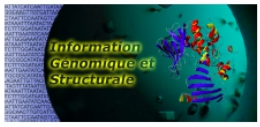
AIX-MARSEILLE UNIVERSITÉ

ÉCOLE DOCTORALE SCIENCES DE LA VIE ET DE LA SANTÉ (ED 62)

LABORATOIRE INFORMATION GÉNOMIQUE ET STRUCTURALE

ETUDE DE VIRUS RÉACTIVÉS À PARTIR D'ÉCHANTILLONS DE CRYOSOL

Thèse présentée pour obtenir le grade universitaire de
Docteur en Sciences de la Vie et de la Santé
Discipline : Biologie



Présentée et soutenue publiquement
le 18 novembre 2020 par :
Eugène CHRISTO-FOROUX

DIRECTION DE THÈSE :

Dr. Chantal ABERGEL : Directeur
Matthieu LEGENDRE : Co-directeur

JURY DE THÈSE :

Dr. Yves DESDEVISES : Rapporteur
Dr. Elisabeth HERNIOU : Rapporteur
Dr. Chantal ABERGEL : Examinateur
Pr. Christophe ROBAGLIA : Examinateur
Dr. Jean-Michel CLAVERIE : Invité
Dr. Emmanuelle GUILLOT-COMBE: Invité
Dr. Matthieu LEGENDRE : Invité

« Si les virus n’avaient pas été découverts, tout se serait bien passé »
(Stanley 1957)

Table des matières

Remerciements

Une thèse étant une étape dans un parcours de vie, l'exercice consistant à remercier l'ensemble des individus m'ayant permis de m'accomplir durant ce cheminement de 3 ans est ardu tant les rencontres et échanges qui ont façonnés ma perception du monde ont été nombreux et riches d'enseignements. En premier lieu, je remercie Chantal Abergel, ma directrice de thèse, pour m'avoir ouvert la porte de l'IGS et de m'avoir offert ce cadre qui m'a permis d'exprimer cette curiosité qui m'habite depuis toujours. Merci d'avoir fait preuve d'empathie et de compréhension à de nombreuses reprises, tu m'as invité à prendre une place juste au sein du laboratoire et c'est la première fois que j'ai eu le sentiment de gagner, au moins en partie, contre ce sentiment d'imposture qui, trop souvent, m'a contraint à des choix de vie par défaut. Je remercie Matthieu Legendre, mon co-directeur, d'avoir su me guider dans la concrétisation de mes idées, en commençant par celles que je souhaitais exprimer dans ce manuscrit. La qualité de tes travaux, l'humilité dont tu fais preuve face à des résultats, sont deux qualités inspirantes pour tous les étudiants qui ont eu la chance travailler avec toi. Je remercie également Jean-Michel Claverie, dont la bienveillance et la culture m'ont conduit à appréhender différemment la complexité du monde. Lors de nos nombreuses discussion, notamment dans les divers avions vers la Sibérie, j'ai pu questionner la portée éminemment politique du statut d'agent producteur de savoir au sein de la société et donner ainsi encore d'avantage de sens à mon travail de thèsard. Il y a 3 ans, je manquais d'équilibre dans l'élaboration de mes idées et je trouvais une forme de satisfaction intellectuelle dans le débordement d'interrogations plutôt que dans le processus d'obtention d'une solution. Je suis heureux d'avoir appris à canaliser cette énergie, cela me permet aujourd'hui d'apprécier le sentiment la concision d'une réponse à une question bien formulée.

Je remercie également l'ensemble des membres du laboratoire. Tout d'abord, merci à vous, mes co-doctorants, Ale, Sofia, d'avoir été de superbes piliers de paillasse et de bar. Merci Ale pour ton imperméabilité à la panique et ton abnégation, tu as été un peu le point d'équilibre de l'open space sur ces aspects. Merci pour les coups de mains lorsqu'il il fallait aller au labo de temps en temps le week-end. Merci Sofia d'avoir été un petit brin de folie. Vous êtes des gens brillants que j'admire énormément et j'espère que nos trajectoires scientifiques futures nous amènerons à nous recroiser.

Merci à Jean-Marie et Audrey de m'avoir coaché pendant mes manip'. Jean-Marie, tu es quelqu'un d'extrêmement cultivé et nos discussions ont profondément marqué mon passage à l'IGS. Je me sens chanceux, honoré, que tu ais partagé avec moi ton expérience en réactivation de virus

géants (sans toi le manuscrit aurait tenu sur peu de pages) et de la vie au sens large. Audrey, merci de m'avoir toujours écouté, surtout dans les galères de ma vie. Souvent ces débats se tenaient entre 18h et 19h, Thomas avait beau t'appeler tu restais pour m'écouter, y repenser me touche énormément. Je pense que, même si tu ne le revendiques pas, tu es quelqu'un d'extrêmement créative. Tu me répondrais probablement, sur un haussement d'épaule, que peindre c'est pas ton truc (j'en conviens) mais toutes tes idées pour pallier aux soucis techniques et scientifiques du quotidien me confortent dans mon ressenti. Merci aux deux Seb ! Vous m'avez permis de franchir le cap du Terminal. LE FAMEUX TERMINAL, cet écran noir que j'ai du apprivoiser durant mes 3 ans de doctorat. Vous êtes tous deux des individus remarquables et dont la sensibilité m'a touché au delà de nos échanges purement scientifiques. Merci à Sebastien Nin pour les parties de jeux de société et à Sebastien Santini pour les conseils, autant en série qu'en musique. Vous avez été d'une patiente extrême (=> demande à Google!). Merci Olivier pour les coups de mains sur Zotero (tu viens encore de me sauver pas plus tard qu'avant hier). Tu mérites amplement ton statut de Awk-master 3ème DAN. Au plaisir de se retrouver pour une sortie photo ! Merci Estelle, pour t'être bagarrée avec les gens méchants des administrations, surtout quand ils voulaient pas me rembourser mon déplacement à la Rochelle ! En espérant avoir la chance de porter un tatouage de ton chéri un jour. Merci à tous ceux dont nos chemins ce sont croisés au laboratoire : Sandra, Adrien, Lionel, Alain, Hugo...

Merci également à Léa, ma compagne, et Claire, ma mère. Léa dans la complexité de ce que tu es, tu me donnes les armes pour aller au bout de ce que j'aimerais être. Merci de m'avoir supporté durant les derniers mois de cette grande expérience qu'on appelle « Thèse ». Maman, tu as traversé durant ces 3 ans des événements difficiles et je te remercie d'avoir compris que j'étais la, avec toi, en permanence. Merci à vous deux d'avoir corrigé mon manuscrit.

Merci à mes amis : Ysa pour tes corrections, Dinh-Long, Zacharie, Laura, Claire, Paul, Anne-Laure, Corentin, Clément, Pierre et Julie pour vos aller-retours entre Marseille et Paris. J'espère que vous avez apprécié nos moments à refaire le monde les pieds dans la mer ou dans la bière. Merci à ces belles rencontres marseillaises : Malvina, Magali, Florian, Jojo, Marie... et tous mes camarades.

Enfin, je tenais à dédier ce manuscrit à Alain, mon oncle, et Angeliki, ma grand mère chypriote, tous deux partis au cours des trois dernières années.

« La vie c'est comme une bicyclette, il faut avancer pour ne pas perdre l'équilibre »

A. Einstein

INTRODUCTION

L'introduction de ce manuscrit a trois objectifs: éclairer sur le contexte historique et scientifique ayant rendu possible la découverte des virus géants, au travers de l'évolution des représentations de la virologie; présenter le contexte scientifique duquel a émergé mon projet de thèse ; définir et présenter les virus géants.

La virologie, histoire d'une discipline en perpétuelle évolution

Si l'émergence de la virologie en tant que discipline apparaît comme récente, un bref travail historiographique de la microbiologie permet de déceler plusieurs étapes ayant conduit à l'émergence de la virologie comme discipline. Dans *Histoires de la virologie, des viroses et des virologues* l'universitaire franco-croate Mirko D. Grmek propose une lecture de l'histoire en 4 grandes «étapes » [1] (Figure 1).

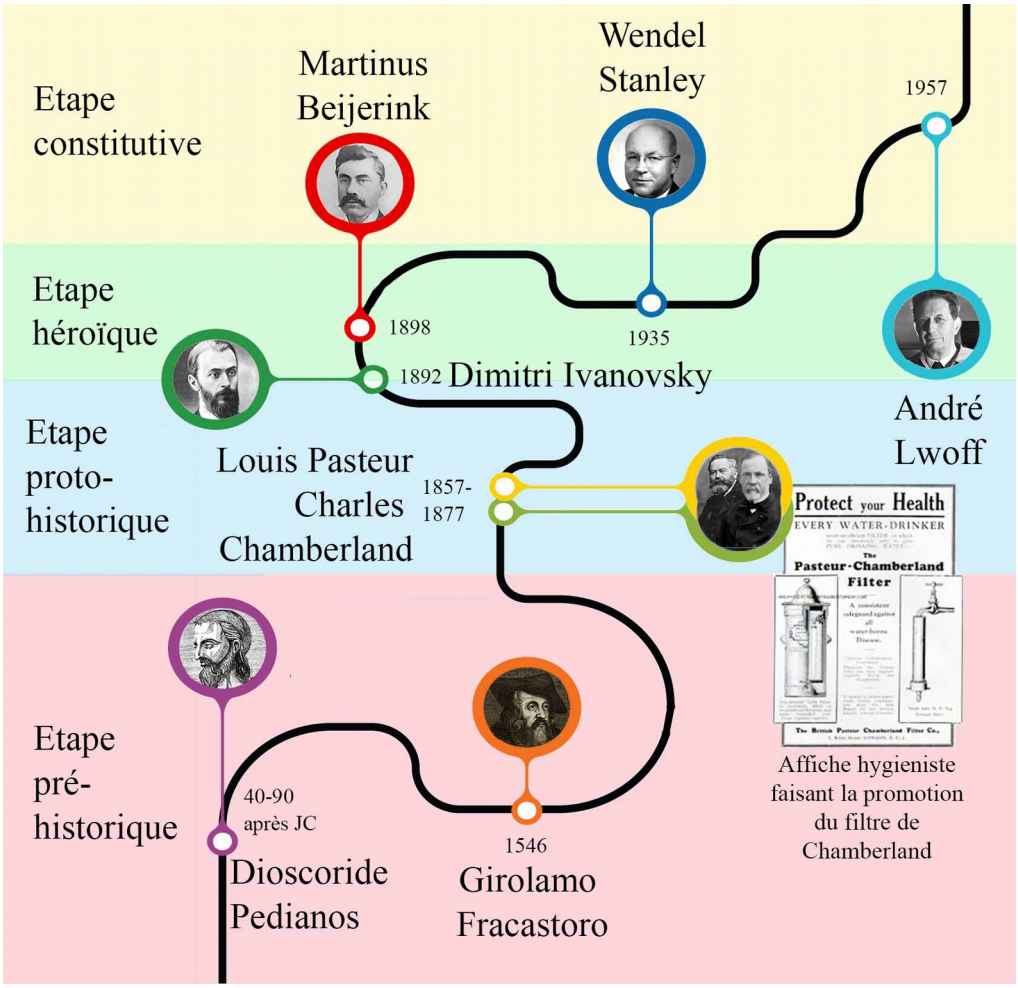


Figure 1: Frise chronologique replaçant les différentes figures des étapes de la virologie.

La première étape, qualifiée de « préhistoire », remonte à l'Antiquité, où les agents

responsables de maladies infectieuses sont appelés « poisons», ou « virus » en latin classique. C'est de cette période que nous proviennent les premières descriptions symptomatiques de maladies, comme la tuberculose, appelée alors par Hippocrate (Vème – IIVème siècle av. J.-C.) « phtisie » ou « tabès », qui font état d'une affection amaigrissante au long cours touchant les poumons. Dans *De Materia medica* le médecin et botaniste grec Dioscoride Pedianos (40-90 après JC) concilie dans cet ouvrage faisant référence jusqu'au début du XVème siècle, l'ensemble des remèdes pharmacologiques connus alors pour lutter contre ces « poisons ». A cette époque les connaissances empiriques en pharmacologie s'accumulent rapidement et permettent un affranchissement des premières pratiques médicales relatives à la sphère sacerdotale [2], il est admis que le corps est constitué de quatre éléments antagonistes fondamentaux : air, feu, eau et terre et que la santé repose sur un équilibre entre ces forces. Cette vision mystique de la santé domine jusqu'à la Renaissance, où elle influence largement les arts et les lettres (Figure 2).

Figure 2: Gravure de Albrecht Dürer, « La Philosophie » (1502). On peut observer à chaque angle de l'oeuvre une allégorie des différents éléments permettant l'équilibre de la pensée.

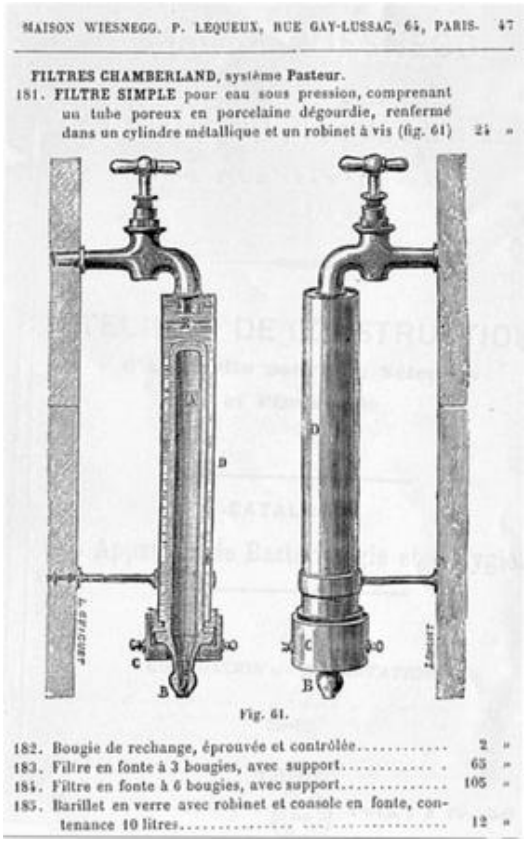
On trouve peu de traces de tentatives d'explication des causes et des facteurs de la maladie. Les premières théories épidémiologiques émergent de façon concomitante et au IXème siècle la théorie des miasmes, attribuant la prolifération des maladies à des brouillards toxiques remplis de particules issues de matière décomposée, s'impose. Cependant, on observe aucune recherche sur la nature des agents responsables de ces affections. Comme souvent, des idées contradictoires ont

émergé, entre autre celle de l'italien Girolamo Fracastoro (Figure 1) qui propose en 1546 que le vecteur des épidémies est lié à la transmission d'entités semblables à des semences, ou celle du médecin réformateur anglais Thomas Southwood Smith qui formule en 1830 l'hypothèse que le toucher permet la transmission des maladies. L'hégémonie de la théorie des miasmes permet, paradoxalement, de nombreuses avancées sociales en terme sanitaire, ainsi, à la suite de la propagation du choléra en 1850, s'entreprend la transformation de Paris sous le second Empire par le préfet Haussmann.

C'est l'étude systématique des épidémies de choléra en France et en Angleterre qui permet de réaliser des avancées en microbiologie, ouvrant une seconde « étape », ou « protohistoire » (Figure 1), de la virologie. Durant cette période, la recherche sur la nature et les causes des maladies s'intensifie, notamment par l'usage de nouveaux instruments : le microscope optique, en particulier les premiers modèles construits par Antoni van Leeuwenhoek à la fin du XVIIème, et de supports d'observation : les milieux de culture. Ainsi, par la mise en place de cultures pures, Pasteur démontre que les « ferments » sont des organismes vivants, qu'il suggère ubiquitaires. Dans ses « Communications sur les fermentations » de 1857-1858, Pasteur met ainsi fin au débat scientifique concernant l'apparition de la vie sans ascendant à partir de matière inanimée, ou génération spontanée. Vingt ans plus tard, son agrégé préparateur Charles Chamberland et lui, en plein développement du courant de pensée hygiéniste, démontrent dans « La théorie des germes et ses applications » [3] que la culture successive d'un micro-organisme dans un milieu de culture stérile peut atténuer la virulence du dit microorganisme. En parallèle, deux mille ans après ses premières descriptions, l'unité nosologique et l'étiologie de la « phtisie », rebaptisée alors « tuberculose », sont établies. De 1865 à 1868, le médecin français Jean-Antoine Villemin reproduit, chez des cobayes, des lésions de la tuberculose humaine par inoculation de tissus infectés, prouvant ainsi qu'il existe bien un lien de causalité entre la présence de l'agent infectieux et l'apparition des symptômes de la tuberculose. C'est dans ce contexte, en 1890, que le médecin allemand Robert Koch définit des critères destinés à établir la relation de cause à effet liant un microbe et une maladie [4] :

- Le micro-organisme doit être présent dans tous les organismes souffrant de la maladie, mais absent des organismes sains.
- Ce micro-organisme doit pouvoir être isolé et croître en milieu de culture pur.
- Le micro-organisme cultivé doit déclencher la même maladie chez un animal de laboratoire sensible.
- Le micro-organisme doit être à nouveau isolé du nouvel organisme hôte rendu malade puis identifié comme étant identique à l'agent infectieux original.

Peu à peu, ce postulat s'impose dans le domaine, alors inédit, de la microbiologie. Si les travaux extrêmement rigoureux de Pasteur et Koch permettent à la « théorie des germes » de supplanter progressivement la « théorie des miasmes », les moyens matériels et l'état des connaissances ne permettent pas encore d'envisager différentes catégories d'agents infectieux. La notion implicite de « taille » des agents pathogènes est introduite lorsque Charles Chamberland (Figure 1), continuant ses travaux initiés avec Louis Pasteur sur la stérilisation, met au point un filtre de porcelaine (Figure 3) dont la granularité des pores (0,1 à 1 µm) permet de stériliser l'eau en retenant les microorganismes permettant ainsi aux habitations de profiter d'une eau courante potable.


MAISON WIESNEGG. P. LEQUEUX, RUE GAY-LUSSAC, 64, PARIS. 47

FILTRES CHAMBERLAND, système **Pasteur**.

181. **FILTRE SIMPLE** pour eau sous pression, comprenant un tube poreux en porcelaine dégourdie, renfermé dans un cylindre métallique et un robinet à vis (fig. 61) 24 »

Fig. 61.

182. Bougie de rechange, éprouvée et contrôlée..... 2 »
183. Filtre en fonte à 3 bougies, avec support..... 65 »
184. Filtre en fonte à 6 bougies, avec support..... 105 »
185. Barillet en verre avec robinet et console en fonte, contenance 10 litres..... 12 »

Figure 3: filtre de Chamberland en porcelaine, fixé à un robinet, permettant de stériliser l'eau courante des habitations.

En 1892, le botaniste russe Dimitri Ivanovsky met en évidence que l'agent infectieux responsable de l'éclaircissement et du dépérissement de plants de tabac n'est pas retenu par le filtre de Chamberland, ouvrant ainsi l'ère « héroïque » de la virologie. Ivanosky, alors encore étudiant, n'ose pas briser le postulat de Koch, et n'envisage pas la découverte d'une nouvelle catégorie d'agents pathogènes. Ivanosky conclut plutôt que le micro-organisme responsable de la mosaïque du tabac est une toxine ou une petite bactérie. En 1898, à la demande d'Adolf Mayer, le chimiste Hollandais Martinus Beijerink reproduit les résultats d'Ivanovsky sans avoir eu connaissance des ses travaux. Par des expériences d'infections croisées de plants de tabac à plusieurs stades de

développement ainsi que par des expériences de diffusions du filtrat en milieu solide [5], Beijerink conclut qu'il existe un micro-organisme d'un genre nouveau, capable de se développer dans des cellules préférentiellement en division. De même, Beijerink [6] propose que l'agent infectieux de la mosaïque du tabac soit baptisé *contagium vivum fluidum,* en opposition au *contagium vivum fixum* qui ne diffuse pas sur les milieux de culture solides nouvellement développés, brisant ainsi le postulat de Koch. De façon concomitante, les deux collaborateurs de Koch : Friedrich Loeffler et Paul Frosch, démontrent que l'agent responsable de la fièvre aphteuse, une maladie du bétail, est un agent filtrable. Cette découverte fait porter la théorie du *contagium vivum fluidum* au delà de la biologie végétale. C'est grâce au développement simultané de la biologie moléculaire et l'apparition de nouveaux outils techniques : ultracentrifugation (1923 par Theodor Svedberg), microscopie électronique (Ernst Ruska 1937), cristallographie, que peuvent s'accumuler de nouvelles observations et questions concernant la nature du *contagium vivum fluidum.* Ainsi, la cristallisation du virus de la mosaïque du tabac (TMV, Figure 4) par Wendel Stanley en 1935 ([7], [8]) la purification du virus par ultracentrifugation puis électrophorèse (Eriksso et Svedberg, 1936) et l'observation du virus par microscopie électronique par Kausche en 1939 (Kausche et al. 1939), permettent de renforcer l'idée que les virus représentent une classe à part entière d'agents infectieux. La découverte des virus infectant les bactéries par Félix d'Hérelle en 1917 à l'Institut Pasteur de Paris ne suffit pas à émanciper la virologie de la parasitologie et l'étude des agents filtrants reste un sous-domaine de la bactériologie clinique. C'est dans ce contexte qu'est proposée la première classification des virus, basée sur les maladies ou les symptômes cliniques qu'ils causaient. En 1940, Francis Holmes reprend le système de nomenclature binomiale de Linné, et sépare comme suit : les *Phaginae* (virus infectant les bactéries), les *Phytophaginae* (virus infectant les plantes) et les *Zoophaginae* (virus infectant les animaux).

