2012) Provirophages and transpovirons as the diverse mobilome of giant viruses. *Proc Natl Acad Sci USA* 109(44):18078–18083.
5. Gaia M, et al. (2014) Zamilon, a novel virophage with Mimiviridae host specificity. *PLoS One* 9(4):e94923.
6. Campos RK, et al. (2014) Samba virus: A novel mimivirus from a giant rain forest, the Brazilian Amazon. *Virol J* 11:95.
7. Yau S, et al. (2011) Virophage control of antarctic algal host–virus dynamics. *Proc Natl Acad Sci USA* 108(15):6163–6168.
8. Fischer MG, Suttle CA (2011) A virophage at the origin of large DNA transposons. *Science* 332(6026):231–234.

9. Santini S, et al. (2013) Genome of *Phaeocystis globosa* virus PgV-16T highlights the common ancestry of the largest known DNA viruses infecting eukaryotes. *Proc Natl Acad Sci USA* 110(26):10800–10805.
10. Zhou J, et al. (2013) Diversity of virophages in metagenomic data sets. *J Virol* 87(8): 4225–4236.
11. Zhou J, et al. (2015) Three novel virophage genomes discovered from Yellowstone Lake metagenomes. *J Virol* 89(2):1278–1285.
12. Yutin N, Raoult D, Koonin EV (2013) Virophages, polintons, and transpovirons: A complex evolutionary network of diverse selfish genetic elements with different reproduction strategies. *Virol J* 10:158.
13. Kapitonov VV, Jurka J (2006) Self-synthesizing DNA transposons in eukaryotes. *Proc Natl Acad Sci USA* 103(12):4540–4545.
14. Pritham EJ (2006) Genes in conflict: The biology of selfish genetic elements. *Am J Hum Biol* 18(5):727–728.
15. Krupovic M, Bamford DH, Koonin EV (2014) Conservation of major and minor jelly-roll capsid proteins in Polinton (Maverick) transposons suggests that they are bona fide viruses. *Biol Direct* 9:6.
16. Gaia M, et al. (2013) Broad spectrum of mimiviridae virophage allows its isolation using a mimivirus reporter. *PLoS One* 8(4):e61912.
17. Wodarz D (2013) Evolutionary dynamics of giant viruses and their virophages. *Ecol Evol* 3(7):2103–2115.
18. Taylor BP, Cortez MH, Weitz JS (2014) The virus of my virus is my friend: Ecological effects of virophage with alternative modes of coinfection. *J Theor Biol* 354:124–136.
19. Maumus F, Epert A, Nogué F, Blanc G (2014) Plant genomes enclose footprints of past infections by giant virus relatives. *Nat Commun* 5:4268.
20. Tanifuji G, Onodera NT, Moore CE, Archibald JM (2014) Reduced nuclear genomes maintain high gene transcription levels. *Mol Biol Evol* 31(3):625–635.
21. Keeling PJ, et al. (2014) The Marine Microbial Eukaryote Transcriptome Sequencing Project (MMETSP): Illuminating the functional diversity of eukaryotic life in the oceans through transcriptome sequencing. *PLoS Biol* 12(6):e1001889.
22. Desnues C, Raoult D (2010) Inside the lifestyle of the virophage. *Intervirology* 53(5):293–303.
23. Katzourakis A, Aswad A (2014) The origins of giant viruses, virophages and their relatives in host genomes. *BMC Biol* 12:51.
24. Wang L, et al. (2014) Endogenous viral elements in algal genomes. *Acta Oceanol Sin* 33(2):102–107.
25. Filée J (2014) Multiple occurrences of giant virus core genes acquired by eukaryotic genomes: The visible part of the iceberg? *Virology* 466-467:53–59.
26. Sharma V, Colson P, Giorgi R, Pontarotti P, Raoult D (2014) DNA-dependent RNA polymerase detects hidden giant viruses in published databanks. *Genome Biol Evol* 6(7):1603–1610.
27. Delaroque N, Boland W (2008) The genome of the brown alga *Ectocarpus siliculosus* contains a series of viral DNA pieces, suggesting an ancient association with large dsDNA viruses. *BMC Evol Biol* 8:110.
28. Keeling PJ (2013) The number, speed, and impact of plastid endosymbioses in eukaryotic evolution. *Annu Rev Plant Biol* 64:583–607.
29. Curtis BA, et al. (2012) Algal genomes reveal evolutionary mosaicism and the fate of nucleomorphs. *Nature* 492(7427):59–65.
30. del Campo J, et al. (2014) The others: Our biased perspective of eukaryotic genomes. *Trends Ecol Evol* 29(5):252–259.
31. Yutin N, Kapitonov VV, Koonin EV (2015) A new family of hybrid virophages from an animal gut metagenome. *Biol Direct* 10:19.
32. Dereeper A, Audic S, Claverie J-M, Blanc G (2010) BLAST-EXPLORER helps you building datasets for phylogenetic analysis. *BMC Evol Biol* 10:8.
33. Edgar RC (2004) MUSCLE: Multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res* 32(5):1792–1797.
34. Guindon S, et al. (2010) New algorithms and methods to estimate maximum-likelihood phylogenies: Assessing the performance of PhyML 3.0. *Syst Biol* 59(3): 307–321.
35. Abascal F, Zardoya R, Posada D (2005) ProtTest: Selection of best-fit models of protein evolution. *Bioinformatics* 21(9):2104–2105.
36. Sun S, et al. (2011) Community cyberinfrastructure for Advanced Microbial Ecology Research and Analysis: The CAMERA resource. *Nucleic Acids Res* 39(Database issue): D546–D551.
37. Langmead B, Salzberg SL (2012) Fast gapped-read alignment with Bowtie 2. *Nat Methods* 9(4):357–359.
38. Jurka J, et al. (2005) Repbase Update, a database of eukaryotic repetitive elements. *Cytogenet Genome Res* 110(1–4):462–467.

Supporting Information

Blanc et al. 10.1073/pnas.1506469112

Fig. S1. Schematic representation of the identified virophage-like elements. Homologous ORFs (>90 codons) are colored with the same color across virophage-like elements. TIRs are represented by yellow semiarrows. Shaded areas between virophage-like elements indicate regions of high nucleotide similarity.

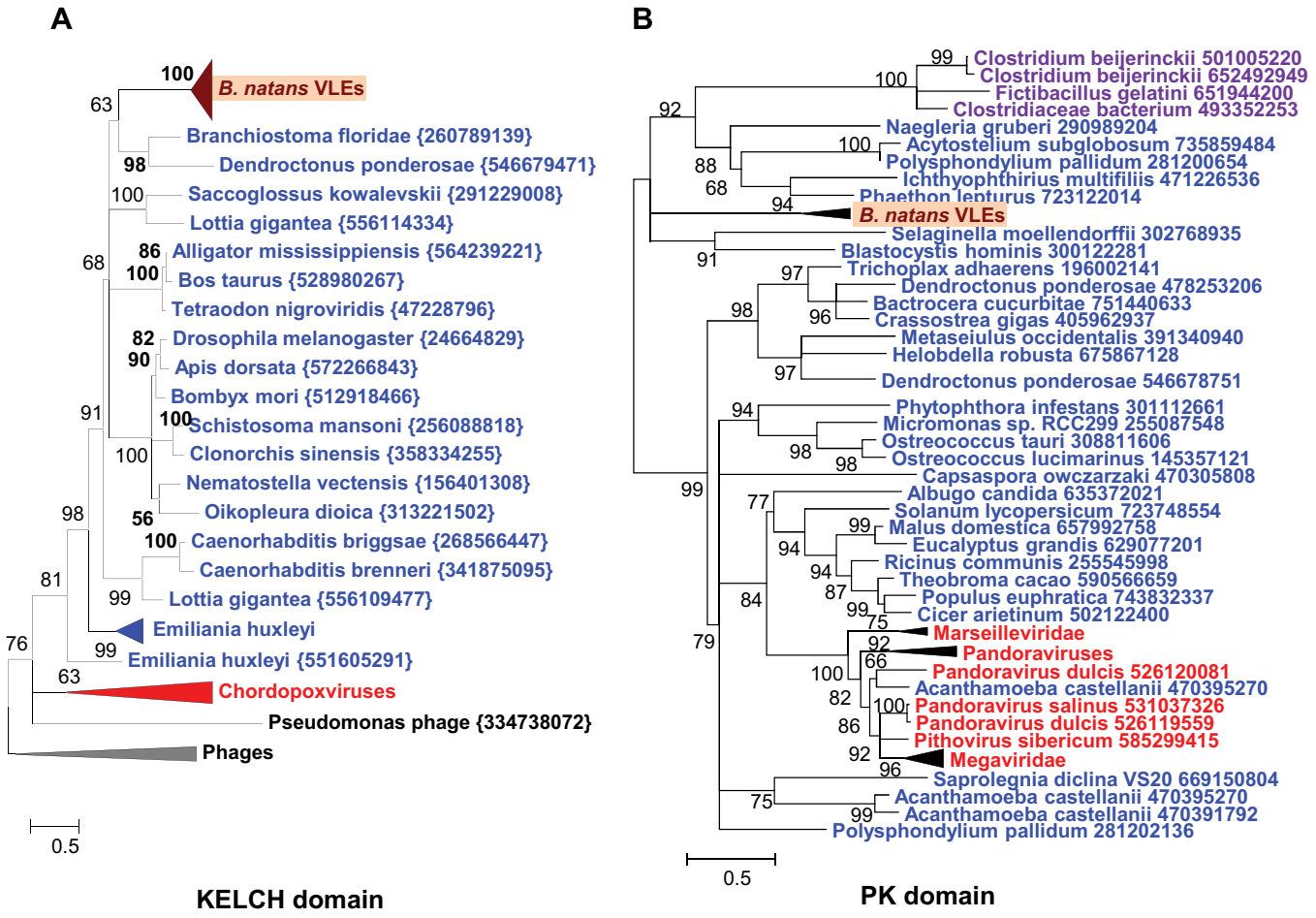


Fig. S2. (Continued)

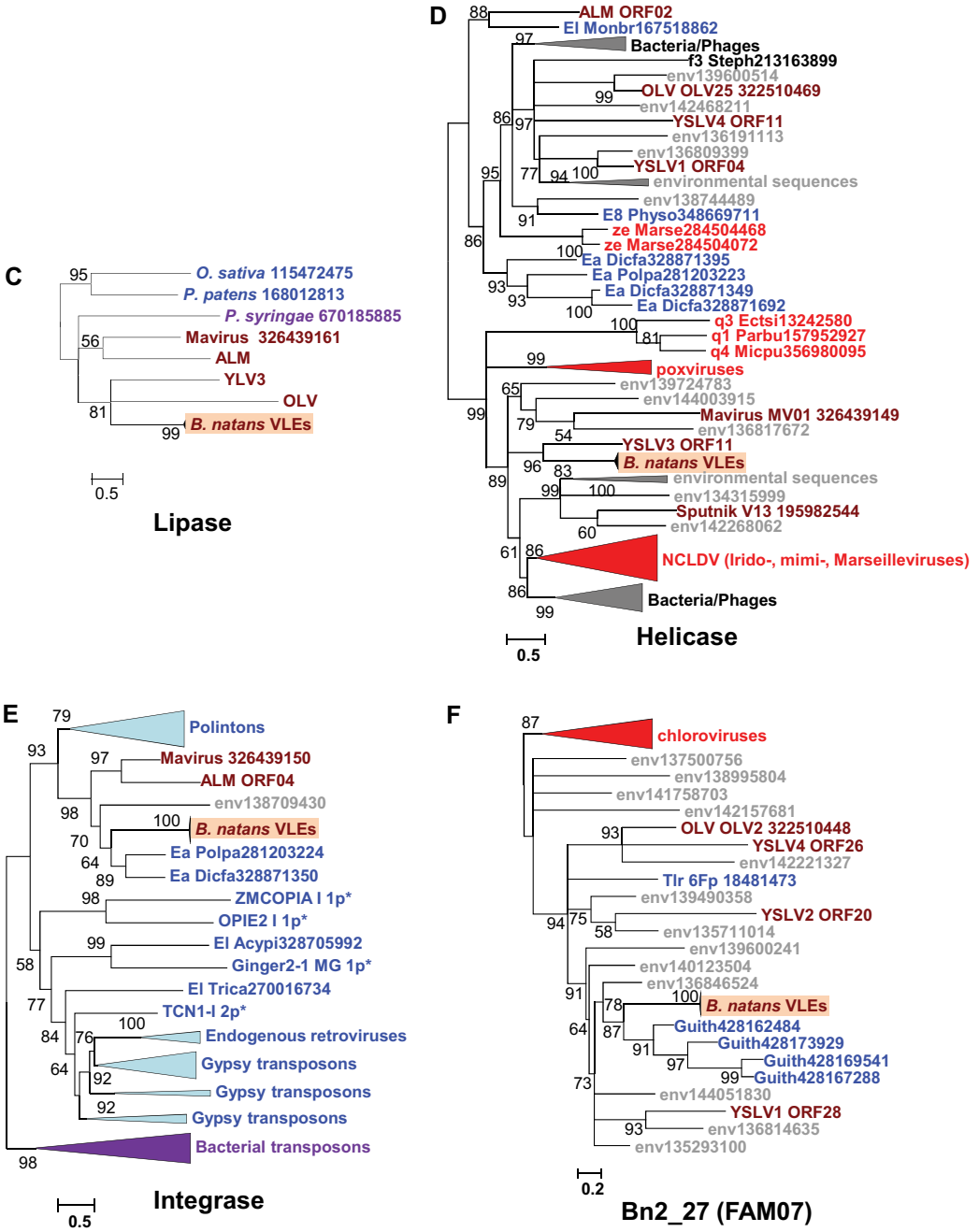


Fig. S2. (Continued)

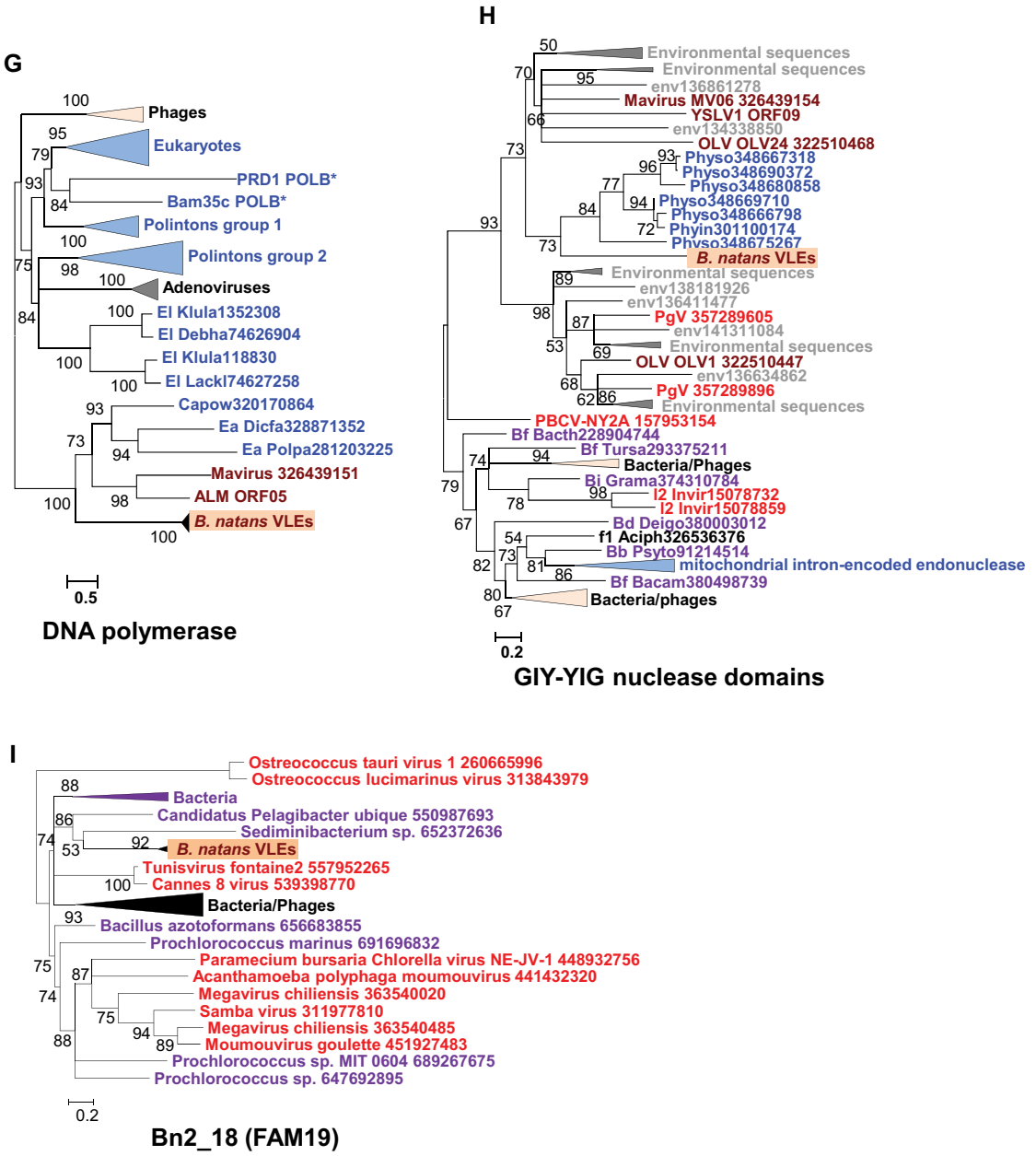
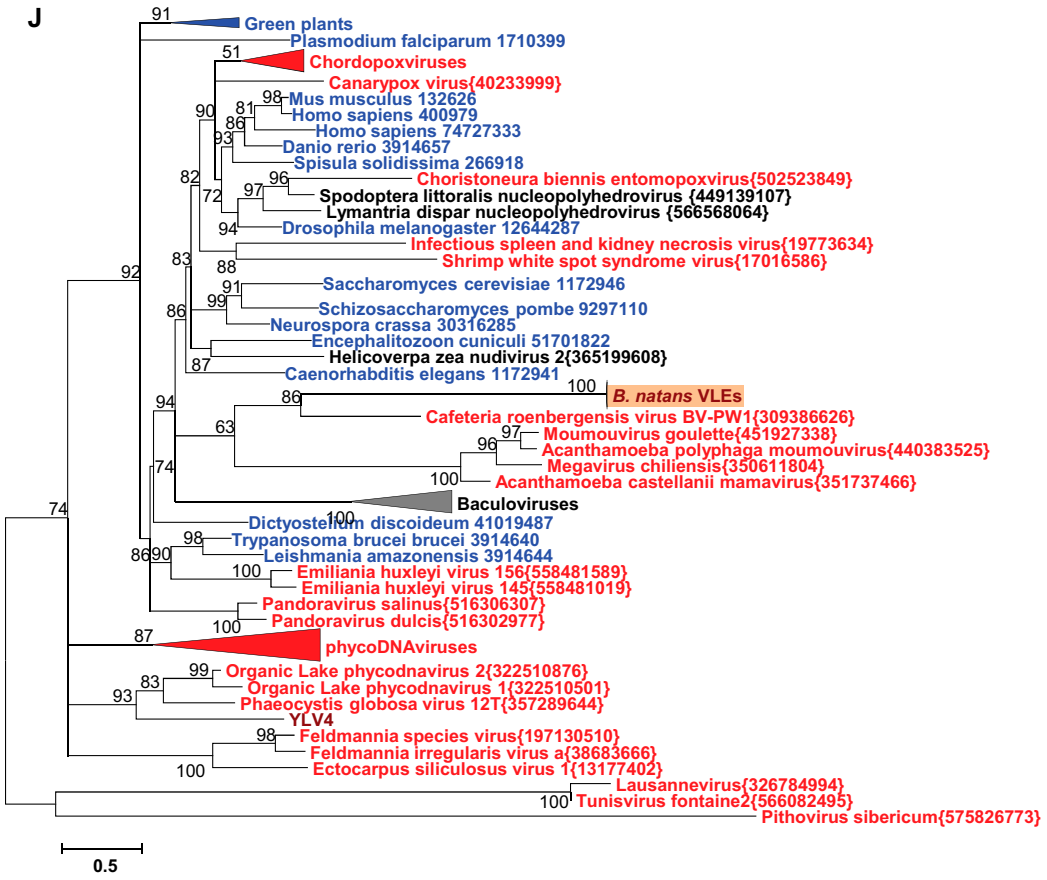


Fig. S2. (Continued)



ribonucleotide reductase small subunit

Fig. S2. Phylogenetic reconstructions of virophage-like proteins. (*A*) ML tree of KELCH domains; (*B*) ML tree of protein kinase (PK) domains contained in KELCH proteins; (*C*) ML tree of lipases; (*D*) ML tree of S3 helicase domains; (*E*) ML tree of rve integrases; (*F*) ML tree of Bn2_27 homologs; (*G*) ML tree of DNA polymerase type B family; (*H*) ML tree of GIY-YIG nuclease domains; (*I*) ML tree of Bn2_18 homologs; and (*J*) ML tree of ribonucleotide reductase small subunits. Virophage, NCLDV, prokaryotic, and eukaryotic sequence names are shown in brown, red, mauve, and blue, respectively. Sequences marked with an asterisk were taken from Repbase (38). For other sequences, species name abbreviation and GenBank accession number are indicated. For clarity, subtrees containing sequences originating from the same taxonomic clade were condensed (colored triangles). Statistical supports for branches in percent (approximate likelihood-ratio test) are shown beside nodes (only branch supports >50%). Branches with statistical support <50% were collapsed. The scale bars represent the number of substitutions per amino acid site. env, marine metagenome. Species: Aciph, *Acinetobacter* phage; Acypi, *Acyrthosiphon pisum*; Bacam, *Bacillus amyloliquefaciens*; Bacth, *Bacillus thuringiensis*; Capow, *Capsaspora owczarzaki*; Debha, *Debaryomyces hansenii*; Deigo, *Deinococcus gobiensis*; Dicfa, *Dictyostelium fasciculatum*; Ectsi, *Ectocarpus siliculosus* virus 1; Grama, *Granulicella mallensis*; Guith, *Guillardia theta*; Klula, *Kluyveromyces lactis*; Lackl, *Lachancea kluyveri*; Marse, Marseillevirus; Micpu, *Micromonas pusilla* virus PL1; Monbr, *Monosiga brevicollis* MX1; Parbu, *Paramecium bursaria* Chlorella virus NY2A; Physo, *Phytophthora sojae*; Popla, *Polysphondylium pallidum* PN500; Psyto, *Psychroflexus torquis*; Steph, *Stenotrophomonas* phage S1; Tlr, *Tetrahymena thermophile*; Trica, *Tribolium castaneum*; Tursa, *Turicibacter sanguinis*. The numbers of validated amino acid positions in cleaned alignments are 340 (*A*), 237 (*B*), 233 (*C*), 333 (*D*), 304 (*E*), 93 (*F*), 472 (*G*), 92 (*H*), 158 (*I*), and 994 (*J*).

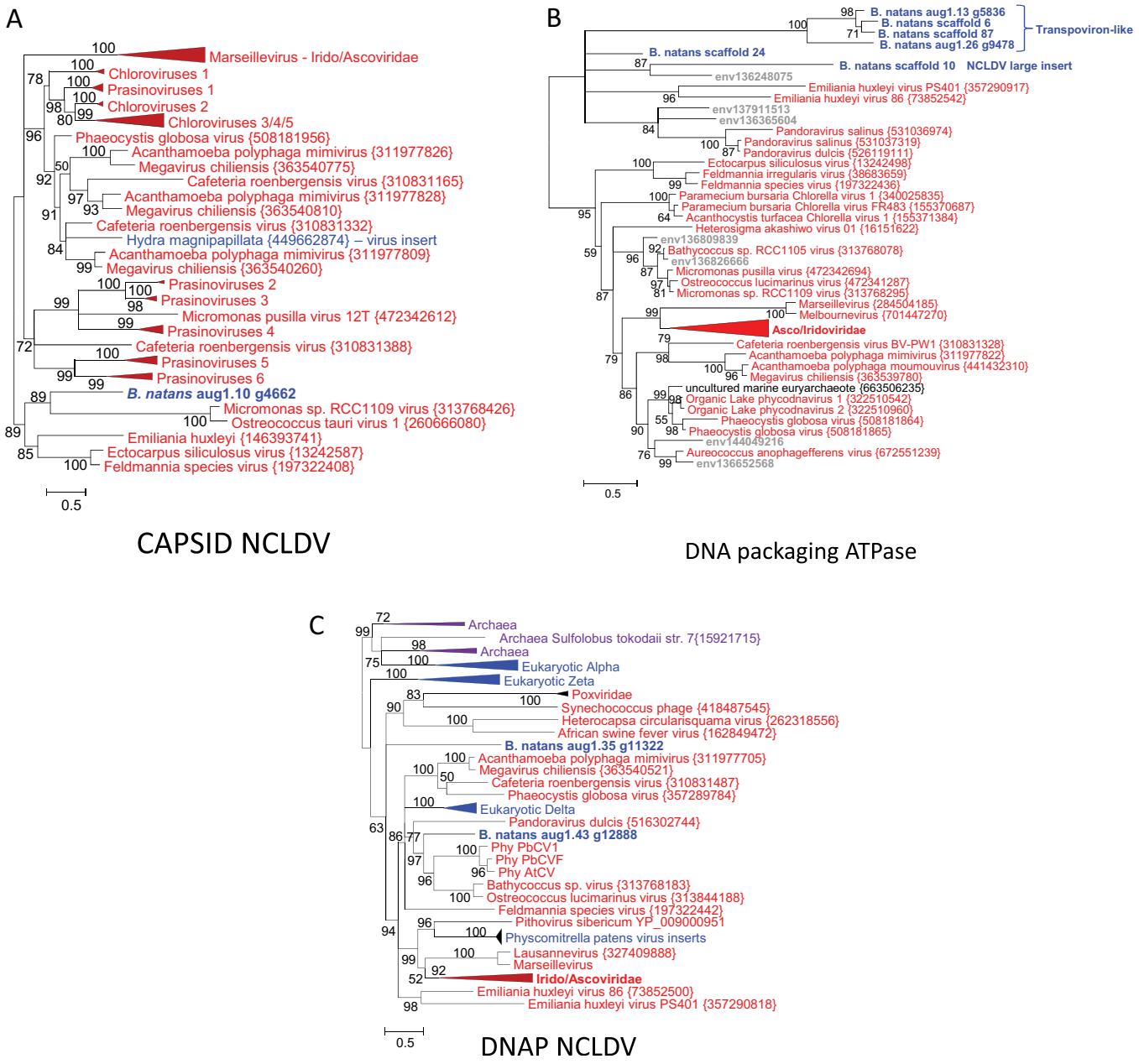


Fig. S3. Phylogenetic reconstructions of NCLDV-like proteins. (*A*) ML tree of capsid proteins; (*B*) ML tree of DNA-packaging ATPases; and (*C*) ML tree of DNA polymerases (DNAPs). Same legend as Fig. S2. The numbers of validated amino acid positions in cleaned alignments are 370 (*A*), 229 (*B*), and 755 (*C*).

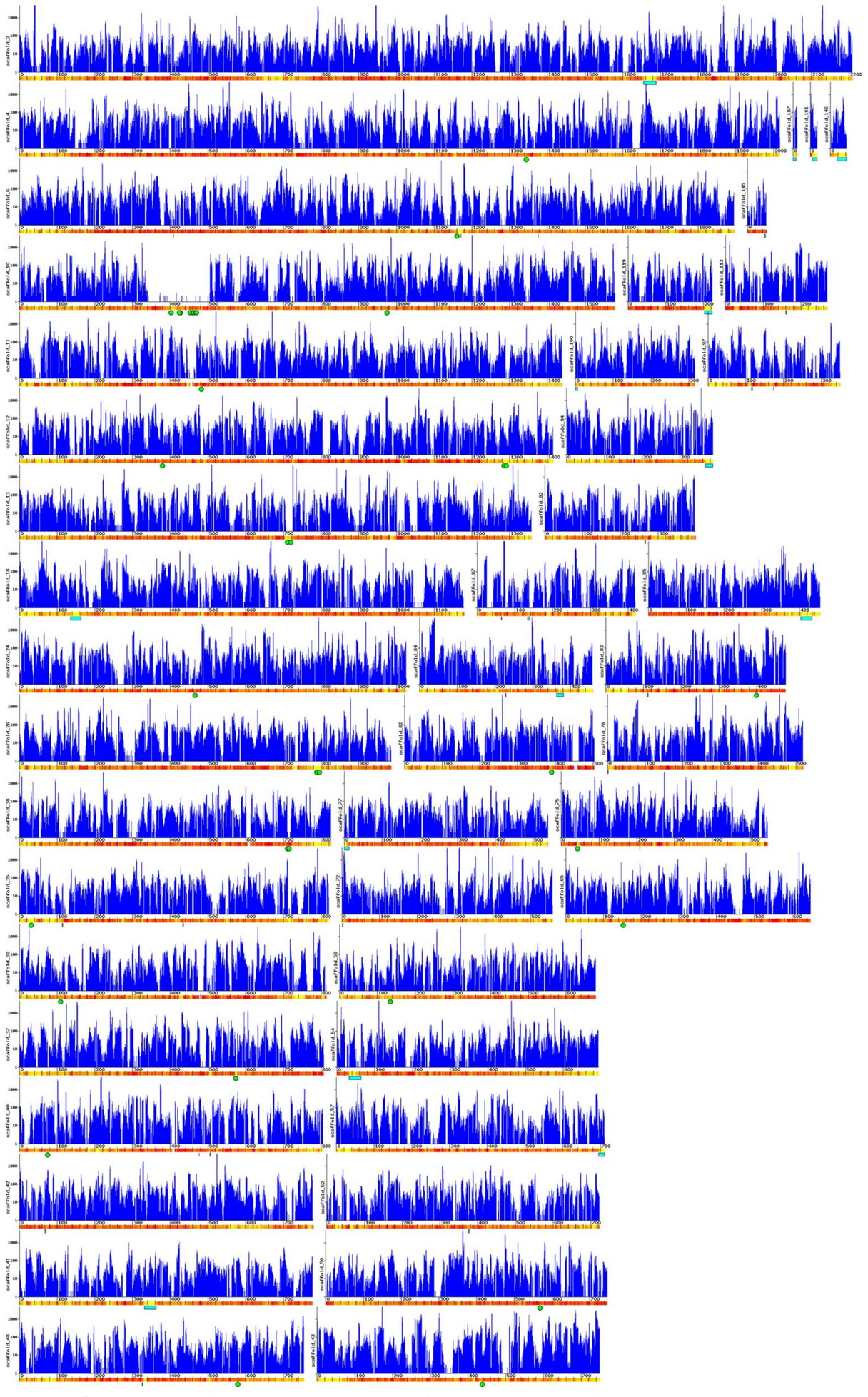


Fig. S4. *B. natans* scaffolds containing virophage- or NCLDV-like inserts. The blue graphs represent the number of transcriptome reads mapped onto contigs. GC content variation in 5-kb sliding windows along the contig sequence is depicted by a yellow-to-red color scale, with yellow indicating low GC content (minimum 30%) and red indicating high GC content (maximum 60%). The distance ruler is expressed in kilobases. Cyan rectangles represent virophage inserts; green circles represent NCLDV-like genes.