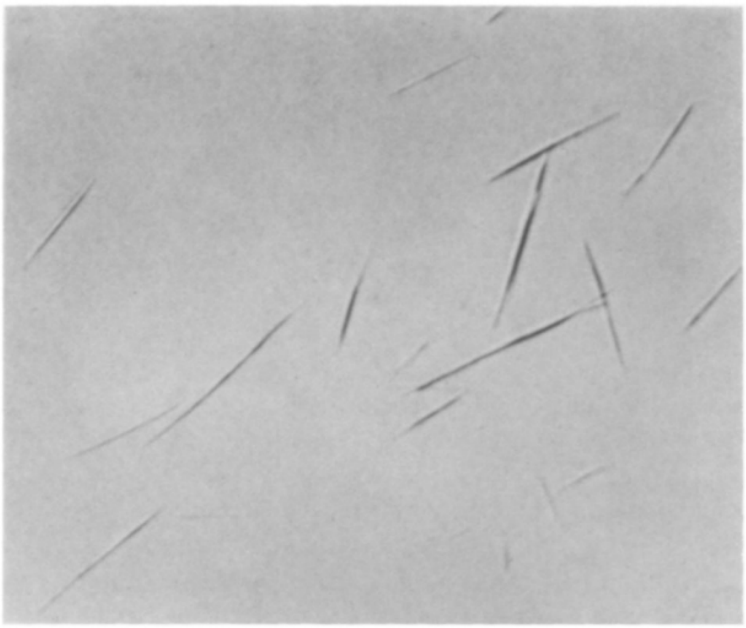


Figure 4: Photos de cristaux de TMV. Figure d'après Stanley & al [8].

Le contexte scientifique est alors propice à un débat institutionnel sur la nature des virus. C'est l'étape « constitutive » de la virologie. Ce débat s'articule autour de trois théories principales. En premier lieu, Wendel Stanley, biochimiste de formation, soutient que les virus sont de nature chimique, issus d'une réaction auto-catalytique. En deuxième lieu, le microbiologiste australien Frank Burnet, dont les travaux ont contribué à mieux comprendre le cycle infectieux de certains phages, soutient que les virus sont des êtres vivants réduits à une forme purement parasitaire [9]. En troisième lieu, Charles Bawden et Norman Wingate Pirie adressent une critique des travaux de cristallisation de Stanley [7] et préfèrent avancer la théorie selon laquelle les virus sont de nature ribonucléoproteiques. Ce débat est alimenté par les expériences d'hémagglutination du virus influenza conduites par George Hirst en 1941 qui conclut alors que les virus peuvent contenir des protéines [10]. De plus, avec la découverte de l'ADN comme support de l'information génétique, s'ouvre un nouveau débat sur la classification des virus, comme chainon manquant entre non-vivant et premières cellules. En 1957, l'accumulation d'expériences d'apparence contradictoires semble propice à une première définition de la notion de « virus ». André Lwoff, microbiologiste français, jette les fondements de la virologie dite « classique » dans une lapalissade devenue célèbre « Les virus sont des virus ». Par un travail aussi éclectique, mêlant biologie moléculaire, génétique et étude morphologiques, que rigoureux [11], André Lwoff propose une nouvelle classification stable du monde viral basée sur une méthode semblable à celle des organismes cellulaires [12]. Fondé sur les travaux de biologie moléculaire les plus récents dans le domaine, Lwoff énonce ce qui distingue les virus des organismes cellulaires avec les propositions suivantes :

- Les virus ne contiennent qu'un seul type d'acide nucléique (ADN ou ARN), support de

l'information génétique virale;

- Les virus sont des parasites intracellulaires obligatoires. Ils dépendent de l'appareil de traduction de l'hôte. N'ayant pas de ribosome, les virus sont également dépendant du métabolisme énergétique de la cellule infectée;
- Les virus se reproduisent directement à partir de leur matériel génétique par réplication (pas de division binaire).

Bien que très évasif sur la notion de taille comme critère de définition des virus, c'est à cette période que la recherche de nouveaux virus dans les fractions virales (filtrat < 0,2 µm) est systématisée. Ces avancées scientifiques et le débat autour de la définition de virus permettent une reconnaissance institutionnelle de la discipline. Ainsi, en 1955, à l'initiative de George Hirst, Lindsay Black et Salvador Luria, la première revue à comité de lecture, dédiée spécifiquement aux publications liées à l'étude des virus, voit le jour sous le nom univoque de *Virology*. De même, en 1970 l'Union Internationale des Sociétés de Microbiologie (IUMS), organisation non gouvernementale dépendante du Conseil international pour la science et ayant vocation à promouvoir la science à l'échelle mondiale, décide sous l'impulsion de son président André Lwoff (1958-1970) de se structurer autour de trois pôles indépendants :

- La division de bactériologie et microbiologie appliquée
- La division de mycologie
- La division de virologie

C'est avec ces événements fondateurs que s'arrête ce premier paragraphe.

Lwoff, de la classification des virus à l'ICTV

Les efforts fournis par Lwoff pour proposer une classification jugée rigoureuse des virus ainsi que son activité à la présidence de l'IUMS, permettent à la division virologie de cette organisation de se doter d'un Comité International de Taxonomie des Virus (ICTV). A sa création en 1966, les prérogatives de l'ICTV sont simples :

- Développer une taxonomie internationalement reconnue.
- Établir des noms convenus au niveau international pour les taxons.
- Communiquer aux virologues les décisions prises concernant la classification et la nomenclature des virus en organisant des réunions et en publiant des rapports annuels.
- Maintenir un index officiel des noms convenus.
- Etudier les effets des virus dans la société moderne et leur comportement.

Suivant la classification des organismes cellulaires, l'ICTV utilise dix niveaux hiérarchiques pour classer les virus:

- Domaine (-viria)
- Règne (-virae)
- Phylum (-viricota)
- Classe (-viricetes)
- Ordre (-virales)
- Famille (-viridae)
- Sous-famille (-virinae)
- Genre (-virus)
- Sous genre et espèce

En accord avec ses prérogatives de maintien d'un index officiel des noms convenus, l'ICTV publie chaque année le détail de l'ensemble de la taxonomie virale validée. Dans son rapport de 2019 l'ICTV fait état de 4 Domaines, 9 Règnes, 16 Phylum, 36 Classes, 55 Ordres, 168 Familles, 103 Sous-familles, 1421 Genres, 68 Sous-genres, 6590 Espèces [13]. Les travaux présentés dans ce manuscrit se focaliseront sur l'étude des virus à ADN double brin infectant les eucaryotes de l'Ordre *Megavirales* incluant les famille :

- *Ascoviridae*
- *Asfarviridae*
- *Iridoviridae*
- *Marseilleviridae*
- *Megaviridae* ou *Mimiviridae*
- *Pandoraviridae*
- *Phycodnaviridae*
- *Pithoviridae*
- *Poxviridae*

Tout au long de ce manuscrit nous verrons les avantages et les limites de cette classification, il est important de noter également que seules les familles de virus géants acceptées par l'ICTV sont les *Marseilleviridae* et les *Mimiviridae*. La référence à d'autres familles de virus géants sous entend

donc que ces dernières ont été proposées mais non validées.

Les grands virus à ADN nucléocytoplasmiques

Le dénominateur commun des *Megavirales*, autrement appelés grands virus à ADN nucléocytoplasmiques (NCLDV), est la grande taille de leur génome et la présence de gènes codants pour tout ou une partie de la machinerie de transcription, la réparation de l'ADN et la réplication. Avant la découverte de mimivirus, on comptait 5 familles de NCLDV [14].

Les *Poxviridae*

La famille des *Poxiviridae* est la famille de NCLDV dont nous connaissons le plus de représentants et dont le spectre d'hôtes connus est le plus large [15]. Ce spectre d'hôte très varié a amené à définir deux sous-familles: les *Entomopoxvirinae,* infectant les arthropodes et les *Chordopoxvirinae* infectant les cellules de vertébrés. L'intérêt historique pour les poxvirus est lié au caractère pathogène de 4 genres de *Chordopoxvirinae*: *Vaccinia virus* (VACV), *Cowpox virus* (CPXV), *Molluscum contagiosum virus* (MCV) et *Monkeypox virus* (MPXV). Les poxvirus expriment l'ensemble de la machinerie de transcription, de maturation des transcrits et de réplication de l'ADN permettant au cycle infectieux de se dérouler exclusivement dans le cytoplasme de la cellule [16]. La capside en forme de donut (Figure 5), d'une taille moyenne de 200 x 300 nm, enveloppée ou non, contient un génome linéaire allant de 128 à 365 kpb (respectivement pour le genre *Parapoxvirus* et *Avipoxvirus*) [17]. Le génome des poxvirus code pour plus de 200 protéines et adopte une structure secondaire particulière à ses extrémités. En effet, les extrémités du génome sont liées de façons covalentes grâce à une structure tige-boucle (*hairpin)* aux extrémités du génome [18].

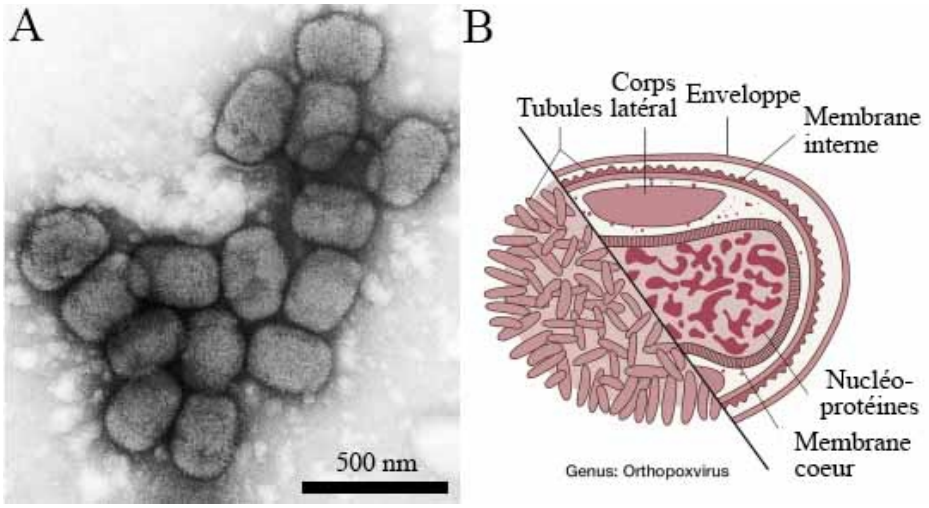


Figure 5: (A) Image de coloration négative de virions du virus de la vaccine. On observe les structures tubulaires de capside caractéristique des poxvirus (excepté les parapoxvirus). (B) Schéma de la capside des poxvirus du genre Orthopoxvirus (et tous les autres genres de poxvirus vertébrés à l'exception du genre Parapoxvirus), montrant le noyau biconcave et les corps latéraux. Figure d'après Burrell & al [17].

Les *Phycodnaviridae*

Les *Phycodnaviridae,* dont la traduction littérale du nom est « virus à ADN infectant les algues » est une famille de NCLDV largement étudiée pour le rôle que ces virus jouent dans la régulation des populations d'algues unicellulaires du genre *Chlorella* [19]. Six genres, au spectre d'hôte spécifique, sont décrits: *Chlorovirus* infectant *Chlorella, Coccolithovirus* infectant les Coccolithophores du genre *Emiliana huxleyi, Prasinovirus* infectant des *Prasinophytes* comme *Micromonas pusilla, Prymnesiovirus* infectant *Prymnesioiphytes, Phaeovirus* infectant des algues multicellulaires de la famille des *Phaeophycae* e t *Raphidovirus* infectant des *Heterosigma akashiwo*. A l'exception des *Coccolithovirus* qui codent pour leur propre ARN polymérases, le cycle infectieux des *Phycodnavirus* nécessite en phase précoce le recrutement des ARN polymérases de l'hôte directement dans le noyau [20]. Autre singularité, les *Phaeovirus* ont la spécificité d'avoir une phase lysogénique [21]. Les particules icosaédriques des *Phycodnaviridae* ont une taille allant de 115nm à 202 nm respectivement pour *Phaeovirus* et *Raphidovirus* (Figure 6). Le support de l'information génétique est un ADN double brin de 160 à 560 kpb linéaire qui peut se circulariser selon plusieurs modèles [22] : tige-boucle (*Coccolithovirus*) ou répétitions en tandem (*Chlorovirus*). Le nombre de protéines prédites varie également beaucoup, de 231 pour EsV-1 (*Phaeovirus)* à 472 pour EhV-86 (*Coccolithovirus).*

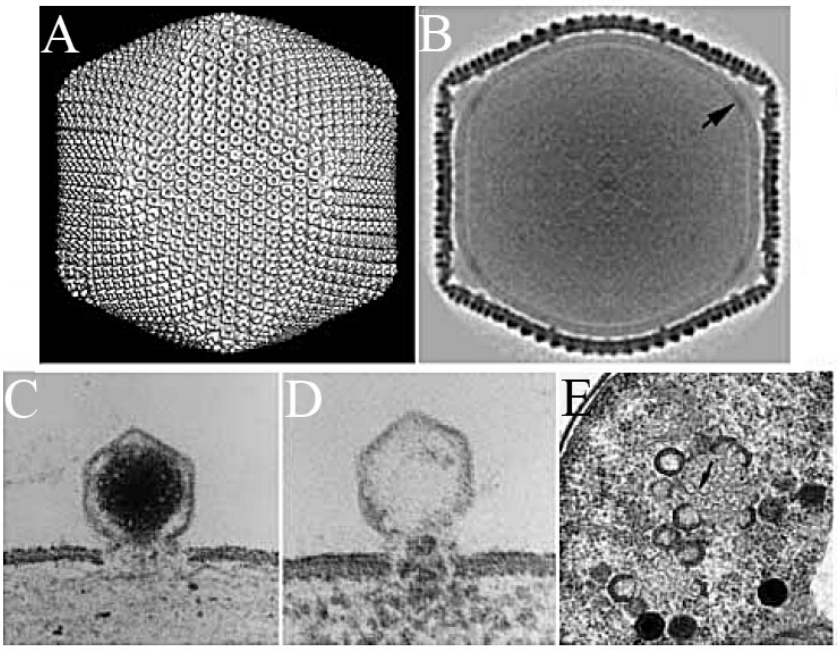


Figure 6: (A) Reconstruction 3-D du virion de Paramecium bursaria chlorella virus (PBCV-1), d'une taille de 190 nm.(B) Coupe transversale de la reconstruction 3-D montrant la membrane interne du virion. (C) Image de microscopie électronique montrant la digestion de la capside au point de contact entre le virion de PBCV-1 et la membrane cellulaire. (D) Image de microscopie éléctronique montrant le relargage du contenu de la capside de PBCV-1 dans le cytoplasme. (E) Zone d'assemblage des virions de PBCV-1 dans le cytoplasme à 3-4h post infection. Figure d'après Van Etten & al [19].

Les *Asfarviridae*

Jusqu'à peu, l'unique représentant de la famille des *Asfarviridae* est l'*African Swine Fever virus* (ASFV), responsable de la peste porcine africaine (ASF). Ce virus provoque des hémorragies chez les porcs domestiques et les épidémies d'ASF se sont rapidement répandues dans l'ensemble de l'Asie du Sud-Est durant la dernière décennie, causant la mort de plus de 1,2 millions de porcs d'élevage en 2018 [23] expliquant ainsi l'intérêt scientifique autour de l'étude de ce virus. Le cycle viral de l'ASFV se déroule dans le cytoplasme de la cellule hôte tout en ayant des interactions avec le noyau. La capside icosahedrique enveloppée de ce virus, d'une taille allant de 170 à 200 nm de diamètre (Figure 7), renferme un génome linéaire d'une taille allant de 170 à 190kpb et codant pour 150 à 167 ORFs [24]. Depuis la première épidémie d'ASFV au Kenya en 1907, d'autres virus apparentés à cette famille ont été découverts : *Faustovirus* [25], un virus infectant les amibes *Vermamoeba vermiformis,* et possédant un génome de 466 kb, *Kaumoebavirus* [26] et *Pacmanvirus* [27]. Très récemment, le plus proche parent d'ASFV connu à ce jour à été identifié [28]. Cet agent

responsable d'une atrophie du pied des ormeaux du genre *Haliotis discus discus*, au génome d'une taille de 155 kpb et codant pour 159 protéines a été baptisé *Abalone asfa-like virus*.

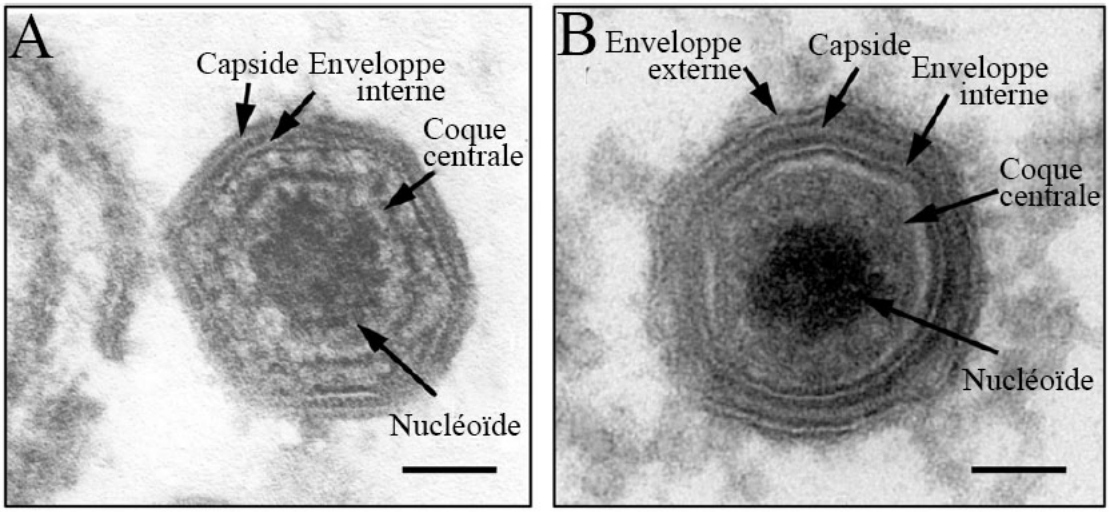


Figure 7: Photographie de microscopie électronique montrant la structure et composition protéique d'une capside d'ASFV. (A) Particule virale intracellulaire complète d'ASFV. (B) Virion d'ASFV mature dans le milieu extracellulaire.FIgure d'après Salas & al [24].

Les *Iridoviridae*

Les *Iridoviridae* consitue une famille de NCLDV très diverse [14]. Deux sous familles sont décrites *Alphairdonovirinae* et *Betairiridovirinae* et six genres: *Megalocytivirus*, *Lymphocystivirus* et *Ranavirus* pour les *Alphairdonovirinae* et *Iridovirus*, *Chloriridovirus* et *Decapodiridovirus* pour les *Betairiridovirinae*. Le spectre d'hôte des *Iridoviridae* est large et comprend à la fois des amphibiens, des poissons et des arthropodes. L'étude des *Iridoviridae* a été motivée par les épidémies de *Ranavirus* responsables de pertes financières importantes pour l'industrie de la pêche [29]. Comme pour les *Poxviridae*, la phase précoce du cycle viral des *Iridoviridae* se déroule dans le noyau de la cellule infectée, permettant l'expression précoce de l'ADN et l'ARN polymérase virale, le reste du cycle infectieux se déroule dans le cytoplasme. Le relargage des virions matures peut se faire par lyse de la cellule hôte ou exocytose. De ce fait, les particules virales icosaédriques des *Iridoviridae*, peuvent être ou enveloppées ou nues. La capside virale, d'une dimension de 120 à 300 nm de diamètre (Figure 8), renferme un génome linéaire de 102 à 303 kpb avec des séquences terminales redondantes et circulairement permutées [30]. On estime que les *Iridoviridae* codent pour 97 à 193 protéines.

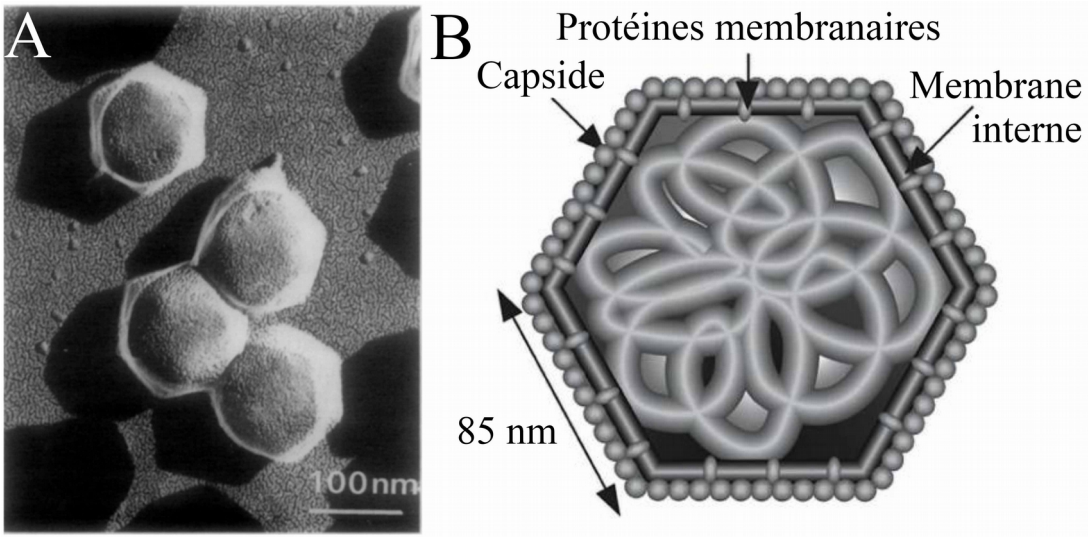


Figure 8: (A) Particule virale de Frog virus3 par microscopie électronique à cryo-fracture. (B) Schéma de la particule virale de Frog virus3. Figure d'après Darcy-Tripier & al [29].

Les *Ascoviridae*

La famille des *Ascoviridae* est la dernière famille de NCLDV à avoir été découverte avant celle des virus géants. A ce jour, elle ne compte qu'un seul genre, *Ascovirus*, capable d'infecter des Lépidoptères. D'une forme ovoïde (du grec *askos « sac »*), la particule virale de 130 nm par 300 nm contient un génome circulaire, d'une taille comprise entre 150 et 186 kpb [31] (Figure 9). Le cycle de réplication de ces virus est nucléocytoplasmique : les premières étapes du cycle s'effectuent dans le noyau de la cellule hôte, avant de se terminer dans le cytoplasme [32]

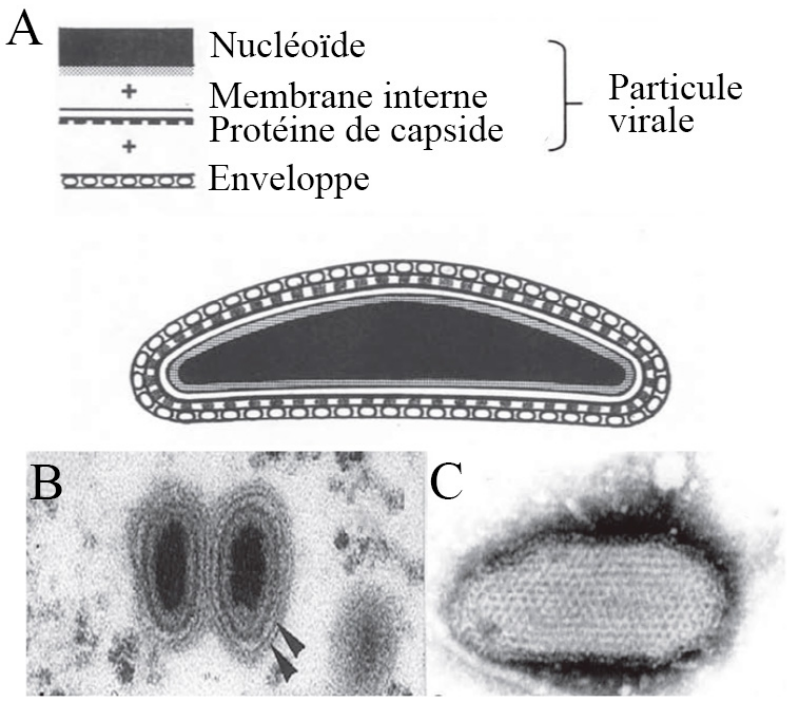


Figure 9: Structure de la particule virale des ascovirus. (A) Schéma des différentes couches composant le virion. (B) Image de microscopie électronique de particules virales d'ascovirus montrant l'enveloppe interne et externe. (C) Image de coloration négative montrant l'enveloppe externe en filet de la particule virale. Figure d'après Zaghloul & al [31].

Nous venons de voir que la classification des virus, permise par les avancées de la génomique, reprend les codes et les usages en vigueur dans la classification des organismes cellulaires. Ainsi, la présente description des différentes familles historiques de NCLDV ouvre le débat sur l'existence ou non d'une origine commune à ces différentes familles virales (*i.e.* constituer un groupe monophylétique) (Figure 10).

Figure 10: Phylogénie des familles historiques de NCLDV. L'arbre a été construit à partir des séquences de protéines majeures de capside. La reconstruction a été faite par méthode du maximum de vraisemblance en utilisant MEGA746 Figure d'après Matsuyama & al [28].

Premièrement, nous pouvons noter que les cycles infectieux ne sont pas les mêmes pour toutes ces familles de NCLDV. Les différences majeures sont à observer dans la dépendance vis à vis du noyau de l'hôte. Les *Poxviridae* sont entièrement indépendants du noyau. Les *Iridoviridae* et *Asfarviridae* en sont partiellement dépendants pour l'expression des gènes précoces dont l'ARN polymérase virale et la ADN polymérase. A contrario, les *Phycodnaviridae* ne possèdent pas d'ARN polymérase [19], et sont dépendants du noyau de l'hôte durant l'intégralité du cycle infectieux (Figure 11).

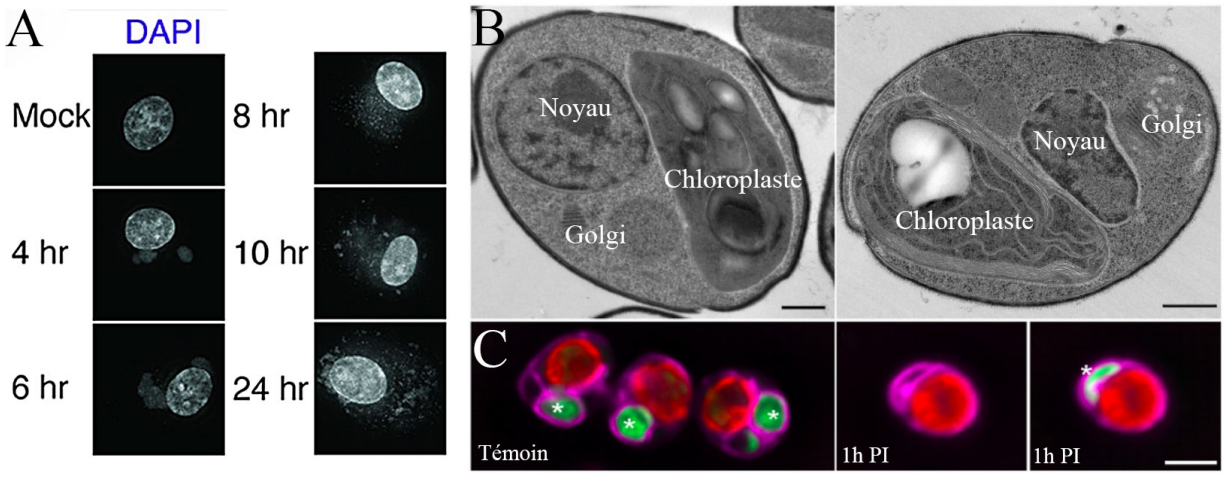


Figure 11: Suivi du noyau de la cellule hôte au cours du cycle infectieux de Paramecium bursaria chlorella virus (PBCV-1) et Vaccinia virus (VACV) par microscopie optique et electronique. (A) Le marquage du noyau cellulaire au DAPI montre que l'intégrité du noyau est maintenu tout au long du cycle infectieux de VACV. (B) Image de microscopie électronique de Paramecium bursaria chlorella non infectée (gauche) et 1h post-infection par PBCV-1. On observe une déformation du noyau. (C) Suivi par fluorescence du noyau (asterix), on observe après 1h d'infection une déformation du noyau. Figure d'après Kieser & al [16] et Milrot & al [19].

Deuxièmement, la structure des virions n'est pas homogène chez tous les NCLDV. La structure icosaédrique semble être un élément commun aux *Iridoviridae, Asfarviridae* et *Phycodnaviridae,* tandis que la structure en sac des *Ascoviridae* et en donut des particules de *Poxviridae* semble singulière. La présence d'une protéine majeure de capside (MCP) ayant un motif double tonneau beta chez l'ensemble de ces virus supporterait une histoire évolutive commune [33].

Troisièmement, malgré la présence de 8 groupes de protéines homologues (orthogroupe) chez tous les représentants des NCLDV historiques (parmi lesquels on retrouve : une hélicase, une *packaging* ATPase, un Facteur de transcription, la MCP et une protéine membranaire) il n'existe pas suffisamment d'arguments clairs validant ou réfutant la possibilité d'une origine commune à ces 5 familles de NCLDV sur la base de l'étude d'un génome coeur aussi restreint.

Nous verrons dans le reste de cette introduction en quoi la découverte des virus géants et leur classification au sein des NCLDV contribue à alimenter le débat en faveur d'une origine polyphylétique de ces derniers. Le travail réalisé durant ma thèse permet également de mettre en lumière pourquoi l'étude comparative de virus géants anciens avec des représentants récents permet d'alimenter ce débat.

Découverte des virus géants : Mimivirus.

C'est avec le postulat de Lwoff et l'avènement de la virologie comme domaine à part entière de la Microbiologie que le travail historiographique de Mirko D. Grmek s'arrête. Depuis, entre les années 60 et 90, l'analyse systématique des fractions virales (*i.e* filtrat de taille inférieur à 200 nm) a permis la découverte de nombreux virus, Ebola (1976), VIH (1983),... Les travaux récents ayant conduits à la découverte des virus géants permettent d'ajouter à cette histoire un cinquième volet. A l'image des conclusions tirées plus haut, la découverte des virus géants a été permise par une conjoncture d'événements, combinant développement technique et avancées des connaissances. L'objectif de ce paragraphe est de décrypter les étapes clefs qui ont conduit à rentrer dans cette nouvelle ère de la virologie.

En 1992, lors d'une épidémie nosocomiale de pneumonie à l'hôpital de Bradford en Angleterre, le microbiologiste du laboratoire de santé publique de Leeds, Tim Rowbotham, se penche sur le système de refroidissement de l'hôpital pour y chercher l'agent infectieux responsable de cette épidémie. Pour identifier les bactéries pathogènes du genre *Legionnella* il utilise en guise d'appât des protozoaires du genre *Acanthamoeba,* sensibles aux infections par les légionelles. C'est ainsi que Tim Rowbotham pense avoir isolé dans une amibe l'agent responsable de la pneumonie. Il observe au microscope optique le parasite intracellulaire et après coloration de Gram en déduit que ce dernier est un coque Gram positif d'un genre nouveau baptisé alors *Bradford coccus* [34] (Figure 12). La caractérisation de *Bradford coccus* est restée longtemps impossible, l'amplification permissive de l'ARN ribosomique 16S en routine, permise par les nouvelles techniques de séquençage, n'y fait rien. Il faut attendre 2003, soit 11 ans plus tard, pour que des images de microscopie électronique soient réalisées qui sèment le doute sur la nature de ce parasite. Ces observations révèlent que l'agent infectant *Acanthamoeba* pourrait ne pas être une bactérie mais un virus icosaédrique à la forme sans équivoque (ref). Les dimensions exceptionnelles de la capside (près de 400nm et 750nm en considérant les fibres entourant la capside), laissent perplexe les chercheurs. Comment envisager l'existence d'un virus suffisamment géant pour être visible au microscope optique ? Qu'en est-il de la définition historique de Lwoff ? Les interrogations soulevées par ces observations trouvent des réponses dans des résultats obtenus par l'utilisation des outils de séquençage de nouvelle génération. En 2004, l'expertise du laboratoire Information Génomique et Structural (IGS) en matière d'analyse des génomes, permet d'apporter la preuve finale sur la nature virale de ce nouveau microorganisme [35]. La taille impressionnante du virion, rendant ce dernier visible au microscope optique, conduit les équipes du projet à baptiser ce virus *Mimivirus* pour *microbe mimicking virus*. Les preuves apportées par les données de séquençage : absence de

gènes du métabolisme énergétique, pas de ribosomes, présence de gènes conservés par tous les NCLDV, satisfont la définition stricte énoncée par Lwoff un demi-siècle auparavant tout en apportant une correction fondamentale : le critère de taille implicite ne semble pas être une propriété intrinsèque des virus.

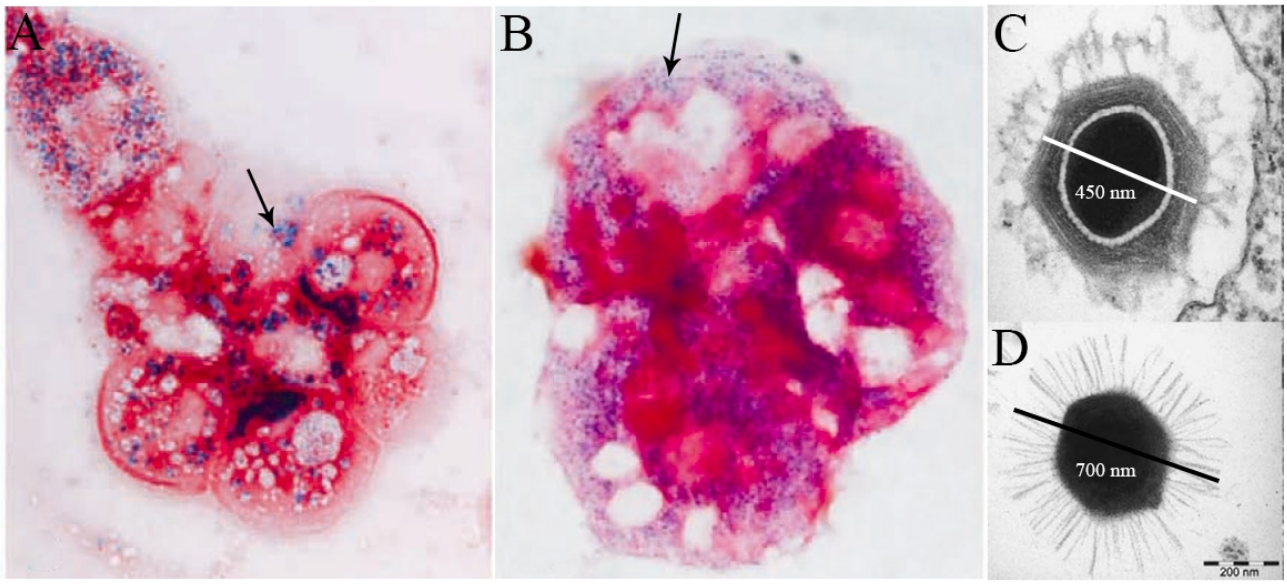


Figure 12:Parasites intra-cellulaires stricts d'Acanthamoeba polyphaga. (A) Parachlamydia acanthamoeba. (B) Mimivirus, initialement décrit comme une bactérie d'un genre nouveau baptisé Bradford coccus, visible par microscopie optique. (C) Image de microscopie électronique de Mimivirus laissant supposer la nature virale de ce parasite intracellulaire. (D) Particule virale de Mimivirus observée par microscopie électronique à transmission, on peut noter la taille des fibres recouvrant la particule. Figure d'après Raoult & al [34].

L'analyse du génome de *Mimivirus* a également ébranlé d'autres dogmes bien établis. Son génome d'une taille avoisinant 1.2Mb, et codant pour plus de 1000 protéines (Figure 13) est largement plus complexe que les plus gros virus connus alors, parmi lesquels les *Chloroviridae*.

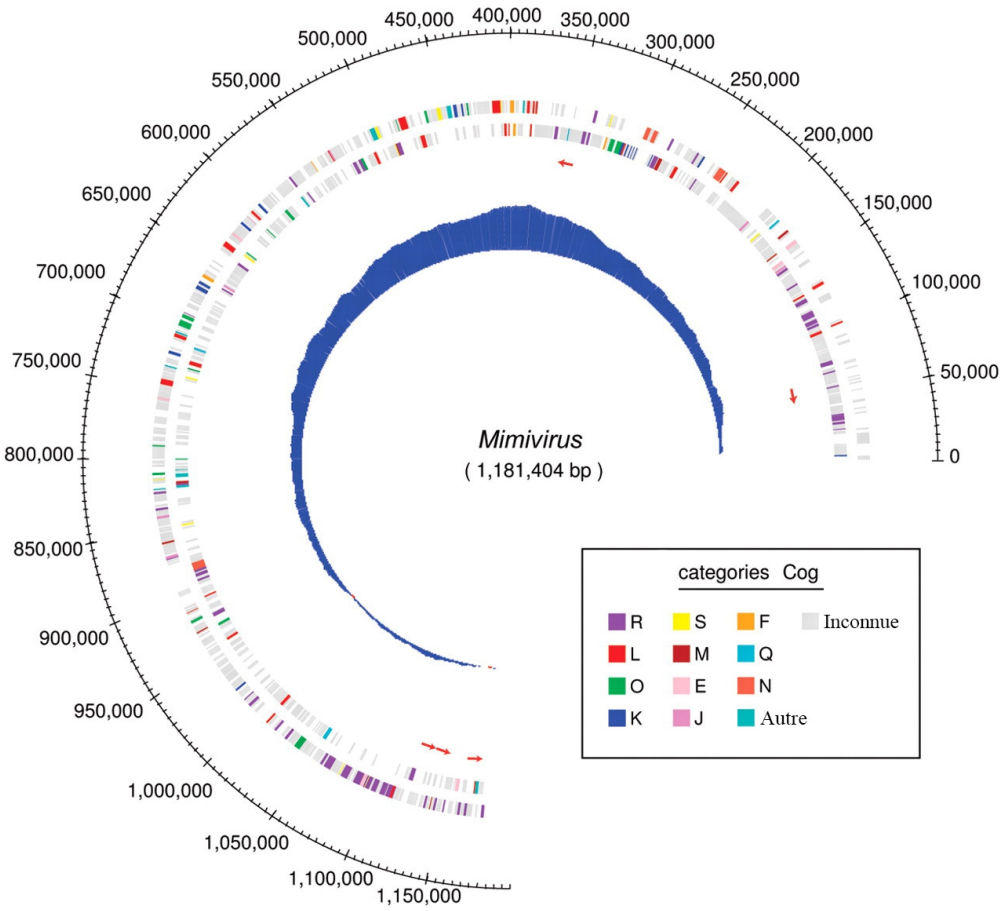


Figure 13:Carte du chromosome de Mimivirus où les séquences qui codent pour des protéines sont indiquées sur les deux brins du génome : les séquences sont colorées selon leurs catégories fonctionnelles. Les séquences colorées en gris sont celles dont la fonction est inconnue. Les flèches rouges indiquent l'emplacement et l'orientation des ARN de transfert (ARNt). Figure d'après Raoult & al [35].