Other Supporting Information Files

Dataset S1 (XLSX)
Dataset S2 (XLSX)
Dataset S3 (XLSX)
Dataset S4 (XLSX)
Dataset S5 (XLSX)
Dataset S6 (XLSX)
Dataset S7 (GZ)

c - Discussion

c.1 - Séquences de virophages

Les virophages se répliquent dans les usines à virions formées par un *Mimiviridae*, dans le cytoplasme d'une cellule hôte. Pour mener à bien leur réplication, les virophages doivent donc infecter (i) une cellule susceptible (ii) qui est par ailleurs déjà infectée par un *Mimiviridae* permissif (i.e., un *Mimiviridae* qui a la capacité de permettre l'expression des gènes du virophage). Bien que nous ayons une connaissance limitée de la diversité des virophages (pour une revue, voir Villain et al 2016 [1]), avec seulement 5 représentants isolés, les stratégies employées pour la co-infection sont diverses. Ainsi, le virophage Sputnik adhère aux fibrilles présentes à la surface de la capside de Mimivirus et nous présumons qu'il entre ainsi dans la cellule hôte en même temps que Mimivirus quand ce dernier est phagocyté par l'amibe (185, 243). De son côté, le virophage Mavirus semblerait pouvoir entrer indépendamment de son virus hôte CroV dans la cellule hôte (161). Une autre stratégie pour le virophage pourrait être d'intégrer le génome de son hôte virus, et ceci pourrait bien être utilisé par Sputnik dont l'intégration dans le génome de Mimivirus a été décrite (166). Dans notre article, nous décrivons la découverte de génomes apparemment complets de virophages dans celui de l'algue *B.natans*. Les gènes du virophage sont de plus exprimés. Ceci nous a mené à proposer qu'il s'agisse là d'une 4ème stratégie : l'intégration du génome du virophage dans celui de l'hôte cellulaire, précédant l'infection de la cellule par un virus géant permissif. La présence de séquences d'origine NCLDVs dans le génome de la même algue suggère qu'en effet, il est/ou a été une proie pour ces virus. Nous avons vu que la réplication du virophage a un effet délétère sur la réplication du virus géant. Des modèles de dynamique de populations hôtes/virus/virophages (160) prédisent que cette co-réplication aurait par conséquent un effet protecteur sur les cellules lors d'infection de NCLDVs. Ceci nous a conduits à proposer que le provirophage servirait d'arme à *B.natans* pour lutter contre la lyse par les virus géants. Une relation mutualiste se serait développée entre la cellule qui permet au virophage de se propager dans la population des algues au fur et à mesure des divisions cellulaires et le virophage qui en échange offre sa protection à la cellule lorsque ces infections se produisent. Voici donc un exemple additionnel (voir introduction) où l'immunité d'un organisme cellulaire est le fruit d'une co-évolution avec les virus.

Fin 2016 paraissait dans la revue Nature une « lettre » de Mathias Fisher et Thomas Hackl de l'institut Max Planck en Allemagne qui valide expérimentalement notre hypothèse de mutualisme entre virus et cellule hôte, et en précise les modalités (244). C'est dans ce laboratoire qu'avait successivement été isolé le virus CroV à partir du stramenopile *Cafetaria roenbergensis,* puis le virophage Mavirus se co-répliquant dans *C.roenbergensis* avec CroV. Dans cette nouvelle étude, ils montrent que (1) Mavirus s'intègre dans le génome de l'hôte cellulaire (2) le provirophage n'est pas exprimé constitutivement, mais spécifiquement lors d'une sur-infection par CroV (3) la lyse des cellules par CroV libère des particules de Mavirus

[1] J'ai eu l'opportunité de contribuer à la production de cette revue pour laquelle j'ai participé à la réalisation de la partie relative aux virophages

infectieuses (4) aux rounds suivants d'infection, Mavirus interfère avec la réplication de CroV ce qui augmente la survie de l'hôte eucaryote *C.roenbergensis* (Figure 11).

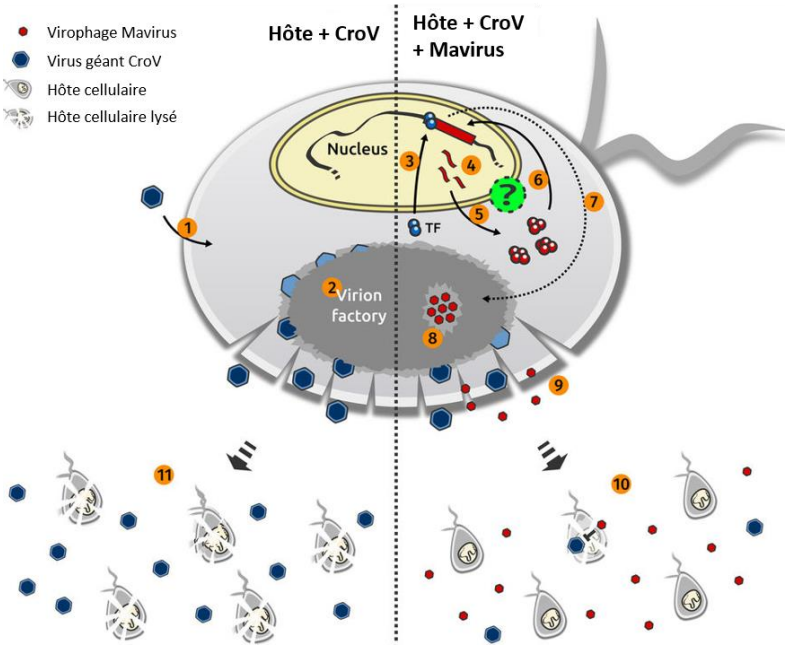


Figure 11 adaptée de la figure *Hypothesis for CroV-induced reactivation of endogenous mavirus* (244). (1) entrée de CroV ; (2) formation de l'usine à virion ; (3, 4) activation de l'expression des gènes du virophage par le facteur de transcription TF codé par CroV ; (5) export et traduction des transcripts de Mavirus, suivit (6) du retour au noyau de certaines d'entre elles pour exciser ou répliquer le génome du provirophage ; (7) le génome de Mavirus est exporté dans l'usine à virion de CroV, où (8) le génome est répliqué et les particules du virophage assemblés ; (9) la lyse de la cellule libère des particules de CroV et de Mavirus ; (10) le virophage réactivé inhibe la réplication de CroV dans les cellules coinfectées, ce qui mène à une augmentation de la survie de la population de l'hôte cellulaire. (11) Lorsque l'hôte cellulaire n'a pas de provirophage, CroV continue d'infecter la population d'hôte.

C'est donc bien une coopération, une relation mutualiste entre le virophage et la cellule. D'autre part, cette relation induit chez la cellule infectée une réaction altruiste. Effectivement, ce n'est pas l'organisme qui a produit les particules qui en bénéficie, mais les autres membres de l'espèce. Il s'agit d'un altruisme mesuré, comparativement au sacrifice de certaines bactéries qui, infectées, se « suicident » avant que le bactériophage ne puisse se répliquer et se propager dans la population (245, 246). Effectivement, par contraste, *C.roenbergensis* n'induit pas lui-même sa mort. Ce comportement pourrait avoir été sélectionné dans le cadre de la sélection de parentèle (kin selection). Effectivement, celle-ci a été invoquée pour expliquer l'existence des comportements altruistes en général (247), et particulier celui précédemment mentionné de la réponse altruiste des bactéries lors d'attaques de phages (248). Ainsi, des caractères apparus à cause d'interactions avec des virus résultent d'une sélection à un autre niveau que celui de l'individu.

c.2 - Séquences de NCLDVs

Au contraire des inserts d'origine virophagiques, les inserts NCLDVs détectés apparaissent pour la majorité silencieux dans les transcriptomes analysés. Cependant, il y a des exceptions. Par exemple, *B.natans* possède 2 unités catalytiques d'ADN polymérase de type Beta (DNApolB) d'origine NCLDVs, en plus de celles typiquement eucaryotes. Deux points sont intéressants concernant ces DNApolB d'origine virales : (i) elles sont significativement exprimées (ii) elles possèdent des introns, une caractéristique des gènes eucaryotes dont sont généralement dépourvus les gènes viraux. Conjointement, ces deux faits sont suggestifs d'un scenario par lequel les gènes auraient été capturés chez les virus puis « domestiqués » par la cellule à travers l'accumulation d'introns.

Une étude réalisée par un laboratoire japonais (249) et publiée au cours de l'année suivant la parution de notre papier suggère que ces gènes sont maintenant dédiés au métabolisme de l'algue. Effectivement, l'équipe japonaise a montré que les DNApolB d'origine NCLDV chez *B. natans* sont adressées au nucléomorphe et pourraient donc être impliquées dans la réplication de son ADN. Les nucléomorphes sont des noyaux vestigiaux trouvés chez certaines algues ayant effectué une endosymbiose secondaire. Il s'agit du noyau de l'algue primitive qui a été endogénisé par le second eucaryote.

3 - Un crible global des eucaryotes

a - Introduction

La découverte des séquences de NCLDVs dans plusieurs génomes eucaryotes nous a fait suspecter que ce phénomène n'est pas anecdotique. C'est ce qui nous a conduits à passer au crible les banques de données de séquences publiques pour mesurer l'étendue du phénomène à l'échelle de tous les eucaryotes séquencés. Pour ceci (Figure 12), nous avons utilisé 5 gènes universellement présents chez les NCLDVs, que nous avons alignés contre les banques de données en utilisant le programme de détection d'homologue BLASTP. 66 génomes ou transcriptomes d'eucaryotes contiennent au moins l'un de ces cinq gènes. J'ai réalisé l'analyse détaillée de l'environnement génomique de ces gènes *cores* viraux, ainsi que l'analyse de l'expression des gènes lorsque les transcriptomes étaient disponibles.

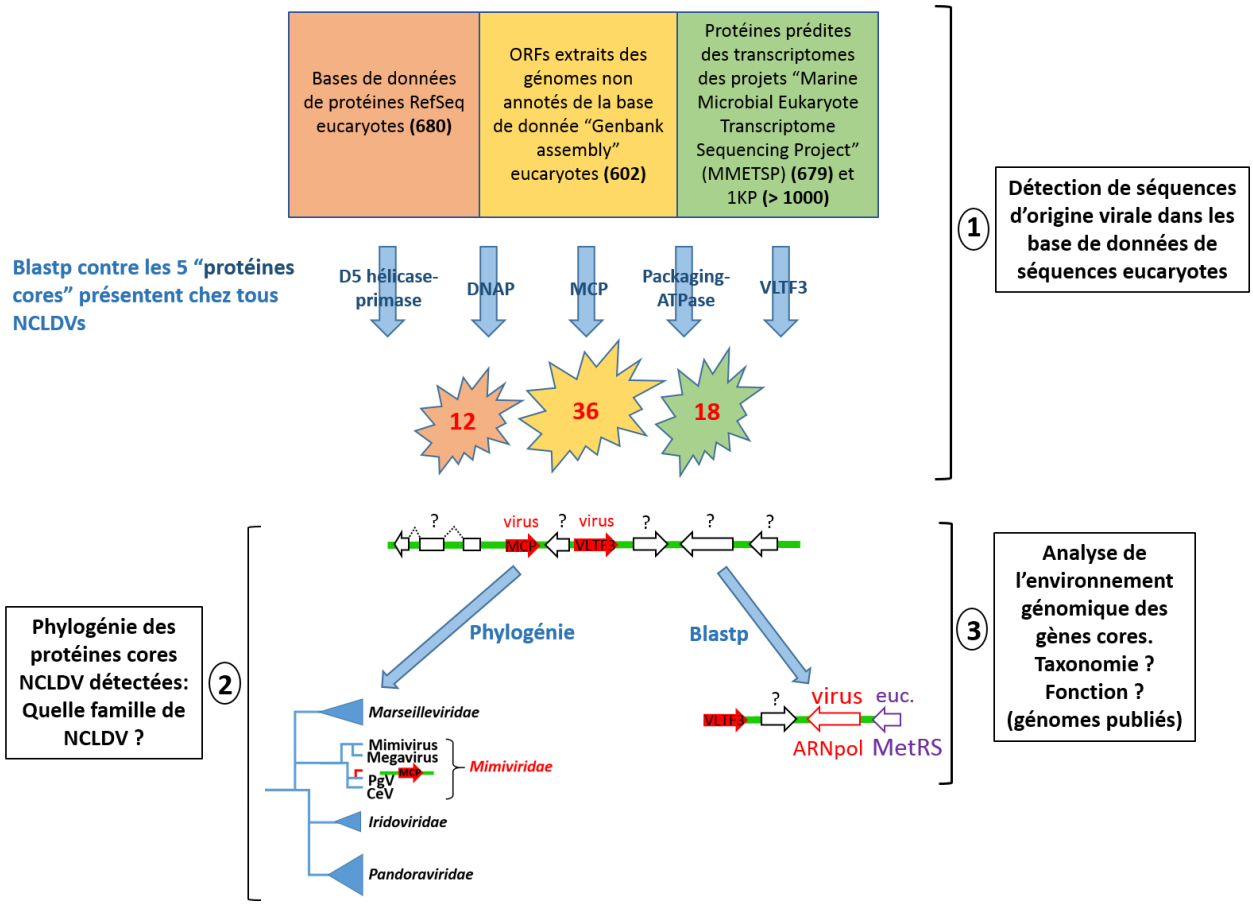


Figure 12: Procédure pour la détection et l'analyse de séquences virales dans les bases de données de séquences eucaryotes. Pour chaque base de données criblée, le nombre de génomes desquels proviennent les protéines analysées est indiqué en gras entre parenthèse. DNAP= ADN polymérase ; MCP= Majeure protéine de capside ; VLTF3= Facteur de transcription très tardif 3; euc.= eucaryote.

b - Article 3

A Glimpse of Nucleo-Cytoplasmic Large DNA Virus Biodiversity through the Eukaryotic Genomics Window. Gallot-Lavallée, L., and Blanc, G. (2017).

viruses

MDPI

Article

A Glimpse of Nucleo-Cytoplasmic Large DNA Virus Biodiversity through the Eukaryotic Genomics Window

Lucie Gallot-Lavallée [1] and Guillaume Blanc [1,2,*]

[1] Structural and Genomic Information Laboratory (IGS), Aix-Marseille Université, CNRS UMR7256 (IMM FR3479), 13288 Marseille cedex 09, France; lucie.gallot-lavallee@igs.cnrs-mrs.fr

[2] Mediterranean Institute of Oceanography (MIO), Aix Marseille Université, Université de Toulon, CNRS/INSU, IRD, UM 110, 13288 Marseille cedex 09, France

* Correspondence: guillaume.blanc@igs.cnrs-mrs.fr

Academic Editor: Bernard La Scola

Received: 30 November 2016; Accepted: 13 January 2017; Published: 20 January 2017

Abstract: The nucleocytoplasmic large DNA viruses (NCLDV) are a group of extremely complex double-stranded DNA viruses, which are major parasites of a variety of eukaryotes. Recent studies showed that certain eukaryotes contain fragments of NCLDV DNA integrated in their genome, when surprisingly many of these organisms were not previously shown to be infected by NCLDVs. We performed an update survey of NCLDV genes hidden in eukaryotic sequences to measure the incidence of this phenomenon in common public sequence databases. A total of 66 eukaryotic genomic or transcriptomic datasets—many of which are from algae and aquatic protists—contained at least one of the five most consistently conserved NCLDV core genes. Phylogenetic study of the eukaryotic NCLDV-like sequences identified putative new members of already recognized viral families, as well as members of as yet unknown viral clades. Genomic evidence suggested that most of these sequences resulted from viral DNA integrations rather than contaminating viruses. Furthermore, the nature of the inserted viral genes helped predicting original functional capacities of the donor viruses. These insights confirm that genomic insertions of NCLDV DNA are common in eukaryotes and can be exploited to delineate the contours of NCLDV biodiversity.

Keywords: nucleo-cytoplasmic large DNA virus; lateral gene transfer; virus insertion

1. Introduction

Viruses have long been viewed only under the angle of human, animal, and plant diseases, which considerably restrained our vision of the viral world and its role in global ecology. In this age of virus discovery, we are beginning to appreciate the enormous diversity of viruses, far beyond what we originally thought. Nucleo-cytoplasmic large DNA viruses (NCLDVs) [1,2] form a monophyletic clade of eukaryotic viruses with a large double-stranded DNA (dsDNA) genome ranging from 100 kbp in the smallest iridoviruses up to 2.50 Mbp in the gigantic pandoraviruses [3]. Their hosts show a remarkably wide taxonomic spectrum from microscopic unicellular eukaryotes to larger animals, including humans [2]. The biodiversity of NCLDVs is thought to be immense however we still do not know how many major clades do exist [4]. Seven taxonomic families have been defined so far including *Ascoviridae*, *Asfarviridae*, *Iridoviridae*, *Marseilleviridae*, *Mimiviridae*, *Phycodnaviridae*, and *Poxviridae*, but new viral isolates, such as pandoravirus, pithovirus, and mollivirus, are likely to become founding members of new families. Historically, isolation of large DNA viruses infecting eukaryotic algae or protists has proceeded by co-culturing a host together with a virus sampled from the environment. In this experimental approach, a eukaryotic host is chosen a priori for its capacity of being infected by

a virus and adapted to lab culture prior to virus isolation. Recently, the metagenomic approach has accelerated the rate at which new viruses are brought to light [5,6]. However, this approach suffers from two main shortcomings: first, viral sequences assembled from metagenomic data are generally short, encompassing often only a few genes at best. Second, the hosts of the identified viruses remain unknown. Yet, host information is an absolutely essential component in the study of viruses, since viral replication is dependent on host organisms [7]. Thus, drawing the contours of virus/host diversities calls for a development of new approaches that can circumvent limitations of the co-culturing and metagenomics methods.

Recent studies have identified NCLDV-related sequences in genomic and transcriptomic datasets generated from eukaryotic organisms [8–13]. Some of these viral sequences were shown to originate from virus genome fragments integrated into the nuclear genome of their presumed eukaryotic hosts, including protists [10,13], land plant [9], and algae [8,11,12,14]. These fragments encompass up to several hundreds of kbp and can contain hundreds of viral genes, including common NCLDV phylogenetic markers. It is currently unclear how and by which mechanisms these viral DNAs became integrated into eukaryotic genomes. DNA integration may result from an active process (i.e., as a result of a virus-encoded integrase activity) or from an accidental incorporation of viral DNA freely floating inside the cell (i.e., as a result of an aborted infection). Phaeoviruses are the only members of NCLDVs to show evidence of a lysogenic cycle. Presumably, they integrate into the genome of their host by means of an integrase encoded by the virus [14,15]. In addition to reports of NCLDV DNA inserts in eukaryotic hosts, NCLDV-like sequences were also found in some algal transcriptomes [11]. These transcripts may originate from viral genes integrated in the host genome or from infected host cells present in the culture from which RNA were extracted. Altogether, these studies suggested that viral DNA insertion in the host genome is a common feature of NCLDVs. However, the frequency at which this phenomenon occurs across eukaryotic lineages, and the short- and long-term evolutions of inserted viral sequences are still poorly understood. Whether they have a potential role in defense mechanisms against infecting viruses based on sequence recognition and/or RNA silencing is also an open question.

Interesting information has come out from the discovery of viral inserts: many of the organisms in which NCLDV sequences were identified were not previously known to be infected by NCLDVs. Moreover, phylogenetic markers harbored by viral genomic inserts or transcripts suggested that certain virus donors were distantly related to known NCLDVs [8,9,16]. Thus, assuming that the NCLDV sequences identified in eukaryotic datasets result from infecting viruses or lateral gene transfers, these sequences may be used as a tool to better describe the realm of NCLDVs. Importantly, identification of NCLDV genes may allow predicting novel virus/host associations and shedding new light on the biodiversity of NCLDVs. With the vertiginous throughput and dropping cost of DNA sequencing, new eukaryotic genomes are nowadays sequenced at an unprecedented pace. Thus, since the pioneering studies performed over the last couple of years [8–13,16], many new eukaryotic genomes have been released in public databases. This prompted us to perform an update survey of NCLDV genes hidden in eukaryotic sequences to measure the incidence of this phenomenon in common public databases. Here, we show that sequences generated from 66 eukaryotes contained NCLDV core genes, most of which have never been reported so far. Phylogenetic reconstruction showed that many of these sequences originated from members of existing NCLDV families, but also possibly from as yet unknown NCLDV clades, thus extending the range of the NCLDV biodiversity.

2. Materials and Methods

Sequences from the five NCLDV core proteins were retrieved from the NCLDV clusters of orthologous gene (NCVOG) database [17,18] and aligned against protein databases using BLASTP (E-value $< 1 \times 10^{-5}$). BLAST searches against RefSeq and 1KP databases were performed on the dedicated website at NCBI and [19]. Unannotated genome assemblies were downloaded from the NCBI Assembly database. We only downloaded the eukaryotic fraction of the assembly database to

the exclusion of very large genomes (i.e., >1 Gbp) to limit computational time; however, annotated proteins of a majority of very large genomes were already available for search in the RefSeq database. Open reading frames >100 codons were extracted from the genome assemblies prior to BLAST searches. Predicted proteins from the Marine Microbial Eukaryote Transcriptome Sequencing Project (MMETSP) transcriptomes were downloaded from the iMicrobe server [20]. The documentation on the experimental conditions used during transcriptome acquisitions was obtained at the following internet adresses: [21] (MMETSP) and [22] (1KP).

Significant similarity and phylogenetic intertwining exist between packaging ATPase of NCLDVs and polintons, a family of large self-synthesizing transposons encoding up to 10 open reading frames [23]. To sort packaging ATPases between NCLDVs and polintons, a phylogenetic tree was constructed with all identified ATPases, NCLDV ATPases and polinton ATPases (reference polinton sequences were retrieved from the relevant supplementary data file of the Yutin et al. paper [23]). Identified ATPases that grouped with reference polinton homologs were removed from further study.

General phylogenetic analyses were performed as follows: additional homologous sequences were first searched in the RefSeq database using the BLAST EXPLORER tool [24]. Multiple-sequence alignment of homologous proteins was then performed using the MAFFT program [25]. We removed alignment positions containing >90% gaps before maximum likelihood phylogenetic reconstruction, which was performed using the FastTree program [26] with the LG + Gamma model of amino acid substitution. Statistical support for branches was estimated with the SH-like local support method. Sequences, alignments and phylogenetic trees are available in Dataset S1. The lengths of multiple-alignments used for phylogenetic reconstruction were 411, 1,198, 2,117, 692, and 565 amino acid positions (including position containing <90% gaps) for the ATPase, D5 primase-helicase, B-family DNA polymerase (DNAP), major capsid protein (MCP), and Very Late Transcription Factor 3 (VLTF3), respectively.

3. Results and Discussion

3.1. NCLDV Protein Markers in Eukaryotes

Although they typically encode hundreds of proteins, NCLDVs were reported to only share five universally-conserved core genes, including genes for MCP, D5 primase-helicase, DNAP, A32-like packaging ATPase, andVLTF3 [17]. These core viral proteins were used as query in BLAST searches against four eukaryotic sequence databases. The databases queried in this study included Genbank Refseq, which contained all annotated proteins from 680 sequenced eukaryotic species. In addition, we screened the Genbank Assembly database which contained 602 raw eukaryotic genome sequences that were not annotated and, therefore, not referenced in RefSeq. Open reading frames >100 codons were extracted from the non-annotated genome assemblies prior to their mining by BLASTP. Altogether, the RefSeq and Assembly databases comprised 1282 fully sequenced eukaryotic genomes. Because preliminary analysis revealed that NCLDV insertions were most frequent in aquatic unicellular eukaryotes, we also downloaded transcriptomic data from sequencing initiatives specifically targeting these organisms. The MMETSP database contained 679 assembled transcriptomes from 413 distinct marine unicellular eukaryotes, including some of the more abundant and ecologically significant species in the oceans such as diatoms [27]. The "1000 plants" (1KP) initiative database contained transcriptomic data from over 1000 plant species, including 214 unicellular eukaryotic algae from the Archaeplastida and Chromista groups [28].

Protein homologous to the five NCLDV core proteins were identified in 48 eukaryotic genomic sequencing projects, including 12 annotated genomes from RefSeq and 36 non-annotated genomes (Table 1). Nine of these genome assemblies contained the five core genes, while 14 genome assemblies contained only one of them. Some of the viral sequences arose from larger viral inserts that have already been described in the genomes of *Ectocarpus siliculosus* [12], *Bigelowiella natans* [8], *Physcomitrella patens* [9], *Acanthamoeba* spp. [13,16], *Hydra vulgaris* [10,13] and *Phytophthora parasitica* [10]. Most of

the genomic datasets associated with NCLDV core protein homologs correspond to organisms living in soil or aquatic environments. The working genome dataset was highly dominated by Metazoa (n = 495), Fungi (n = 392) and land plants (n = 110), collectively representing 80% of the analyzed genomes. However, only 15% (7/48) of the eukaryotes positive for NCLDV proteins belonged to one of these groups, including three metazoans (*Daphnia pulex*, *H. vulgaris*, *Echinacea pallida*), three fungi (*Gonapodya prolifera*, *Rhizophagus irregularis*, *Allomyces macrogynus*), and a land plant (moss *P. patens*). Thus, a majority of eukaryotes associated with NCLDV proteins have a unicellular or simple multicellular structure and are members of less studied clades. The most impacted eukaryotic groups in terms of frequency are (i) brown algae (*Phaeophyceae*) for which all three genomes contained NCLDV homologs, Amoebozoa (11 out of 32 genomes = 34%), green algae (9/28 = 32%; i.e., *Chlorophyta* + *Streptophyta*), and *Oomycetes* (10/40 = 25%). As a matter of fact, small eukaryotes living either constantly (aquatic) or transiently (soil or swimming gametes (i.e., moss *P. patens*)) in waters appear to more frequently have integrated NCLDV sequences in their genome. This host bias may be a consequence of the relative large size of NCLDV particles, which makes their propagation more difficult out of fluidic environments. Also, the large virion size may limit propagation in complex multicellular organisms that have thick cell walls (i.e., terrestrial plants), giving them less chance to access the germ-line cells where lateral gene transfers must occur to be transmitted to the next host generation.

In addition, 18 transcriptomes generated from eukaryotic microalgae or aquatic protistans encode homologs to at least one of the NCLDV core proteins. In contrast to genomic datasets, none of the transcriptomes encode the five core proteins; 14 contained only one core protein sequence. DNAP was the most frequently identified NCLDV core gene among transcriptomes (10 species). This observation is consistent with a previous study reporting that a supernumerary DNAP subunit of possible NCLDV origin was transcribed in the rhizarian alga *B. natans*, whereas most of the other inserted NCLDV-like genes were transcriptionally silent [8]. The NCLDV-like DNAP of *B. natans* has been shown to be targeted to the nucleomorph where it might be involved in the nucleomorph genome replication [29]. Thus, some of the NCLDV-like DNAPs identified in these transcriptomes might also originate from lateral gene transfer from viruses and have acquired a functional role in their respective eukaryotes. Another possibility is that these transcripts were produced by viruses replicating in infected cells in the cultures used for sequencing. This is most likely the case for the *Emiliania huxleyii* viral transcripts because the corresponding transcriptomes have been reportedly acquired during a viral infection experiment (see information on the experimental conditions in Materials and Methods). In addition, the *Pleurochrysis carterae* strain sequenced in the MMETSP was suspected to contain a persistent virus, and our analysis gives credit to this hypothesis.

All in all, our study reveals many more potential NCLDV hosts than previously thought. Out of the 66 sequence datasets positive for NCLDV core genes, only four were generated from species already known for being infected by NCLDVs (i.e., *Acanthamoeba castellanii*, *Acanthamoeba polyphaga*, *E. siliculosus*, *E. huxleyii*). Two other species are closely related to organisms hosting NCLDVs. This is the case for the marine flagellate *Halocafeteria seosinensis* that is closely related to *Cafeteria roenbergensis* [30], a host for giant viruses and virophage [31,32]. The freshwater green alga *Chlorella vulgaris* is also closely related to *Craspedia variabilis* infected by *Paramecium bursaria* Chlorella viruses [33]. Overall NCLDV core proteins were identified in virtually all major groups of algae, including Chlorophyta, Streptophyta, Stramenopiles, *Cryptophyta*, Euglenozoa, Haptophyceae, and Rhizaria. Remarkably, NCLDV core proteins were identified from multiple species of a same genus such as *Acanthamoeba* spp., *Sphaeroforma* spp., *Phytophthora* spp., *Pythium* spp., *Klebsormidium* spp. and *Chlamydomonas* spp.

Table 1. Nucleocytoplasmic large DNA viruses (NCLDV) core protein homologs in eukaryotic sequence datasets.

Eukaryotic Clade	Species	Habitat	Database	DNAP *	MCP *	ATPase *	D5 *	VLTF3 *
Genomic datasets								
Amoebozoa (Discosea)	*Acanthamoeba astronyxis*	terrestrial and aquatic	Assembly	√		√		√
	Acanthamoeba castellanii		RefSeq		√	√	√	√
	Acanthamoeba divionensis		Assembly	√		√		√
	Acanthamoeba healyi		Assembly		√			
	Acanthamoeba lenticulata		Assembly		√	√	√	√
	Acanthamoeba lugdunensis		Assembly		√	√		
	Acanthamoeba mauritaniens		Assembly	√	√	√		√
	Acanthamoeba pearcei		Assembly			√	√	√
	Acanthamoeba polyphaga		Assembly			√	√	√
	Acanthamoeba quina		Assembly		√	√	√	√
	Acanthamoeba rhysodes		Assembly		√			
Cryptophyta (Pyrenomonadales)	*Guillardia theta*	sea	RefSeq					MI
Euglenozoa	*Euglena gracilis*	freshwater	Assembly		√			
Fungi (Chytridiomycota)	*Gonapodya prolifera*	freshwater	RefSeq	Phy	√	√	Phy	√
Fungi (Glomeromycota)	*Rhizophagus irregularis*	terrestrial	Assembly	Asf				
Fungi (Blastocladiomycota)	*Allomyces macrogynus*	freshwater	RefSeq				Phy	
Metazoa (Arthropoda)	*Daphnia pulex*	freshwater	RefSeq		√			
Metazoa (Cnidaria)	*Exaiptasia pallida*	sea	RefSeq					Asf
	Hydra vulgaris	freshwater	RefSeq		Mi	√	Mi	√
Opisthokonta (Ichthyosporea)	*Sphaeroforma arctica*	sea	RefSeq	Irma		Irma		
	Sphaeroforma sirkka	sea	Assembly	Irma	Irma	Irma	√	Irma
Rhizaria (Cercozoa)	*Bigelowiella natans*	sea	RefSeq	Phy + √	√	√	Phy	√
Stramenopiles (Bicosoecida)	*Halocafeteria seosinensis*	saltern pond	Assembly					√
Stramenopiles (Eustigmatophyceae)	*Nannochloropsis limnetica*	freshwater	Assembly		Pha			Pha
Stramenopiles (Hyphochytriomycetes)	*Hyphochytrium catenoides*	terrestrial	Assembly	Asf	Asf	Asf	Asf	Asf
Stramenopiles (Oomycetes)	*Phytophthora* sp. *totara*	soilborne plant pathogen	Assembly		Asf			
	Phytophthora agathidicida		Assembly		Asf			
	Phytophthora alni		Assembly					Asf
	Phytophthora cambivora		Assembly					Asf
	Phytophthora cryptogea		Assembly		Asf			
	Phytophthora nicotianae		Assembly		Asf			Asf
	Phytophthora parasitica		RefSeq		Asf			Asf
	Pythium irregulare		Assembly					Asf
	Pythium oligadrum		Assembly		Asf			
	Pythium ultimum		Assembly		Asf			Asf

Table 1. *Cont.*

Eukaryotic Clade	Species	Habitat	Database	DNAP *	MCP *	ATPase *	D5 *	VLTF3 *
Stramenopiles (Phaeophyceae)	*Cladosiphon okamuranus*	sea	Assembly	Pha	Pha	Pha	Pha	Pha
	Ectocarpus siliculosus		RefSeq	Pha	Pha	Pha	Pha	Pha
	Saccharina japonica		RefSeq		Pha			
Viridiplantae (Chlorophyta)	*Asterochloris glomerata*	lichen photobiont	Assembly	Phy	Phy	Phy	Phy	
	Chlamydomonas applanata	terrestrial	Assembly	Mi				
	Chlamydomonas asymmetrica	freshwater	Assembly	Mi	Mi	Mi	Mi + Phy	Mi
	Chlamydomonas sphaeroides	freshwater	Assembly	Mi	Mi	Mi	Mi + Phy	Mi + √
	Chlorella vulgaris	freshwater	Assembly		√			
	Coccomyxa sp. *LA000219*	unknown	Assembly	Mi	Mi	Mi	Mi	Mi
	Cymbomonas tetramitiformis	sea	Assembly	Phy	Phy	Phy	Phy	Phy
	Haematococcus pluvialis	freshwater	Assembly	Mi	Mi	Mi	Mi + Phy	Mi
Viridiplantae (Streptophyta)	*Klebsormidium flaccidum*	terrestrial	RefSeq	Phy	Phy	Phy	√	Phy
Viridiplantae (Streptophyta)	*Physcomitrella patens*	terrestrial	RefSeq	Pitho			Pitho	
Transcriptomic datasets								
Cryptophyta (Cryptomonadales)	*Hemiselmis andersenii*	sea	MMETSP	√				
Cryptophyta (Pyrenomonadales)	*Hanusia phi*	sea	MMETSP	√				
Haptophyceae (Coccolithales)	*Pleurochrysis carterae*	sea	MMETSP		√	√		√
Haptophyceae (Isochrysidales)	*Chrysochromulina polylepis*	sea	MMETSP			√		
	Isochrysis galbana	sea	MMETSP			√		
Haptophyceae (Phaeocystales)	*Phaeocystis antarctica*	sea	MMETSP		√			
Haptophyceae (Prymnesiales)	*Emiliania huxleyi*	sea	MMETSP	Coc	Coc		Coc	
Rhizaria (Cercozoa)	*Lotharella globosa*	sea	MMETSP	Phy				
Stramenopiles (Labyrinthulomycetes)	*Aurantiochytrium limacinum*	sea	MMETSP	Pha				
	Schizochytrium aggregatum	sea	MMETSP		√			√
	Thraustochytrium sp.	sea	MMETSP		√			√
Undescribed Strain	*CCMP2135*	sea	MMETSP					√
Undescribed Strain	*CCMP2436*	sea	MMETSP				√	
Viridiplantae (Chlorophyta)	*Carteria crucifera*	freshwater	1KP	Mi			Mi	
	Cylindrocapsa geminella	freshwater	1KP	Mi				
Viridiplantae (Streptophyta)	*Entransia fimbriat*	freshwater	1KP	Phy				
	Interfilum paradoxum	terrestrial	1KP	Phy				
	Klebsormidium subtile	terrestrial	1KP	Phy				

* putative phylogenetic grouping of the NCLDV core protein homologs based on the phylogenetic trees presented in Figure 1 and Figure S1–S4: √ = unknown clade, Phy = *Phycodnaviridae*, Mi = *Mimiviridae*, Pha = phaeoviruses, Coc = coccolithoviruses, Pitho = putative Pithoviridae, Asf = *Asfarviridae* and IrMa = *Iridoviridae*/*Marseilleviridae* cluster. Column names: DNAP, DNA polymerase; MCP, major capsid protein; ATPase, DNA packaging ATPase; D5, D5 helicase; VLTF3, very late transcription factor 3.

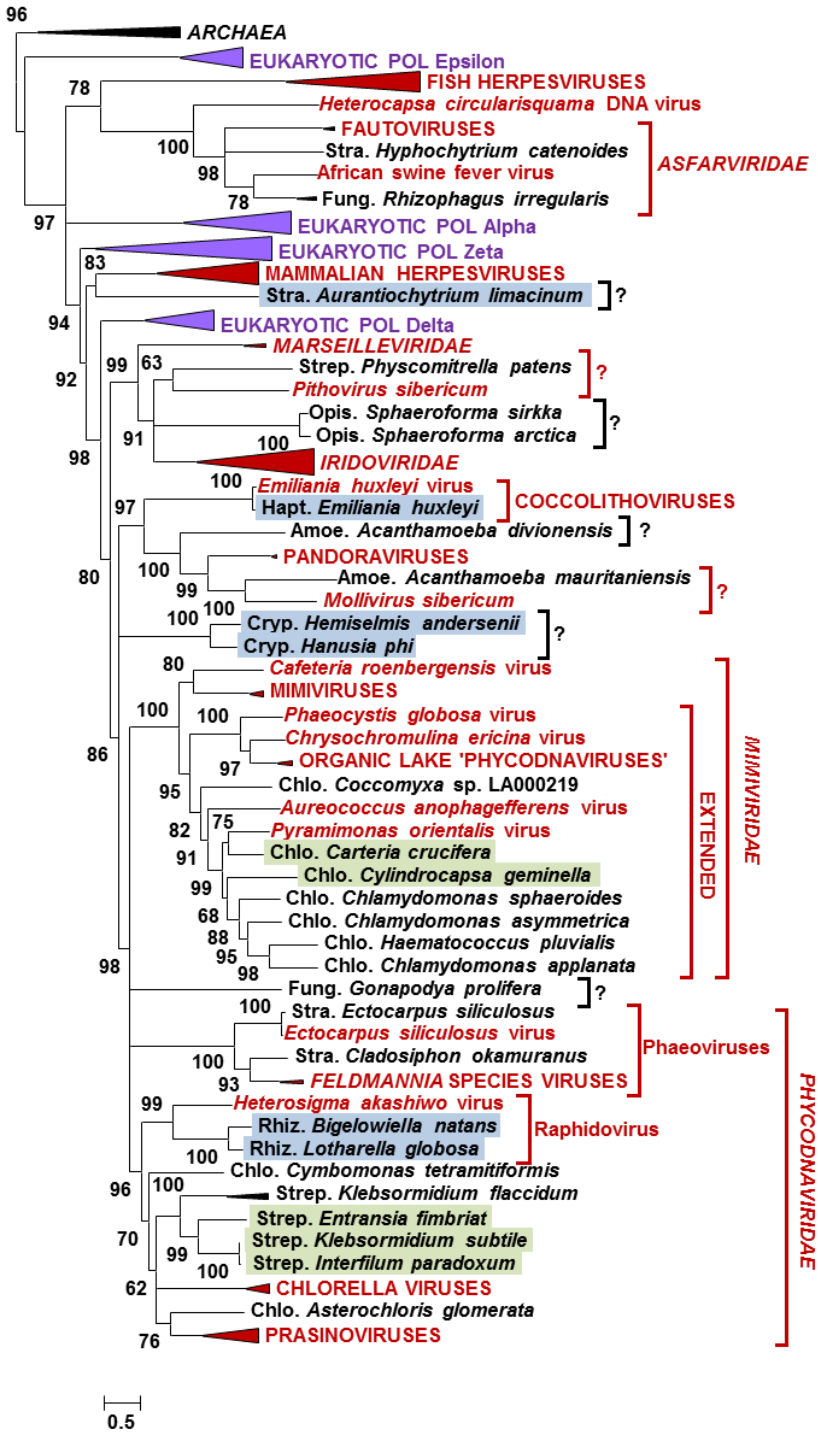


Figure 1. Maximum likelihood phylogenetic tree of DNA polymerase proteins. Statistical supports for branch (SH-like local support test) are given above or below nodes in percent. Branches with support less than 50% were collapsed. Species names with colored background indicate transcribed genes: green, 1KP transcriptomes; blue, MMETSP transcriptomes. Red and black question marks show potential extension of recognized viral groups or new viral clades, respectively. The scale bar indicates the number of substitution per site. Sequences, alignments, and phylogenetic trees are available in Dataset S1.

3.2. Phylogeny of Eukaryotic NCLDV-Like Proteins

To investigate the phylogenetic relationships between NCLDV-like proteins identified in eukaryotic datasets and their homologs in extant viruses, maximum likelihood phylogenetic trees were constructed for each of the five NCLDV core proteins (Figure 1 and Figure S1–S4). Overall, the resulting phylogenetic trees revealed several general characteristics of the NCLDV-like sequences. First, most of the NCLDV core proteins identified in eukaryotic datasets branched close to or within existing viral clades, further supporting the hypothesis of their viral origin. Second, 20 eukaryotes listed in Table 1, containing two or more NCLDV core genes, occupied consistent positions across phylogenetic trees (i.e., grouped within the same viral clade in each phylogenetic tree). This observation suggests that the viral sequences in each eukaryote arose from a single unique virus rather than multiple unrelated viral sources. Lastly, closely related eukaryotes tended to share closely-related viral sequences. This involved organisms beyond the genus rank such as for example chlorophytan or streptophytan species which had sequences forming subtrees within the *Phycodnaviridae* or *Mimiviridae* clades, or Stramenopile species branching within the *Asfarviridae* clade. This phylogenetic "correlation" between virus sequences and potential hosts can occur if closely related virus-like sequences originate from a single viral genome integration event in an ancestral eukaryotic host—the transferred genes could then spread across the host progeny up to the extant species. It has been suggested that most *Acanthamoeba* inserted viral genes became nonfunctional and decayed by accumulation of mutations [16]; it is, therefore, possible that sequence homology can no longer be recognized between viral inserts after a sufficiently long period of divergence. Alternatively, some of these sequences may originate from closely related contaminating viruses. In fact, some viral clades are apparently specific to certain eukaryotic groups, such as for example chloroviruses that infect *Chlorella* species [33], prasinoviruses that infect prasinophyte algae [34], or phaeoviruses that infect brown algae [35], possibly because their speciation and diversification co-occurred with those of their hosts. Under this co-speciation scenario, closely related eukaryotes may be infected by closely related viral species, which can also result in the observations made in our phylogenetic trees.

The B-family DNAP gene is a core NCLDV gene traditionally used as a reference phylogenetic marker to establish the taxonomy of large DNA viruses [36,37]. The DNAPs of NCLDVs are phylogenetically related to those of eukaryotes [38]. Our phylogenetic tree is largely in agreement with previous studies and shows that the eukaryotic DNAP delta emerged as a sister group to most NCLDV DNAPs (Figure 1). Furthermore, *Asfarviridae*, together with fish herpesviruses and *Heterocapsa circularisquama* DNA virus, form a separate group from the other NCLDV DNAPs [38,39]. As expected, some of the virus-like DNAPs identified in algae branched within the *Phycodnaviridae* family, which is a group of NCLDVs exclusively infecting phytoplanktonic species [35]. Among these sequences were proteins from streptophytan and rhizarian species, two major algal clades for which no NCLDV has ever been reported so far. Moreover, virus-like DNAPs from seven chlorophytan species grouped within the "extended *Mimiviridae*" clade, another group of large DNA viruses infecting a variety of microalgae but more closely related to giant mimiviruses [40]. Interestingly these chlorophytes include three species of *Chlamydomonas.* This genus of green algae also contains *Chlamydomonas reinhardtii,* a model organism for molecular and chloroplast biology. Although 25 putative viral genes were identified in this species [11], none of the five NCLDV core protein genes were found in the *C. reinhardtii* genome. Finally, the non-photosynthtic stramenopile *Hyphochytrium catenoides* and the fungus *Rhizophagus irregularis* contained DNAP sequences branching within the *Asfarviridae*. This result suggests that viruses from the *Asfarviridae* family have a much wider host range than currently thought. Fungi have never been reported as a potential host for NCLDVs and the presence of a viral DNAP in *R. irregularis* and in another fungus *Gonapodya prolifera* (as well as a D5 primase-helicase in the fungus *Allomyces macrogynus;* Figure S3 and Table 1) suggest that these organisms may have been infected by members of NCLDVs. Note that *G. prolifera* and *A. macrogynus* are members of two ancestral fungal lineages, which contain species feeding on algae. Furthermore, the fungal D5 primase-helicases branched close to the *Heterosigma akashiwo* virus, which is a member of the *Phycodnaviridae*. Thus

we cannot rule out that the NCLDV-like sequences identified in these two lower fungi may in fact originate from viruses infecting their algal preys.

Interestingly a number of virus-like DNAPs branched outside recognized taxonomic clades [41] suggesting that they belong to yet unknown taxa (e.g., indicated by black question marks in Figure 1). Other sequences grouping close to single, unclassified viral isolates (e.g., indicated by red question marks in Figure 1), might originate from members of the extended putative *Pithoviridae* family (e.g., represented by a sequence from the moss *P. patens*) and *Molliviridae* family (e.g., represented by a sequence from the amoeba *A. mauritaniensis*). Thus, our data, and more generally the approach of searching viral sequences in eukaryotic sequence data, make it possible to consolidate and even improve our knowledge of the NCLDV biodiversity. It is likely that some of the donor viruses encoded original functions and have developed new ways of interacting with their host that are radically different from the mechanisms already characterized.

3.3. Genomic Context around NCLDV-Like Genes

The discovery of viral sequences in eukaryotic genomes naturally poses the question of their origin, which can be a viral contamination, a provirus or a lateral gene transfer. A viral infection was suspected to be at the origin of the viral transcripts identified in the *E. huxleyii* and *P. carterae* transcriptomes. Filée suggested to examine the genomic environment around the virus-like genes to decipher whether they come from inserted viral DNA or contamination with free viral DNA during sequencing [13]. According to the author, insertion is the most likely hypothesis when virus-like genes are surrounded by intron-rich genes highly similar to eukaryotic homologs. We investigated the nature of genes surrounding NCLDV-like sequences in genomes that have a publicly available annotation, to the exception of organisms for which viral inserts have already been studied in details (i.e., *E. siliculosus* [12], *B. natans* [8], *P. patens* [9], *Acanthamoeba* spp. [13,16], *H. vulgaris* [10,13] and *P. parasitica* [10]). Figure S5 show the gene organization in contigs containing NCLDV core genes in seven eukaryotes, namely *A. macrogynus*, *D. pulex* [42], *E. pallida* [43], *Sphaeroforma arctica*, *Klebsormidium flaccidum* [44], *Saccharina japonica* [45], and *G. prolifera* [46].

Contigs with homologs to NCLDV core genes had sizes ranging from 1.2 kbp to 1.4 Mbp, so most of them contained more than one gene. The origin of the neighboring genes, inferred from the taxonomic information of their best match, generally indicates that other viral genes are present in the immediate vicinity of NCLDV core genes (Figure S5). Thus, most of the NCLDV core genes do not seem to result from horizontal transfer of a single isolated gene. However, this is not the case for the two metazoans, *D. pulex* and *E. pallida*, that each carries copies of a single NCLDV core gene (respectively, MCP and VLTF3) isolated in the midst of typical metazoan genes (Figure S5). Furthermore, the viral genes of the fungi *G. prolifera and A. macrogynus,* the brown alga *S. japonica* and the protist *S. arctica* are grouped in small genomic islands amid genes of eukaryotic origin, suggesting that they result from an insertion of a larger viral genome fragment. The viral sequences of *S. japonica* are closely related to phaeoviruses, which have a lysogenic reproduction and exist as provirus elements incorporated into the genomes of the brown algae *Ectocarpus siliculosus* and *Feldmannia* species [12,14,15]. Thus, the viral genes identified in the *S. japonica* genome might be remnants of an ancient provirus.

In contrast, NCLDV core genes of *K. flaccidum* are contained in contigs that have a dominance of viral genes, intermingled with a minority of genes most closely related to bacterial or eukaryotic homologs. Such a gene mosaicism is typical of NCLDV genomes [47]. Except for contig DF237168, there is no apparent juncture between a eukaryotic genomic region and a viral genomic region (i.e., a region containing a majority of eukaryotic genes followed by a region containing a majority of viral genes). Although this observation is compatible with a contaminating virus, this hypothesis is unlikely because a single viral genotype would be expected in the case of an infected culture. In contrast, we found five contigs containing a DNAP gene while NCLDVs only contain a single DNAP gene per genome. We also found six packaging ATPase genes, six MCP genes, and five VLTF3 genes. The levels of protein similarity between the DNAPs ranged from 40% to 92%, which excluded the possibility that

the homologous regions originate from variants of a same initial viral genotype. Furthermore the gene order was highly rearranged between homologous regions, further refuting the hypothesis of a single viral genotype. These data suggest the *K. flaccidum* genome contains distinct viral insertions. These inserts may result from duplication of an original viral insert followed by sequence divergence and rearrangements of the duplicated copies. Alternatively, they may result from independent acquisitions from multiple viruses.

Altogether, our analysis of seven annotated eukaryotic genomes supports the hypothesis of lateral gene transfer from viruses rather than contamination with free viral DNA during sequencing. The same conclusion was drawn in other studies of various organisms including *Acanthamoeba* spp., *P. patens* and *B. natans* [8,9,13,16]. Thus, there is a general consensus indicating that when viral sequences are identified in a given eukaryotic genome assembly, they are likely to result from bona fide viral genomic insertions rather than an alternative source. Given the substantial number of eukaryotic genomes concerned by inserted viral sequences (Table 1), viruses, and especially NCLDVs, may soon take center stage in our understanding of eukaryotic genome evolution. Viral insertions may turn out to be a major force driving lateral gene transfers between viruses and eukaryotes or between eukaryotes. The wide phylogenetic spectrum of eukaryotes containing viral sequences also suggest that these inserts might serve as DNA template in an evolutionarily conserved defense mechanism against viruses based on sequence recognition such as the RNA interference (RNAi) pathway [48].

3.4. Hints on Viral Functions

Another interesting aspect of viral inserts is that genes contained within viral regions can provide hints on unexpected functional capabilities of the original viruses. For instance, we found two highly similar expansin genes in two NCLDV-like contigs of the *K. flaccidum* genome assembly (DF237168.1 and DF237869.1; Figure S5). Although most similar to plant homologs, these expansin genes are both surrounded by a VLTF3 gene and a hypothetical protein gene that only has homologs in phycodnaviruses. This suggests that the expansin genes had been captured by the original donor virus from a plant or algal cell, before lending to the *K. flaccidum* genome through integration of viral DNA. Expansins mediate cell wall extension in plants by disrupting non-covalent binding of wall polysaccharides [49,50]. Lateral transfers of expansin genes from plants toward their fungal and bacterial parasites have been described, and their functional similarity suggests that these proteins mediate plant-microbial interaction [51]. Thus, in analogy to cellular plant parasites, this gene could also have a role during viral infection by enabling the virus to cross the host cell wall barrier. This would consist in a case of functional convergence between eukaryotic, bacterial and viral plant pathogens.

Additionally, two *K. flaccidum* viral regions contained a gene encoding a U-type cyclin domain (DF237607.1 and DF237785.1; Figure S5). In the available *K. flaccidum* genome annotation the cyclin domain is predicted to be fused with a MCP domain. However, this protein structure is likely an annotation error resulting from the merging of two independent exons each containing one of the two domains. In fact, no transcript sequence supports the junction between the two introns in a *K. flaccidum* RNAseq study [44]. On the other side of the cyclin gene, we found a viral DNA packaging ATPase gene. Although the cyclin domain is more closely related to plant homologs, the viral origin of the surrounding genes suggests a cyclin gene captured from a plant cell was present in the original donor viral genome. Viral encoded cyclins have been identified in several viral families including herpesviruses, retroviruses, and baculoviruses, where they drive cell cycle transitions of the host [52]. Many DNA viruses induce quiescent cells to enter the cell cycle; this is thought to increase pools of deoxynucleotides and, thus, facilitate viral replication. In contrast, some viruses can arrest cells in a particular phase of the cell cycle that is favorable for replication of the specific virus [53]. If the existence of the predicted *K. flaccidum* virus is confirmed in future studies, it would represent the first instance of a cyclin gene in a NCLDV.

A contig of the fungus *G. prolifera* (KQ965906.1; Figure S5) contained a chitinase gene, together with three other viral genes encoding a MCP, a VATPase_H domain containing protein and a protein

of unknown function. Chitinases are enzymes that degrade chitin, which is one of the most abundant biopolymers in nature. Chitin occurs in various contexts across a broad range of species and is the main constituent of fungal cell wall [54]. Remarkably, the *G. prolifera* chitinase is most closely related to homologs in chloroviruses (phycodnavirus) which are presumably involved in degradation of the cell wall of green algae of the *Chlorella* genus [55]. Interestingly the same *G. prolifera* viral region contains another chitinase-like protein surrounded by 2 viral genes, but this one is more similar to bacterial homologs. Thus, the putative *G. prolifera* virus might use a similar enzyme apparatus as chloroviruses to pass through the chitin-rich fungal cell wall.

4. Conclusions

We are only beginning to appreciate the extraordinary diversity of NCLDVs, which are among the most intriguing viruses on the planet. Here we show that substantial progress in the description of the NCLDV biodiversity can be made by mining potential host sequences in order to identify genetic markers of NCLDVs. Using this approach, both a virus and its putative host can be brought to light, a significant advantage over metagenomics, which cannot directly identify the two partners. This approach takes advantage of what appears to be an important, but as yet poorly understood, feature of NCLDVs: they leave footprints of their passage in the cell in the form of viral inserts in the host genome. Here we chose to search virus-like sequences using the five most consistently conserved genes in NCLDVs. Others suggested to use the RNA polymerase subunit 2 to identify giant viruses sequences in (meta)genomic data [10]. However, some of these genes are not universally conserved among NCLDVs. For instance, a RNA polymerase gene is absent in most *Phycodnaviridae* genomes, whereas a MCP gene could not be detected in the genomes of pandoraviruses and *P. sibericum* [3,56,57]. *P. sibericum* is also apparently lacking a gene for packaging ATPase. Thus additional combinations of NCLDV reference markers may lead to an increasing number of eukaryotic datasets positive for NCLDV sequences [11]. It is also worth noting that the abundance of viruses in environmental samples is sometimes estimated by quantitative PCR using primers specific for virus genes [58] or by the number of metagenomic reads overlapping viral genes [6]. However, given the apparent ease with which NCLDV genes find themselves integrated into host genomes, these approaches may lead to over estimating the viral abundances if the surveyed samples also contain hosts harboring viral HGTs.

In this study we could predict the existence of new members of NCLDVs, some of which are apparently distantly related from already characterized viruses and may define new viral clades. Examination of the gene content of viral regions also helped us predicting some potential functional capabilities of the original viruses. We also predicted a wide range of potential hosts, most of whom have never had an association described with NCLDVs. The validity of all these predictions must now be evaluated through experimental approaches. Most of the organisms in which viral sequences were found are cultivable in laboratory conditions. This offers a favorable experimental framework to prospect environmental samples in order to isolate viral strains by co-culture with a eukaryote. A potential host may be chosen according to the phylogenetic reconstruction of its viral sequences in order to target the isolation of novel NCLDVs of special scientific interest. If such an approach proves to be successful, it may help in improving our understanding of the NCLDV world.

Supplementary Materials: The following are available online at www.mdpi.com/1999-4915/9/1/17/s1, Figure S1: Maximum likelihood phylogenetic tree of MCPs, Figure S2: Maximum likelihood phylogenetic tree of packaging ATPases, Figure S3: Maximum likelihood phylogenetic tree of D5 helicase-primase, Figure S4: Maximum likelihood phylogenetic tree of VLTF3, Figure S5: Genomic context of the NCLDV core genes identified in eukaryotic annotated genomes, Dataset S1: Alignments and phylogenetic trees of the 5 NCLDV core proteins and their eukaryotic homologs.

Acknowledgments: We thank Daniele Armaleo and Olivier Vallon for providing the *A. glomerata* and *H. pluvialis* sequences ahead of publication.