L'analyse des gènes a d'une part révélé la présence des 9 gènes du génome coeur des NCLDV, conduisant de fait à proposer les *Mimiviridae* comme une nouvelle famille de grands virus nucléocytoplasmiques. Parmi les protéines prédites détectées chez *Mimivirus,* la moitié n'ont pas d'homologues connus dans les bases de données (ORFan [36]), laissant envisager des structures tertiaires de protéines et des fonctions encore inconnues. Parmi les protéines ayant des fonctions connues on retrouve, en plus des protéines impliquées dans divers voies métabolique (glycosylation, lipides, acides aminés), la réplication, la réparation de l'ADN et la transcription. Il s'est avéré que *Mimivirus* code également pour 4 aminoacyl-ARNt ligases (Arg, Tyr, Cys, Met), des facteurs d'initiation, d'élongation et de terminaison de la traduction. La présence de gènes impliqués dans la traduction a alors surpris la communauté scientifique. Cette découverte fait de *Mimivirus* le premier virus à coder pour une partie de la machinerie de traduction, étendant largement le répertoire des gènes viraux connus jusqu'alors et venant poser une nouvelle limite à la définition de Lwoff

concernant l'absence de traduction chez les virus. Enfin, en présentant une complexité génétique proche de celle de certaines bactéries intracellulaires comme les *Rickettsia* ou les *Mycoplasma, Mimivirus* remet en question la frontière arbitraire, relative à la complexité génétique, qui existait entre le monde viral et monde cellulaire (Figure 14).

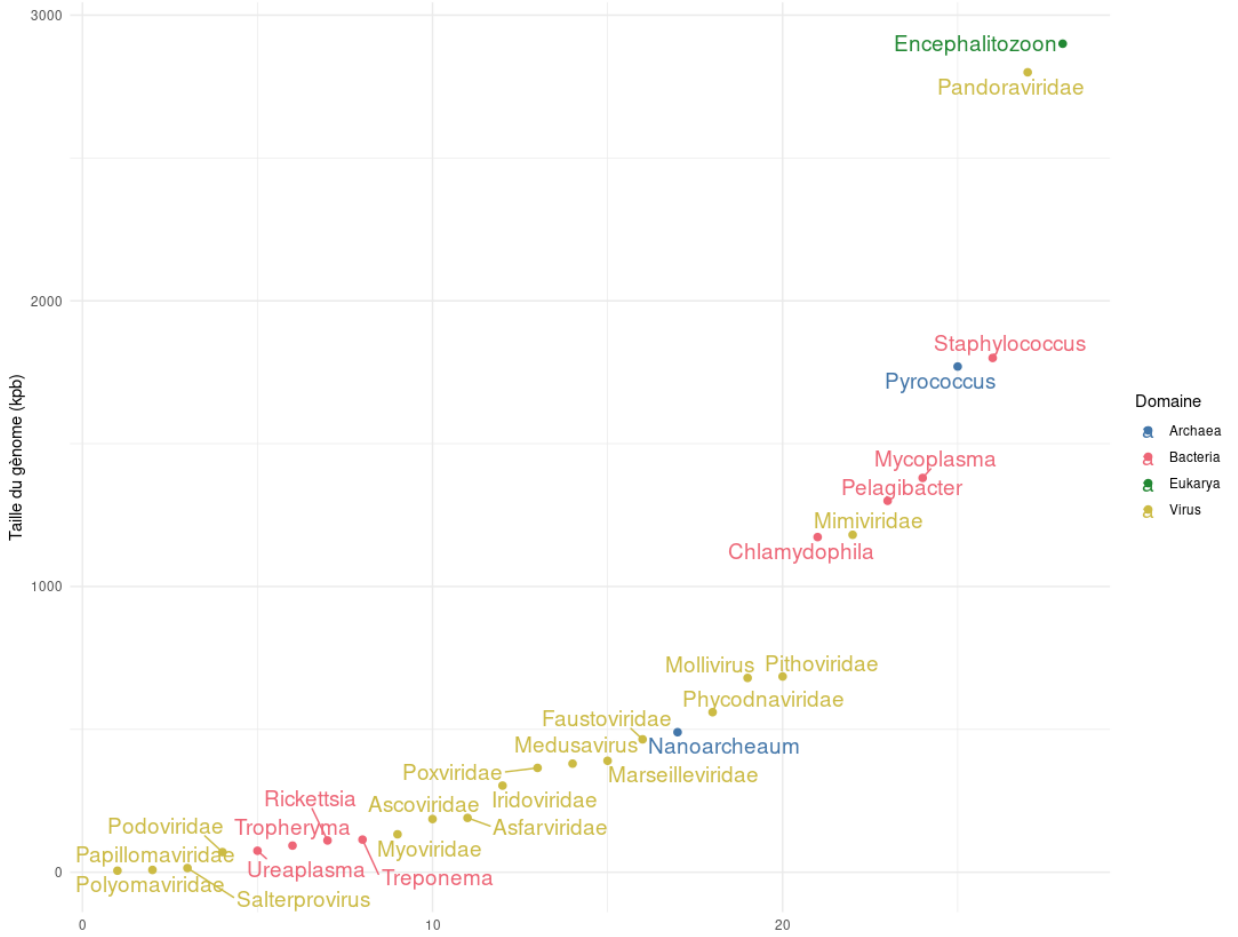


Figure 14: Continuum entre les différentes tailles de génome de plusieurs entités biologiques. Cette figure démontre qu'il n'existe plus de barrière basée sur la complexité des génomes entre ces différentes classes d'entités.

Pour finir, cette découverte ouvre une porte sur un nouveau pan du monde viral inconnu : les virus géants infectant des eucaryotes unicellulaires.

Ecologie et virus géants, enjeux et étude du permafrost

Comme nous avons vu dans les paragraphes précédents, la recherche de nouveaux virus a été, jusqu'au début du 21ème siècle, largement liée à des questions sanitaires et économiques. Ce sont les avancées de la génomique environnementale permises par les nouvelles techniques dites « omiques », qui sortent la virologie progressivement du carcan épidémiologique. Le caractère révolutionnaire de la découverte de *Mimivirus* pousse Jean-Michel Claverie et Chantal Abergel à

explorer ce champ inconnu ouvert par *Mimivirus* en explorant divers environnement via la collecte d'échantillons d'eau et de sédiments notamment. Avec le temps, l'orientation du laboratoire IGS vers la découverte de nouvelles familles de virus géants a permis de développer un savoir-faire unique dans la collecte d'échantillons environnementaux, notamment d'eau de mer, d'eau douce, de sédiment, de terre et enfin de pergélisol, ainsi que dans les protocoles de réactivation de virus géants. Nous allons voir dans ce paragraphe que l'étude de la diversité virale de milieux inexplorés répond à divers enjeux scientifiques.

La mer couvrant 70% de la surface terrestre, l'engouement pour la virologie environnementale a commencé par l'étude de cet environnement. En 2004, par une campagne très médiatisée de collecte et d'analyse d'échantillons d'eau de mer par métagénomiqe baptisée *Global Ocean Sampling*, le biologiste Craig Venter se lance dans la première étude du « microbiome » marin à grande échelle. En guise de test, une première série échantillonnage pilote a été réalisée par les équipes de Craig Venter en mer des Sargasses. L'analyse des données de séquençage avait alors permis de caractériser *in* silico la présence de séquences proches de *Mimivirus* dans les échantillons d'eau de mer [37]. C'est avec les prélèvements réalisés au cours de la mission *Tara Oceans*, rassemblant un consortium international, qu'est lancé la première analyse et caractérisation fonctionnelle de la biodiversité des écosystèmes planctoniques des océans à grande échelle. L'analyse récente des échantillons de séquences ADN provenant des échantillons de *Tara Ocean* a confirmé que les *Mimiviridae,* infectant probablement des haptophytes, sont largement representés dans les océans et jouent un rôle clef dans le processus par lequel les organismes océaniques transfèrent le carbone des eaux de surface à l'intérieur des océans et aux sédiments des fonds marins[38] (Figure 15).

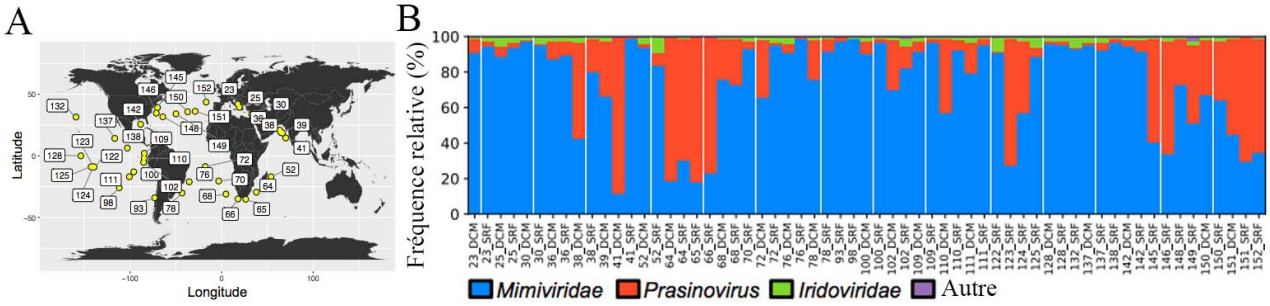


Figure 15: Abondance des NCLDV associés à l'exportation du carbone océanique. (A) Localisation des 40 points d'échantillonnage de surface de la mission Tara Ocean à l'origine des 61 métagénomes. (B) Abondance relative de diverses familles de NCLDV dans les échantillons. On observe que les Mimiviridae sont largement majoritaires dans les échantillons. Figure d'après Blanc-Mathieu & al [38].

En parallèle, en 2011, une équipe de l'IGS, ayant bénéficié du programme européen ASSEMBLE ("Association of European Marine Biological Laboratories"), se lance dans une

campagne d'échantillonnage d'eau de mer sur les côtes chiliennes, première expédition exclusivement dédiée à la recherche de nouveaux virus géants marins.

Dans la base de référence mondiale pour les ressources en sol, le cryosol est le sol des régions froides contenant du pérgelisol. Le pergélisol est une couche de sol, de sable ou de roche plus ou moins profonde dont la température est inférieure à 0°C durant plusieurs années, on en retrouve de trois types : continu, discontinu ou sporadique. L'intérêt pour l'étude de ces cryosols et celle des virus infectieux qu'ils renferment est double. La recherche de nouvelles familles virales et l'étude de virus géants anciens, la crainte grandissante dans la société vis à vis du réchauffement climatique et des conséquences de la fonte du pergélisol [39] [40]. S'intéresser aux virus du permafrost c'est contribuer à percer le secret de l'origine des virus géants et anticiper la nature des potentiels agents infectieux potentiellement libérés par la fonte du pergélisol.

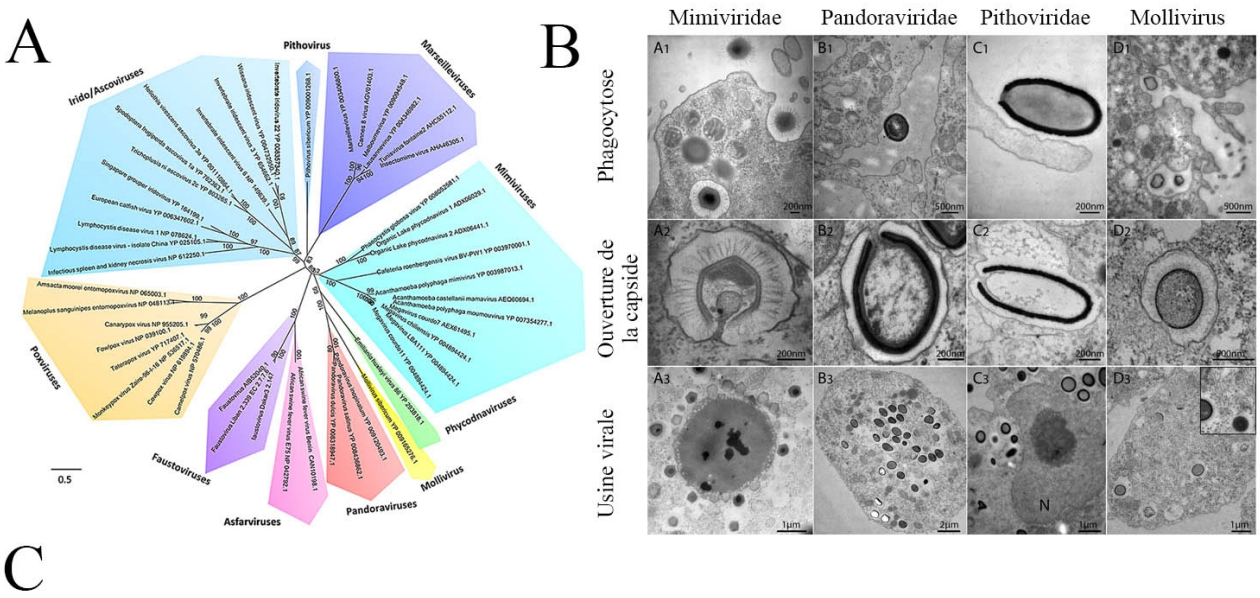
En 2010, le laboratoire s'est penché sur les travaux conduits par les équipes du laboratoire de David Gilichinsky de l'Académie des Sciences de Russie. Ces études commencées il y a plus 30 ans, et peu connues du grand public, ont fourni de multiples démonstrations que les sols gelés de Sibérie contiennent une grande diversité de microorganismes cultivables, conservant leur capacité infectieuse après plusieurs dizaines de milliers d'années passés dans le pergélisol [41] [42]. Point d'orgue de ces recherches, les équipes de David Gilichinsky ont réactivé en 2012 une plante fertile entière du genre *Silene* (famille des *Caryophyllaceae*) à partir de tissus de fruits maternels immatures datant du Pléistocène supérieur (126000-11700 ans) [43]. Sur le plan fondamental ces travaux ont conduits le laboratoire à entamer une collaboration avec les équipes du Pr Elizaveta Rivkina du laboratoire d'expertise des cryosol de l'Académie des Sciences de Russie basé à Pushino, via l'obtention d'un appel conjoint à un projet de recherche collaboratif (PRC) avec la Russie. L'objectif est clair : savoir si, au même titre que des eucaryotes et des procaryotes, il existe dans le pergélisol des virus infectieux.

En plus du caractère exploratoire que constitue la recherche de nouveaux virus géants, l'étude du pergélisol constitue un enjeu sanitaire majeur. L'amplification par PCR de gènes (B7R, A30L et E9L) du virus de la variole (*Siberian smallpoxvirus)* dans des échantillons pulmonaires de momies gelése enterrées il y a 300 ans en Yakoutie, Sibérie [44] confirme que les capacités cryoprotectrices du pergélisol permettent le maintien de fragments d'ADN viraux dans le sol gelé. Les épidémies récentes d'anthrax provoquées par des souches de *Bacillus anthracis* dont les spores conservées dans le pergélisol des régions de Yakoutie (2015) et de la péninsule de Yamal (2016) [45], le pergélisol constitue donc un réservoir encore largement inconnu d'espèces microbiennes cultivables dont certaines pathogènes et conservant leur potentiel infectieux après plusieurs milliers

d'années passés dans le pergélisol. En plus d'être un réservoir de micro-organismes pathogènes, l'étude et le séquençage de souches de *Staphylococcaccae* psychrophiles (*Staphylococcus warneri* et *hominis*) réactivées à partir d'échantillons de pergélisol du Miocène (15,97 millions d'années à 11,6 millions d'années) ont permis la découverte de nouveaux gènes de résistance à plusieurs familles d'antibiotiques : aminoglycosides, beta-lactimes, macrolide, lincosamide, streptogramine B et chloramphénicol [46]. Ces récentes découvertes font donc du pergélisol un réservoir de gènes encore inconnus dont certains pourraient présenter un intérêt médical. C'est dans la continuité de ces questionnements que s'inscrit le présent sujet de thèse.

Les différentes familles de virus géants

La découverte historique de *Mimivirus* a permis d'entrevoir un pan entier du monde viral jusqu'alors masqué par une pratique arbitraire deux fois centenaire, à savoir la recherche de virus uniquement à partir de la fraction virale. Le travail exploratoire réalisé au laboratoire IGS, grâce à la lignée de protozoaires du genre *Acanthamoeba castellanii,* a contribué à la découverte de 5 nouveaux types de virus géants, *i.e* de virus visibles par microscopie optique, conduisant à proposer 4 nouvelles familles de NCLDV aux caractéristiques propres variées : structure de la capside, taille du génome, mode de réplication [47] [48] (Figure 16)... Pour la suite du manuscrit, je tiens à préciser un élément important : je ne parlerai ici que des virus isolés, laissant donc volontairement de côté les génomes viraux assemblés par méta-génomique.



C

Famille	Virion	Taille de la particule (nm)	Isolats	Taille du génome (kb)	GC%	Cycle
Mimiviridae	Icosaédrique	755	>20	1600-370	25-41	Cytoplasmique
Marseilleviridae	Icosaédrique	200	>20	390-360	43	Nucleaire et cytoplasmique
Pandoraviridae	Amphore	1000x500	>10	2800-1850	59-60	Nucleaire et cytoplasmique
Pithoviridae	Amphore	2000x500	5	685-575	38	Cytoplasmique
Mollivirus	Spherique	600	1	650	60	Nucleaire et cytoplasmique

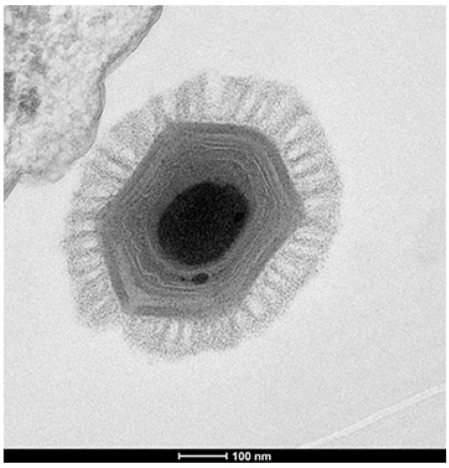
Figure 16 : Les différentes familles de virus géants. (A) Reconstruction phylogénique basée sur l'ARN polymérase ARN-dépendante des NCLDV. Figure d'après Aherfi & al. (B) Images de microscopie électronique présentant différentes étapes des cycles infectieux de différentes familles de virus géants. Figure d'après Abergel & al [47]. (C) Caractéristiques des grandes familles de virus géants connus [48].

La famille des *Mimiviridae* s'agrandit

A la suite de la campagne d'échantillonnage à la station marine de Las Cruces au Chili, la mise en culture en présence d'amibes du genre *Acanthamoeba castellanii* du retentât d'un litre d'eau de mer filtrée sur membrane 0,2 µm a permis la réactivation d'un nouveau virus géant icosaédrique, cousin éloigné de *Mimivirus,* baptisé *Megavirus chilensis* [49]. La découverte de ce nouveau virus géant tout aussi exceptionnel a fini de convaincre la communauté scientifique que *Mimivirus* n'était pas un «phénomène de foire» isolé mais au contraire que les *Mimiviridae* étaient vraisemblablement une nouvelle famille virale abondante dans différents types d'environnements. Ainsi, *Megavirus chilensis* légèrement plus complexe que *Mimivirus* code pour 7 aminoacyl-ARNt-ligase, soit un héritage de gènes impliqués dans la traduction encore plus important que pour *Mimivirus.* Plus étonnant encore, bien que les deux génomes soient colinéaires en leurs centres, le contenu génique des deux virus diffère. Ils contiennent 20% de gènes qui leurs sont propres, 80% desquels n'ayant pas d'homologues dans les bases de données publiques. Ces gènes, propres à chaque familles virale

et n'ayant pas d'homologues disponibles dans les bases de données publiques sont appelés ORFans (Figure 17). Ces gènes codant pour des protéines inconnues constituent donc un réservoir de nouvelles fonctions potentielles.