Author Contributions: G.B. conceived and designed the experiments; L.G.-L. and G.B. analyzed the data; L.G.-L. and G.B. wrote the paper.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Iyer, L.M.; Balaji, S.; Koonin, E.V.; Aravind, L. Evolutionary genomics of nucleo-cytoplasmic large DNA viruses. *Virus Res.* **2006**, *117*, 156–184. [CrossRef] [PubMed]
2. Koonin, E.V.; Yutin, N. Nucleo-cytoplasmic Large DNA Viruses (NCLDV) of Eukaryotes. *eLS* **2012**.
3. Philippe, N.; Legendre, M.; Doutre, G.; Couté, Y.; Poirot, O.; Lescot, M.; Arslan, D.; Seltzer, V.; Bertaux, L.; Bruley, C.; et al. Pandoraviruses: Amoeba Viruses with Genomes Up to 2.5 Mb Reaching That of Parasitic Eukaryotes. *Science* **2013**, *341*, 281–286. [CrossRef] [PubMed]
4. Fischer, M.G. Giant viruses come of age. *Curr. Opin. Microbiol.* **2016**, *31*, 50–57. [CrossRef] [PubMed]
5. Wommack, K.E.; Nasko, D.J.; Chopyk, J.; Sakowski, E.G. Counts and sequences, observations that continue to change our understanding of viruses in nature. *J. Microbiol. Seoul Korea* **2015**, *53*, 181–192. [CrossRef] [PubMed]
6. Hingamp, P.; Grimsley, N.; Acinas, S.G.; Clerissi, C.; Subirana, L.; Poulain, J.; Ferrera, I.; Sarmento, H.; Villar, E.; Lima-Mendez, G.; et al. Exploring nucleo-cytoplasmic large DNA viruses in Tara Oceans microbial metagenomes. *ISME J.* **2013**, *7*, 1678–1695. [CrossRef] [PubMed]
7. Mihara, T.; Nishimura, Y.; Shimizu, Y.; Nishiyama, H.; Yoshikawa, G.; Uehara, H.; Hingamp, P.; Goto, S.; Ogata, H. Linking Virus Genomes with Host Taxonomy. *Viruses* **2016**, *8*, 66. [CrossRef] [PubMed]
8. Blanc, G.; Gallot-Lavallée, L.; Maumus, F. Provirophages in the Bigelowiella genome bear testimony to past encounters with giant viruses. *Proc. Natl. Acad. Sci. USA* **2015**, *112*, E5318–E5326. [CrossRef] [PubMed]
9. Maumus, F.; Epert, A.; Nogué, F.; Blanc, G. Plant genomes enclose footprints of past infections by giant virus relatives. *Nat. Commun.* **2014**, *5*. [CrossRef] [PubMed]
10. Sharma, V.; Colson, P.; Giorgi, R.; Pontarotti, P.; Raoult, D. DNA-Dependent RNA Polymerase Detects Hidden Giant Viruses in Published Databanks. *Genome Biol. Evol.* **2014**, *6*, 1603–1610. [CrossRef] [PubMed]
11. Wang, L.; Wu, S.; Liu, T.; Sun, J.; Chi, S.; Liu, C.; Li, X.; Yin, J.; Wang, X.; Yu, J. Endogenous viral elements in algal genomes. *Acta Oceanol. Sin.* **2014**, *33*, 102–107. [CrossRef]
12. Delaroque, N.; Boland, W. The genome of the brown alga *Ectocarpus siliculosus* contains a series of viral DNA pieces, suggesting an ancient association with large dsDNA viruses. *BMC Evol. Biol.* **2008**, *8*, 110. [CrossRef] [PubMed]
13. Filée, J. Multiple occurrences of giant virus core genes acquired by eukaryotic genomes: The visible part of the iceberg? *Virology* **2014**, *466–467*, 53–59. [CrossRef] [PubMed]
14. Meints, R.H.; Ivey, R.G.; Lee, A.M.; Choi, T.-J. Identification of Two Virus Integration Sites in the Brown Alga Feldmannia Chromosome. *J. Virol.* **2008**, *82*, 1407–1413. [CrossRef] [PubMed]
15. Delaroque, N.; Maier, I.; Knippers, R.; Müller, D.G. Persistent virus integration into the genome of its algal host, *Ectocarpus siliculosus* (Phaeophyceae). *J. Gen. Virol.* **1999**, *80*, 1367–1370. [CrossRef] [PubMed]
16. Maumus, F.; Blanc, G. Study of gene trafficking between *Acanthamoeba* and giant viruses suggests an undiscovered family of amoeba-infecting viruses. *Genome Biol. Evol.* **2016**. [CrossRef] [PubMed]
17. Yutin, N.; Wolf, Y.I.; Raoult, D.; Koonin, E.V. Eukaryotic large nucleo-cytoplasmic DNA viruses: Clusters of orthologous genes and reconstruction of viral genome evolution. *Virol. J.* **2009**, *6*, 223. [CrossRef] [PubMed]
18. Available online: ftp://ftp.ncbi.nih.gov/pub/wolf/COGs/NCVOG/ (accessed on 30 November 2016).
19. Available online: https://www.bioinfodata.org/Blast4OneKP (accessed on 30 November 2016).
20. Available online: ftp://ftp.imicrobe.us/projects/104/CAM_P_0001000.pep.fa.gz (accessed on 30 November 2016).
21. Available online: ftp://ftp.imicrobe.us/projects/104/Callum_FINAL_biosample_ids.xls (accessed on 30 November 2016).
22. Available online: http://www.onekp.com/samples/list.php (accessed on 30 November 2016).
23. Yutin, N.; Raoult, D.; Koonin, E.V. Virophages, polintons, and transpovirons: a complex evolutionary network of diverse selfish genetic elements with different reproduction strategies. *Virol. J.* **2013**, *10*, 158. [CrossRef] [PubMed]
24. Dereeper, A.; Audic, S.; Claverie, J.-M.; Blanc, G. BLAST-EXPLORER helps you building datasets for phylogenetic analysis. *BMC Evol. Biol.* **2010**, *10*, 8. [CrossRef] [PubMed]
25. Katoh, K.; Kuma, K.; Toh, H.; Miyata, T. MAFFT version 5: improvement in accuracy of multiple sequence alignment. *Nucleic Acids Res.* **2005**, *33*, 511–518. [CrossRef] [PubMed]

26. Price, M.N.; Dehal, P.S.; Arkin, A.P. FastTree 2 – Approximately Maximum-Likelihood Trees for Large Alignments. *PLoS ONE* **2010**, *5*, e9490. [CrossRef] [PubMed]
27. Keeling, P.J.; Burki, F.; Wilcox, H.M.; Allam, B.; Allen, E.E.; Amaral-Zettler, L.A.; Armbrust, E.V.; Archibald, J.M.; Bharti, A.K.; Bell, C.J.; et al. The Marine Microbial Eukaryote Transcriptome Sequencing Project (MMETSP): Illuminating the Functional Diversity of Eukaryotic Life in the Oceans through Transcriptome Sequencing. *PLoS Biol* **2014**, *12*, e1001889. [CrossRef] [PubMed]
28. Matasci, N.; Hung, L.-H.; Yan, Z.; Carpenter, E.J.; Wickett, N.J.; Mirarab, S.; Nguyen, N.; Warnow, T.; Ayyampalayam, S.; Barker, M.; et al. Data access for the 1,000 Plants (1KP) project. *GigaScience* **2014**, *3*, 17. [CrossRef] [PubMed]
29. Suzuki, S.; Ishida, K.-I.; Hirakawa, Y. Diurnal transcriptional regulation of endosymbiotically derived genes in the chlorarachniophyte *Bigelowiella natans*. *Genome Biol. Evol.* **2016**. [CrossRef] [PubMed]
30. Park, J.S.; Cho, B.C.; Simpson, A.G.B. *Halocafeteria seosinensis* gen. et sp. nov. (Bicosoecida), a halophilic bacterivorous nanoflagellate isolated from a solar saltern. *Extrem. Life Extreme Cond.* **2006**, *10*, 493–504. [CrossRef] [PubMed]
31. Fischer, M.G.; Allen, M.J.; Wilson, W.H.; Suttle, C.A. Giant virus with a remarkable complement of genes infects marine zooplankton. *Proc. Natl. Acad. Sci. USA* **2010**, *107*, 19508–19513. [CrossRef] [PubMed]
32. Fischer, M.G.; Suttle, C.A. A virophage at the origin of large DNA transposons. *Science* **2011**, *332*, 231–234. [CrossRef] [PubMed]
33. Van Etten, J.L.; Dunigan, D.D. Chloroviruses: not your everyday plant virus. *Trends Plant Sci.* **2012**, *17*, 1–8. [CrossRef] [PubMed]
34. Clerissi, C.; Grimsley, N.; Ogata, H.; Hingamp, P.; Poulain, J.; Desdevises, Y. Unveiling of the Diversity of Prasinoviruses (*Phycodnaviridae*) in Marine Samples by Using High-Throughput Sequencing Analyses of PCR-Amplified DNA Polymerase and Major Capsid Protein Genes. *Appl. Environ. Microbiol.* **2014**, *80*, 3150–3160. [CrossRef] [PubMed]
35. Wilson, W.H.; Van Etten, J.L.; Allen, M.J. The *Phycodnaviridae*: The Story of How Tiny Giants Rule the World. *Curr. Top. Microbiol. Immunol.* **2009**, *328*, 1–42. [PubMed]
36. Hanson, L.A.; Rudis, M.R.; Vasquez-Lee, M.; Montgomery, R.D. A broadly applicable method to characterize large DNA viruses and adenoviruses based on the DNA polymerase gene. *Virol. J.* **2006**, *3*, 28. [CrossRef] [PubMed]
37. Chen, F.; Suttle, C.A. Evolutionary relationships among large double-stranded DNA viruses that infect microalgae and other organisms as inferred from DNA polymerase genes. *Virology* **1996**, *219*, 170–178. [CrossRef] [PubMed]
38. Takemura, M.; Yokobori, S.; Ogata, H. Evolution of Eukaryotic DNA Polymerases via Interaction Between Cells and Large DNA Viruses. *J. Mol. Evol.* **2015**, *81*, 24–33. [CrossRef] [PubMed]
39. Yutin, N.; Koonin, E.V. Hidden evolutionary complexity of Nucleo-Cytoplasmic Large DNA viruses of eukaryotes. *Virol. J.* **2012**, *9*, 161. [CrossRef] [PubMed]
40. Yutin, N.; Colson, P.; Raoult, D.; Koonin, E.V. *Mimiviridae*: clusters of orthologous genes, reconstruction of gene repertoire evolution and proposed expansion of the giant virus family. *Virol. J.* **2013**, *10*, 106. [CrossRef] [PubMed]
41. King, A.M.Q.; Adams, M.J.; Carstens, E.B.; Lefkowitz, E.J. *Virus Taxonomy: Classification and Nomenclature of Viruses: Ninth Report of the International Committee on Taxonomy of Viruses*; Elsevier Academic Press: San Diego, CA, USA, 2011.
42. Colbourne, J.K.; Pfrender, M.E.; Gilbert, D.; Thomas, W.K.; Tucker, A.; Oakley, T.H.; Tokishita, S.; Aerts, A.; Arnold, G.J.; Basu, M.K.; et al. The ecoresponsive genome of *Daphnia pulex*. *Science* **2011**, *331*, 555–561. [CrossRef] [PubMed]
43. Baumgarten, S.; Simakov, O.; Esherick, L.Y.; Liew, Y.J.; Lehnert, E.M.; Michell, C.T.; Li, Y.; Hambleton, E.A.; Guse, A.; Oates, M.E.; et al. The genome of Aiptasia, a sea anemone model for coral symbiosis. *Proc. Natl. Acad. Sci. USA* **2015**, *112*, 11893–11898. [CrossRef] [PubMed]
44. Hori, K.; Maruyama, F.; Fujisawa, T.; Togashi, T.; Yamamoto, N.; Seo, M.; Sato, S.; Yamada, T.; Mori, H.; Tajima, N.; et al. *Klebsormidium flaccidum* genome reveals primary factors for plant terrestrial adaptation. *Nat. Commun.* **2014**, *5*, 3978. [CrossRef] [PubMed]
45. Ye, N.; Zhang, X.; Miao, M.; Fan, X.; Zheng, Y.; Xu, D.; Wang, J.; Zhou, L.; Wang, D.; Gao, Y.; et al. Saccharina genomes provide novel insight into kelp biology. *Nat. Commun.* **2015**, *6*, 6986. [CrossRef] [PubMed]

46. Chang, Y.; Wang, S.; Sekimoto, S.; Aerts, A.L.; Choi, C.; Clum, A.; LaButti, K.M.; Lindquist, E.A.; Yee Ngan, C.; Ohm, R.A.; et al. Phylogenomic Analyses Indicate that Early Fungi Evolved Digesting Cell Walls of Algal Ancestors of Land Plants. *Genome Biol. Evol.* **2015**, *7*, 1590–1601. [CrossRef] [PubMed]
47. Filée, J.; Chandler, M. Gene Exchange and the Origin of Giant Viruses. *Intervirology* **2010**, *53*, 354–361. [CrossRef] [PubMed]
48. Stram, Y.; Kuzntzova, L. Inhibition of viruses by RNA interference. *Virus Genes* **2006**, *32*, 299–306. [CrossRef] [PubMed]
49. McQueen-Mason, S.; Durachko, D.M.; Cosgrove, D.J. Two endogenous proteins that induce cell wall extension in plants. *Plant Cell* **1992**, *4*, 1425–1433. [CrossRef] [PubMed]
50. Yennawar, N.H.; Li, L.-C.; Dudzinski, D.M.; Tabuchi, A.; Cosgrove, D.J. Crystal structure and activities of EXPB1 (Zea m 1), a beta-expansin and group-1 pollen allergen from maize. *Proc. Natl. Acad. Sci. USA* **2006**, *103*, 14664–14671. [CrossRef] [PubMed]
51. Nikolaidis, N.; Doran, N.; Cosgrove, D.J. Plant expansins in bacteria and fungi: evolution by horizontal gene transfer and independent domain fusion. *Mol. Biol. Evol.* **2014**, *31*, 376–386. [CrossRef] [PubMed]
52. Hardwick, J.M. Cyclin' on the viral path to destruction. *Nat. Cell Biol.* **2000**, *2*, E203–E204. [CrossRef] [PubMed]
53. Bagga, S.; Bouchard, M.J. Cell cycle regulation during viral infection. *Methods Mol. Biol. Clifton NJ* **2014**, *1170*, 165–227.
54. Bowman, S.M.; Free, S.J. The structure and synthesis of the fungal cell wall. *BioEssays* **2006**, *28*, 799–808. [CrossRef] [PubMed]
55. Yamada, T.; Onimatsu, H.; Van Etten, J.L. *Chlorella* viruses. *Adv. Virus Res.* **2006**, *66*, 293–336. [PubMed]
56. Klose, T.; Rossmann, M.G. Structure of large dsDNA viruses. *Biol. Chem.* **2014**, *395*, 711–719. [CrossRef] [PubMed]
57. Legendre, M.; Bartoli, J.; Shmakova, L.; Jeudy, S.; Labadie, K.; Adrait, A.; Lescot, M.; Poirot, O.; Bertaux, L.; Bruley, C.; et al. Thirty-thousand-year-old distant relative of giant icosahedral DNA viruses with a pandoravirus morphology. *Proc. Natl. Acad. Sci. USA* **2014**, *111*, 4274–4279. [CrossRef] [PubMed]
58. Short, S.M. The ecology of viruses that infect eukaryotic algae. *Environ. Microbiol.* **2012**, *14*, 2253–2271. [CrossRef] [PubMed]

Supplementary Materials: A Glimpse of Nucleo-Cytoplasmic Large DNA Virus Biodiversity through the Eukaryotic Genomics Window

Lucie Gallot-Lavallée [1] and Guillaume Blanc [1,2,*]

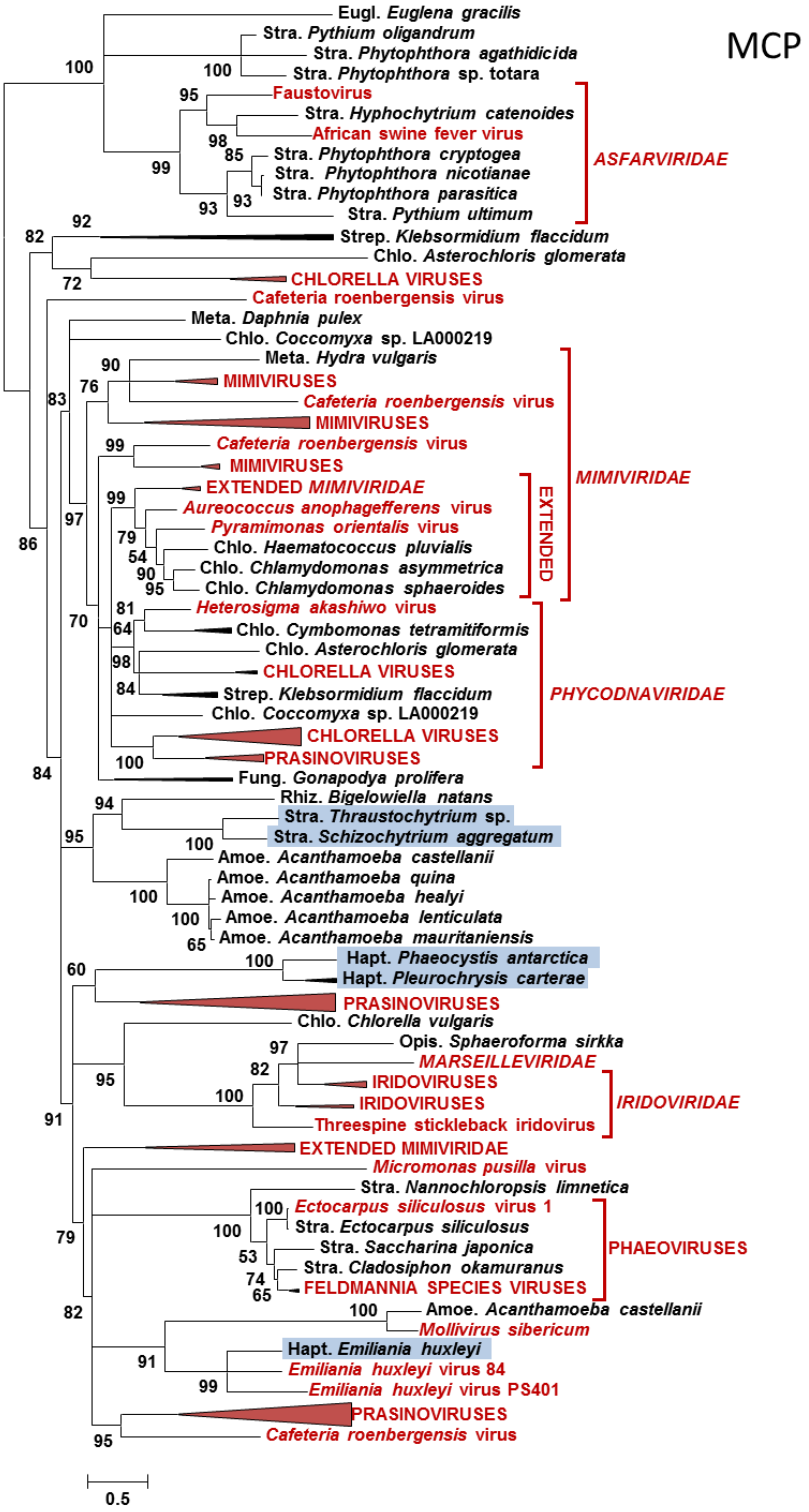


Figure S1. Maximum likelihood phylogenetic tree of MCPs. The figure legend is same as in Figure 1.

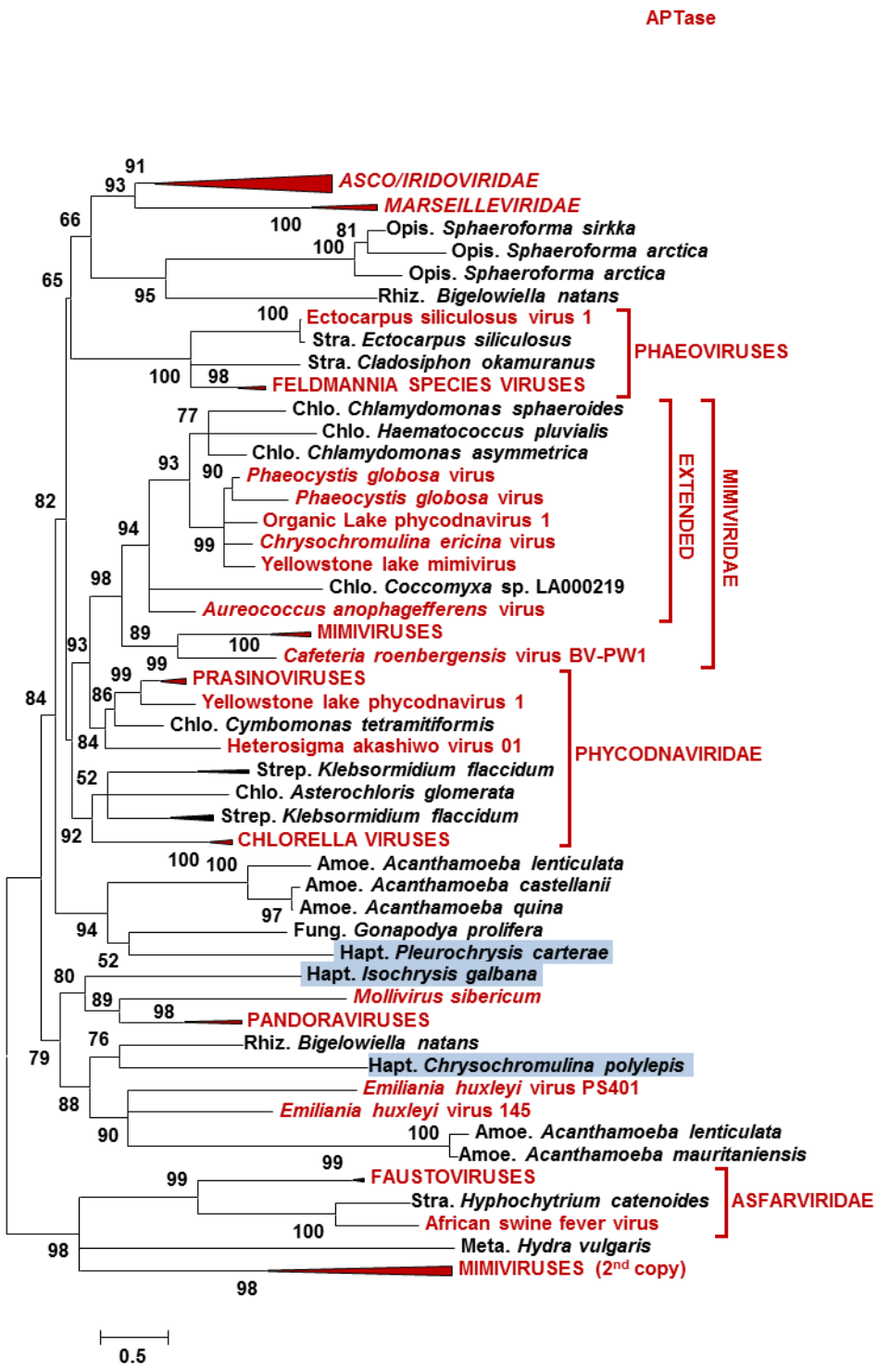


Figure S2. Maximum likelihood phylogenetic tree of packaging ATPases. The figure legend is same as in Figure 1.

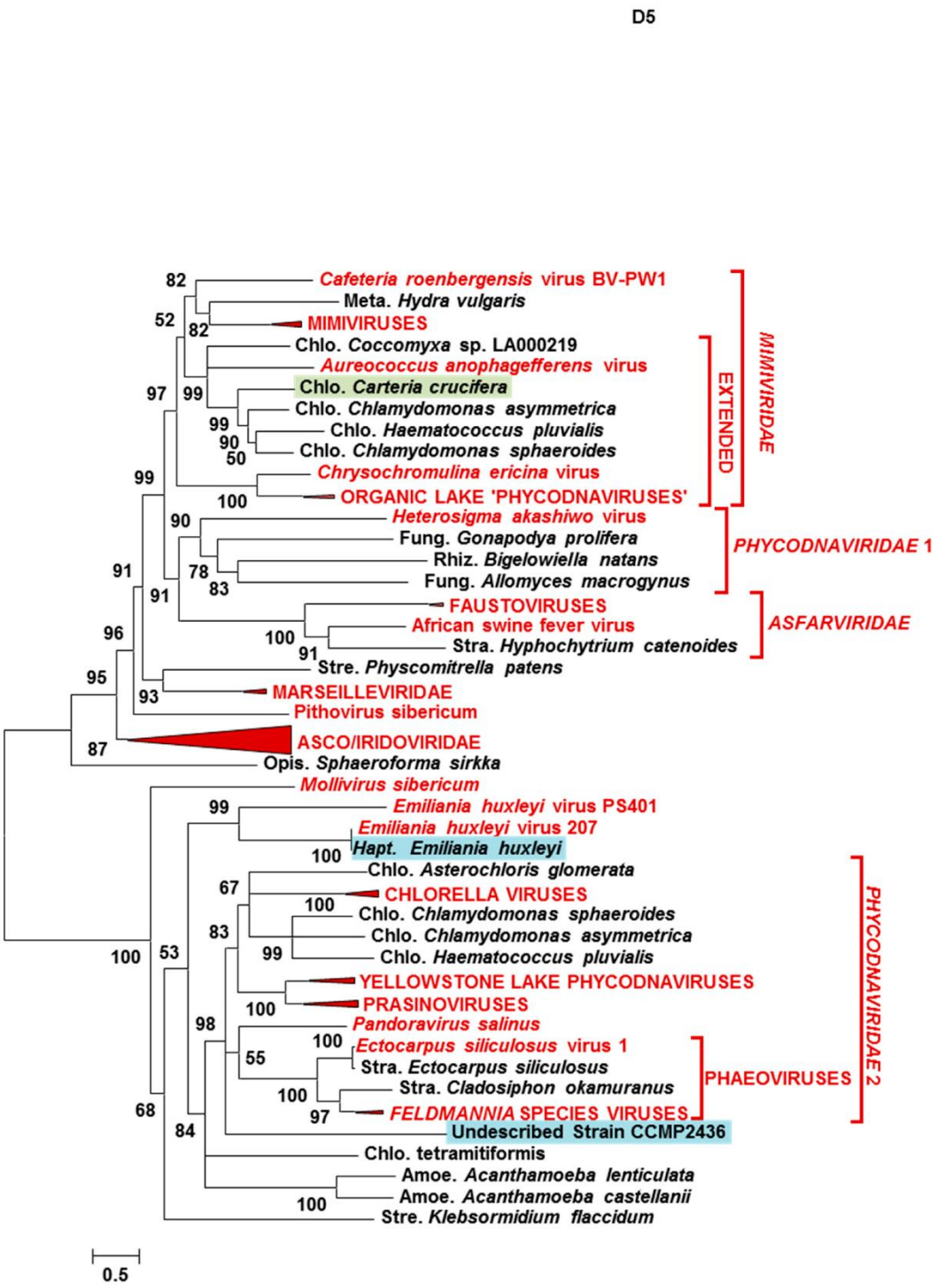


Figure S3. Maximum likelihood phylogenetic tree of D5 helicase-primase. The figure legend is same as in Figure 1.

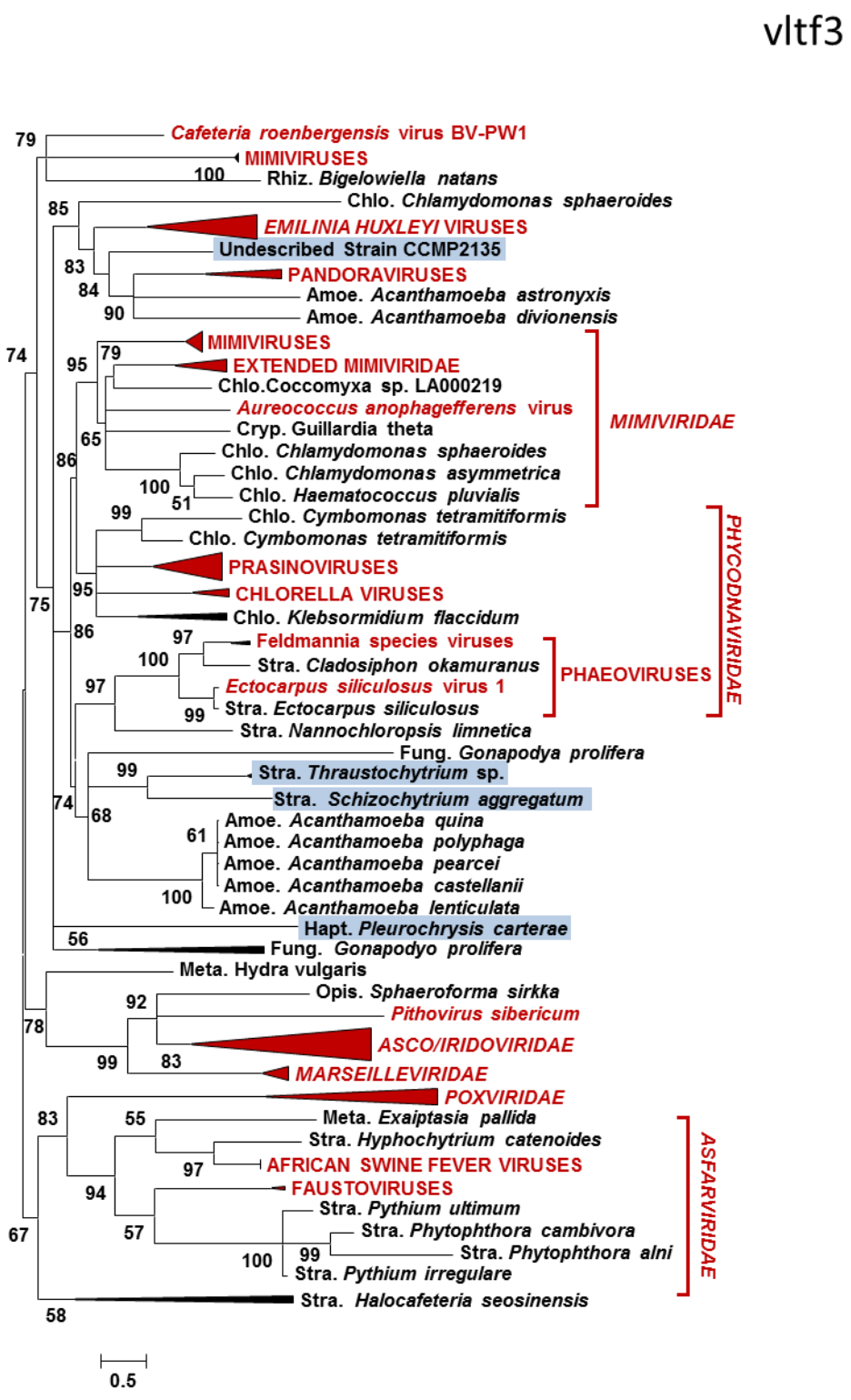


Figure S4. Maximum likelihood phylogenetic tree of VLTF3. The figure legend is same as in Figure 1.

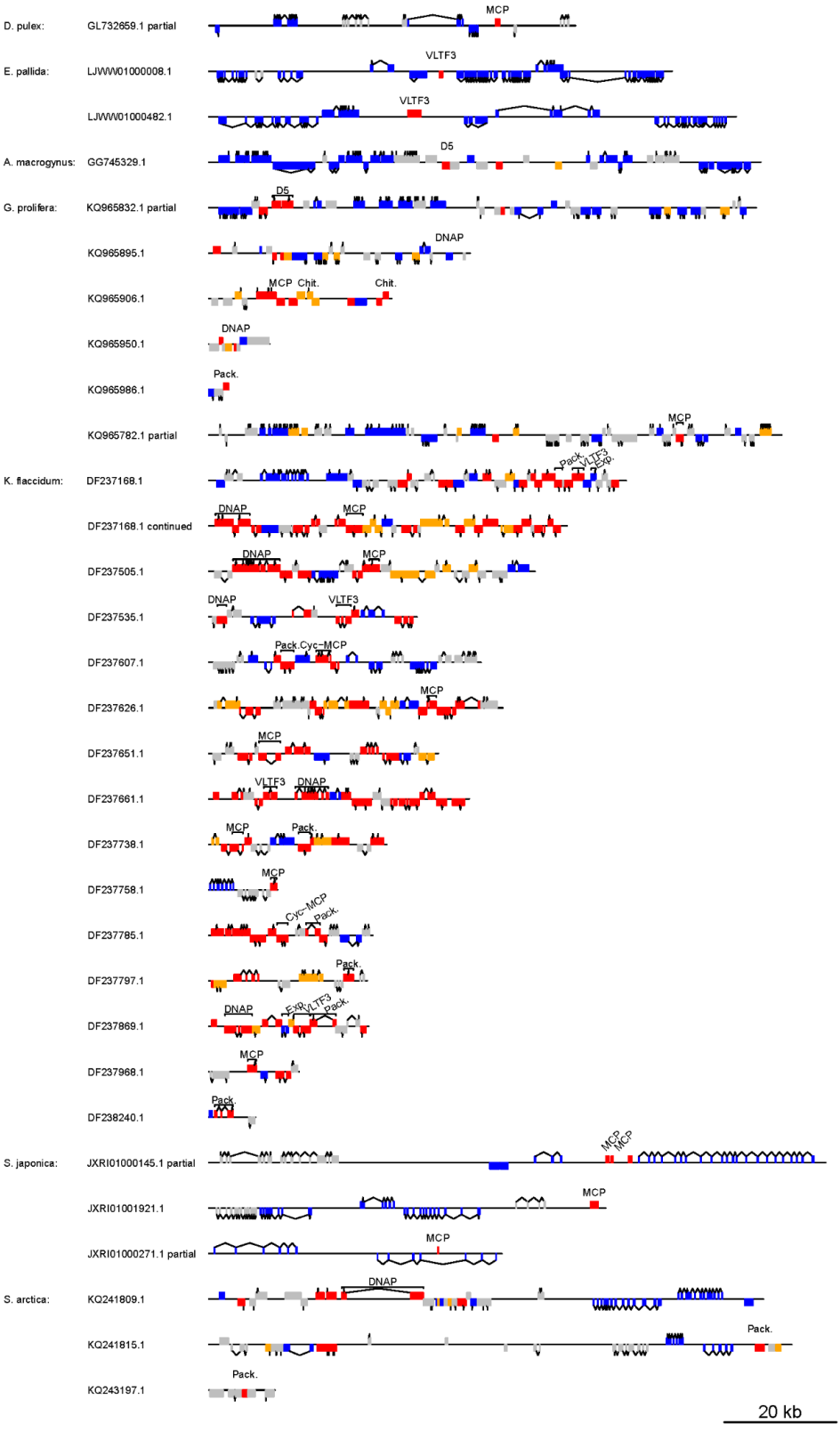


Figure S5. Genomic context of the NCLDV core genes identified in eukaryotic annotated genomes. Genbank gene annotations and viral genes detected in this study are represented by colored rectangles along the contigs. The colors indicate the taxonomic assignation of the gene's best match: blue, eukaryotes; orange, prokaryote; red, NCLDV; gray, no match. Interesting genes discussed in the manuscript are flagged as follows: MCP: major capsid protein; DNAP: DNA polymerase; Pack.: DNA packaging ATPase; D5: D5 helicase-primase; VLTF3: very late transcription factor 3; Cyc: cyclin; Exp.: expansin; Chit.: chitinase.

c - Résultats supplémentaires et discussion

Nous avons montré ici que les NCLDVs peuvent laisser des empreintes de leur passage dans les cellules sous la forme d'inserts génétiques. Ces inserts d'origine virale nous permettent de progresser dans notre compréhension de la diversité des NCLDVs. Effectivement, nous avons pu prédire l'existence de nouveaux NCLDVs, dont certains semblent phylogénétiquement éloignés des virus connus et pourraient donc appartenir à de nouvelles lignées virales. L'examen du contenu en gènes de ces inserts viraux nous a également permis de prédire de potentielles fonctionnalités originales de ces virus. D'autre part, cette approche fournit pour chaque virus putatif l'information de son hôte correspondant – l'organisme dans lequel est trouvée la séquence d'origine virale. Ceci est un avantage par rapport à la métagénomique où l'identité des hôtes associés aux virus présents dans un échantillon est plus difficilement accessible. Ainsi, des inserts viraux sont présents dans les génomes d'hôtes déjà connus, mais également dans de potentiels nouveaux hôtes de NCLDVs. Parmi ces derniers, certains sont proches d'hôtes connus ; d'autres appartiennent à des clades d'eucaryotes éloignés de ceux-ci. Concernant les virus d'autre part, certaines familles de NCLDVs ont pu être associées à des clades d'eucaryotes extrêmement divergents des hôtes à partir desquels ont été isolés les représentants de cette famille de virus. C'est le cas pour les *Mimiviridae* (c.1), les *Asfarviridae* (c.2), et les *Phaeovirus* (c.3). Ceci sera mis en relief dans les paragraphes suivants correspondants. D'autre part, la présence répétée du gène VLTF3 chez l'anémone *Exaiptasia pallida* sera discuté (c.4). Ensuite, les limitations des méthodes de paléovirologie au regard de la qualité et de la curation des bases de données sera abordée (c.5); Enfin une ouverture sera proposée (c.6) quant à l'intérêt de ces résultats pour les méthodes classiques d'isolement de virus.

c.1 - Retour sur les *Mimiviridae*

J'avais conclu la partie concernant CeV sur la prédiction que le spectre d'hôtes des *Mimiviridae* pourrait comprendre des clades eucaryotes éloignés de ceux qui sont connus comme tels. Nous avons observé lors de notre criblage que des séquences apparentées aux *Mimiviridae* sont effectivement présentes dans les génomes d'eucaryotes appartenant à des clades d'eucaryotes additionnels. La figure 13 synthétise nos connaissances actuelles quant aux hôtes des *Mimiviridae.* Sont apposés sur l'arbre eucaryote : (i) les hôtes des *Mimiviridae* isolés, (ii) les hôtes potentiels déduits de notre étude, (iii) les hôtes potentiel déduits de transfert de gènes décrits ailleurs (101), ou encore (iv) déduits d'analyses transcriptomiques (250). Les prédictions ont donc été validées.

c.2 - Les *Asfarviridae*

Une autre famille semble avoir un spectre d'hôte particulièrement large : Les *Asfarviridae*. Le membre fondateur de cette famille African *Swine* Fever virus a été isolé à partir du porc, tandis que Faustovirus a été isolé tout récemment à partir d'une amibe (251). Ces eucaryotes font partie

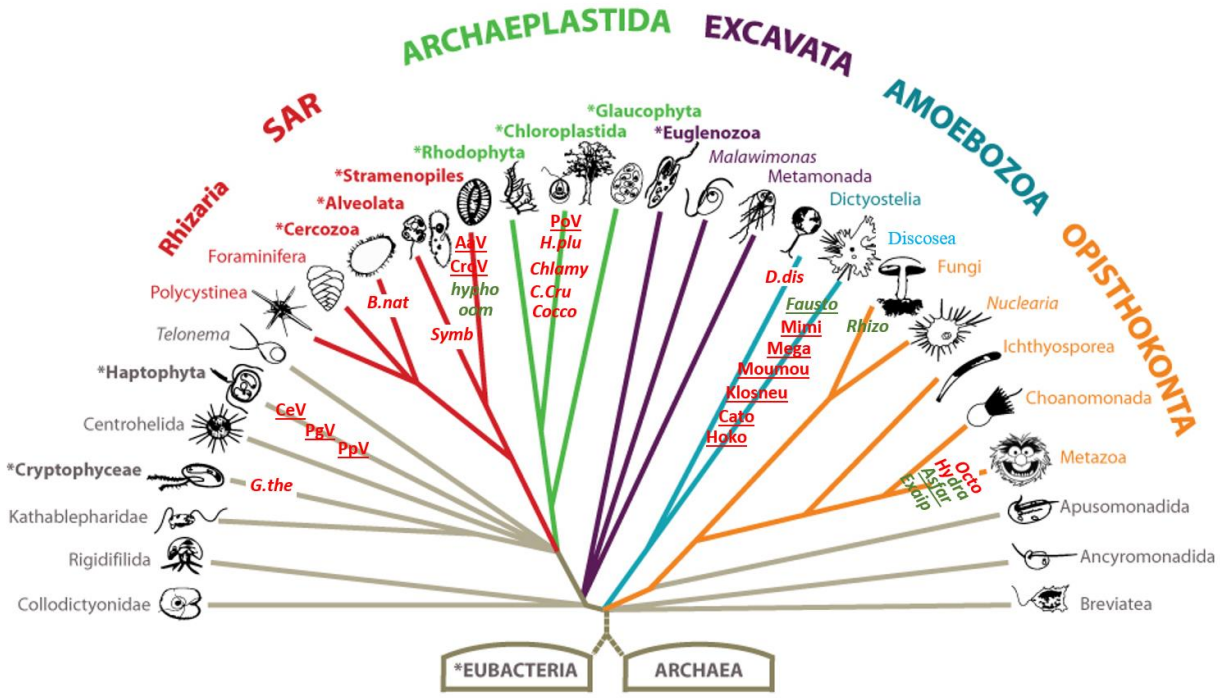


Figure 13: Diversité des hôtes eucaryotes connus ou supposés des Mimiviridae (en rouge) et des Asfarviridae (en vert). Les noms des virus isolés sont soulignés. Les noms écrits en rouge et vert correspondent à des eucaryotes dont l'assemblage du génome a révélé la présence de gènes originaire de Mimiviridae et Asfarviridae respectivement, ce qui suggère une association passée avec des virus de ces familles. *G.the=Guillardia theta, B.nat=Bigelowiella natans, Symb=Symbiodinium, H.plu=Haematococcus pluvialis, Chlamy=Chlamydomonas, C.Cru=Carteria crucifera, Cocco=Coccomyxa sp. LA000219, D.dis=Dictyostellium discoidum, Hydra=Hydra vulgaris, Octo=Octocorrallia, hypho=hyphochytrium catenoides, oom=oomycètes, Rhizo=Rhizofahus irregularis, Exaip=Exaiptasia pallida.* L'arbre du domaine eucaryote sur lequel sont apposés les noms des virus est tiré de (252).

des opisthocontes et des amœbozoaires respectivement. Nous avons ici détecté des inserts d'origine *Asfarviridae* dans les assemblages d'autres opistokonts : le cnidaire *Exaiptasia Pallida* et le fungus *Rhizofagus irregularis*. Nous en avons également détectés d'autres dans *Hydra vulgaris*, également un cnidaire (ils n'ont pas été intégrés au papier car ils ne possèdent aucun des 5 gènes « cores » utilisés pour le criblage des banques de données) (Figure 14). Plus important, des inserts *Asfarviridae*-like sont présents dans l'assemblage d'*Hyphochytrium catenoides* ainsi que dans plusieurs assemblages d'oomycètes. Or, ceux-ci font partie d'un troisième embranchement majeur d'eucaryotes, le super-groupe SAR (figure 13). Tout récemment, d'autres *Asfarviridae*-like capables de se répliquer dans des amibes ont été décrits et séquencés : Kaumoeba (253) et Pacmanvirus (254). La taille de leur génome est respectivement 1,95 et 2,19 fois supérieure à celle d'African Swine fever virus, un chiffre qui atteint 2,59 pour Faustovirus. De façon intéressante, ce

caractère de plus grande taille des virus amibiens comparativement à celle des virus infectant d'autres organismes est reminiscent du cas des *Mimiviridae*. Le fait que ces deux familles de NCLDVs possèdent à la fois la caractéristique de se répliquer dans des amibes et d'avoir un grand génome et celle d'avoir un large spectre d'hôte pourrait indiquer un lien entre ces caractéristiques. Notons que ces deux familles ont apparemment une autre catégorie d'hôte en commun : les cnidaires. Ceci sera développé par la suite.

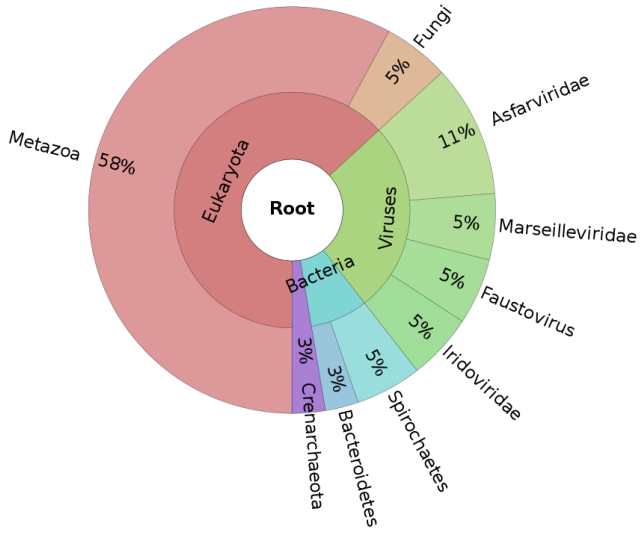


Figure 14: Répartition taxonomique des ORFs prédites à partir de 4 séquences génomiques (GL021887.1, GL022445.1, GL026077.1, GL030555.1) provenant de l'assemblage du génome de Hydra Vulgaris (GCA_000004095.1). Ces quatre séquences génomiques semblent présenter des inserts originaires d'Asfarviridae.

c.3 - *Les Phaeovirus*

Les *Phaeovirus* connus ont été nommés ainsi car ils infectent les phaeophycae (aussi appelées algues brunes qui appartiennent au super-groupe straménopile). *Ectocarpus siliculosus* virus et *feldmania sp* virus et *Feldmannia irregularis virus* sont les trois *Phaeovirus* connus et infectent deux algues brunes de l'ordre des ectocarpales. 5 autres *Phaeovirus* ont été identifiés sur la base de leur morphologie et de leur cycle réplicatif (8). Ici, nous en détectons également dans *Saccharina japonica*, la seule algue brune séquencée en dehors des ectocarpales (elle fait partie des laminariales, un clade ayant divergé avant les ectocarpales). Le séquençage de phaeophycea additionnels révèlera si la présence soit d'inserts, soit de provirus, est une caractéristique commune de cette famille d'algues. Il est d'ailleurs possible que le groupe viral contenant les *Phaeovirus* soit associé de manière plus générale aux stramenopiles, voir au clade regroupant les stramenopiles et les alveolata. En effet, dans notre article nous rapportons l'existence de plusieurs gènes apparentés aux *Phaeovirus* dans le stramenopile Nannochloropsis limnetica (eustigmatophyceae). Par ailleurs, nous trouvons dans la littérature un élément suggérant que les alveolata ont également été des proies de ces virus. J'en ai parlé précédemment : il s'agit du cas des protéines d'origine *Phaeovirus* qui ont remplacé les histones chez les dinoflagellés (alveolata). On peut imaginer que les dinoflagellés aient trouvé chez leur parasites *Phaeovirus* le moyen même de s'en prévenir. Effectivement, les

Phaeovirus isolés sont connus comme étant lysogéniques et encodent un mécanisme d'intégration dans le génome de leur hôte. Or, l'ADN des dinokaryons (les noyaux des dinoflagellés) est fortement et constamment compacté, dû à ce succédané d'histone *Phaeovirus*-like. Cette compaction constante pourrait être un frein à l'intégration des virus.

Avec les points c.1 à c.3, nous avons trois exemples où le spectre d'hôte d'une famille de virus semble plus vaste qu'initialement supposé. Ceci montre les limites qu'il y a à nommer les familles de virus à partir des hôtes des virus fondateurs de ces familles ; Ces réserves avaient également été émises dans le papier sur CeV. En effet, CeV et ses deux plus proches parents *Phaeocystis globosa* virus (PgV) et *Aureoccoccus anaphogeferens* virus (AaV) ont longtemps été classé au sein des *Phycodnaviridae*, un nom qui signifie virus à ADN infectant les algues. Cette famille regroupe plusieurs genres de virus NCLDVs infectant des algues. Certes, comme nous l'avons vu, CeV, PgV et AaV infectent aussi des algues, mais sont plus proche des *Mimiviridae* que des virus appartenant aux *Phycodnaviridae.*

c.4 - VLTF3

Les inserts viraux détectés dans ces génomes cellulaires sont autant de matériel génétique qui potentiellement pourrait être réutilisé par les eucaryotes pour leur fournir un avantage sélectif original.

Chez l'anémone *Exaiptasia pallida*, un seul gène core est détecté : le facteur de transcription VLTF3, *Asfarviridae*-like. Une analyse complémentaire suggère qu'il s'agit probablement du seul gène d'origine NCLDV présent dans l'assemblage du génome du cnidaire (i.e., sur la base des gènes de NCLDVs déjà connus). De façon intéressante, le gène VLTF3 est présent en de multiples copies chez *E.pallida*. 17 copies ont été dénombrées, présents sur 17 contigs différents. Les 17 protéines prédites se groupent en un clade lorsque nous construisons la phylogénie de VLTF3 (figure 15). Ceci peut être le résultat d'un évènement unique d'insertion virale, suivit de multiples duplications et divergence de la séquence originelle. Alternativement, l'explication d'évènements répétés d'intégration par des virus apparentés peut également satisfaire aux observations. Cependant, que 17 évènements d'intégration résultent en la conservation du même et unique gène VLTF3 semble peu probable. Ainsi, la première des deux hypothèses est favorisée. Une possibilité est qu'*E.pallida* ait domestiqué l'un de ces gènes pour son propre métabolisme. L'analyse des transcriptomes de l'anémone réalisée par Wolfowicz I. et al. (255) est compatible avec ce scénario. Pendant leur stade larvaire, certains coraux dont *E.pallida* établissent une symbiose avec des dinoflagellés (alveolata). Dans cette étude, les auteurs ont comparé les transcriptomes de larves symbiotiques avec ceux de larves aposymbiotiques, dans le but de mettre en évidence les facteurs génétiques de *E. pallida* potentiellement impliqués dans l'établissement de la symbiose. L'une des copies du gène VLTF3 (fusionné à un domaine zinc finger) (numéro d'accession de la protéine corespondante KXJ09396.1) est classée par les auteurs dans les gènes différentiellement exprimés. Plus précisément, ce gène est surexprimé d'un facteur 10 environ (logFC=3.3). Ceci est compatible avec une utilisation de ce gène pour l'établissement des

symbioses. Cependant, ce classement dans les gènes surexprimés pourrait être un artefact due a la possibilité que des transcrits provenant de plusieurs des 17 copies aient été tous alignés sur le même gène KXJ09396.1.

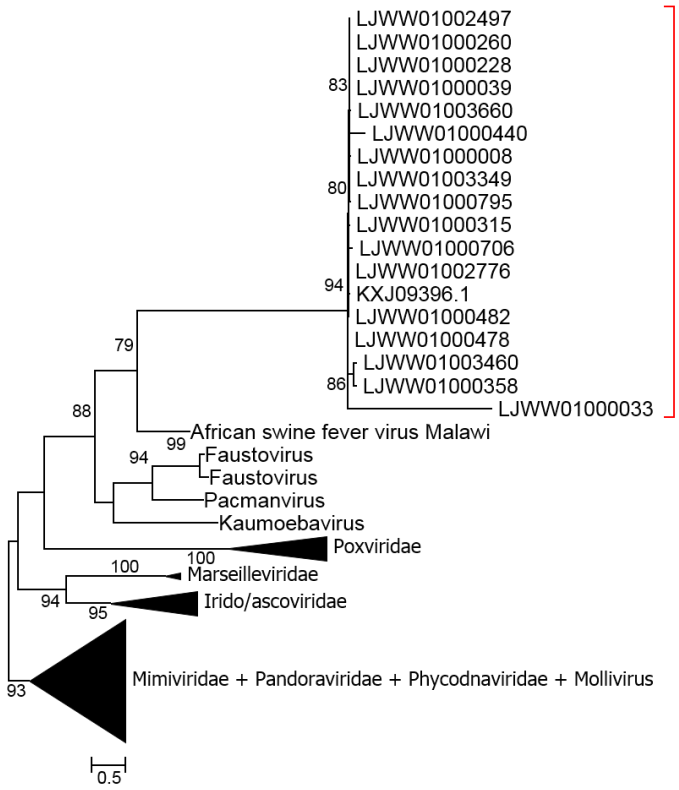


Figure 15: Arbre phylogénétique des protéines VLTF3 des NCLDVs et de *E.pallida.* Les séquences de *E.pallida* se regroupent en un clade qui émerge au sein des *Asfarviridae*.

c.5 - Les limitations de la méthode.

Nous avons détecté des séquences NCLDVs dans 48 assemblages de génomes eucaryotes. Il y a plusieurs possibilités quant à l'origine de ces séquences : il peut s'agir (i) de provirus (ii) de séquences issue de transferts latéraux d'un ou quelques gènes viraux dans le génome eucaryote (iii) de séquences contaminantes d'un *bona fide* virus séquencé en même temps que le génome de l'eucaryote étudié. Effectivement, cette dernière possibilité est liée à la qualité parfois médiocre des génomes assemblés exclusivement à partir de données générées par des méthodes de séquençage haut débit nouvelle génération. Un exemple notoire est la controverse du génome du tardigrade (256), un groupe d'animaux capables de survivre dans des environnements extrêmes. Lors de l'étude initiale du génome (257), il avait été avancé que le tardigrade aurait acquis une

proportion conséquente de ses gènes par HGT depuis des bactéries (17,5 %). Une seconde analyse doublée de séquençage additionnel à révéler que ~0,4 % seulement des gènes du tardigrade pouvaient être identifiés comme HGT de façon certaine (258), et que le reste était dû à des contaminations de bactéries. De telles contaminations peuvent *a fortiori* se produire dans le cas de virus, qui ont un stade intracellulaire et peuvent donc facilement être isolés en même temps que la cellule ciblée. Par exemple, les docteurs Aris Katzourakis et Amr Aswad, pionniers de la paléovirologie, montrent que tel a été le cas pour le génome du saumon. Effectivement, leur étude réalisée sur les génomes de 54 poissons téléostéens, montre d'une part que 19 d'entre eux possèdent des séquences virales endogènes (résultant d'un évènement ancien d'insertion), mais également qu'un contig de l'assemblage constitue en fait un génome complet d'un *bona fide* virus, représentant d'ailleurs une nouvelle lignée virale (259) (ce contig était présent dans la première version de l'assemblage et a ensuite été retiré des versions ultérieures ; la version obsolète pouvait néanmoins encore être trouvée ce qui a permis d'analyser ce contig en détail). Lors de notre crible global des génomes eucaryotes, nous avons voulu trancher entre les trois possibilités présentées précédemment. Nous avons donc analysé en détail l'environnement génomique des gènes cœurs NCLDVs détectés, analyse qui suggère qu'il s'agisse bien de séquences virales endogènes (Gallot-Lavallée et Blanc, 2017). Cependant, cette analyse n'a pu être réalisée que sur les génomes publiés ; la possibilité d'une contamination n'a donc pas été éliminée pour les autres génomes. Notons que dans les deux cas, insertion ou contamination, la prédiction hôte/virus qui peut être effectuée garde sa pertinence. D'autre part, une étude publiée par Béliveau et al (260) montre que le génome de la guêpe *Glypta fumiferanae* présente deux gènes coeurs NCLDVs que nous n'avons pas détectés lors de notre crible: une ADN polymérase et une D5 hélicase-primase. Concernant l'ADN polymérase, ceci s'explique par le fait que la protéine de la guêpe présente un meilleur score de similarité blast pour des protéines bactériennes que pour des protéines virales, nous ne l'avons donc pas conservé pour la suite de l'analyse. La D5 hélicase-primase présente par contre une meilleure affinité blast pour des protéines NCLDVs, et montre que certaines protéines cœurs NCLDVs présentent dans les génomes eucaryotes ont pu échapper à notre analyse.

c.6 - Une ouverture : quels hôtes utilisés pour isoler de nouveaux virus ?

Nous avons effectué un certains nombres de prédictions de couples hôtes/virus inconnus. Leur validité doit maintenant être évaluée grâce à des approches expérimentales. La plupart des organismes dans lesquels des séquences virales ont été trouvées sont cultivables en conditions de laboratoire. Ceci offre un terrain favorable pour tenter d'isoler des virus d'échantillons environnementaux en co-culture avec ces eucaryotes. En se basant sur les reconstructions phylogénétiques des séquences virales présentées dans le papier, l'hôte putatif pourra être choisi en fonction du clade de NCLDVs, connu ou inconnu, qui présente pour nous un intérêt scientifique particulier. Si cette approche est couronnée de succès, ceci pourrait nous aider à améliorer notre compréhension du monde NCLDV.

Discussion générale

A - Les NCLDVs, des acteurs de l'évolution

Nous avons vu en introduction que nombre d'études ont révélé que les virus sont des acteurs de l'évolution des cellules. Ces études s'étaient principalement concentrées sur les phages et les rétrovirus. Les travaux présentés ici montrent que ceci est également vrai pour les NCLDVs. Effectivement, nous avons mis en évidence qu'ils laissent des traces de leur passage dans les cellules sous la forme d'inserts génomiques. Or, les ORFs absentes du monde cellulaire abondent dans les génomes de ces grands virus. Ce matériel pourrait être recruté par les hôtes eucaryotes et représente donc une source potentielle d'innovation pour les cellules. Les NCLDVs infectent plusieurs des grandes branches eucaryotes, ils ont potentiellement co-évolué avec les eucaryotes depuis leur émergence, et donc également influencé leur évolution. Les protéines communes aux virus et aux cellules peuvent également être l'objet d'évolution médiée par les virus, si le gène viral est transféré dans des génomes cellulaires. La fusion DNAPolX-DNA ligase de CeV n'a pas – encore – été détectée chez des cellules. Par contre, nous avons vu qu'une DNApol provenant d'un insert NCLDV est adressée au nucléomorphe chez *B.natans*, et pourrait donc être responsable de la réplication du génome contenu dans ce compartiment cellulaire. D'autres cas pourraient être mis en lumière au fur et à mesure du séquençage de nouveaux génomes – viraux et cellulaires –, et d'analyses fonctionnelles.

Ensuite, on a vu que d'une part les virus peuvent servir de navettes de gènes entre cellules, et que d'autre part certains contiennent des AMGs hérités des cellules (voir introduction et résultats). Ils pourraient donc avoir un rôle dans la propagation des innovations génétiques cellulaires. Une étude a montré que les protéines virales acquises de leur hôte « convergent vers une architecture simplifiée » (261), c'est-à-dire sont plus courtes, possèdent moins de domaines et/ou des domaines de liaison plus courts, leurs gènes ont par ailleurs perdu leurs introns. Cette simplification pourrait faciliter le transfert subséquent vers de nouveaux hôtes. D'autre part, elle pourrait également faciliter la réutilisation de ces domaines, dans un cadre différent de celui de la protéine cellulaire originelle. Un exemple fameux de ce genre d'exaptation (262) au niveau moléculaire se produit lorsque les duplications de gènes sont suivies de néofonctionnalisations (263). De la même façon, certains domaines protéiques sont présents dans des protéines aux fonctions diverses. En effet, il a été suggéré que le « shuffling » de domaines constitue une caractéristique clé de l'évolution des protéomes (264, 265). Ainsi les virus (en plus de créer de nouveaux domaines), grâce à cette simplification, pourrait remettre à la disposition des cellules un matériel de base pouvant potentiellement être recruté pour une variété de fonctions.