A



B

Virus	Gènes	Gènes codants protèines homologues	Gènes spécifiques	ORFan strict	ARNt ligase
Mimivirus	979	594	199	186	4
Mégavirus chilensis	1120	594	268	258	7

Figure 17: Caractéristiques de Megavirus, second virus géant découvert et proche parent de Mimivirus. (A) Image de microscopie électronique montrant la structure icosahedrique de Megavirus chilensis (©Chantal Abergel, IGS). (B) Tableau présentant les différences entre Mimivirus et Megavirus chilensis. On observe que plus de la moitié des gènes sont partagés entre les deux virus (594), parmis les gènes strictement uniques à chacun des deux virus 186 gènes de Mimivirus n'ont pas d'homologues dans les bases de données (ORFan strict) et 258 pour Megavirus chilensis.

Megavirus chilensis a ouvert la voie à la découverte de nombreux autres virus géants de la famille des *Mimiviridae,* élargissant d'autant le spectre d'hôte de ces virus. Comme pour les autres familles de NCLDV, il a été proposé de subdiviser les *Mimiviridae* en plusieurs sous-familles dépendant de leurs hôtes respectifs : les *Megavirinae* infectant des protozoaires du genre *Amoebozoa,* les *Mesomimivirinae* infectant des algues uni-cellulaires et les *Klosneuvirinae* dont l'unique représentant est *Bodo saltans virus* et infectant un protozoaire du même nom [50] (Figure 18).

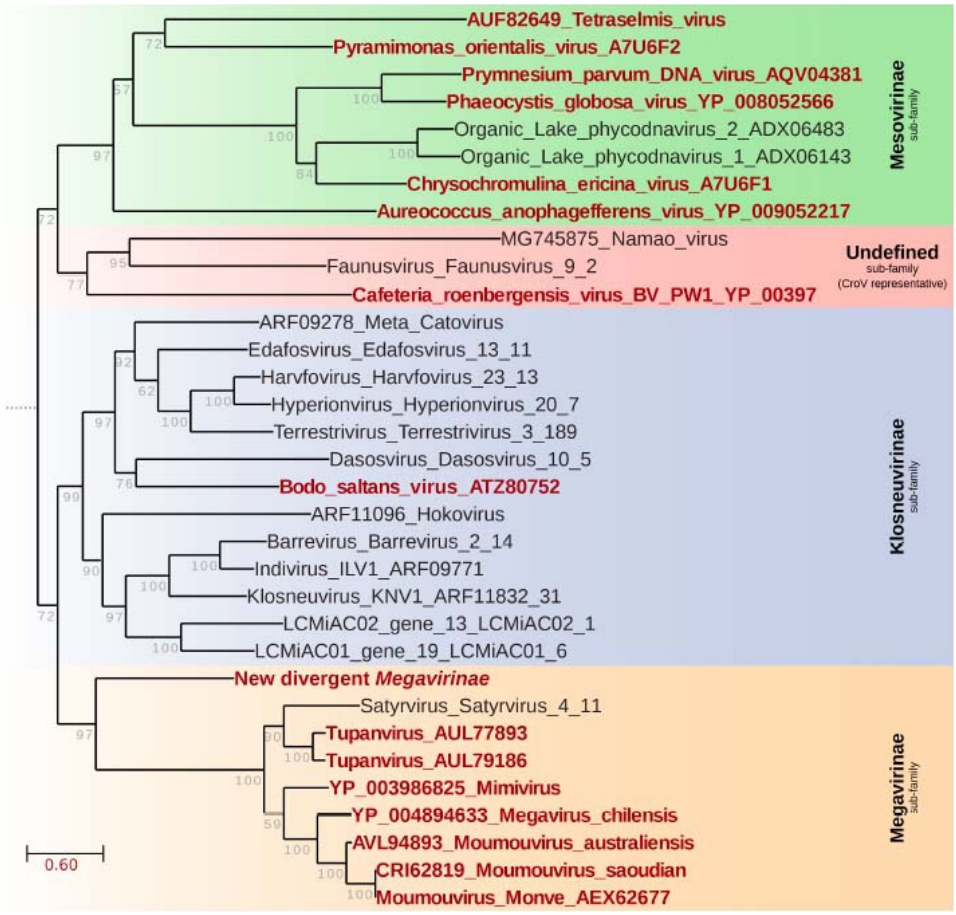


Figure 18: Phylogénie des Mimiviridae. L'arbre phylogénétique est basé sur la séquence de l'ADN polymérase (modèle de substitution LG+F+R5). Dans ce manuscrit nous parlerons des virus isolés et dont le génome est séquencé (en rouge), les autres virus sont issus d'assemblages de séquences de métagénomique.

On s'intéressera ici aux 3 lignées de *Megavirinae* infectant des protozoaires du genre *Acanthamoeba* :*Mimivirus* (A), *Moumouvirus* (B) *Megavirus* (C), et auxquelles ont peu ajouter depuis 2018, *Tupanvirus* [51] *(Figure 19)*.

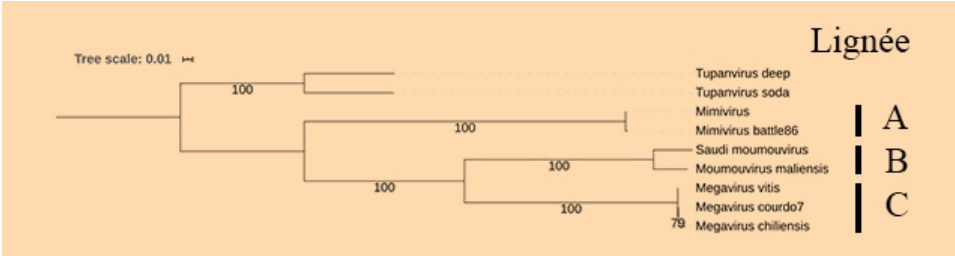


Figure 19: Phylogénie des Megavirinae. L'arbre phylogénétique est basé sur la séquence de l'ADN polymérase (modèle de substitution JTTDCMut+F+G4).

La structure des particules virales de ces *Mimiviridae* est similaire, à l'exception de *Tupanvirus* qui prolonge sa capside d'une queue tubulaire de 500 nm (Figure 20). On retrouve une

capside de symétrie icosaédrique d'un diamètre avoisinant 450 nm à 500 nm et recouverte d'une couche dense de fibres mesurant 125 nm à 150 nm de long (Figure 20). La structure de la particule virale est donnée par des protéines homologues à la protéine majeure de capside (MCP) L425 de *Mimivirus*. On retrouve deux membranes lipidiques entourant un nucléoïde de 340 nm de diamètre [52]. La faible densité de compaction de l'ADN viral (0,006nm^3/bp) suggère que ce nucléoïde pourrait contenir plusieurs copies du génome viral et/ou des protéines en abondance. A l'heure actuelle, la caractérisation de cette structure interne permettant la compactions de l'ADN viral est un axe de recherche prioritaire pour le laboratoire.

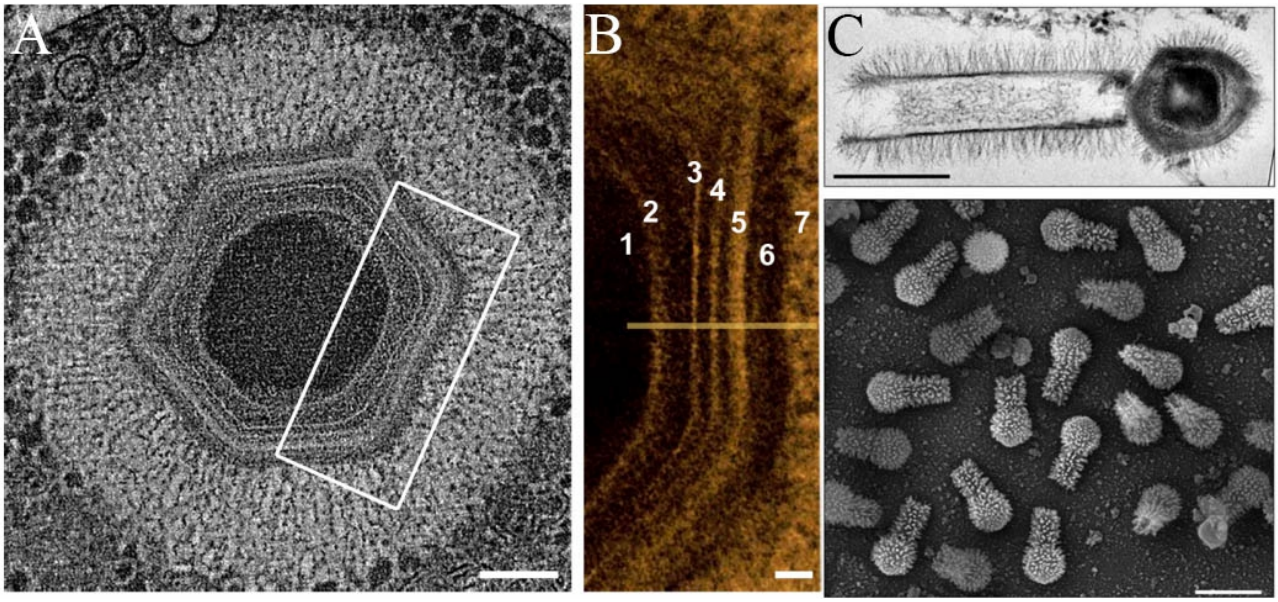


Figure 20: (A) Coupe de microscopie électronique à transmission d'une particule virale mature de Mimivirus dans le cytoplasme de l'amibe. (B) Détail des structures membranaires de la particule virale. Figure d'après Mutsafi & al [52]. (C) Deux photographies par microscopie électronique de particules virales de Tupanvirus, présentant une capside similaire à Mimivirus mais ayant une queue tubulaire de 500 nm, en haut microscopie à transmission et en bas microscopie à balayage.Figure d'après Abrahão & al [51].

Les fibres qui entourent la capside virale sont essentiellement composées de trois protéines, R135, L725 et L829, dont deux sont glycosylées (R135 et L829), et deux sont de fonction inconnue (L725 et L829). La caractérisation récente de viosamine acétylée dans la composition des fibres laisse supposer que ce sucre, composant notamment la couche lipopolysaccharidique de certain *Escherichia coli*, permet à la particule virale d'adhérer à *Acanthamoeba castellanii* et d'être phagocyté par l'amibe au même titre qu'une bactérie [53] [54].

Le cycle viral d'une durée avoisinant 10h se déroule en plusieurs étapes [55] : après avoir été phagocyté la particule virale s'ouvre grâce à une structure tétragramme, baptisée « startgate » et permettant la fusion de la membrane interne du virion avec la membrane du phagosome [56] (Figure 21). Le canal ainsi crée relie l'intérieur de la particule virale avec le cytoplasme permettant ainsi au

« nucleoide » d'être transféré dans le cytoplasme de l'amibe. La transcription précoce des gènes viraux, ainsi que la maturation des transcrits viraux, est initiée par les protéines contenues dans la particule. Une fois cette phase précoce terminée, une zone d'exclusion des organites de l'hôte se forme et permet la mise en place d'une « usine virale ». Cette zone non membranaire concentre l'activité des protéines virales et amibiennes, permettant la réplication du génome viral [57]. C'est également le lieu d'un trafic important de membranes lipidiques, le recyclage du réticulum endoplasmique de la cellule hôte est ainsi utilisé pour la synthèse des membranes internes des particules neosynthétisées. Une fois les capsides assemblées, le génome viral est compacté et chargé dans la particule via une ouverture dans le virion [52 58]. La dernière étape d'assemblage, permettant la maturation des virions, est la fixation de la couche fibrillaire en périphérie des capsides [59]. On estime la productivité des cycles infectieux de *Mimivirus* à 1/1000. La dissémination dans le milieu extra-cellulaire des virions est ensuite permise par la rupture de la membrane cellulaire amibienne.

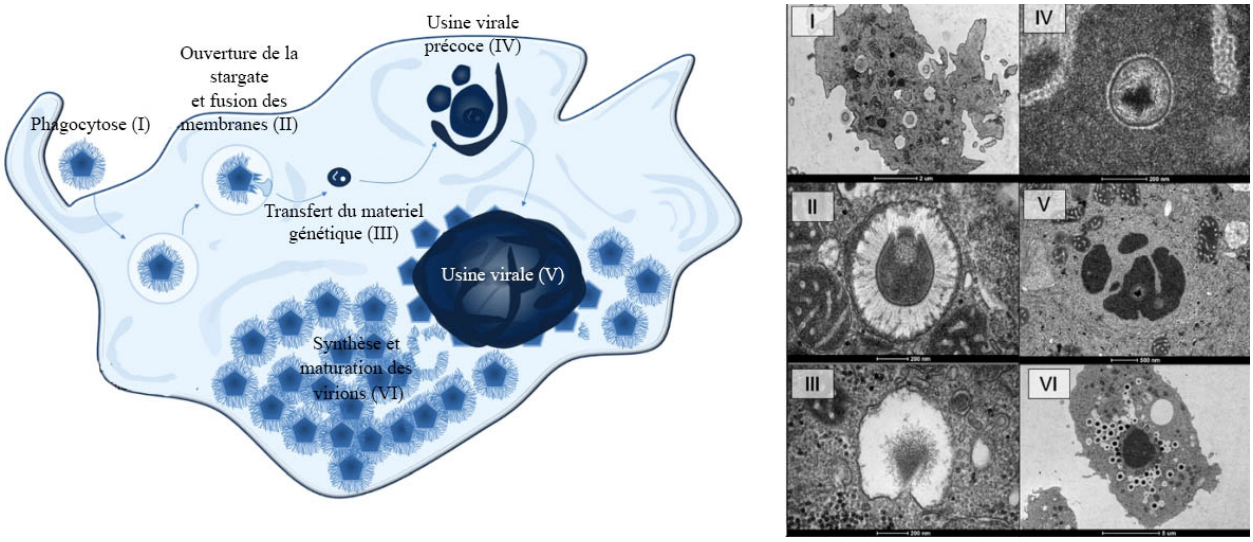


Figure 21 : Schéma du cycle infectieux de Mimivirus. (I) Phagocytose. (II) Entrée du virus dans l'amibe, suivie de l'ouverture de la stargate (III). (VI) Le contenu de la capside virale est libéré dans le cytoplasme de l'amibe, permettant la mise en place de l'usine virale précoce (V). Après quelques heures, l'usine virale permet l'assemblage de virions (VI). A droite, images de microscopie électronique illustrant les diverses étapes. Figure d'après Abrahão & al [55].

La découverte de nouveaux *Mimiviridae* a permis en 2008 d'isoler au cours d'une infection d'*Acanthamoeba castellanii,* un virus possédant une capside icosaédrique d'environ 50 nm de diamètre qui coexistait aves les capsides de virus géant. C'est ainsi qu'a été découvert le premier virus de virus infectant l'usine virale [60 61]. Son ADN double-brin circulaire de 18 kb, code pour 21 protéines dont 13 ORFans. Ce virophage, baptisé *Sputnik*, entraine des changements phénotypiques des capsides de *Mimivirus* ou *Mamavirus :* virions déformés, infectiosité diminuée. Depuis, la recherche de virophage est systématique lors de la découverte de nouveaux *Mimiviridae* (Figure 22). Ainsi, a été découvert *Zamilon*, un virophage infectant uniquement les *Mimiviridae* de lignée B

et C. En 2012, un nouvel élément mobile a été découvert. D'une taille de 7 kpb ce fragment d'ADN linéaire codant pour 6 à 8 protéines a été baptisé transpoviron. Cet élément génétique mobile se réplique dans les usines virales de *Mimiviridae* et, en plus d'une forme libre utilisant le virophage comme outil de dissémination, est capable de s'intégrer dans le chromosome du virophage. Cette découverte vient à nouveau flouter les frontières entre monde vivant et monde viral. Les effets de dominance entre réplication du virus géant, réplication du virophage et réplication du transpoviron, ont été récemment décrits et il apparaît que le trio transpovirion/virophage/virus géant constitue le premier cas de commensalisme dans le monde viral [62].

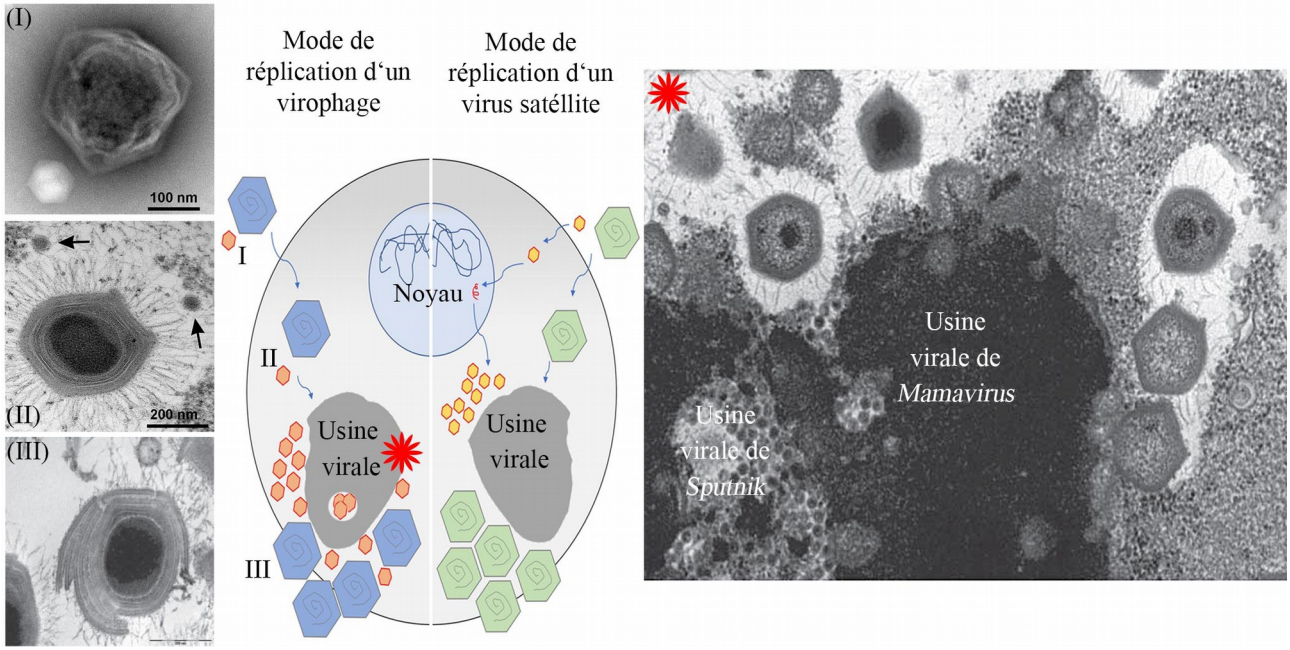


Figure 22 : Détails du cycle infectieux du virophage Sputnik infectant l'usine virale de Mamavirus. (I) Entrée du virophage en même temps que le virus géant auquel il est associé. L'image de microscopie par coloration négative indique les différences de tailles entre les deux virus. (II) Virus géant et son virophage à l'intérieur de l'amibe. D'après Duponchel & al. (III) Particule virale de Mamavirus aberrante suite à une infection de l'usine virale par Sputnik.. Le schéma permet de se rendre compte des différences fondamentales entre virus satellite et virophage. Le virus satellite ne dépend pas de la machinerie cellulaire et virale et n'est donc pas sensus-stricto un virus de virus. D'après La Scola & al [60].

Pour conclure, la découverte de nouveaux représentants des *Mimiviridae* a permis de faire accepter que certaines familles de virus géants étaient ubiquitaires. La découverte d'une partie de la machinerie de traduction chez les *Megamimiviridae,* et la découverte récente de Yasminevirus (ref) codant pour un jeu quasi-complet de 20 aminoacyl-ARNt ligases implique une gradation dans le concept de parasitisme. De plus, si Sputnik répond aux critères de Lwoff, comment considérer l'usine virale, sensible aux infections d'une classe de virus plus petits ?

Marseilleviridae, le plus petit des géants

Au sens strict, les *Marseilliviridae* sont à la frontière de la définition des virus géants. La taille de leur capside icosaédrique d'environ 250nm les rendent à peine visibles par microscopie optique. Le premier prototype de cette famille virale a été isolé en 2007, à partir d 'échantillons en provenance d'une tour de refroidissement parisienne. Depuis la découverte du prototype baptisé *Marseillevirus*, 5 lignées ont été proposées :A, B, C, D et E (Figure 23).

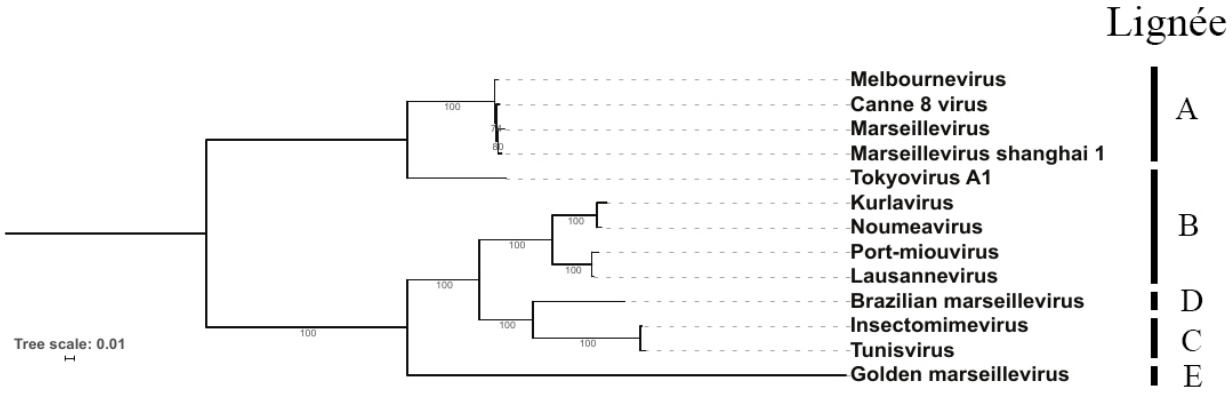


Figure 23 : Phylogénie des Marseilleviridae. L'arbre phylogénétique est basé sur la séquence de l'ADN polymérase (modèle de substitution LG+G4).

Le cycle infectieux des *Marseilleviridae* d'une durée avoisinant 8h se déroule en plusieurs étapes. La petite taille des particules conduit à l'internalisation des virions par trois moyens différents : phagocytose de vésicules contenant jusqu'à 1000 particules, phagocytose de particules groupées et endocytose de virions isolés [63] (Figure 24). Après 30 min, les particules virales internalisées sont probablement digérées dans des endosomes. Par rupture probable de la membrane de l'endosome le nucléoïde est transféré dans le cytoplasme de l'amibe. On observe durant les deux premières heures de l'infection une désorganisation du noyau et du nucléole de la cellule hôte [64]. Durant cette phase d'éclipse, se forme une zone d'exclusion dans le cytoplasme où est recrutée l'ARN polymérase amibienne. Ce recrutement en phase précoce du cycle est nécessaire pour permettre l'expression de l'ARN polymérase virale, absente de la particule et au delà de 2h post infection, l'intégrité du noyau est rétablie et l'usine virale concentre l'activité de transcription, traduction et de réplication du génome virale. L'intense recyclage membranaire permet la synthèse de nouvelles particules virales d'abord vides avant que soit embarqué le nucléoïde. Les virions neosynthetisés se retrouvent ensuite enveloppés dans des membranes issues du réticulum endoplasmique ou nus, entrainant un relargage des virions matures par lyse ou exocytose après 10h.

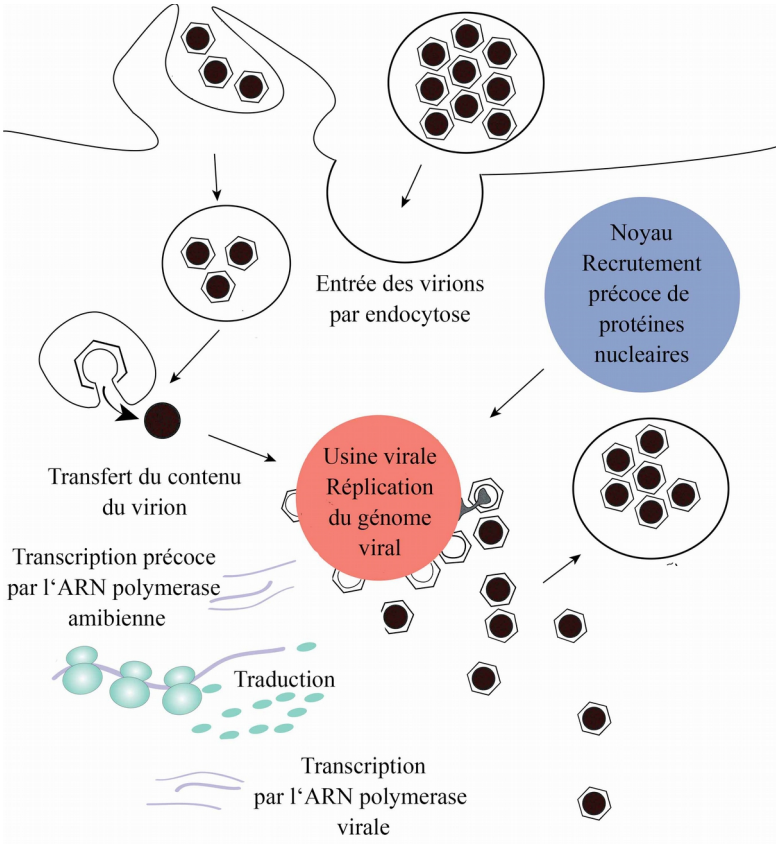


Figure 24:Réplication virale des Marseilleviridae. Figure d'après Abergel & al [47].

En 2005 a été observé dans des amibes isolées à partir d'eau de la Seine des parasites intracellulaires ronds et positifs à la coloration de Gimenez. Il a fallu attendre 6 ans pour que la nature virale de ce parasite intracellulaire soit finalement révélée. L'étude comparée de ce nouveau *Marseilleviridae* baptisé *Lausannevirus,* avec *Marseillevirus* a permis de constater une répartition non aléatoire de différentes classes de gènes le long du génome viral [65], en amont et en aval des régions conservées du génome de *Lausannevirus* et *Marseillevirus* (150kb en aval et en amont) semblent s'accumuler d'un coté les protéines hypothétiques, ORFan inclus, et de l'autre les protéines ayant un domaine prédit par analogies avec des protéines présentes dans des bases de données publiques (Pfam, Swissprot, TIGRfam et KEGG).

Sur le plan génétique, le génome des *Marseilleviridae* est plus petit que celui des *Mimiviridae,* d'une taille allant de 360 à 390kb et il a été récemment prouvé expérimentalement au laboratoire que les génomes de *Marseilleviridae* étaient circulaires, faisant donc de marseillevirus la première famille de virus géants au génome non linéaire. Ces virus codent environ 450 gènes parmi lesquels on retrouve les gènes impliqués dans le métabolisme de l'ADN, la transcription ainsi que 2 doublets d'histones : H2B-H2A, H3-H4-*like* (avec H des histones homologues à des histones eucaryotes). L'analyse phylogénétique de ses histones laisse penser que l'acquisition des histones

H2, H3 et H4-*like* par marseillevirus est antérieure à la duplication et la néo-fonctionalisation des variants eucaryotes H2A.Z et cenH3, spécialisés dans l'assemblage de l'hétérochromatine et dans le maintien des brins de chromatide. Laissant penser ainsi que les histones de marseillevirus ont prédaté l'apparition de la diploïdie chez les eucaryotes et les innovations génétiques qui ont suivi. [66].

Pour conclure, les marseillevirus apportent un éclairage particulier sur les potentielles origines des virus géants permettant d'expliquer la présence d'histone. Bien que très superficielle, l'analyse de la répartition des gènes viraux laisse supposer qu'il existe des mécanismes génétiques entrainant une répartition non aléatoire des gènes le long du génome.

Pandoraviridae, géants parmi les géants

Il faut attendre 2013 pour que ce soit découvert le premier virus géant non icosaédrique. De façon concomitante, deux nouveaux virus ont été isolés en 2013, l'un à partir d'un échantillon d'eau douce en provenance d'un étang de Melbourne (Australie) et le second à partir de sédiments prélevés à l'embouchure de la rivière Tunken au Chili. La mise en culture d'amibes du genre *Ancanthamoeba castellanii* en présence de ces deux échantillons a conduit à l'observation de parasites intracellulaires ressemblant à des amphores et visibles par microscopie optique. Des observations similaires avaient déjà été faites en 2005 dans des amibes du genre *Acanthamoeba* isolées chez un patient atteint de kératite [67] et avaient conduits les auteurs à la conclusion que cette particule était un nouvel endosymbionte de nature inconnue [68]. C'est donc l'expérience acquise par la découverte de *Mimivirus* qui permet au laboratoire IGS d'identifier la nature virale de ce nouveau parasite intracellulaire infectant *Ancanthamoeba castellanii*. Baptisés *Pandoravirus dulcis* et *Pandoravirus salinus* en référence à la boite de Pandore (historiquement une amphore dans la mythologie grecque) et à la salinité de l'eau à partir desquels ils ont été isolés. Depuis, ce sont onze représentants supplémentaires qui ont été découverts, conduisant à proposer une nouvelle famille de NLCDV, les *Pandoraviridae*. La particularité commune à l'ensemble des pandoravirus est le gigantisme de leur capside ainsi que celui de leur génome (Figure 16). Si *Mimivirus* avait contribuer à flouter les frontières entre monde viral et monde cellulaire, *Pandoravirus* fini de les effacer. Ainsi, avec un génome d'une taille comprise entre 2 et 3 Mpb et une capside d'une taille de 1000 x 500 nm les pandoravirus dépassent largement la taille et la complexité des virus connus jusqu'alors et concurrençant même certains eucaryotes comme les Fungi du genre *Encephalitozoon (*Figure 14).

La forme et la structure des particules des *Pandoraviridae* soulèvent beaucoup d'interrogations. En effet, ce sont les premiers virus géants découverts non icosaédriques. La capside, en forme d'amphore est composée de plusieurs couches. La partie la plus externe est un

tégument de composition inconnue, vient ensuite une couche dense aux électrons d'aspect lamellaire, composé d'un réseau de fibres parallèles. Sous cette couche externe dense aux électrons apparaissent des structures ressemblant à un réseau membranaires qui viennent délimiter la zone interne du virion. Les capsides sont polarisées avec à l'une des extrémité un pore apical, scellé par un bouchon qui apparaît diffus par microscopie électronique. Étonnamment, les images de microscopie électroniquesne permettent pas de déterminer si il existe des structures internes à la capside. Seul élément interne de la particule notable, un point plus foncé à l'opposé du bouchon apical.

Le cycle infectieux des pandoravirus est d'une durée avoisinant quinze heures (Figure 25). La taille des virions laisse penser que les particules virales sont internalisées par phagocytose. On observe ensuite la perte du bouchon apical, entrainant une fusion de la membrane interne de la capside avec la membrane du phagosome, ouvrant un passage entre l'intérieur de la particule et le cytoplasme de l'hôte. En phase précoce de l'infection virale on observe des événements nucléaires dont une disparition progressive de la structure nucléaire. Une zone se forme alors dans le cytoplasme, où toutes les structures sub-cellulaires, comme les mitochondries, sont exclues : c'est l'usine virale. La formation des nouveaux virions est étonnante : l'intérieur, ainsi que l'enveloppe entourant la particule, sont assemblés simultanément, comme s'ils étaient "tricotés". Des particules à différents stades de maturité sont observées dans l'usine virale, dans le cytoplasme de l'hôte. Des particules matures peuvent être libérées dans le milieu extra-cellulaire par exocytose, mais la fin du cycle est marquée par la lyse des cellules remplies de particules, libérant environ une centaine de virions dans le milieu.

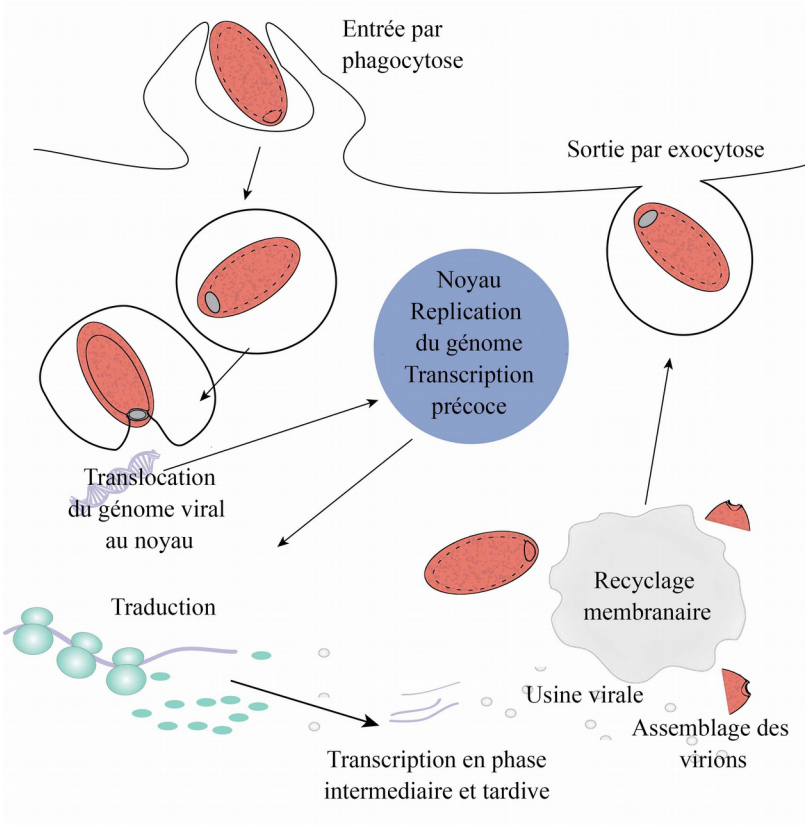


Figure 25 : Réplication virale des Pandoraviridae. Figure d'après Abergel & al [47].

Pithovirus sibericum et *Mollivirus sibericum*, les premiers virus anciens

Grâce à la collaboration avec les équipes russes du professeur Rivkina les équipes de l'IGS ont recu leurs premiers échantillons de pergélisols anciens en 2013. C'est à partir d'un échantillon de pergélisol datant de 30000 ans et en provenance de la rivière Anouï dans le grand est sibérien que deux nouveaux virus géants jusqu'alors inconnus ont été découverts. Ce résultat démontre, dans un premier temps, que le pergélisol est un environnement permettant de conserver des virus à ADN double brins infectieux durant plusieurs dizaines de milliers d'années. Même si la dangerosité pour les humains des virus géants n'a jamais été démontrée clairement, il n'est pas à exclure que d'autres virus, notamment certains NCLDV pathogènes du Pleistocène, puissent également être cryo-conservés dans le pergélisol durant plusieurs milliers d'années. La découverte de ces deux nouveaux virus, baptisés *Pithovirus sibericum* et *Mollivirus sibericum* ont réservé leur lot de surprises.

Si on pouvait difficilement imaginer une particule virale plus grande que celles des pandoravirus, la découverte de *Pithovirus sibericum* est venue repousser encore plus loin le gigantisme de ces virus [69]. Avec une taille de 1,5 par 0,5 µm les particules virales de pithovirus dépassent la taille de toutes celles des NCLDV connus. De plus, la forme de la capside est elle aussi hors du commun. La couche externe de la capside apparait dense aux électrons et striée. On trouve à

l'apex de la capside un bouchon caractéristique en forme de nid d'abeille. Depuis la découverte de *Pithovirus sibericum* plusieurs autres représentants de cette famille ont été activés, en commençant par *Pithovirus massiliensis*, découvert à la Ciotat dans des échantillons d'eau usées et représentant moderne à 84% identique au niveau nucléotidique à son homologue du pleistocène [70]. En 2016, un virus à la morphologie tres similaire des pithovirus a été découvert. Seule différence phénotypique notable, la présence systématique de deux bouchons aux deux extrémités de la particule virale. C'est par l'analyse génomique de ces nouveaux virus qu'il a été possible de dessiner deux lignées distinctes dans la famille des *Pithoviridae*: d'un côté nous retrouvons les P*ithovirus*, dont la morphologie la plus commune ne présente qu'un seul bouchon apical et de l'autre les *Cedratvirus* ayant quant à eux quasi systématiquement deux bouchons apicaux [71]. En 2018 un nouveau *Pithoviridae*, nommé *Orphéovirus* [72], a été découvert. Bien qu'ayant un génome trois fois plus gros que les autres pithovirus (1,4 Mpb contre 0,6 Mpb pour les autres pithovirus), la présence de 60 gènes communs à l'ensemble de ces virus suggère qu'ils partagent une histoire évolutive commune. La classification des *Pithoviridae* est actuellement à l'étude au laboratoire. En attendant de trouver des arguments retracer l'histoire évolutive d' *Orphéovirus*, il a été convenu que ce nouveau virus constituait une troisième lignée divergente de la famille des *Pithoviridae* (Figure 26). De façon anecdotique nous pouvons noter qu'à nouveau c'est la découverte de *Mimivirus* qui a ouvert la voie à la découverte des *Pithoviridae,* en effet, en 2002 un endosymbionte non cultivable baptisé KC5/2 avait décrit chez des amibes du genre *Acanthamoeba* [73]. Ce n'est que 12 ans plus tard que sa nature virale a été révélée [74].

Figure 26 : Phylogénie des Pithoviridae. L'arbre phylogénétique est basé sur la séquence de l'ADN polymérase (modèle de substitution LG+G4). La longueur de la branche menant à Orpheovirus souligne la divergence de ce virus vis à vis des autre Pithoviridae.

Le cycle infectieux des pithovirus, d'une quinzaine d'heure en moyenne, démarre par la phagocytose des virions (Figure 27). La perte du bouchon apical permet une fusion de la membrane interne de la capside avec la membrane de la vacuole. Ce canal permet donc de connecter le compartiment interne du virion avec le cytoplasme de la cellule hôte. L'ensemble du cycle infectieux se déroule dans le cytoplasme et aucun changement dans la structure du noyau de l'amibe

est à noter au cours du cycle infectieux. C'est au bout de 6h d'infection qu'on observe une zone d'exclusion se former dans le cytoplasme de la cellule infectée, suggérant la mise en place de l'usine virale. La formation des capsides virales est similaire à l'assemblage des particules de pandoravirus, on observe ainsi un assemblage simultané des structures membranaires internes et un « tricotage » de la couche externe plus dense.

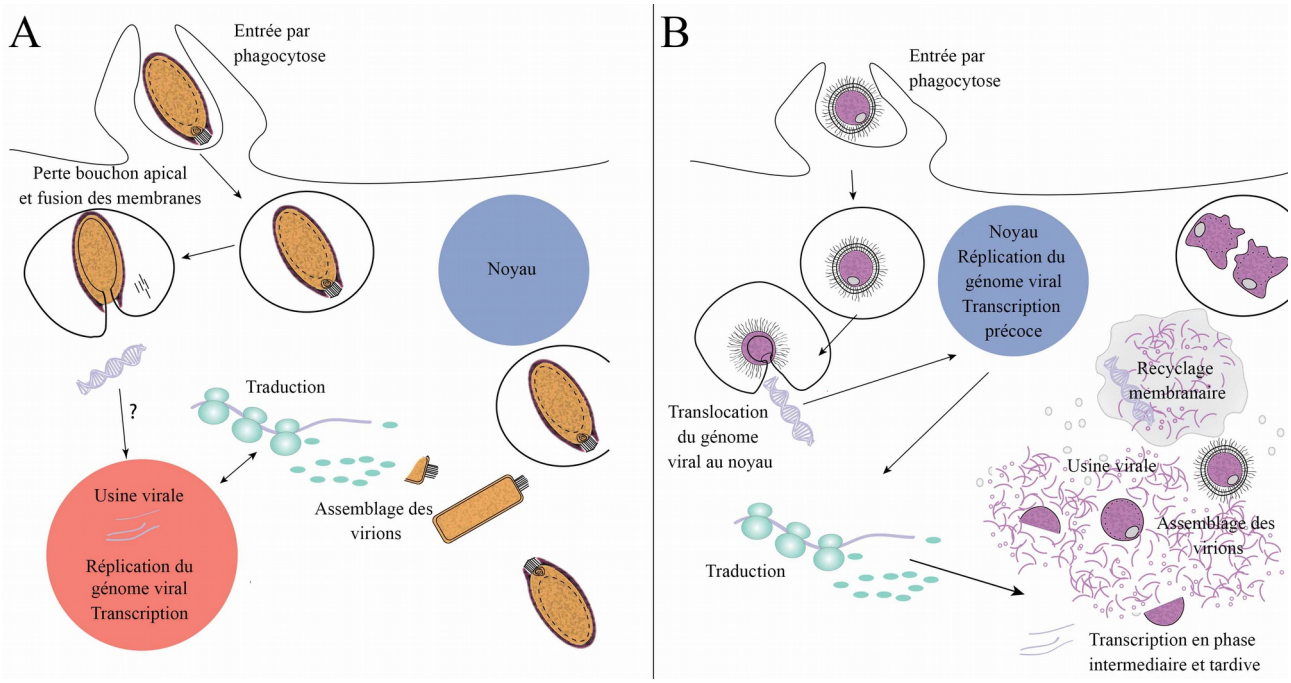


Figure 27 : Réplication virale des Pithoviridae (A) et de Mollivirus (B). Figure d'après Abergel & al [47].