Notons qu'une telle « remise à disposition » produite par les virus se produit également au niveau des cycles des nutriments et du carbone, lorsque le « viral shunt » permet le recyclage des

constituants de la cellule. Ceci n'agit pas directement sur l'évolution des cellules, mais indirectement en modifiant les niches écologiques du plancton hétérotrophe et phototrophe qui réutilise cette matière première.

B - NCLDVs acteurs d'établissement de symbioses/endosymbioses

Au cours de ce travail de thèse, j'ai pu accumuler plusieurs exemples de relations de symbiose dans lesquelles les virus joueraient, ou auraient joué un rôle de catalyseur. Effectivement, des gènes originaires de virus, ou même des virus proprement dit, pourraient être impliqués dans l'établissement ou la conservation de relations symbiotiques entre des organismes cellulaires. Ces exemples comprennent à la fois des symbioses « classiques », transitoires, qui seront abordés en 1/ et des endosymbioses pérennes que constituent l'acquisition de la mitochondrie ou l'acquisition de chloroplaste chez les eucaryotes. Elles seront traitées en 2/.

1 - Symbioses « transitoires »

Tableau 2 : Implication de virus dans des symbioses

Exosymbionte	**Endosymbionte/parasite**	**Mécanisme supposé**	**Références**
D. lanuginosum (plante)	*C. protuberata* (fungus)	Infection du fungus par le virus CThTV serait nécessaire pour que la symbiose fonctionne.	(80)
***E.Pallida* (anémone)**	**Dinoflagellés**	**Expression chez *E.pallida* de VLTF3, gène d'origine virale, lors de l'établissement de la symbiose.**	**(Cette thèse,** (255)

Deux cas de symbioses entre eucaryotes nécessitant un virus ou des gènes d'origine virale ont été collectés et sont résumés dans le tableau 2. Les aspects « mécanistiques » sont variés : le cas de l'anémone implique un transfert horizontal de gènes depuis un virus vers un organisme cellulaire. Concernant l'anémone, nous avons montré dans l'article « A Glimpse of Nucleo-Cytoplasmic Large DNA Virus Biodiversity through the Eukaryotic Genomics Window » (266) que le gène VLTF3 avait été acquis depuis un virus NCLDV ; des données transcriptomiques indiquent par ailleurs qu'il est surexprimé lors de l'établissement de la symbiose anémone/dinoflagellé (255). Ce gène n'a pas d'équivalent cellulaire : chez les virus, le facteur de transcription est impliqué dans l'initiation de la transcription des gènes tardifs. Il s'agit donc possiblement d'un gène viral qui aurait été domestiqué par l'anémone et pourrait aujourd'hui être impliqué dans l'établissement de la symbiose. Le cas de la symbiose entre la plante *D. lanuginosum*

et le fungus *C. protuberata*, de son côté, nécessite la présence d'un troisième partenaire, le virus CThTV qui infecte le champignon.

L'établissement de symbioses semble être un principe général de l'évolution des eucaryotes (267). Effectivement, ces derniers sont généralement associés à une communauté de microorganismes symbiotiques au sein du microbiome, principalement constituée de procaryotes mais également d'autres eucaryotes, voire de virus. Nous avons réuni 3 exemples où des virus jouent un rôle dans des relations symbiotiques eucaryote/eucaryote, édifiant encore les virus comme acteurs de l'évolution des eucaryotes. S'agit-il de cas anecdotiques ? Des études fonctionnelles d'organismes symbiotiques pourraient apporter des éléments de réponse. Lors de notre crible des insertions de NCLDVs dans les génomes eucaryotes, nous avons détecté des séquences d'origine virale dans les génomes de plusieurs organismes engagés dans des symbioses entre eucaryotes. *E.Pallida* a déjà été citée ; nous avons d'autre part le photobionte (symbiote photosynthétique d'un lichen) *Asterochloris glomerata* (268); *Chlorella vulgaris* (269), Coccomyxa sp. (270), *Hydra vulgaris* (271); le fungus mycorhizien (qui forme une association symbiotique avec des racines de plante) *Rhizofagus irregularis* (272)… Ceux-ci pourraient être de bons objets d'étude pour caractériser le rôle potentiel des virus dans des symbioses.

2 - Le cas des organelles

Tableau 3: Des gènes provenant de virus sont aujourd'hui des « gènes domestiques » dédiés au métabolisme de trois organelles cellulaires différents.

Organelle	**Virus**	**enzyme**	**fonction**	**références**
Mitochondrie	*Mimiviridae*	MutS7	Réparation de l'ADN + correction des erreurs d'appariement	(101, 273)
	Phages	Primase-hélicase	réplication	(104)
	Phages (part du génome de l'endosymbionte ?)	ARN polymérase	transcription	(105)
Chloroplaste	Phages	ARN polymérase	transcription	(110, 274)
Nucléomorphe	NCLDV	ADN polymérase	réplication	(249, 275)

Nous avons détecté dans le génome de l'algue *Bigelowiella natans* des séquences d'origine NCLDV, et notamment une ADN polymérase (275). Cette protéine est adressée au nucléomorphe de l'algue (249) et pourrait donc être impliquée dans la réplication du génome de l'organelle (249). Ceci – une protéine d'origine virale participant au métabolisme d'une organelle – n'est pas un cas isolé (tableau 3).

En effet, la primase-hélicase réplicative mitochondriale codée par le noyau présente une affinité phylogénétique avec celle de phages T7 (105), ce qui suggère qu'elle est également héritée de virus. Ce pourrait également être le cas pour l'ARN polymérase qui transcrit les gènes mitochondriaux. Celle-ci est également codée par le noyau et les homologues les plus proches que nous connaissons appartiennent à des phages de la famille T-odd (105).

Enfin, chez les plantes, le noyau encode deux copies de ces ARN polymérases originaires de phages. Elles proviennent d'une duplication récente. Le produit de l'une des copies est adressé à la mitochondrie, comme c'est le cas chez tous les eucaryotes (à l'exception des jakobids). Le produit de la seconde copie est adressé au chloroplaste, suggérant que celui-ci prend part à la transcription des gènes chloroplastiques (274).

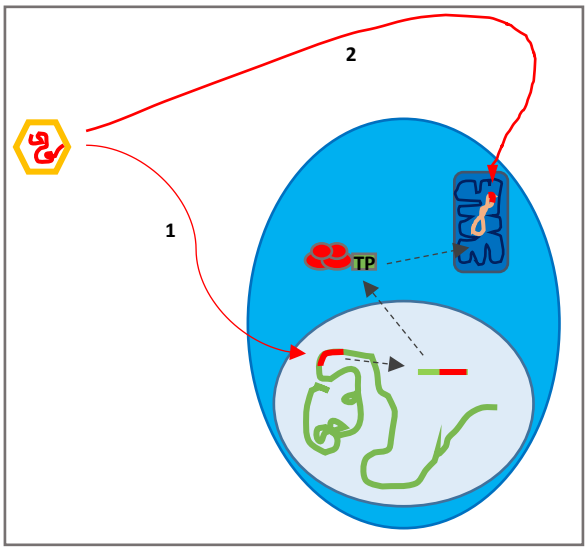


Figure 16: 1 - Des gènes viraux ont été transférés au noyau. Après transcription puis traduction, ils sont ensuite adressés à la mitochondrie (ou à un autre organelle). C'est ce qui est supposé pour la primase-hélicase (mitochondrie), la RNA polymérase (chloroplaste) ou encore la DNA polymérase (nucléomorphe). TP: « *transit peptid* », peptide signal permettant l'adressage au compartiment cellulaire cible. 2- Des gènes viraux sont transférés directement au génome mitochondrial. C'est ce qui est supposé pour MutS7.

Ainsi, des gènes nucléaires provenant de virus sont aujourd'hui des « gènes domestiques » dédiés au métabolisme de trois organelles cellulaires différentes (Figure 16, « voie » 1). Ce pourrait être plus qu'anecdotique. Effectivement, conceptuellement, les virus pourraient constituer une source très adaptée pour remplacer un homologue organellaire (voir Box 1, sur l'évolution du protéome mitochondrial). Nous avons vu que certains virus ciblent la mitochondrie et le chloroplaste parce qu'ils ont besoin d'énergie pour effectuer leur cycle réplicatif et qu'eux même n'encodent pas le matériel nécessaire à cette production. Nous avons parlé du cas, par exemple, de la protéine de photosystème LHC codé par CeV et des prasinovirus. Ces protéines présentent le type de « peptide signal » pour l'adressage au chloroplaste de leur hôte algue (primaire, une seule membrane, pour les prasinophytes; secondaire, donc un compartiment cellulaire à traverser en plus, pour les haptophytes). Ainsi, certains gènes de virus sont « prêts à l'utilisation », ce qui pourrait faciliter leur maintien une fois intégrés au noyau, comme nous avons vu qu'il peut arriver.

Box 1 : Evolution du protéome mitochondrial

Les mitochondries sont issues de l'endosymbiose d'un ancêtre des alpha-protéobactéries actuelles (276). Bien que cette alpha-protéobactérie ait vraisemblablement possédé au moins 1000 gènes (277), les génomes mitochondriaux séquencés n'encodent que 3 à 66 protéines (278). Il apparait donc que la majorité des gènes ait été éliminée du génome de l'endosymbionte ancestral au cours de la symbiogénèse (fusion de deux symbiontes créant une nouvelle entité). Parmi ces gènes, ceux qui devaient ne pas être essentiels dans un contexte de vie intracellulaire ont été perdus ; Toutefois, le protéome mitochondrial (PM) est beaucoup plus important que les soixantes protéines au plus qui sont codées par l'ADN mitochondrial. En effet, il a par exemple été estimé qu'au moins 1,500 protéines contribuent à la maintenance et au fonctionnement des mitochondries de mammifères (277). Ce PM est aujourd'hui majoritairement codé par le noyau. Une part de ce protéome affiche une claire ascendance α-proteobactérienne: un mécanisme pouvant expliquer cela est qu'au cours de la domestication de l'endosymbionte-mitochondrie, la désintégration par les vacuoles de mitochondries ou de bactéries apparentées ait libéré leur ADN dans la cellule qui aurait « transformé » le génome nucléaire (279). Cependant, un résultat surprenant des analyses phylogénétiques du PM est que seul 10-20% des protéines présentent une affinité phylogénétique avec des α-proteobactéries (278). Une seconde partie du PM possède des homologies détectables avec d'autres eucaryotes uniquement et aurait donc émergé spécifiquement au sein des eucaryotes (278). Une autre fraction significative consiste en des protéines présentant des homologies avec des protéines procaryotes mais sans affinités pour une lignée spécifique (278). Enfin, les virus semblent également être une source pour le PM. Ceci a été mis en évidence lorsque le professeur émérite M.W. Gray et professeur assistant T. Shutt. ont révélé que la primase-hélicase réplicative mitochondriale ainsi que l'ARN polymérase qui transcrit les gènes mitochondriaux, toutes deux codées par le noyau, présentent des affinités phylogénétiques avec celles de bactériophages.

Notons que le protéome chloroplastique présente une histoire similaire : l'acquisition originelle du chloroplaste résulte également d'une symbiogénèse entre une cellule (un *bona fide* eucaryote cette fois) et une bactérie (une cyanobactérie photosynthétique). Le protéome chloroplastique, à l'instar du protéome mitochondrial, est aujourd'hui en partie codé par le

Je n'ai pas encore évoqué les cas des gènes qui auraient été transmis directement au génome mitochondrial postérieurement à sa « domestication ». Les exemples connus sont rares. Effectivement, peu de gènes codés par les mitochondries seraient spécifiquement « apparus » dans des lignées d'eucaryotes en particulier. Les différences observées seraient plutôt dûes à des pertes différentielles à partir du génome mitochondrial de l'ancêtre des eucaryotes LECA (last eukaryotic common ancestor) (277). En effet, les gènes mitochondriaux de chaque lignée eucaryote constituent un sous-set des gènes codés par le génome des mitochondries des jakobids, les

eucaryotes qui auraient divergé les premiers après LECA (277). MutS7, exclusif aux mitochondries d'octocoralliaires, est l'une des rares exceptions décrite (100). Nous avons vu que l'hypothèse qu'il ait été transféré par un virus a été soulevée (101) (Figure 16, « voie » 2). Si d'autres cas similaires existent, ils pourraient être détectés en criblant les génomes d'organelles séquencés.

C - Les cnidaires, un autre lieu de « melting pot » : *Mimiviridae*, *Asfarviridae*, bactéries, symbiontes eucaryotes

Durant la dernière décénie, les amibes du genre *Acanthamoeba* ont été les hôtes les plus prolifiques pour isoler de nouveaux virus géants. Il y a tout d'abord une raison historique à cela : le tout premier virus géant Mimivirus a effectivement été découvert à partir de l'amibe, qui est probablement son hôte naturel. C'est donc l'amibe qui a été utilisée par la suite pour tester la présence d'autres virus géants dans des échantillons environnementaux. Ceci a permis de mettre à jour deux nouvelles familles de virus géants (*Marseilleviridae* et *Pandoraviridae*) ; deux virus appartenant potentiellement à deux nouvelles familles (Pithovirus et Mollivirus), ainsi que d'isoler Faustovirus apparenté aux *Asfarviridae* (il faut cependant nuancer pour ce dernier : effectivement Faustovirus à été isolé dans une amibe du genre *Vermamoeba*). Ces virus sont tous capable de se répliquer dans des amibes. Les virus géants ne sont pas les seuls parasites intracellulaires amibiens. En effet, des bactéries de taxa variés font également partie du microbiome amibien ainsi que leurs phages associés (280). Or, de multiples évènements d'échanges de gènes entre ces micro-organismes, ou bien impliquant les amibes, ont été décrits (280, 281). D'autre part, l'ADN des microorganismes digéré par les amibes est libéré dans le cytoplasme suivant leur phagocytose. Cet ADN peut également être intégré aux génomes de l'amibe ou des microoragnismes associés (280). Ceci a mené à qualifier l'amibe de « melting pot » de gènes ou encore « melting pot » de l'évolution (117, 183).

Plusieurs animaux du groupe des cnidaires représentent un autre foyer potentiel de « melting pot » favorisant les échanges génétiques entre les NCLDVs et d'autres organismes appartenant à leur holobionte. Effectivement, plusieurs indices suggèrent que certains coraux, des hydres ou encore des anémones de mer, sont des hôtes potentiels des NCLDVs (voir tableau 4 pour des indices allant dans ce sens), entre autres virus (250). D'autres parts, plusieurs cnidaires font des symbioses avec des micro-algues, comme des Chlorelles (algues vertes) pour Hydra ou des dinoflagellés (alveolata) pour les coraux (269, 282). Or, ces symbiontes sont également des proies (ou proies supposées) de NCLDVs. Enfin, des bactéries de taxa variées viennent compléter cet holobionte (283–285).

Tableau 4 : Indices sur la possibilité que des Cnidaires et des Dinoflagellés soient des proies de NCLDVs

NCLDVs impliqués	**Organisme hôte supposé**	**Type d'indices**	**références**
Mimiviridae	Octocoralliaires	Echange du gène MutS7	(101, 273)
	Hydra vulgaris	Inserts viraux	Gallot-Lavallée et Blanc, 2017
	Hydra Magnipillata	Inserts viraux	(242, 286)
	Symbiodinium (symbionte des coraux)	Transcriptomique	(250)
	Dinoflagellés (symbionte)	Cooccurrence metatranscriptomique	(287)
Asfarviridae	E.pallida	Inserts viraux	Gallot-Lavallée et Blanc, 2017
	Hydra vulgaris	Inserts viraux	Cette thèse
	Heterocapsa circularisquama	Virus « isolé »	(288)
Phaeovirus	Dinoflagellés (symbionte des coraux)	Présence d'un gène *Phaeovirus* (histone-like)	(108)

Y a-t-il des indices que des transferts de matériel génétiques se produisent entres ces différents acteurs ? Nous nous sommes ici concentrés sur les échanges potentiels impliquant les virus. Nous avons vu et décrit dans Gallot-Lavallée et Blanc (2017) que des transferts virus vers cnidaires ont eu lieu. Un autre indice réunissant deux familles de NCLDVs et les coraux est le cas du gène MutS7. Effectivement, les protéines MutS7 des *Mimiviridae*, de HcDNAV (*Asfarviridae* supposé) et des octocoralliaires (cnidaires), forment un clade monophylétique (229). Ceci suggère que ces différents acteurs se sont échangés ce gène (229). D'autre part, des échanges virus/virus sont probables. Prenons par exemple les *Asfarviridae*-like infectant les amibes. Parmi les protéines de Faustovirus, Kaumoebavirus et Pacmanvirus dont le meilleur score d'alignement blast contre la base de données nr (une base de données regroupant les séquences de l'entièreté des protéines connues) est contre une protéine virale, 20, 30 et 19 % respectivement ont un meilleur score d'alignement avec une protéine de *Mimiviridae*, ce qui représente à chaque fois la seconde famille après les *Asfarviridae*. Rien ne nous permet toutefois d'affirmer que ces échanges ont eu lieu lors d'une infection d'un cnidaire. Nous avons vu en effet que ces deux familles ont également l'amibe en hôte commun. Je mentionne cependant cela car l'absence d'échanges entre ces deux familles aurait été un argument allant contre une qualité de lieu de « melting-pot » pour les cnidaires. Une

étude génomique des différents organismes constituants ces holobiontes variés est nécessaire pour apporter du crédit à cette hypothèse et en dégager les potentielles implications évolutives.

Conclusion générale

Avec mon travail de thèse, j'ai pu contribuer à la caractérisation de la famille des *Mimiviridae*, à la compréhension de l'évolution de leurs génomes, ainsi qu'à l'identification de fonctions nouvelles chez ces virus. Les « *Mesomimivirinae* » sont identifiés chez les algues. Nous n'avons pas pu expliquer cette spécificité. Afin d'apporter des éléments de réponses, il faudrait être en mesure de caractériser le métabolisme de ces virus lors d'une infection. Pour ceci, des pathosystèmes pourraient par exemple être développés, en choisissant des algues modèles pour lesquels des outils génétiques sont disponibles.

J'ai pu également contribuer à nos connaissances sur les échanges génétiques depuis les virus géants vers l'hôte. En effet, nous ne connaissions précédemment que le cas des *Phaeovirus* qui semblait anecdotique. Nous savons à présent que plusieurs familles de virus géants laissent des traces chez leur hôte sous la forme d'inserts génomiques. Ces traces peuvent révéler des aspects inédits de l'écologie et l'évolution des virus. Elles constituent par ailleurs des indices quant à la qualité d'hôte des propriétaires des génomes dans lesquelles elles sont trouvées. Ainsi, utiliser ces eucaryotes comme appâts pour isoler de nouvelles familles de virus pourrait être une stratégie fructueuse. Enfin, ces inserts génomiques d'origine virale peuvent parfois être recrutés par l'eucaryote récipiendaire. J'ai donc également contribué à la compréhension de cet aspect particulier de l'évolution des cellules : celui qui dépend de leurs interactions avec le monde des virus.

Matériel et Méthodes

La majorité des méthodes que j'ai utilisées pendant ma thèse ont été décrites dans les articles. Cette section comprendra donc seulement les procédures appliquées pour les résultats supplémentaires.

A - CeV et les *Mimiviridae*

Reconstruction du profil de gains/pertes de gènes des *Mimiviridae*

Les protéomes de 8 *Mimiviridae* ont été utilisés. Ces *Mimiviridae* appartiennent à chacun des différents clades connus (à l'époque). Les 3 « *Mesomimivirinae* » CeV, PgV et AaV d'une part, CroV du genre *Cafeteriavirus* ensuite, et les représentants des 3 clades A, B et C du genre *Mimivirus :* Megavirus Chiliensis, Moumouvirus, *Acanthamoeba polyphaga* Mimivirus et *Acanthamoeba polyphaga* Mamavirus. Le regroupement en familles de protéines de ces protéomes a été réalisé en utilisant le logiciel OrthoMCL (178), avec un index d'inflation (paramêtre qui régule la « granularités des clusters) de 1,5. Dans l'article présentant OrthoMCL (178), cette valeur a été testée et apparait comme un bon compromis entre la sensibilité et la sélectivité du regroupement en groupe de protéine. A partir de cette classification des protéines homologues, le schéma suivant était appliqué : si une protéine était présente chez au moins un « *Mesomimivirinae* » et au moins un « *Megavirinae* » (i.e., groupe comprenant CroV, et les virus du genre *Mimivirus*), alors ce gène était considéré comme présent dans le génome de l'ancêtre commun des *Mimiviridae*. Ce principe était appliqué à chaque nœud de l'arbre phylogénétique des *Mimiviridae* (i.e., chaque nœud représente l'ancêtre commun des virus descendants). Ensuite, à partir des protéomes reconstruits de chacun des nœuds de l'arbre, les gains et pertes de gènes étaient inférés comme suit : si un gène était présent chez l'ancêtre, et absent dans l'un de ses nœuds descendants, alors il était comptabilisé comme une perte dans la branche menant à ce descendant (i.e., chaque branche représente l'évolution de la lignée entre deux nœuds). A l'inverse, si un gène était absent d'un ancêtre mais présent dans l'un de ses nœuds descendants, alors il était comptabilisé comme un gain dans la branche menant à ce nœud.

Phylogénie des ATCs

Les protéines virales (i.e., celle de CeV, numéro d'accession : YP_009173409 et celle d'*Ostreococcus lucimarus* virus 7, numéro d'accession : YP_009173244.1) ont été alignées contre la base de donnée nr au NCBI grâce à l'outil Blastp (289). Pour chacune des deux ATCs virales, les protéines présentant les meilleurs scores d'alignement ont été sélectionnées pour la suite. D'autre part, des protéines représentantes de chacune des branches de l'arbre des ATC décrit dans (192) ont été ajoutées à la sélection précédante, ainsi que les meilleurs score blastp contre nr de chacune de ces protéines. Les séquences de la sélection finale ont été alignées avec le programme Muscle (290), les sites de l'alignement présentant moins de 30% de *gaps* conservés. Le logiciel

MrBAYES (291) a été utilisé pour la reconstruction phylogénétique avec le modèle de substitution « mixed ». L'analyse a été stoppée après que l'indice de convergence (« average standard deviation of split frequencies ») soit passé en dessous de 0,05.

Phylogénie des AsnRSs et AspRSs

Ont été sélectionnées pour la reconstruction : les protéines de la base de données nr au NCBI présentant les meilleurs scores d'alignement BLAST contre la protéine de CeV tout d'abord ; afin de placer l'histoire de la protéine virale au sein de l'histoire globale de cette protéine, les LysRS, AspRS et AsnRS de représentants de chacun des trois domaines archaea, bactérien et eucaryote ont été récupérées sur la base de données de protéines du NCBI d'autre part. Les protéines ont été alignées avec le logiciel PSICOFFEE (292), l'alignement a été validé visuellement, puis les sites de l'alignement présentant une proportion de « gaps » (en anglais les trous dans l'alignement) inférieure à 15% ont été sélectionnés. Le logiciel MrBAYES (291) a été utilisé pour la reconstruction phylogénétique avec le modèle de substitution « mixed ». L'analyse a été stoppée après que l'indice de convergence (« average standard deviation of split frequencies ») soit passé en dessous de 0,05.

B - Séquences de NCLDVs et virophages dans les génomes eucaryotes

- Pas de méthodes supplémentaires pour le sous-chapitre *Bigelowiella natans* – un génome analysé en détail.

- Un crible global des eucaryotes :

Taxonomie des ORFs présentent sur 4 contigs de l'assemblage de *Hydra Vulgaris*

Les contigs GL021887.1, GL022445.1, GL026077.1, GL030555.1 provenant de l'assemblage du génome de Hydra Vulgaris (GCA_000004095.1) ont été analysés grâce à l'outil seqtinizer, développé au laboratoire IGS. Pour inférer la taxonomie liée a chaque ORF prédite, Seqtinizer se base sur ses meilleurs scores d'alignement blast contre la base de données nr. Cet outil génère également le graphique comme présenté en figure 15.

Reconstruction phylogénétique de VLTF3

Les séquences des protéines VLTF3 des NCLDVs ont été récupérées sur la base de données de familles de gènes orthologues des NCLDVs (*NCVOG database* pour *NCLDV clusters of orthologous gene database)* (293). Les séquences en acides-aminés des VLTF3 de *E.pallida*, ont été extraites grâce à l'outil Exonerate (294). Les séquences des deux lots ont été alignées avec MCOFFEE (295). La reconstruction phylogénétique a été effectuée grâce au programme PhyML, modèle LG, à partir des sites contenants moins de 30% de *gaps*.

Bibliographie

1. Claverie J-M, Abergel C. 2016. Les virus géants - État des connaissances, énigmes, controverses et perspectives. médecine/sciences 32:1087–1096.

2. Lecoq H. 2001. [Discovery of the first virus, the tobacco mosaic virus: 1892 or 1898?]. C R Acad Sci III 324:929–933.

3. Norrby E. 2008. Nobel Prizes and the emerging virus concept. Arch Virol 153:1109–1123.

4. Bândea CI. 1983. A new theory on the origin and the nature of viruses. J Theor Biol 105:591–602.

5. Claverie J-M. 2006. Viruses take center stage in cellular evolution. Genome Biol 7:110.

6. Forterre P. 2011. Manipulation of cellular syntheses and the nature of viruses: The virocell concept. Comptes Rendus Chim 14:392–399.

7. Forterre P. 2006. The origin of viruses and their possible roles in major evolutionary transitions. Virus Res 117:5–16.

8. Stevens K, Weynberg K, Bellas C, Brown S, Brownlee C, Brown MT, Schroeder DC. 2014. A Novel Evolutionary Strategy Revealed in the Phaeoviruses. PLOS ONE 9:e86040.

9. Nasir A, Forterre P, Kim KM, Caetano-Anollés G. 2014. The distribution and impact of viral lineages in domains of life. Front Microbiol 5:194.

10. Krupovic M, Koonin EV. 2017. Multiple origins of viral capsid proteins from cellular ancestors. Proc Natl Acad Sci U S A 114:E2401–E2410.

11. Lynch M. 2010. Evolution of the mutation rate. Trends Genet TIG 26:345–352.

12. Trus BL, Cheng N, Newcomb WW, Homa FL, Brown JC, Steven AC. 2004. Structure and polymorphism of the UL6 portal protein of herpes simplex virus type 1. J Virol 78:12668–12671.

13. Benson SD, Bamford JK, Bamford DH, Burnett RM. 1999. Viral evolution revealed by bacteriophage PRD1 and human adenovirus coat protein structures. Cell 98:825–833.

14. Koonin EV, Senkevich TG, Dolja VV. 2006. The ancient Virus World and evolution of cells. Biol Direct 1:29.

15. Nasir A, Kim KM, Caetano-Anolles G. 2012. Giant viruses coexisted with the cellular ancestors and represent a distinct supergroup along with superkingdoms Archaea, Bacteria and Eukarya. BMC Evol Biol 12:156.

16. Prangishvili D, Forterre P, Garrett RA. 2006. Viruses of the Archaea: a unifying view. Nat Rev Microbiol 4:837–848.

17. Forterre P, Prangishvili D. 2013. The major role of viruses in cellular evolution: facts and hypotheses. Curr Opin Virol 3:558–565.

18. Nasir A, Sun F-J, Kim KM, Caetano-Anollés G. 2015. Untangling the origin of viruses and their impact on cellular evolution: Origin and evolution of viruses. Ann N Y Acad Sci 1341:61–74.

19. Nasir A, Caetano-Anollés G. 2015. A phylogenomic data-driven exploration of viral origins and evolution. Sci Adv 1:e1500527.

20. Jalasvuori M, Bamford JKH. 2008. Structural co-evolution of viruses and cells in the primordial world. Orig Life Evol Biosphere J Int Soc Study Orig Life 38:165–181.

21. Forterre, P (1992) New hypotheses about the origins of viruses, prokaryotes and eukaryotes. "Frontiers of Life": 221-234. Editions Frontières, Gif-sur-Yvette –France

22. Yang S, Doolittle RF, Bourne PE. 2005. Phylogeny determined by protein domain content. Proc Natl Acad Sci U S A 102:373–378.

23. Harish A, Abroi A, Gough J, Kurland C. 2016. Did Viruses Evolve As a Distinct Supergroup from Common Ancestors of Cells? Genome Biol Evol 8:2474–2481.

24. Hendrix RW, Lawrence JG, Hatfull GF, Casjens S. 2000. The origins and ongoing evolution of viruses. Trends Microbiol 8:504–508.

25. Prangishvili D, Stedman K, Zillig W. 2001. Viruses of the extremely thermophilic archaeon Sulfolobus. Trends Microbiol 9:39–43.

26. Suttle CA. 2005. Viruses in the sea. Nature 437:356–361.

27. Weinbauer MG. 2004. Ecology of prokaryotic viruses. FEMS Microbiol Rev 28:127–181.

28. Falkowski PG, Fenchel T, Delong EF. 2008. The microbial engines that drive Earth's biogeochemical cycles. Science 320:1034–1039.