L'analyse du protéome de la particule des pithovirus a démontré que ces virus embarquaient la majorité des protéines nécessaires à la réplication et la transcription du génome viral. Ces observations sont similaires à celles faites chez les *Megaviridae* et les *Poxviridae,* permettant ainsi d'expliquer la plus grande indépendance de ces virus vis à vis du noyau de l'hôte.

De façon paradoxale, bien que les *Pithoviridae* aient la capside la plus grande en terme de taille, leur génome d'une taille de 600 kpb est plus petit que celui des *Mimiviridae*. De plus, le génome des pithovirus est composé à 20% de zones répétées non codantes. Cette découverte bien que n'ayant pas d'explication pour l'instant laisse envisager la présence d'éléments transposables encore inconnus.

Mollivirus sibericum, le second virus activé à partir d'un échantillon de pergélisol du Pléistocène est, quant à lui, sphérique et d'une taille de 600 µm [75]. La structure fine de la capside de mollivirus a été étudiée par microscopie électronique. De façon étonnante, la couche dense au électrons semble avoir une structure lamellaire rappelant la couche dense aux électrons des

pandoravirus. Plus étonnant encore, des expériences de transcriptomique menées au laboratoire ont montrées que les amibes *A. castellanii* sur-expriment des gènes codants pour des celluloses synthases au cours de la phase d'assemblage des particules de mollivirus et pandoravirus. A l'inverse, lors d'une infection par mimivirus ou megavirus, ces même gènes amibiens semblent régulés négativement. Des résultats préliminaires ont également démontrés la présence de glucose et de cellobiose, produits caractéristiques de la dégradation de la cellulose, après digestion et lyse des particules virales de pandoravirus et mollivirus. Ces résultats préliminaires permettent de formuler l'hypothèse que la couche lamellaire de la capside des mollivirus et pandoravirus est en partie composée de cellulose.

Le séquencage et l'analyse des 600 kpb du génome de *M. sibericum* a révélé que mollivirus est le seul virus géant ne codant pas pour certaines protéines clefs de la biosynthèse de l'ADN comme la ribonucleotide réductase. De même l'analyse du protéome indique que l'ARN polymérase virale n'est pas embarquée dans la capside. D'autant plus surprenant, *M. sibericum* code pour une MCP dont l'homologue le plus proche se retrouve dans le génome de *Acanthamoeba castellanii* [76] (Figure 28). Cette donnée est surprenante pour plusieurs raisons, comment envisager que des virus codant pour une MCP puissent avoir une capside sphérique ; comment les pandoravirus, dont une partie de la structure de la capside est proche de celle de *M. sibericum,* n'ont pas de MCP ; quelle est l'histoire évolutive des mollivirus pour que ces derniers aient gardé une MCP partagée avec leur seul hôte contemporain connu tout en ayant une capside sphérique ? Nous verrons dans la suite de ce manuscrit comment les travaux présentés ici ont contribué à alimenter ce débat.

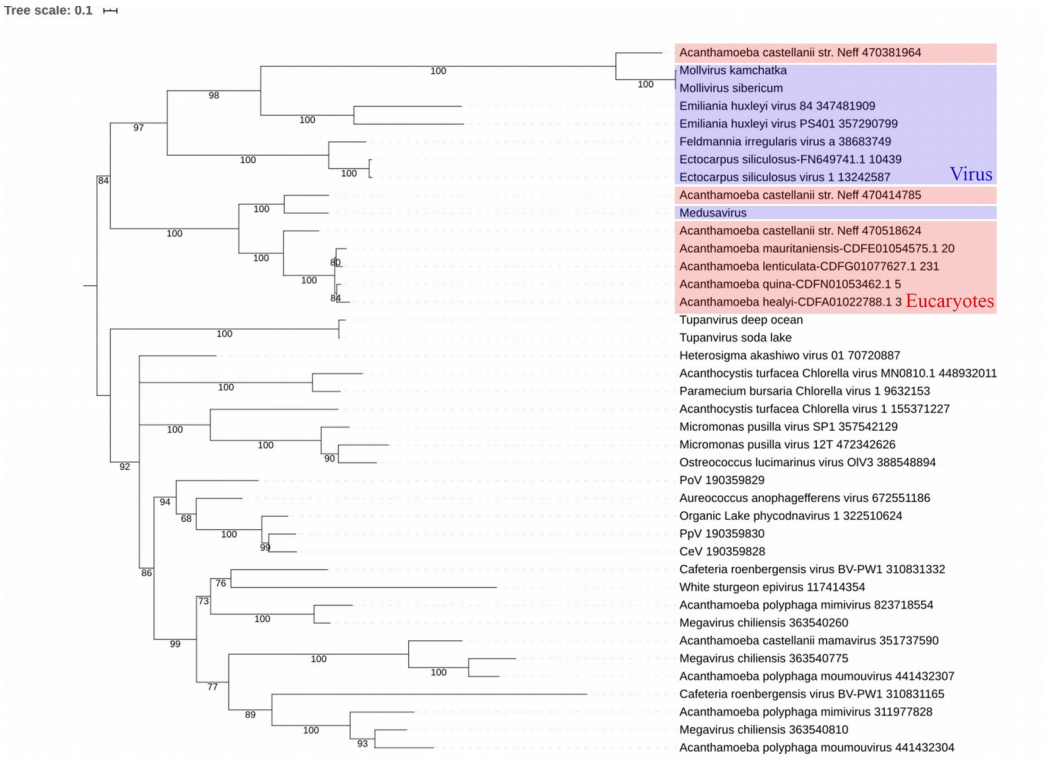


Figure 28 : Phylogénie de la protéine majeure de capside (MCP). L'arbre phylogénétique est basé sur la séquence de la MCP (modèle de substitution LG+F+I+G4).

Ainsi, depuis la découverte de *Mimivirus*, de nouveaux virus géants ont été découverts. A plusieurs égards, leur caractérisation a permis de repousser les limites de la virologie actuelle.

Hôte des virus géants

Les amibes sont des protozoaires ubiquitaires. On les retrouve dans l'ensemble des environnements et plus particulièrement dans les sols. Les travaux récents de nos collaborateurs du laboratoire de Puschino ont permis de découvrir 6 nouvelles souches d'amibes du genre *Acanthamoeba* dans des strates de pergélisol allant du Pleistocène à l'Holocène, prouvant ainsi que les paléocryosols abritent des protistes réactivables [77]. Les amibes ont longtemps été étudiées pour leur intérêt médical, en plus d'être des chevaux de Troie capables de transporter des endosymbiontes d'intérêt médical, les amibes du genre *Acanthamoeba* sont des pathogènes opportunistes provocant notamment des lésions cutanées, des infections des sinus, des kératites, ou encore des encéphalites granulomateuses. Pour ces raisons, les amibes sont classés comme pathogènes de classe 2 et leur manipulation nécessite des équipements de protection individuelle (EPI) de rigueur pour la manipulation d'agents de classe 2: gants, blouse, poste de sécurité microbiologique de type II (PSM II) et lunettes.

La présence de cette diversité eucaryotes dans le pergélisol ancien laisse donc supposer qu'il

existe une grande diversité de micro-organismes parasitant ces cellules encore largement inconnue. L'utilisation de la souche Neff ATCC 30010™ comme espèce modele pour la réactivation de virus géants a permis au laboratoire d'acquérir une expertise dans un ensemble de techniques incluant : expression de gènes viraux, édition du gènome... C'est la forme « végétative libre » ("Free Living Amoeba" ou FLA) des amibes qui est sensible à l'infection virale. Cette forme FLA, ou trophozoïte, permet l'expansion clonale de l'amibe ; en plus de la forme FLA, il existe une forme dormante, permettant la survie de l'espèce dans des conditions défavorables [78] (Figure 29).

Les deux formes prises par *Acanthamoeba* ont des morphologies distinctes qui les caractérisent:

- Le trophozoïte, d'une taille allant de 10 à 25 μm, est couvert de spicules et contient de nombreuses vacuoles digestives et lysosomes. La mobilité des trophozoïtes est permise par de longs pseudopodes, qui permettent à l'amibe de se déplacer et phagocyter les bactéries dont elle se nourrie.
- Le kyste, forme dormante dont l'activité métabolique est réduite, est de forme sphérique ou en étoile. La résistance des kystes à l'environnement est permise par la présence de deux parois composées de protéines et de polysaccharides. La nature des composés polysaccharidiques formant la paroi des amibes a été caractérisée comme contenant de la cellulose [79]. Lors de l'enkystement, la production de cette paroi est permise par la sur-expression de cellulose synthases amibienne.

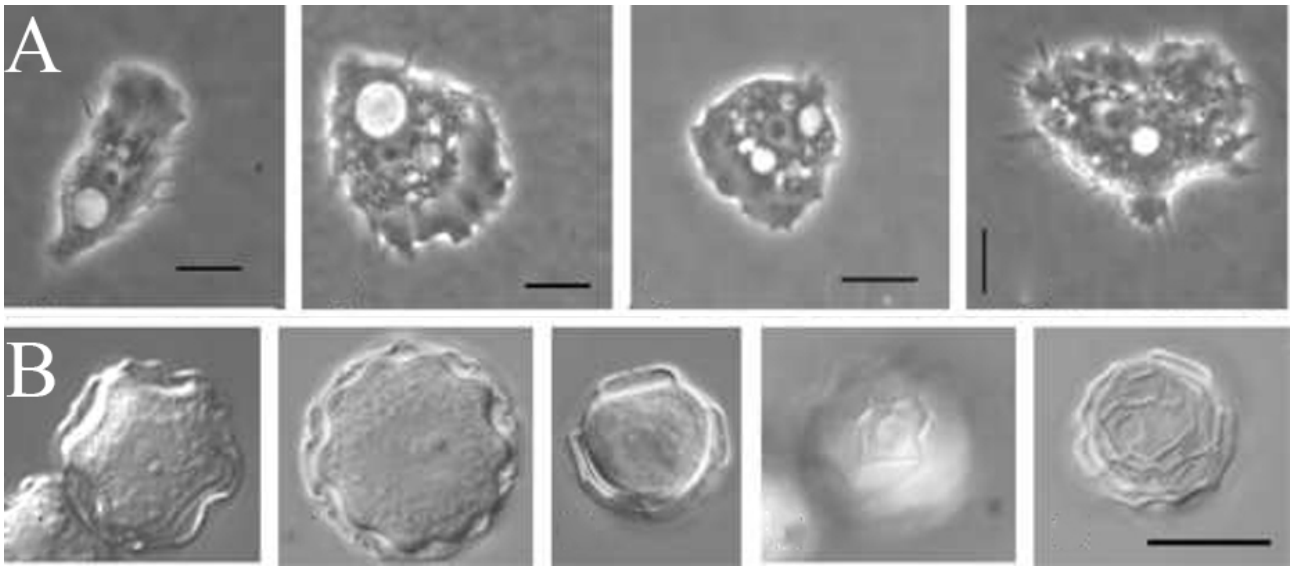


Figure 29 : Images de microscopie optique des souches d'Acanthamoeba issues de strates plus ou moins anciennes de pérgélisol sibérien. (A) Formes trophozoites. (B) Formes kystes. Barres d'echelle : 10 μm. Figure d'après Malavin & al [77].

Comme nous l'avons vu dans l'introduction, la présence de *Mimiviridae* dans l'océan suggère que certains représentants de cette famille virale puissent infecter différents types d'organismes marins. Parmi ces organismes marins potentiellement infectés par des *Mimiviridae*, il apparaît que

les micro-algues constitueraient des hôtes potentiels pour de nouvelles espèces de virus appartenant à la famille des *Mimiviridae*. La recherche de virus d'algues a permis d'isoler trois nouveaux *Mimiviridae* : *Cafeteria roenbergensis virus* (CroV), isolé en 2010, ce virus icosaedrique de 300 nm infecte un biflagellé hétérotrophe du genre *Cafeteria roenbergensis* ; *Phaeocystis globosa virus* (PgV), isolé en 2013 et infectant des algues du genre *Phaeocystis globosa* ; enfin un virus infectant l'algue unicellulaire *Haptolina ericina*, a été isolé en 1998, à partir d'échantillons marins. Ce virus, *Chrysochromulina ericina virus* (CeV) possède une capside icosaédrique de 160 nm de diamètre et ne rentre donc pas strictement dans la définition de « virus géant ».

A ce jour, l'ensemble des virus découverts au laboratoire infectent la souche modèle d'*Acanthamoeba castellannii.* Bien qu'ont été explorés d'autres modèles : *Dictiostelium, Paramecium*... et que certains laboratoires ont fait mention de virus géants capables d'infecter *Dictiostelium* [51], aucun virus géant n'a, à ce jour, été réactivé sur des cellules autres que des amibes du genre *Acanthamoeba* au sein du laboratoire. Sur le plan évolutif ce constat interpelle sur l'origine des virus géants. Au regard des spectres d'hôtes des autres familles de NCLDV, il apparaît que les virus géants connus à l'heure actuelle sont une fraction d'une diversité encore largement inexplorée. Cependant, les amibes étant ubiquitaire, il reste à ce jour de nombreux environnements encore inexplorés où il est parfaitement envisageable de découvrir des nouveaux parasites à ces organismes unicellulaires [80].

OBJECTIFS DU PRÉSENT MANUSCRIT

Mon sujet de thèse s'inscrit dans ce nouveau champ qu'est la virologie environnementale. Le projet porte sur l'exploration de la biodiversité microbienne, incluant les virus géants à ADN double brins, dans des échantillons de cryosols russes. Pour se faire deux approches sont combinées. D'une part la détermination de profils métagénomiques, par le séquençage massif d'ADN extrait de la partie centrale (coeur) d'échantillons de cryosols en provenance de zones diverses : Sibérie, Kamchatka et Arctique. D'autre part, la co-culture de ces mêmes échantillons de pérgélisol en présence du système modèle utilisé au laboratoire : *A. castellanii,* utilisé comme appât cellulaire pour les virus géants à ADN. La découverte de nouveaux virus géants répond à deux objectifs distincts. Le caractère exploratoire conduisant à l'isolement et la caractérisation de nouveaux virus géants permet, par des approches bioinformatiques et expérimentales, d'ouvrir différents champs d'étude : caractériser les différentes spécificités des nouvelles familles virales ainsi constituées et contribuer au débat sur la classification du monde viral, mieux comprendre les mécanismes évolutifs en cours au sein d'une même famille, comprendre l'origine et la diversité des virus géants... Au-delà de cet aspect exploratoire, les cryosols de type pergélisol constituent à la fois un réservoir

de microorganismes conservant un potentiel infectieux et de gènes d'intérêt médical. Ainsi, les données de métagénomique permettent d'évaluer le risque lié à ces deux aspects dans un contexte de fonte accrue du pérgélisol et de préoccupation croissante concernant le réchauffement climatique dans la société.

L'acquisition, l'analyse et la conservation d'échantillons de cryosol sont des enjeux clefs du développement du laboratoire. A ce jour, nous conservons ainsi quatre séries d'échantillons de cryosols collectés dans le cadre de plusieurs collaborations et de provenances diverses. La série d'échantillon la plus récente a été collectée lors d'une expédition conjointe entre l'IGS et nos collaborateurs russes de Puschino sur la rivière Kolyma, dans la région de Tchersky dans l'extrême nord est russe. La première série (A) est exclusivement composée d'échantillons de surfaces, ou couche active, de cryosols (*i.e.* non gelés tout au long de l'année). Ces 5 échantillons ont été ramenés par un collaborateur et explorateur amateur, Alexander Morawitz, du Kamchatka. La seconde série d'échantillons (B) est issue d'une campagne menée par nos collaborateurs du laboratoire de Pushchino. Ces 7 échantillons proviennent d'expositions de pergélisol aux abords des rivières sibériennes suivantes : la Bolshaya Chukochya (70.0911°N 159.9269°E), l'Alazeya (70.5142°N 153.4046°E), l'Omolon (68.7042°N 158.6972°E) et la Kolyma (69.5477°N 161.3641°E), ainsi qu'un échantillon de cryosol en provenance de l'oasis de Schirmacher en Antarctique (-70.7633°N 011.7810°E). Les 15 fragments de cryosol de la troisième série d'échantillon (C) ont été prélevés dans le cadre d'une collaboration avec le Wegner Institute dans une zone proche de la rivière Lena baptisée Yukechi (61.7608°N 130.4746°E), aux abords de la ville de Yakutsk. Enfin, la dernière série d'échantillon (D) est en provenance à nouveau de la zone de la rivière Kolyma et a été collectée durant ma thèse, en août 2019 au cours d'une campagne d'échantillonnage conjointe avec nos collaborateurs russes du laboratoire de Pushchino (Figure 30). Ces 4 séries d'échantillons ont donc constitué le matériel de départ pour mes travaux de thèse et nous verrons comment ces derniers ont permis de faire avancer nos connaissances sur les microorganismes présents dans le cryosol de ces régions.

La caractérisation de *Pandoravirus Y4* a permis de soulever diverses questions : quels sont les mécanismes de diversité génétique en cours chez les *Pandoraviridae*, comment l'architecture des génomes permet-elle de formuler des hypothèses sur les mécanismes probables de réplication. Ensuite, nous verrons comment la découverte d'un second mollivirus, *Mollivirus kamchatka,* nous a permis de proposer un premier génome cœur de cette nouvelle famille virale. L'étude de ces gènes a permis de tester plusieurs scénarios évolutifs différents ainsi que tracer les contours des liens évolutifs entre mollivirus et les autres familles de virus géants. Sur le plan expérimental, nous

verrons que l'analyse comparée des cycles infectieux de *M. kamchatka* et *M. sibericum* par différentes techniques de microscopie électronique a permis de définir des mécanismes encore inconnus d'assemblage des particules virales.