29. Thingstad T., Lignell R. 1997. Theoretical models for the control of bacterial growth rate, abundance, diversity and carbon demand. Aquat Microb Ecol 13:19–27.

30. Thingstad TF. 2000. Elements of a theory for the mechanisms controlling abundance, diversity, and biogeochemical role of lytic bacterial viruses in aquatic systems. Limnol Oceanogr 45:1320–1328.

31. Wilson WH (William H), Tarran GA, Schroeder D, Cox MJ, Oke J, Malin G. 2002. Isolation of viruses responsible for the demise of an Emiliania huxleyi bloom in the English Channel. J Mar Biol Assoc UK Vol.82:369–377.

32. Nagasaki K, Yamaguchi M. 1997. Isolation of a virus infectious to the harmful bloom causing microalga Heterosigma akashiwo (Raphidophyceae). Aquat Microb Ecol 13:135–140.

33. Keppler CJ, Hoguet J, Smith K, Ringwood AH, Lewitus AJ. 2005. Sublethal effects of the toxic alga Heterosigma akashiwo on the southeastern oyster (Crassostrea virginica). Harmful Algae 4:275–285.

34. Black EA, Whyth J n. C, Bagshaw JW, Ginther NG. 1991. The effects of Heterosigma akashiwo on juvenile Oncorhynchus tshawytscha and its implications for fish culture. J Appl Ichthyol 7:168–175.

35. Wilhelm SW, Suttle CA. 1999. Viruses and Nutrient Cycles in the SeaViruses play critical roles in the structure and function of aquatic food webs. BioScience 49:781–788.

36. Weitz JS, Wilhelm SW. 2012. Ocean viruses and their effects on microbial communities and biogeochemical cycles. F1000 Biol Rep 4:17.

37. Shelford E, Middelboe M, Møller E, Suttle C. 2012. Virus-driven nitrogen cycling enhances phytoplankton growth. Aquat Microb Ecol 66:41–46.

38. Poorvin L, Rinta-Kanto JM, Hutchins DA, Wilhelm SW. 2004. Viral release of iron and its bioavailability to marine plankton. Limnol Oceanogr 49:1734–1741.

39. Bonnain C, Breitbart M, Buck KN. 2016. The Ferrojan Horse Hypothesis: Iron-Virus Interactions in the Ocean. Front Mar Sci 3.

40. Wilson W, Mann N. 1997. Lysogenic and lytic viral production in marine microbial communities. Aquat Microb Ecol 13:95–100.

41. McDaniel L, Houchin LA, Williamson SJ, Paul JH. 2002. Lysogeny in marine Synechococcus. Nature 415:496.

42. De Smet J, Zimmermann M, Kogadeeva M, Ceyssens P-J, Vermaelen W, Blasdel B, Bin Jang H, Sauer U, Lavigne R. 2016. High coverage metabolomics analysis reveals phage-specific alterations to Pseudomonas aeruginosa physiology during infection. ISME J 10:1823–1835.

43. Ankrah NYD, May AL, Middleton JL, Jones DR, Hadden MK, Gooding JR, LeCleir GR, Wilhelm SW, Campagna SR, Buchan A. 2014. Phage infection of an environmentally relevant marine bacterium alters host metabolism and lysate composition. ISME J 8:1089–1100.

44. Rosenwasser S, Mausz MA, Schatz D, Sheyn U, Malitsky S, Aharoni A, Weinstock E, Tzfadia O, Ben-Dor S, Feldmesser E, Pohnert G, Vardi A. 2014. Rewiring Host Lipid Metabolism by Large Viruses Determines the Fate of Emiliania huxleyi, a Bloom-Forming Alga in the Ocean. Plant Cell 26:2689–2707.

45. Goodwin CM, Xu S, Munger J. 2015. Stealing the Keys to the Kitchen: Viral Manipulation of the Host Cell Metabolic Network. Trends Microbiol 23:789–798.

46. Diamond DL, Syder AJ, Jacobs JM, Sorensen CM, Walters K-A, Proll SC, McDermott JE, Gritsenko MA, Zhang Q, Zhao R, Metz TO, Ii DGC, Waters KM, Smith RD, Rice CM, Katze MG. 2010. Temporal Proteome and Lipidome Profiles Reveal Hepatitis C Virus-Associated Reprogramming of Hepatocellular Metabolism and Bioenergetics. PLOS Pathog 6:e1000719.

47. Samsa MM, Mondotte JA, Iglesias NG, Assunção-Miranda I, Barbosa-Lima G, Poian ATD, Bozza PT, Gamarnik AV. 2009. Dengue Virus Capsid Protein Usurps Lipid Droplets for Viral Particle Formation. PLOS Pathog 5:e1000632.

48. Reinero A, Beachy RN. 1989. Reduced Photosystem II Activity and Accumulation of Viral Coat Protein in Chloroplasts of Leaves Infected with Tobacco Mosaic Virus. Plant Physiol 89:111–116.

49. Breitbart M. 2012. Marine viruses: truth or dare. Annu Rev Mar Sci 4:425–448.

50. Gao E-B, Huang Y, Ning D. 2016. Metabolic Genes within Cyanophage Genomes: Implications for Diversity and Evolution. Genes 7.

51. Hurwitz BL, U'Ren JM. 2016. Viral metabolic reprogramming in marine ecosystems. Curr Opin Microbiol 31:161–168.

52. Rosenwasser S, Ziv C, Creveld SG van, Vardi A. 2016. Virocell Metabolism: Metabolic Innovations During Host-Virus Interactions in the Ocean. Trends Microbiol 24:821–832.

53. Mann NH, Cook A, Millard A, Bailey S, Clokie M. 2003. Marine ecosystems: bacterial photosynthesis genes in a virus. Nature 424:741.

54. Lindell D, Jaffe JD, Johnson ZI, Church GM, Chisholm SW. 2005. Photosynthesis genes in marine viruses yield proteins during host infection. Nature 438:86–89.

55. Millard AD, Zwirglmaier K, Downey MJ, Mann NH, Scanlan DJ. 2009. Comparative genomics of marine cyanomyoviruses reveals the widespread occurrence of Synechococcus host genes localized to a hyperplastic region: implications for mechanisms of cyanophage evolution. Environ Microbiol 11:2370–2387.

56. Dammeyer T, Bagby SC, Sullivan MB, Chisholm SW, Frankenberg-Dinkel N. 2008. Efficient Phage-Mediated Pigment Biosynthesis in Oceanic Cyanobacteria. Curr Biol 18:442–448.

57. Millard A, Clokie MRJ, Shub DA, Mann NH. 2004. Genetic organization of the psbAD region in phages infecting marine Synechococcus strains. Proc Natl Acad Sci U S A 101:11007–11012.

58. Martiny AC, Huang Y, Li W. 2009. Occurrence of phosphate acquisition genes in Prochlorococcus cells from different ocean regions. Environ Microbiol 11:1340–1347.

59. Kelly L, Ding H, Huang KH, Osburne MS, Chisholm SW. 2013. Genetic diversity in cultured and wild marine cyanomyoviruses reveals phosphorus stress as a strong selective agent. ISME J 7:1827–1841.

60. Anantharaman K, Duhaime MB, Breier JA, Wendt KA, Toner BM, Dick GJ. 2014. Sulfur oxidation genes in diverse deep-sea viruses. Science 344:757–760.

61. Sullivan MB, Coleman ML, Weigele P, Rohwer F, Chisholm SW. 2005. Three Prochlorococcus cyanophage genomes: signature features and ecological interpretations. PLoS Biol 3:e144.

62. Huang S, Zhang S, Jiao N, Chen F. 2015. Comparative Genomic and Phylogenomic Analyses Reveal a Conserved Core Genome Shared by Estuarine and Oceanic Cyanopodoviruses. PloS One 10:e0142962.

63. Zeng Q, Chisholm SW. 2012. Marine Viruses Exploit Their Host's Two-Component Regulatory System in Response to Resource Limitation. Curr Biol 22:124–128.

64. Thompson LR, Zeng Q, Kelly L, Huang KH, Singer AU, Stubbe J, Chisholm SW. 2011. Phage auxiliary metabolic genes and the redirection of cyanobacterial host carbon metabolism. Proc Natl Acad Sci 108:E757–E764.

65. Sullivan MB, Lindell D, Lee JA, Thompson LR, Bielawski JP, Chisholm SW. 2006. Prevalence and evolution of core photosystem II genes in marine cyanobacterial viruses and their hosts. PLoS Biol 4:e234.

66. Lindell D, Sullivan MB, Johnson ZI, Tolonen AC, Rohwer F, Chisholm SW. 2004. Transfer of photosynthesis genes to and from Prochlorococcus viruses. Proc Natl Acad Sci U S A 101:11013–11018.

67. Hurwitz BL, Brum JR, Sullivan MB. 2015. Depth-stratified functional and taxonomic niche specialization in the 'core'and 'flexible'Pacific Ocean Virome. ISME J 9:472–484.

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

69. Sanjuán R, Nebot MR, Chirico N, Mansky LM, Belshaw R. 2010. Viral Mutation Rates. J Virol 84:9733–9748.

70. Caroll L. 1872. Through the looking glass And what Alice found there. MacMillan and Co. London.

71. VanValen L. 1973. A new evolutionary law. Evolutionary theory.

72. Caroll L. Alice's Adventures in Wonderland. MacMillan and Co. London.

73. Frada M, Probert I, Allen MJ, Wilson WH, Vargas C de. 2008. The "Cheshire Cat" escape strategy of the coccolithophore Emiliania huxleyi in response to viral infection. Proc Natl Acad Sci 105:15944–15949.

74. Yau S, Hemon C, Derelle E, Moreau H, Piganeau G, Grimsley N. 2016. A Viral Immunity Chromosome in the Marine Picoeukaryote, Ostreococcus tauri. PLoS Pathog 12:e1005965.

75. Grimsley N, Yau S, Piganeau G, Moreau H. 2015. Typical Features of Genomes in the Mamiellophyceae, p. 107–127. *In* Marine Protists. Springer, Tokyo.

76. Penadés JR, Christie GE. 2015. The Phage-Inducible Chromosomal Islands: A Family of Highly Evolved Molecular Parasites. Annu Rev Virol 2:181–201.

77. Avrani S, Wurtzel O, Sharon I, Sorek R, Lindell D. 2011. Genomic island variability facilitates Prochlorococcus-virus coexistence. Nature 474:604–608.

78. Roossinck MJ. 2011. The good viruses: viral mutualistic symbioses. Nat Rev Microbiol 9:99–108.

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

80. Márquez LM, Redman RS, Rodriguez RJ, Roossinck MJ. 2007. A virus in a fungus in a plant: three-way symbiosis required for thermal tolerance. Science 315:513–515.

81. Desjardins CA, Gundersen-Rindal DE, Hostetler JB, Tallon LJ, Fadrosh DW, Fuester RW, Pedroni MJ, Haas BJ, Schatz MC, Jones KM, Crabtree J, Forberger H, Nene V. 2008. Comparative genomics of mutualistic viruses of Glyptapanteles parasitic wasps. Genome Biol 9:R183.

82. Strand MR, Burke GR. 2014. Polydnaviruses: Nature's Genetic Engineers. Httpdxdoiorg101146annurev-Virol-031413-085451. review-article.

83. Hacker J, Kaper JB. 2000. Pathogenicity islands and the evolution of microbes. Annu Rev Microbiol 54:641–679.

84. Koonin EV, Wolf YI. 2012. Evolution of microbes and viruses: a paradigm shift in evolutionary biology? Front Cell Infect Microbiol 2:119.

85. Perna NT, Plunkett G, Burland V, Mau B, Glasner JD, Rose DJ, Mayhew GF, Evans PS, Gregor J, Kirkpatrick HA, Pósfai G, Hackett J, Klink S, Boutin A, Shao Y, Miller L, Grotbeck EJ, Davis NW, Lim A, Dimalanta ET, Potamousis KD, Apodaca J, Anantharaman TS, Lin J, Yen G, Schwartz DC, Welch RA, Blattner FR. 2001. Genome sequence of enterohaemorrhagic Escherichia coli O157:H7. Nature 409:529–533.

86. Jain R, Rivera MC, Moore JE, Lake JA. 2002. Horizontal Gene Transfer in Microbial Genome Evolution. Theor Popul Biol 61:489–495.

87. Kondo N, Nikoh N, Ijichi N, Shimada M, Fukatsu T. 2002. Genome fragment of Wolbachia endosymbiont transferred to X chromosome of host insect. Proc Natl Acad Sci U S A 99:14280–14285.

88. Richards TA, Soanes DM, Jones MDM, Vasieva O, Leonard G, Paszkiewicz K, Foster PG, Hall N, Talbot NJ. 2011. Horizontal gene transfer facilitated the evolution of plant parasitic mechanisms in the oomycetes. Proc Natl Acad Sci 108:15258–15263.

89. Ku C, Martin WF. 2016. A natural barrier to lateral gene transfer from prokaryotes to eukaryotes revealed from genomes: the 70 % rule. BMC Biol 14:89.

90. Canchaya C, Fournous G, Chibani-Chennoufi S, Dillmann ML, Brüssow H. 2003. Phage as agents of lateral gene transfer. Curr Opin Microbiol 6:417–424.

91. Balcazar JL. 2014. Bacteriophages as vehicles for antibiotic resistance genes in the environment. PLoS Pathog 10:e1004219.

92. Varga M, Kuntová L, Pantůček R, Mašlaňová I, Růžičková V, Doškař J. 2012. Efficient transfer of antibiotic resistance plasmids by transduction within methicillin-resistant Staphylococcus aureus USA300 clone. FEMS Microbiol Lett 332:146–152.

93. Schaack S, Gilbert C, Feschotte C. 2010. Promiscuous DNA: horizontal transfer of transposable elements and why it matters for eukaryotic evolution. Trends Ecol Evol 25:537–546.

94. Gilbert C, Chateigner A, Ernenwein L, Barbe V, Bézier A, Herniou EA, Cordaux R. 2014. Population genomics supports baculoviruses as vectors of horizontal transfer of insect transposons. Nat Commun 5:ncomms4348.

95. Pierce SK, Maugel TK, Rumpho ME, Hanten JJ, Mondy WL. 1999. Annual Viral Expression in a Sea Slug Population: Life Cycle Control and Symbiotic Chloroplast Maintenance. Biol Bull 197:1–6.

96. Pierce SK, Massey SE, Hanten JJ, Curtis NE. 2003. Horizontal transfer of functional nuclear genes between multicellular organisms. Biol Bull 204:237–240.

97. Bhattacharya D, Pelletreau KN, Price DC, Sarver KE, Rumpho ME. 2013. Genome analysis of Elysia chlorotica Egg DNA provides no evidence for horizontal gene transfer into the germ line of this Kleptoplastic Mollusc. Mol Biol Evol 30:1843–1852.

98. Rauch C, Vries J de, Rommel S, Rose LE, Woehle C, Christa G, Laetz EM, Wägele H, Tielens AGM, Nickelsen J, Schumann T, Jahns P, Gould SB. 2015. Why It Is Time to Look Beyond Algal Genes in Photosynthetic Slugs. Genome Biol Evol 7:2602–2607.

99. de Vries J, Habicht J, Woehle C, Huang C, Christa G, Wägele H, Nickelsen J, Martin WF, Gould SB. 2013. Is ftsH the key to plastid longevity in sacoglossan slugs? Genome Biol Evol 5:2540–2548.

100. Bilewitch JP, Degnan SM. 2011. A unique horizontal gene transfer event has provided the octocoral mitochondrial genome with an active mismatch repair gene that has potential for an unusual self-contained function. BMC Evol Biol 11:228.

101. Ogata H, Ray J, Toyoda K, Sandaa R-A, Nagasaki K, Bratbak G, Claverie J-M. 2011. Two new subfamilies of DNA mismatch repair proteins (MutS) specifically abundant in the marine environment. ISME J 5:1143–1151.

102. Claverie J-M, Abergel C. 2013. Open Questions About Giant Viruses. Adv Virus Res 85:25–56.

103. Yoshida T, Claverie J-M, Ogata H. 2011. Mimivirus reveals Mre11/Rad50 fusion proteins with a sporadic distribution in eukaryotes, bacteria, viruses and plasmids. Virol J 8:427.

104. Shutt TE, Gray MW. 2006. Twinkle, the mitochondrial replicative DNA helicase, is widespread in the eukaryotic radiation and may also be the mitochondrial DNA primase in most eukaryotes. J Mol Evol 62:588–599.

105. Shutt TE, Gray MW. 2006. Bacteriophage origins of mitochondrial replication and transcription proteins. Trends Genet TIG 22:90–95.

106. Cornelis G, Heidmann O, Bernard-Stoecklin S, Reynaud K, Véron G, Mulot B, Dupressoir A, Heidmann T. 2012. Ancestral capture of syncytin-Car1, a fusogenic endogenous retroviral envelope gene involved in placentation and conserved in Carnivora. Proc Natl Acad Sci U S A 109:E432-441.

107. Fédry J, Liu Y, Péhau-Arnaudet G, Pei J, Li W, Tortorici MA, Traincard F, Meola A, Bricogne G, Grishin NV, Snell WJ, Rey FA, Krey T. 2017. The Ancient Gamete Fusogen HAP2 Is a Eukaryotic Class II Fusion Protein. Cell 168:904–915.e10.

108. Gornik SG, Ford KL, Mulhern TD, Bacic A, McFadden GI, Waller RF. 2012. Loss of Nucleosomal DNA Condensation Coincides with Appearance of a Novel Nuclear Protein in Dinoflagellates. Curr Biol 22:2303–2312.

109. Abroi A, Gough J. 2011. Are viruses a source of new protein folds for organisms? - Virosphere structure space and evolution. BioEssays News Rev Mol Cell Dev Biol 33:626–635.

110. Koonin EV. 2016. Viruses and mobile elements as drivers of evolutionary transitions. Philos Trans R Soc Lond B Biol Sci 371.

111. Szathmáry E. 2015. Toward major evolutionary transitions theory 2.0. Proc Natl Acad Sci 112:10104–10111.

112. Smith JM, Szathmary E. 1997. The Major Transitions in Evolution. OUP Oxford.

113. Forterre P. 2002. The origin of DNA genomes and DNA replication proteins. Curr Opin Microbiol 5:525–532.

114. Takahashi I, Marmur J. 1963. Replacement of thymidylic acid by deoxyuridylic acid in the deoxyribonucleic acid of a transducing phage for Bacillus subtilis. Nature 197:794–795.

115. Iyer LM, Aravind L, Koonin EV. 2001. Common origin of four diverse families of large eukaryotic DNA viruses. J Virol 75:11720–11734.

116. Gould SJ. 1990. Wonderful Life: The Burgess Shale and the Nature of History. W. W. Norton.

117. Boyer M, Yutin N, Pagnier I, Barrassi L, Fournous G, Espinosa L, Robert C, Azza S, Sun S, Rossmann MG, Suzan-Monti M, La Scola B, Koonin EV, Raoult D. 2009. Giant Marseillevirus highlights the role of amoebae as a melting pot in emergence of chimeric microorganisms. Proc Natl Acad Sci U S A 106:21848–21853.

118. Philippe N, Legendre M, Doutre G, Couté Y, Poirot O, Lescot M, Arslan D, Seltzer V, Bertaux L, Bruley C, Garin J, Claverie J-M, Abergel C. 2013. Pandoraviruses: amoeba viruses with genomes up to 2.5 Mb reaching that of parasitic eukaryotes. Science 341:281–286.

119. Legendre M, Bartoli J, Shmakova L, Jeudy S, Labadie K, Adrait A, Lescot M, Poirot O, Bertaux L, Bruley C, Couté Y, Rivkina E, Abergel C, Claverie J-M. 2014. Thirty-thousand-year-old distant relative of giant icosahedral DNA viruses with a pandoravirus morphology. Proc Natl Acad Sci U S A 111:4274–4279.

120. Legendre M, Lartigue A, Bertaux L, Jeudy S, Bartoli J, Lescot M, Alempic J-M, Ramus C, Bruley C, Labadie K, Shmakova L, Rivkina E, Couté Y, Abergel C, Claverie J-M. 2015. In-depth study of Mollivirus sibericum, a new 30,000-y-old giant virus infecting Acanthamoeba. Proc Natl Acad Sci U S A 112:E5327-5335.

121. Monier A, Larsen JB, Sandaa R-A, Bratbak G, Claverie J-M, Ogata H. 2008. Marine mimivirus relatives are probably large algal viruses. Virol J 5:12.

122. Larsen JB, Larsen A, Bratbak G, Sandaa R-A. 2008. Phylogenetic Analysis of Members of the Phycodnaviridae Virus Family, Using Amplified Fragments of the Major Capsid Protein Gene. Appl Environ Microbiol 74:3048–3057.

123. Hingamp P, Grimsley N, Acinas SG, Clerissi C, Subirana L, Poulain J, Ferrera I, Sarmento H, Villar E, Lima-Mendez G, Faust K, Sunagawa S, Claverie J-M, Moreau H, Desdevises Y, Bork P, Raes J, de Vargas C, Karsenti E, Kandels-Lewis S, Jaillon O, Not F, Pesant S, Wincker P, Ogata H. 2013. Exploring nucleo-cytoplasmic large DNA viruses in Tara Oceans microbial metagenomes. ISME J 7:1678–1695.

124. Allen LZ, Ishoey T, Novotny MA, McLean JS, Lasken RS, Williamson SJ. 2011. Single Virus Genomics: A New Tool for Virus Discovery. PLOS ONE 6:e17722.

125. Harris TD, Buzby PR, Babcock H, Beer E, Bowers J, Braslavsky I, Causey M, Colonell J, DiMeo J, Efcavitch JW, Giladi E, Gill J, Healy J, Jarosz M, Lapen D, Moulton K, Quake

SR, Steinmann K, Thayer E, Tyurina A, Ward R, Weiss H, Xie Z. 2008. Single-Molecule DNA Sequencing of a Viral Genome. Science 320:106–109.

126. Aherfi S, Colson P, La Scola B, Raoult D. 2016. Giant Viruses of Amoebas: An Update. Front Microbiol 7.

127. Claverie J-M, Abergel C. 2009. Mimivirus and its virophage. Annu Rev Genet 43:49–66.

128. Moss B. Poxviridae: The viruses and their replication Chapter 83. p2637-2672. Fields BN Knipe DM Howley PM Fields Virol Third Ed Edn 2.

129. Abergel C, Legendre M, Claverie J-M. 2015. The rapidly expanding universe of giant viruses: Mimivirus, Pandoravirus, Pithovirus and Mollivirus. FEMS Microbiol Rev 39:779–796.

130. Van Etten JL, Meints RH. 1999. Giant viruses infecting algae. Annu Rev Microbiol 53:447–494.

131. García-Beato R, Salas ML, Viñuela E, Salas J. 1992. Role of the host cell nucleus in the replication of African swine fever virus DNA. Virology 188:637–649.

132. Goorha R. 1982. Frog virus 3 DNA replication occurs in two stages. J Virol 43:519–528.

133. Fabre E, Jeudy S, Santini S, Legendre M, Trauchessec M, Couté Y, Claverie J-M, Abergel C. 2017. Noumeavirus replication relies on a transient remote control of the host nucleus. Nat Commun 8:15087.

134. Wilson WH, Van Etten JL, Allen MJ. 2009. The Phycodnaviridae: the story of how tiny giants rule the world. Curr Top Microbiol Immunol 328:1–42.

135. Colson P, De Lamballerie X, Yutin N, Asgari S, Bigot Y, Bideshi DK, Cheng X-W, Federici BA, Van Etten JL, Koonin EV, La Scola B, Raoult D. 2013. "Megavirales", a proposed new order for eukaryotic nucleocytoplasmic large DNA viruses. Arch Virol 158:2517–2521.

136. Netherton CL, Wileman T. 2011. Virus factories, double membrane vesicles and viroplasm generated in animal cells. Curr Opin Virol 1:381–387.

137. Yutin N, Wolf YI, Koonin EV. 2014. Origin of giant viruses from smaller DNA viruses not from a fourth domain of cellular life. Virology 466–467:38–52.

138. Legendre M, Audic S, Poirot O, Hingamp P, Seltzer V, Byrne D, Lartigue A, Lescot M, Bernadac A, Poulain J, Abergel C, Claverie J-M. 2010. mRNA deep sequencing reveals 75 new genes and a complex transcriptional landscape in Mimivirus. Genome Res 20:664–674.

139. Schulz F, Yutin N, Ivanova NN, Ortega DR, Lee TK, Vierheilig J, Daims H, Horn M, Wagner M, Jensen GJ, Kyrpides NC, Koonin EV, Woyke T. 2017. Giant viruses with an expanded complement of translation system components. Science 356:82–85.

140. Koonin EV, Yutin N. 2010. Origin and evolution of eukaryotic large nucleo-cytoplasmic DNA viruses. Intervirology 53:284–292.

141. Boyer M, Madoui M-A, Gimenez G, Scola BL, Raoult D. 2010. Phylogenetic and Phyletic Studies of Informational Genes in Genomes Highlight Existence of a 4th Domain of Life Including Giant Viruses. PLOS ONE 5:e15530.

142. Williams TA, Embley TM, Heinz E. 2011. Informational gene phylogenies do not support a fourth domain of life for nucleocytoplasmic large DNA viruses. PloS One 6:e21080.

143. Lwoff A. 1957. The concept of virus. Microbiology 17:239–253.

144. Lartigue A, Burlat B, Coutard B, Chaspoul F, Claverie J-M, Abergel C. 2015. The megavirus chilensis Cu,Zn-superoxide dismutase: the first viral structure of a typical cellular copper chaperone-independent hyperstable dimeric enzyme. J Virol 89:824–832.

145. Jeudy S, Lartigue A, Claverie J-M, Abergel C. 2009. Dissecting the unique nucleotide specificity of mimivirus nucleoside diphosphate kinase. J Virol 83:7142–7150.

146. Hakim M, Ezerina D, Alon A, Vonshak O, Fass D. 2012. Exploring ORFan domains in giant viruses: structure of mimivirus sulfhydryl oxidase R596. PloS One 7:e50649.

147. Piacente F, De Castro C, Jeudy S, Molinaro A, Salis A, Damonte G, Bernardi C, Abergel C, Tonetti MG. 2014. Giant Virus Megavirus chilensis Encodes the Biosynthetic Pathway for Uncommon Acetamido Sugars. J Biol Chem 289:24428–24439.

148. Piacente F, Bernardi C, Marin M, Blanc G, Abergel C, Tonetti MG. 2014. Characterization of a UDP-N-acetylglucosamine biosynthetic pathway encoded by the giant DNA virus Mimivirus. Glycobiology 24:51–61.

149. Piacente F, Gaglianone M, Laugieri ME, Tonetti MG. 2015. The Autonomous Glycosylation of Large DNA Viruses. Int J Mol Sci 16:29315–29328.

150. Piacente F, De Castro C, Jeudy S, Gaglianone M, Laugieri ME, Notaro A, Salis A, Damonte G, Abergel C, Tonetti MG. 2017. The rare sugar N-acetylated viosamine is a major component of Mimivirus fibers. J Biol Chem 292:7385–7394.

151. La Scola B, Desnues C, Pagnier I, Robert C, Barrassi L, Fournous G, Merchat M, Suzan-Monti M, Forterre P, Koonin E, Raoult D. 2008. The virophage as a unique parasite of the giant mimivirus. Nature 455:100–104.

152. Villain A, Gallot-Lavallée L, Blanc G, Maumus F. 2016. Giant viruses at the core of microscopic wars with global impacts. Curr Opin Virol 17:130–137.

153. Gaia M, Pagnier I, Campocasso A, Fournous G, Raoult D, La Scola B. 2013. Broad spectrum of mimiviridae virophage allows its isolation using a mimivirus reporter. PloS One 8:e61912.

154. Campos RK, Boratto PV, Assis FL, Aguiar ER, Silva LC, Albarnaz JD, Dornas FP, Trindade GS, Ferreira PP, Marques JT, Robert C, Raoult D, Kroon EG, La Scola B, Abrahão JS. 2014. Samba virus: a novel mimivirus from a giant rain forest, the Brazilian Amazon. Virol J 11:95.

155. Palukaitis P. 2015. Satellite RNAs and Satellite Viruses. Mol Plant Microbe Interact 29:181–186.

156. Krupovic M, Cvirkaite-Krupovic V. 2011. Virophages or satellite viruses? Nat Rev Microbiol 9:762–763.

157. Fischer MG. 2011. Sputnik and Mavirus: more than just satellite viruses. Nat Rev Microbiol 10:78; author reply 78.

158. Krupovic M, Kuhn JH, Fischer MG. 2016. A classification system for virophages and satellite viruses. Arch Virol 161:233–247.

159. Wodarz D. 2013. Evolutionary dynamics of giant viruses and their virophages. Ecol Evol 3:2103–2115.

160. Taylor BP, Cortez MH, Weitz JS. 2014. The virus of my virus is my friend: Ecological effects of virophage with alternative modes of coinfection. J Theor Biol 354:124–136.

161. Fischer MG, Suttle CA. 2011. A virophage at the origin of large DNA transposons. Science 332:231–234.

162. Kapitonov VV, Jurka J. 2006. Self-synthesizing DNA transposons in eukaryotes. Proc Natl Acad Sci U S A 103:4540–4545.

163. Pritham EJ, Putliwala T, Feschotte C. 2007. Mavericks, a novel class of giant transposable elements widespread in eukaryotes and related to DNA viruses. Gene 390:3–17.

164. Krupovic M, Bamford DH, Koonin EV. 2014. Conservation of major and minor jelly-roll capsid proteins in Polinton (Maverick) transposons suggests that they are bona fide viruses. Biol Direct 9:6.

165. Yutin N, Raoult D, Koonin EV. 2013. Virophages, polintons, and transpovirons: a complex evolutionary network of diverse selfish genetic elements with different reproduction strategies. Virol J 10:158.

166. Desnues C, La Scola B, Yutin N, Fournous G, Robert C, Azza S, Jardot P, Monteil S, Campocasso A, Koonin EV, Raoult D. 2012. Provirophages and transpovirons as the diverse mobilome of giant viruses. Proc Natl Acad Sci U S A 109:18078–18083.

167. Santini S, Jeudy S, Bartoli J, Poirot O, Lescot M, Abergel C, Barbe V, Wommack KE, Noordeloos AAM, Brussaard CPD, Claverie J-M. 2013. Genome of Phaeocystis globosa virus PgV-16T highlights the common ancestry of the largest known DNA viruses infecting eukaryotes. Proc Natl Acad Sci U S A 110:10800–10805.

168. Gogarten JP, Senejani AG, Zhaxybayeva O, Olendzenski L, Hilario E. 2002. Inteins: structure, function, and evolution. Annu Rev Microbiol 56:263–287.

169. Ogata H, Raoult D, Claverie J-M. 2005. A new example of viral intein in Mimivirus. Virol J 2:8.

170. Pietrokovski S. 1998. Identification of a virus intein and a possible variation in the protein-splicing reaction. Curr Biol CB 8:R634-635.

171. Nagasaki K, Shirai Y, Tomaru Y, Nishida K, Pietrokovski S. 2005. Algal Viruses with Distinct Intraspecies Host Specificities Include Identical Intein Elements. Appl Environ Microbiol 71:3599–3607.

172. Arslan D, Legendre M, Seltzer V, Abergel C, Claverie J-M. 2011. Distant Mimivirus relative with a larger genome highlights the fundamental features of Megaviridae. Proc Natl Acad Sci U S A 108:17486–17491.

173. Yoosuf N, Yutin N, Colson P, Shabalina SA, Pagnier I, Robert C, Azza S, Klose T, Wong J, Rossmann MG, La Scola B, Raoult D, Koonin EV. 2012. Related giant viruses in distant locations and different habitats: Acanthamoeba polyphaga moumouvirus represents a third lineage of the Mimiviridae that is close to the megavirus lineage. Genome Biol Evol 4:1324–1330.

174. Fischer MG, Allen MJ, Wilson WH, Suttle CA. 2010. Giant virus with a remarkable complement of genes infects marine zooplankton. Proc Natl Acad Sci U S A 107:19508–19513.

175. Moniruzzaman M, LeCleir GR, Brown CM, Gobler CJ, Bidle KD, Wilson WH, Wilhelm SW. 2014. Genome of brown tide virus (AaV), the little giant of the Megaviridae, elucidates NCLDV genome expansion and host–virus coevolution. Virology 466–467:60–70.

176. Parfrey LW, Lahr DJG, Knoll AH, Katz LA. 2011. Estimating the timing of early eukaryotic diversification with multigene molecular clocks. Proc Natl Acad Sci U S A 108:13624–13629.

177. Sandaa R-A, Heldal M, Castberg T, Thyrhaug R, Bratbak G. 2001. Isolation and Characterization of Two Viruses with Large Genome Size Infecting Chrysochromulina ericina (Prymnesiophyceae) and Pyramimonas orientalis (Prasinophyceae). Virology 290:272–280.

178. Li L, Stoeckert CJ, Roos DS. 2003. OrthoMCL: identification of ortholog groups for eukaryotic genomes. Genome Res 13:2178–2189.

179. Yutin N, Koonin EV. 2009. Evolution of DNA ligases of Nucleo-Cytoplasmic Large DNA viruses of eukaryotes: a case of hidden complexity. Biol Direct 4:51.

180. Maruyama F, Ueki S. 2016. Evolution and Phylogeny of Large DNA Viruses, Mimiviridae and Phycodnaviridae Including Newly Characterized Heterosigma akashiwo Virus. Front Microbiol 7.

181. Blanc G, Ogata H, Robert C, Audic S, Suhre K, Vestris G, Claverie J-M, Raoult D. 2007. Reductive Genome Evolution from the Mother of Rickettsia. PLOS Genet 3:e14.

182. Wernegreen JJ. 2005. For better or worse: genomic consequences of intracellular mutualism and parasitism. Curr Opin Genet Dev 15:572–583.

183. Moliner C, Fournier P-E, Raoult D. 2010. Genome analysis of microorganisms living in amoebae reveals a melting pot of evolution. FEMS Microbiol Rev 34:281–294.

184. Doutre G, Philippe N, Abergel C, Claverie J-M. 2014. Genome Analysis of the First Marseilleviridae Representative from Australia Indicates that Most of Its Genes Contribute to Virus Fitness. J Virol 88:14340–14349.

185. Boyer M, Azza S, Barrassi L, Klose T, Campocasso A, Pagnier I, Fournous G, Borg A, Robert C, Zhang X, Desnues C, Henrissat B, Rossmann MG, Scola BL, Raoult D. 2011. Mimivirus shows dramatic genome reduction after intraamoebal culture. Proc Natl Acad Sci 108:10296–10301.

186. Yutin N, Koonin EV. 2012. Hidden evolutionary complexity of Nucleo-Cytoplasmic Large DNA viruses of eukaryotes. Virol J 9:161.

187. Yutin N, Colson P, Raoult D, Koonin EV. 2013. Mimiviridae: clusters of orthologous genes, reconstruction of gene repertoire evolution and proposed expansion of the giant virus family. Virol J 10:106.

188. Forterre P. 2010. Giant viruses: conflicts in revisiting the virus concept. Intervirology 53:362–378.

189. Ogata H, Claverie J-M. 2007. Unique genes in giant viruses: regular substitution pattern and anomalously short size. Genome Res 17:1353–1361.

190. Carvunis A-R, Rolland T, Wapinski I, Calderwood MA, Yildirim MA, Simonis N, Charloteaux B, Hidalgo CA, Barbette J, Santhanam B, Brar GA, Weissman JS, Regev A, Thierry-Mieg N, Cusick ME, Vidal M. 2012. Proto-genes and de novo gene birth. Nature 487:370–374.

191. Reinhardt JA, Wanjiru BM, Brant AT, Saelao P, Begun DJ, Jones CD. 2013. De novo ORFs in Drosophila are important to organismal fitness and evolved rapidly from previously non-coding sequences. PLoS Genet 9:e1003860.

192. Vinella D, Brochier-Armanet C, Loiseau L, Talla E, Barras F. 2009. Iron-sulfur (Fe/S) protein biogenesis: phylogenomic and genetic studies of A-type carriers. PLoS Genet 5:e1000497.

193. Sullivan MB, Huang KH, Ignacio-Espinoza JC, Berlin AM, Kelly L, Weigele PR, DeFrancesco AS, Kern SE, Thompson LR, Young S, Yandava C, Fu R, Krastins B, Chase M, Sarracino D, Osburne MS, Henn MR, Chisholm SW. 2010. Genomic analysis of oceanic cyanobacterial myoviruses compared with T4-like myoviruses from diverse hosts and environments. Environ Microbiol 12:3035–3056.

194. Monier A, Welsh RM, Gentemann C, Weinstock G, Sodergren E, Armbrust EV, Eisen JA, Worden AZ. 2012. Phosphate transporters in marine phytoplankton and their viruses: cross-domain commonalities in viral-host gene exchanges. Environ Microbiol 14:162–176.

195. Clerissi C, Grimsley N, Subirana L, Maria E, Oriol L, Ogata H, Moreau H, Desdevises Y. 2014. Prasinovirus distribution in the Northwest Mediterranean Sea is affected by the environment and particularly by phosphate availability. Virology 466–467:146–157.

196. Hamacher K, Greiner T, Ogata H, Van Etten JL, Gebhardt M, Villarreal LP, Cosentino C, Moroni A, Thiel G. 2012. Phycodnavirus potassium ion channel proteins question the virus molecular piracy hypothesis. PloS One 7:e38826.

197. Thiel G, Moroni A, Blanc G, Van Etten JL. 2013. Potassium ion channels: could they have evolved from viruses? Plant Physiol 162:1215–1224.

198. Roux S, Brum JR, Dutilh BE, Sunagawa S, Duhaime MB, Loy A, Poulos BT, Solonenko N, Lara E, Poulain J, Pesant S, Kandels-Lewis S, Dimier C, Picheral M, Searson S, Cruaud C, Alberti A, Duarte CM, Gasol JM, Vaqué D, Tara Oceans Coordinators, Bork P, Acinas SG, Wincker P, Sullivan MB. 2016. Ecogenomics and potential biogeochemical impacts of globally abundant ocean viruses. Nature 537:689–693.

199. Krejcík Z, Denger K, Weinitschke S, Hollemeyer K, Paces V, Cook AM, Smits THM. 2008. Sulfoacetate released during the assimilation of taurine-nitrogen by Neptuniibacter caesariensis: purification of sulfoacetaldehyde dehydrogenase. Arch Microbiol 190:159–168.

200. Weinitschke S, Denger K, Cook AM, Smits THM. 2007. The DUF81 protein TauE in Cupriavidus necator H16, a sulfite exporter in the metabolism of C2 sulfonates. Microbiol Read Engl 153:3055–3060.

201. Biedlingmaier S, Schmidt A. 1987. Uptake and metabolism of taurine in the green alga Chlorella fusca. Physiol Plant 70:688–696.

202. Luther M. 1990. Nitro- and amino- substituted aromatic compounds as nitrogen sources for the green alga Scenedesmus obliquus. Chemosphere 21:231–241.

203. Costantini A, Doria F, Saiz J-C, Garcia-Moruno E. 2017. Phage-host interactions analysis of newly characterized Oenococcus oeni bacteriophages: Implications for malolactic fermentation in wine. Int J Food Microbiol 246:12–19.

204. Bauer R, Dicks LMT. 2017. Control of Malolactic Fermentation in Wine. A Review. South Afr J Enol Vitic 25:74–88.

205. Favier M, Bilhère E, Lonvaud-Funel A, Moine V, Lucas PM. 2012. Identification of pOENI-1 and Related Plasmids in Oenococcus oeni Strains Performing the Malolactic Fermentation in Wine. PLOS ONE 7:e49082.

206. Sozzi T, Maret R, Poulin JM. 1976. [Observation of bacteriophages in wine (author's transl)]. Experientia 32:568–569.

207. Raoult D. 2004. The 1.2-Megabase Genome Sequence of Mimivirus. Science 306:1344–1350.

208. Colson P, Fournous G, Diene SM, Raoult D. 2013. Codon usage, amino acid usage, transfer RNA and amino-acyl-tRNA synthetases in Mimiviruses. Intervirology 56:364–375.

209. Abergel C, Rudinger-Thirion J, Giegé R, Claverie J-M. 2007. Virus-Encoded Aminoacyl-tRNA Synthetases: Structural and Functional Characterization of Mimivirus TyrRS and MetRS. J Virol 81:12406–12417.

210. Roy H, Becker HD, Reinbolt J, Kern D. 2003. When contemporary aminoacyl-tRNA synthetases invent their cognate amino acid metabolism. Proc Natl Acad Sci 100:9837–9842.

211. Francklyn C. 2003. tRNA synthetase paralogs: Evolutionary links in the transition from tRNA-dependent amino acid biosynthesis to de novo biosynthesis. Proc Natl Acad Sci 100:9650–9652.

212. Guo M, Yang X-L, Schimmel P. 2010. New functions of tRNA synthetases beyond translation. Nat Rev Mol Cell Biol 11:668–674.

213. Salazar JC, Ambrogelly A, Crain PF, McCloskey JA, Söll D. 2004. A truncated aminoacyl-tRNA synthetase modifies RNA. Proc Natl Acad Sci U S A 101:7536–7541.

214. Sissler M, Delorme C, Bond J, Ehrlich SD, Renault P, Francklyn C. 1999. An aminoacyl-tRNA synthetase paralog with a catalytic role in histidine biosynthesis. Proc Natl Acad Sci 96:8985–8990.

215. Blaise M, Fréchin M, Oliéric V, Charron C, Sauter C, Lorber B, Roy H, Kern D. 2011. Crystal Structure of the Archaeal Asparagine Synthetase: Interrelation with Aspartyl-tRNA and Asparaginyl-tRNA Synthetases. J Mol Biol 412:437–452.

216. Nakatsu T, Kato H, Oda J. 1998. Crystal structure of asparagine synthetase reveals a close evolutionary relationship to class II aminoacyl-tRNA synthetase. Nat Struct Mol Biol 5:15–19.

217. Doolittle RF, Handy J. 1998. Evolutionary anomalies among the aminoacyl-tRNA synthetases. Curr Opin Genet Dev 8:630–636.

218. Enright AJ, Iliopoulos I, Kyrpides NC, Ouzounis CA. 1999. Protein interaction maps for complete genomes based on gene fusion events. Nature 402:86–90.

219. Marcotte EM, Pellegrini M, Ng HL, Rice DW, Yeates TO, Eisenberg D. 1999. Detecting protein function and protein-protein interactions from genome sequences. Science 285:751–753.

220. Suhre K. 2007. Inference of gene function based on gene fusion events: the rosetta-stone method. Methods Mol Biol Clifton NJ 396:31–41.

221. Bailey S, Sedelnikova SE, Blackburn GM, Abdelghany HM, Baker PJ, McLennan AG, Rafferty JB. 2002. The Crystal Structure of Diadenosine Tetraphosphate Hydrolase from Caenorhabditis elegans in Free and Binary Complex Forms. Structure 10:589–600.

222. Goerlich O, Foeckler R, Holler E. 1982. Mechanism of Synthesis of Adenosine(5′)tetraphospho(5′)adenosine (AppppA) by Aminoacyl-tRNA Synthetases. Eur J Biochem 126:135–142.

223. Thompson AW, Foster RA, Krupke A, Carter BJ, Musat N, Vaulot D, Kuypers MMM, Zehr JP. 2012. Unicellular Cyanobacterium Symbiotic with a Single-Celled Eukaryotic Alga. Science 337:1546–1550.

224. Villareal TA. 1991. Nitrogen-fixation by the cyanobacterial symbiont of the diatom genus Hemiaulus. Mar Ecol Prog Ser Oldendorf 76:201–204.

225. Park E, Hwang D-S, Lee J-S, Song J-I, Seo T-K, Won Y-J. 2012. Estimation of divergence times in cnidarian evolution based on mitochondrial protein-coding genes and the fossil record. Mol Phylogenet Evol 62:329–345.

226. Filée J, Siguier P, Chandler M. 2007. I am what I eat and I eat what I am: acquisition of bacterial genes by giant viruses. Trends Genet TIG 23:10–15.

227. Filée J, Chandler M. 2010. Gene exchange and the origin of giant viruses. Intervirology 53:354–361.

228. Filée J, Pouget N, Chandler M. 2008. Phylogenetic evidence for extensive lateral acquisition of cellular genes by Nucleocytoplasmic large DNA viruses. BMC Evol Biol 8:320.

229. Moreira D, López-García P. 2005. Comment on "The 1.2-megabase genome sequence of Mimivirus." Science 308:1114; author reply 1114.

230. Ogata H. 2005. Response to Comment on "The 1.2-Megabase Genome Sequence of Mimivirus." Science 308:1114b–1114b.

231. Raoult D, Forterre P. 2008. Redefining viruses: lessons from Mimivirus. Nat Rev Microbiol 6:315–319.

232. Forterre P, Prangishvili D. 2009. The great billion-year war between ribosome- and capsid-encoding organisms (cells and viruses) as the major source of evolutionary novelties. Ann N Y Acad Sci 1178:65–77.

233. Emerman M, Malik HS. 2010. Paleovirology--modern consequences of ancient viruses. PLoS Biol 8:e1000301.

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

235. Bill CA, Summers J. 2004. Genomic DNA double-strand breaks are targets for hepadnaviral DNA integration. Proc Natl Acad Sci U S A 101:11135–11140.

236. Delaroque N, Maier I, Knippers R, Müller DG. 1999. Persistent virus integration into the genome of its algal host, Ectocarpus siliculosus (Phaeophyceae). J Gen Virol 80 (Pt 6):1367–1370.

237. Meints RH, Ivey RG, Lee AM, Choi T-J. 2008. Identification of two virus integration sites in the brown alga Feldmannia chromosome. J Virol 82:1407–1413.

238. Read BA, Kegel J, Klute MJ, Kuo A, Lefebvre SC, Maumus F, Mayer C, Miller J, Monier A, Salamov A, Young J, Aguilar M, Claverie J-M, Frickenhaus S, Gonzalez K, Herman EK, Lin Y-C, Napier J, Ogata H, Sarno AF, Shmutz J, Schroeder D, de Vargas C, Verret F, von Dassow P, Valentin K, Van de Peer Y, Wheeler G, Emiliania huxleyi Annotation Consortium, Dacks JB, Delwiche CF, Dyhrman ST, Glöckner G, John U, Richards T, Worden AZ, Zhang X, Grigoriev IV. 2013. Pan genome of the phytoplankton Emiliania underpins its global distribution. Nature 499:209–213.

239. Blanc G, Duncan G, Agarkova I, Borodovsky M, Gurnon J, Kuo A, Lindquist E, Lucas S, Pangilinan J, Polle J, Salamov A, Terry A, Yamada T, Dunigan DD, Grigoriev IV, Claverie J-M, Van Etten JL. 2010. The Chlorella variabilis NC64A genome reveals adaptation to photosymbiosis, coevolution with viruses, and cryptic sex. Plant Cell 22:2943–2955.

240. Maumus F, Epert A, Nogué F, Blanc G. 2014. Plant genomes enclose footprints of past infections by giant virus relatives. Nat Commun 5:4268.

241. Kapp M. 1998. Viruses Infecting Marine Brown Algae. Virus Genes 16:111–117.

242. Filée J. 2014. Multiple occurrences of giant virus core genes acquired by eukaryotic genomes: The visible part of the iceberg? Virology 466–467:53–59.

243. Desnues C, Raoult D. 2010. Inside the lifestyle of the virophage. Intervirology 53:293–303.

244. Fischer MG, Hackl T. 2016. Host genome integration and giant virus-induced reactivation of the virophage mavirus. Nature 540:288–291.

245. Labrie SJ, Samson JE, Moineau S. 2010. Bacteriophage resistance mechanisms. Nat Rev Microbiol 8:317–327.

246. Allocati N, Masulli M, Di Ilio C, De Laurenzi V. 2015. Die for the community: an overview of programmed cell death in bacteria. Cell Death Dis 6:e1609.

247. Foster KR, Wenseleers T, Ratnieks FLW. 2006. Kin selection is the key to altruism. Trends Ecol Evol 21:57–60.

248. Lewis K. 1998. Pathogen Resistance as the Origin of Kin Altruism. J Theor Biol 193:359–363.

249. Suzuki S, Ishida K-I, Hirakawa Y. 2016. Diurnal Transcriptional Regulation of Endosymbiotically Derived Genes in the Chlorarachniophyte Bigelowiella natans. Genome Biol Evol 8:2672–2682.

250. Correa AMS, Ainsworth TD, Rosales SM, Thurber AR, Butler CR, Vega Thurber RL. 2016. Viral Outbreak in Corals Associated with an In Situ Bleaching Event: Atypical Herpes-Like Viruses and a New Megavirus Infecting Symbiodinium. Front Microbiol 7:127.

251. Reteno DG, Benamar S, Khalil JB, Andreani J, Armstrong N, Klose T, Rossmann M, Colson P, Raoult D, La Scola B. 2015. Faustovirus, an asfarvirus-related new lineage of giant viruses infecting amoebae. J Virol 89:6585–6594.

252. Adl SM, Simpson AGB, Lane CE, Lukeš J, Bass D, Bowser SS, Brown MW, Burki F, Dunthorn M, Hampl V, Heiss A, Hoppenrath M, Lara E, le Gall L, Lynn DH, McManus H, Mitchell EAD, Mozley-Stanridge SE, Parfrey LW, Pawlowski J, Rueckert S, Shadwick L, Schoch CL, Smirnov A, Spiegel FW. 2012. The Revised Classification of Eukaryotes. J Eukaryot Microbiol 59:429–514.

253. Bajrai LH, Benamar S, Azhar EI, Robert C, Levasseur A, Raoult D, La Scola B. 2016. Kaumoebavirus, a New Virus That Clusters with Faustoviruses and Asfarviridae. Viruses 8.

254. Andreani J, Khalil JYB, Sevvana M, Benamar S, Di Pinto F, Bitam I, Colson P, Klose T, Rossmann MG, Raoult D, La Scola B. 2017. Pacmanvirus, a New Giant Icosahedral Virus at the Crossroads between Asfarviridae and Faustoviruses. J Virol 91.

255. Wolfowicz I, Baumgarten S, Voss PA, Hambleton EA, Voolstra CR, Hatta M, Guse A. 2016. Aiptasia sp. larvae as a model to reveal mechanisms of symbiont selection in cnidarians. Sci Rep 6:32366.

256. Richards TA, Monier A. 2016. A tale of two tardigrades. Proc Natl Acad Sci U S A 113:4892–4894.

257. Boothby TC, Tenlen JR, Smith FW, Wang JR, Patanella KA, Nishimura EO, Tintori SC, Li Q, Jones CD, Yandell M, Messina DN, Glasscock J, Goldstein B. 2015. Evidence for extensive horizontal gene transfer from the draft genome of a tardigrade. Proc Natl Acad Sci U S A 112:15976–15981.

258. Koutsovoulos G, Kumar S, Laetsch DR, Stevens L, Daub J, Conlon C, Maroon H, Thomas F, Aboobaker AA, Blaxter M. 2016. No evidence for extensive horizontal gene transfer in the genome of the tardigrade Hypsibius dujardini. Proc Natl Acad Sci U S A 113:5053–5058.

259. Aswad A, Katzourakis A. 2017. A novel viral lineage distantly related to herpesviruses discovered within fish genome sequence data. Virus Evol 3.

260. Béliveau C, Cohen A, Stewart D, Periquet G, Djoumad A, Kuhn L, Stoltz D, Boyle B, Volkoff A-N, Herniou EA, Drezen J-M, Cusson M. 2015. Genomic and Proteomic Analyses Indicate that Banchine and Campoplegine Polydnaviruses Have Similar, if Not Identical, Viral Ancestors. J Virol 89:8909–8921.

261. Rappoport N, Linial M. 2012. Viral proteins acquired from a host converge to simplified domain architectures. PLoS Comput Biol 8:e1002364.

262. Gould SJ, Vrba ES. 1982. Exaptation-A Missing Term in the Science of Form. Paleobiology 8:4–15.

263. Magadum S, Banerjee U, Murugan P, Gangapur D, Ravikesavan R. 2013. Gene duplication as a major force in evolution. J Genet 92:155–161.

264. Doolittle RF. 1995. The Multiplicity of Domains in Proteins. Annu Rev Biochem 64:287–314.

265. Basu MK, Poliakov E, Rogozin IB. 2009. Domain mobility in proteins: functional and evolutionary implications. Brief Bioinform 10:205–216.

266. Gallot-Lavallée L, Blanc G. 2017. A Glimpse of Nucleo-Cytoplasmic Large DNA Virus Biodiversity through the Eukaryotic Genomics Window. Viruses 9.

267. Douglas AE. 2014. Symbiosis as a General Principle in Eukaryotic Evolution. Cold Spring Harb Perspect Biol 6.

268. Řídká T, Peksa O, Rai H, Upreti DK, Škaloud P. 2014. Photobiont Diversity in Indian Cladonia Lichens, with Special Emphasis on the Geographical Patterns, p. 53–71. *In* Rai, H, Upreti, DK (eds.), Terricolous Lichens in India. Springer New York, New York, NY.

269. Kawaida H, Ohba K, Koutake Y, Shimizu H, Tachida H, Kobayakawa Y. 2013. Symbiosis between hydra and chlorella: molecular phylogenetic analysis and experimental study provide insight into its origin and evolution. Mol Phylogenet Evol 66:906–914.

270. Trémouillaux-Guiller J, Rohr T, Rohr R, Huss VAR. 2002. Discovery of an endophytic alga in Ginkgo biloba. Am J Bot 89:727–733.

271. Ishikawa M, Yuyama I, Shimizu H, Nozawa M, Ikeo K, Gojobori T. 2016. Different Endosymbiotic Interactions in Two Hydra Species Reflect the Evolutionary History of Endosymbiosis. Genome Biol Evol 8:2155–2163.

272. Zhang H, Wei S, Hu W, Xiao L, Tang M. 2017. Arbuscular Mycorrhizal Fungus Rhizophagus irregularis Increased Potassium Content and Expression of Genes Encoding Potassium Channels in Lycium barbarum. Front Plant Sci 8.

273. Claverie J-M, Grzela R, Lartigue A, Bernadac A, Nitsche S, Vacelet J, Ogata H, Abergel C. 2009. Mimivirus and Mimiviridae: Giant viruses with an increasing number of potential hosts, including corals and sponges. J Invertebr Pathol 101:172–180.

274. Filée J, Forterre P. 2005. Viral proteins functioning in organelles: a cryptic origin? Trends Microbiol 13:510–513.

275. Blanc G, Gallot-Lavallée L, Maumus F. 2015. Provirophages in the Bigelowiella genome bear testimony to past encounters with giant viruses. Proc Natl Acad Sci U S A 112:E5318-5326.

276. Sagan L. 1967. On the origin of mitosing cells. J Theor Biol 14:255–274.

277. Kannan S, Rogozin IB, Koonin EV. 2014. MitoCOGs: clusters of orthologous genes from mitochondria and implications for the evolution of eukaryotes. BMC Evol Biol 14:237.

278. Gray MW. 2015. Mosaic nature of the mitochondrial proteome: Implications for the origin and evolution of mitochondria. Proc Natl Acad Sci 112:10133–10138.

279. Kurland CG, Andersson SG. 2000. Origin and evolution of the mitochondrial proteome. Microbiol Mol Biol Rev MMBR 64:786–820.

280. Thomas V, Greub G. 2010. Amoeba/amoebal symbiont genetic transfers: lessons from giant virus neighbours. Intervirology 53:254–267.

281. Raoult D, Boyer M. 2010. Amoebae as Genitors and Reservoirs of Giant Viruses. Intervirology 53:321–329.

282. Shinzato C, Mungpakdee S, Satoh N, Shoguchi E. 2014. A genomic approach to coral-dinoflagellate symbiosis: studies of Acropora digitifera and Symbiodinium minutum. Front Microbiol 5.

283. Dinsdale EA, Pantos O, Smriga S, Edwards RA, Angly F, Wegley L, Hatay M, Hall D, Brown E, Haynes M, Krause L, Sala E, Sandin SA, Thurber RV, Willis BL, Azam F, Knowlton N, Rohwer F. 2008. Microbial ecology of four coral atolls in the Northern Line Islands. PloS One 3:e1584.

284. Rohwer F, Seguritan V, Azam F, Knowlton N. 2002. Diversity and distribution of coral-associated bacteria. Mar Ecol Prog Ser 243:1–10.

285. Di Camillo CG, Luna GM, Bo M, Giordano G, Corinaldesi C, Bavestrello G. 2012. Biodiversity of prokaryotic communities associated with the ectoderm of Ectopleura crocea (Cnidaria, Hydrozoa). PloS One 7:e39926.

286. Sharma V, Colson P, Giorgi R, Pontarotti P, Raoult D. 2014. DNA-Dependent RNA Polymerase Detects Hidden Giant Viruses in Published Databanks. Genome Biol Evol 6:1603–1610.

287. Moniruzzaman M, Wurch LL, Alexander H, Dyhrman ST, Gobler CJ, Wilhelm SW. 2017. Virus-host relationships of marine single-celled eukaryotes resolved from metatranscriptomics. Nat Commun 8:16054.

288. Tarutani K, Nagasaki K, Itakura S, Yamaguchi M. 2001. Isolation of a virus infecting the novel shellfish-killing dinoflagellate Heterocapsa circularisquama. Aquat Microb Ecol 23:103–111.

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

290. Edgar RC. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res 32:1792–1797.

291. Huelsenbeck JP, Ronquist F. 2001. MRBAYES: Bayesian inference of phylogenetic trees. Bioinforma Oxf Engl 17:754–755.

292. Taly J-F, Magis C, Bussotti G, Chang J-M, Di Tommaso P, Erb I, Espinosa-Carrasco J, Kemena C, Notredame C. 2011. Using the T-Coffee package to build multiple sequence alignments of protein, RNA, DNA sequences and 3D structures. Nat Protoc 6:1669–1682.

293. Yutin N, Wolf YI, Raoult D, Koonin EV. 2009. Eukaryotic large nucleo-cytoplasmic DNA viruses: clusters of orthologous genes and reconstruction of viral genome evolution. Virol J 6:223.

294. Slater GSC, Birney E. 2005. Automated generation of heuristics for biological sequence comparison. BMC Bioinformatics 6:31.

295. Notredame C, Higgins DG, Heringa J. 2000. T-Coffee: A novel method for fast and accurate multiple sequence alignment. J Mol Biol 302:205–217.