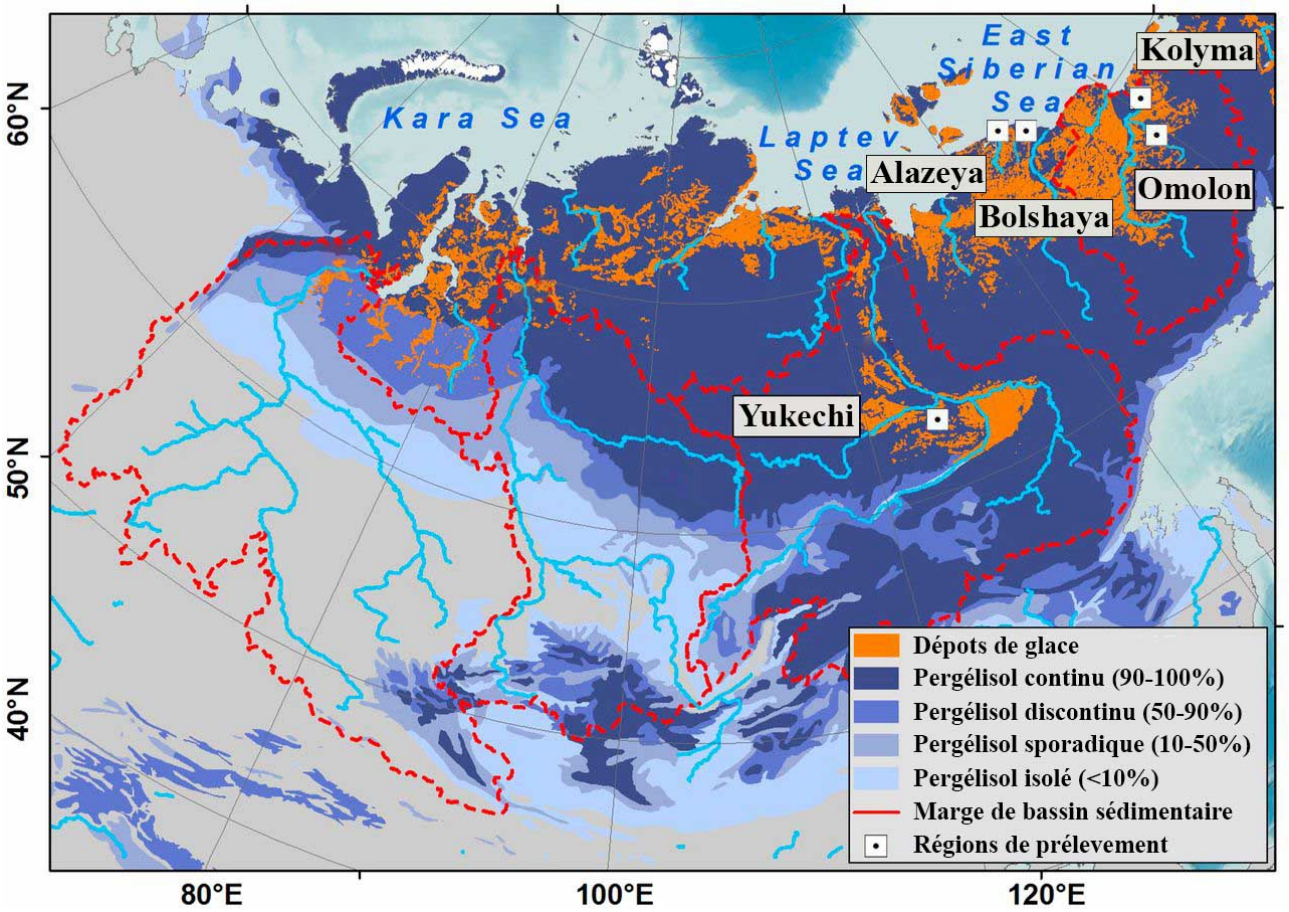


Figure 30 : Carte des différentes zones de prélèvements des échantillons de cryosol conservés au laboratoire. La nature du pergélisol présents à ces divers points de prélèvement est donné par la teinte de la carte à ces endroits.

Mon sujet de thèse couvrait initialement l'étude des virus activés à partir des échantillons de cryosols ainsi que l'étude des métagénomes associés. En raison de la réactivation très rapide de quatre nouveaux virus, mon travail de thèse s'est axé sur l'étude de deux d'entre eux. Dans ce cadre, l'exploitation des données de métagénomique a servi à confirmer la présence des virus activés dans les échantillons. Diverses pistes pour exploiter ces données sont à l'étude, la partie de ce manuscrit dédiée à l'étude de ces métagénome, peu approfondie, fera état des différentes pistes poursuivies à ce jour.

MATERIEL ET METHODES

Echantillons environnementaux et échantillonnage

La particularité du cryosol russe est que ce dernier contient un type de pergélisol riche en

matière organique et en eau sous forme de glace (50-90%). Ce sol sédimentaire gelée, appelée yedoma, s'est formé durant le Pleistocène au cours de la dernière glaciation (il y a 110000 à 10000 ans) par accumulation de sédiments, le recul de la glace qui a suivi cette dernière période glaciaire a rendu accessible ces zones de cryosol.

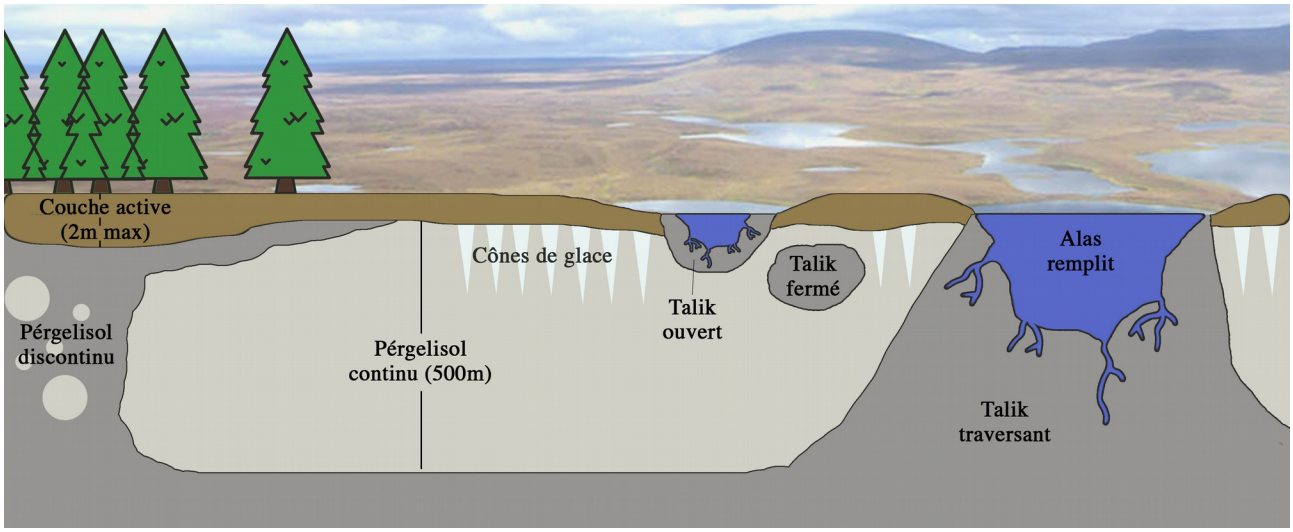


Figure 31 : Schéma en coupe des différentes structures de pergélisol. Les alas sont des dépressions formées par subsidence du pérgélisol subissant des cycles de réchauffement/refroidissement, le lac thermokarstique ainsi formé (Alas humide) peut éventuellement s'assecher (Alas sec).

Dans les zones subarctiques le yedoma peut être « continu », 90 à 100% du sol, « discontinu », 50 à 90% du sol ou sporadique (10 à 50%). On observe aussi des structures de sols partiellement dégelés, la principale, appelée talik, est un sol non gelé dans une zone de permafrost. Par un phénomène de subsidence les « talik » provoquent un affaissement du terrain et souvent se remplissent d'eau pour former des lacs thermokarstiques ou « alas ». Les « alas » subissent des cycles de remplissage/assèchement entrainant des niveaux d'eau variables au cours du temps (Figure 31). Du fait de l'érosion, le yedoma est particulièrement accessible aux abords des fleuves. J'ai eu la chance de pouvoir participer à la campagne d'échantillonnage de la série C, durant le mois d'août 2019. Pour cette expédition nous avons rejoint les équipes du laboratoire de Puschino sur le site de Duvany Yar situé sur les bords de la rivière Kolyma dans l'extrême est Sibérien (68.6282°N 159.1948°E). Le passage du fleuve Kolyma dans cette zone permet l'érosion des sols gelés, créant ainsi des gorges où le pergélisol est accessible depuis la berge. Ce relief d'érosion permet d'avoir accès directement à toutes les strates de yedoma, sans besoin de réaliser des forages verticaux, permettant donc des carottages sans contamination des strates supérieures à la strate étudiée. Les zones d'excavation sont choisies en fonction de la structure géologique étudiée. Pour la découverte d'organisme et de microorganisme, les zones de pergélisol riche en carbone de types résidus de tourbes, sont favorisés (Figures 32, 33).

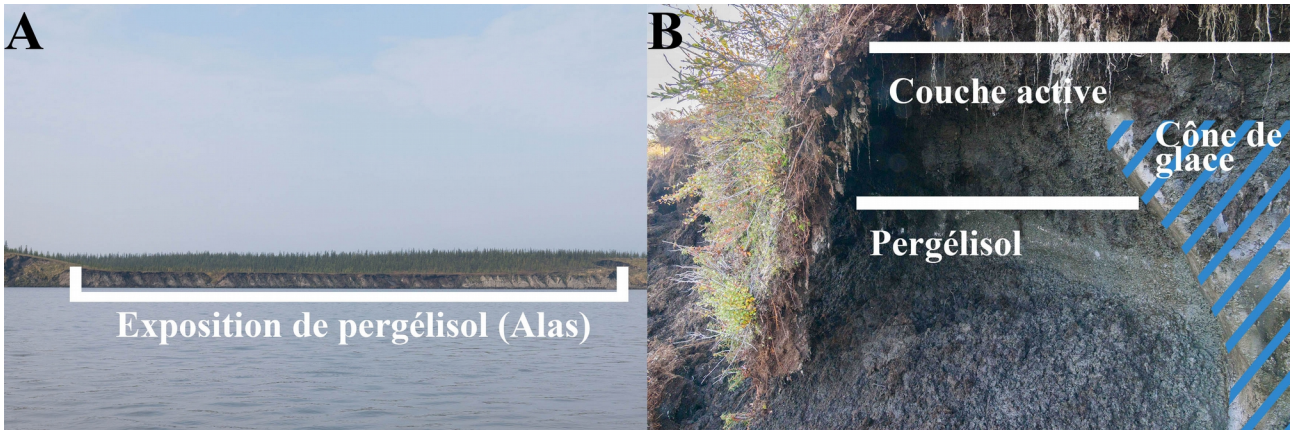


Figure 32 : Zone de prélevement des échantillons de la serie D, sur les bords de la rivière Kolyma. (A) Exposition de pérgélisol, du à l'érosion lié à l'activité de la rivière, on peut observer en coupe la dépression caractéristique d'un alas sec. (B) Détail des différentes structures du pergélisol : Cône de glace, couche active et pérgélisol continu.

Une fois une zone choisie, de préférence entre deux cônes, ou coins de glace, la surface non gelée est retirée à l'aide d'une pelle. Une fois la surface propre de tout sol fondu, une brève description de la structure du sol est réalisée. Le forage est conduit à l'horizontale à l'aide d'une perceuse électrique à cloche. Une première carotte est excavée puis jetée. La perceuse est ensuite enduite de microsphères fluorescentes de polystyrène (ThermoFisher). Une fois la seconde carotte excavée cette dernière est conservée dans un sachet plastique stérile et maintenu à une température inférieure à zéro degré (Figure 33). Une fois l'échantillon récupéré au laboratoire, sa surface est grattée à l'aide d'un scalpel stérile sous un PSM de type II et le retrait de la couche externe est contrôlé sous lampe UV. Une fois l'ensemble des billes fluorescentes disparues on considère l'échantillon stérile prêt à être étudié.

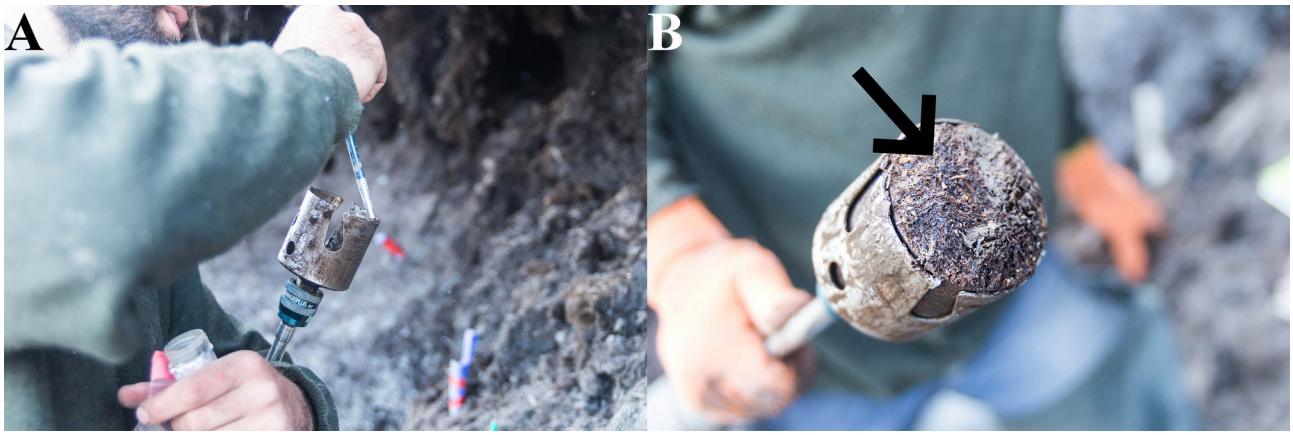


Figure 33 : Détail du processus de carottage horizontal des échantillons de pergélisol. (A) Dépôt des billes fluorescentes sur la cloche de la perceuse électrique. (B) Détail de la seconde carotte, on observe la présence débris végétaux et une couleur plus foncée du pergélisol à cet endroit. Ce changement local de composition et de couleur du pergélisol indique que la carotte a été faite dans un dépôt sédimentaire riche en matière carbonée (type marécage). Ce genre d'environnement est particulièrement favorable à l'activation de kystes de protozoaires et donc à l'activation de parasites de ces eucaryotes unicellulaires.

Réactivation de virus géants

Les amibes du genre *Acanthamoeba* sont utilisées comme organismes modèles au laboratoire, c'est l'amibe *Acanthamoeba castellanii* (Douglas, Neff ATCC 30010™) qui sert à l'heure actuelle à l'ensemble des études menées au laboratoire : réactivation de virus, productions virales et études des interactions hôtes-pathogènes. Les deux formes prises par l'amibe conférent des avantages. La forme enkystée permet une conservation de longue durée de lignées clonales à -80°C et la remise en culture de kystes permet l'obtention de cellules trophozoites. Ces trophozoites sont cultivés dans un milieu axénique appelé "PPYG" ou Proteose-Peptone-Yeast extract-Glucose dont la préparation s'effectue en deux temps : le milieu basal est composé de 15 g de proteose-peptone et d'1 g d'extrait de levure dans 900 mL d'eau ultra-pure puis autoclavé. Les oligo-éléments et le glucose, thermo-sensibles, sont ajoutés extemporanément : 4 mM de MgSO4, 0,4 mM de CaCl2, 0,05 mM de Fe(NH4)2(SO4)2, 2,5 mM de Na2HPO4, 2,5 mM de KH2PO4, et 100 mM de glucose dans 1 L d'eau ultra-pure, le pH est ajusté à 6,5 et le mélange est stérilisé par filtration sur membranes filtrantes de 0,22 µm. Les conditions optimales de croissance des trophozoites se situent à une température de 32°C, à l'obscurité et à une pression partielle en dioxygène ambiante. Ces conditions de cultures font de ce protozoaire un organisme facilement cultivable et conservable.

Les cultures cellulaires sont repiquées régulièrement, dès que le tapis cellulaire couvre 80 à 90% de la surface de la flasque de culture. Le tapis cellulaire est gratté à l'aide d'un râteau stérile, la concentration en amibe est calculée par comptage en cellule de Malassez du nombre de cellules contenu dans un aliquot de la suspension d'amibe. Les cellules ont ensuite été ensemencées en faible quantité, entre 500 et 1 000 cellules/cm^2, dans du milieu de culture frais auquel est ajoutée une concentration en antibiotique de 100 µg/mL (Chloramphénicol, Ampicilline et Kanamycine), et placé à 32°C pour incubation. On observe après un certain nombre de repiquages, au delà de 80, que le temps de doublement des cellules augmente et la présence croissante de cellules présentant un phénotype différent des trophozoites.

Pour augmenter la probabilité d'une rencontre entre un virus et une amibe, les échantillons environnementaux utilisés durant ce projet de thèse ont été traités de deux façon différentes. La première méthode consiste à re-suspendre une fraction de l'échantillon de sol dans un milieu liquide contenant de l'amidon de riz. Une fois la suspension réalisée, elle est placée à l'obscurité durant 1 mois. Initialement conçu pour les échantillons d'eau, ce traitement permet d'enrichir la fraction d'eucaryotes hétérotrophes présents dans l'échantillon (parmi lesquels on retrouve les hôtes potentiels des virus géants) par croissance de ces populations et réduction des populations

d'eucaryotes autotrophes. Ce milieu riz est préparé en ajoutant 40 grains de riz à un litre d'eau ultra-pure, puis stérilisé par autoclave et stocké à 4°C. Après 1 mois d'incubation à température ambiante de l'échantillon dans du milieux riz à 1% en PBS, la suspension est vortexée durant 2 minutes, 2 ml de la suspension est ensuite centrifugé à 800 x g pendant 5 minutes et le culot est ensuite re-suspendu dans 200 µL de PBS pour constituer *l'innoculum*. Le second traitement, plus direct et permettant de traiter plus rapidement les échantillons nouvellement reçus, consiste à prendre un fragment d 'échantillon brut et le re-suspendre directement dans un volume approximatif de 30 ml de PBS, l'échantillon est ensuite vortéxé 10 minutes puis laissé à décanter durant 5 min. Ensuite, 2 ml de la suspension non sédimentée est centrifugée à 1000 g durant 2 min. Le culot formé est ensuite re-suspendu dans 200 µL de PBS constituant la suspension utilisée pour la réactivation.
Deux souches d' *Acanthamoeba castellanii* sont utilisées en parallèle pour les expériences d'activation de virus. La première souche est le modèle Neff ATCC 30010™ en conditions de culture standard. La seconde souche est la souche modèle mais adaptée à une concentration de 2,5 µg/mL d'amphotéricine B (Fungizone). La présence de nombreux spores fongiques dans le permafrost entraine la croissance souvent rapide de champignons qui appauvrissent le milieu de culture et rendent les observations en microscopie optique impossibles. Cependant les effets de la Fungizone sur le cycle cellulaire ainsi que sur le cycle infectieux sont inconnus ; des observations de capsides virales aberrantes après production en milieux contenant de la Fungizone laisse penser qu'il existe un effet délétère de l'antifongique sur le cycle infectieux et il est donc important de réaliser les expériences de réactivation en utilisant les deux souches en parallèle. De façon empirique, il est observé que le milieu riz n'apparait pas plus efficace que la méthode directe lorsque les échantillons sont récemment prélevés.

Les deux types *d'inoculum* sont co-cultivés avec les deux souches d'amibe et pour chaque *inoculum* est ensemencé: deux flasques T25 contenant 10 000 cellules/cm^2, une plaque 6 puit contenant 10 000 cellules/cm^2. Le PPYG utilisé contient le cocktail d'antibiotiques à 100 µg/mL et, pour les amibes adaptées, de la Fungizone à 2,5 µg/mL final. Entre 5 µL et 20 µL d*'inoculum* est ajouté dans chaque puit et entre 30 µL et 100 µL est ajouté dans chaque flasque. C'est la nature de l'échantillon (granulométrie, densité en matière organique...) qui conditionne le volume d'*inoculum* afin que la turbidité dans le milieu de culture ne soit pas trop importante et permette de réaliser des observations des amibes en co-culture par microscopie optique. Quotidiennement les cellules sont observées au microscope optique à divers grossissements, afin de contrôler le tapis cellulaire (proportion de cellules décollées, pourcentage d'enkystement...) et à fort grossissement pour contrôler le phénotype des cellules (état de vacuoles, morphologie du noyaux...). En cas de changement phénotypique observé, le tapis cellulaire est décollé, 2 ml sont centrifugés à 500 g

pendant 2 minutes. Le surnageant est centrifugé à 10 000 g durant 45 min et un état frais est réalisé sur une fraction des cellules décollées. Le culot issu du surnageant, censé contenir les virus réactivés est re-suspendu dans du milieux frais et inoculé sur un tapis cellulaire frais sans Fungizone à une concentration de 40 000 cellules/cm^2 .

Une fois le nouveau virus amplifié par repiquages successifs vient l'étape d'identification de ce nouveau parasite. A ce stade, deux méthodes sont déployées pour répondre à cette question. En premier lieu une observation directe du culot du dernier surnageant de culture par microscopie électronique puis une méthode moléculaire par PCR. Un millitre du surnageant du dernier milieu de culture est centrifugé à 10 000 g durant 45 min. Le culot est lavé dans 1 ml d'eau ultra-pure pour débarrasser l'échantillon du phosphate et du glucose contenue dans le PPYG puis re-suspendu dans 25 µL à 30 µL d'eau distillée. Une goutte de 7 µL de suspension est mise à sédimenter 5 minutes sur une grille de microscopie électronique recouverte de thermoplastique formvar (pour permettre à l'échantillon de se fixer à la grille) et carbonée (permettant ainsi un meilleur contraste). Le surplus de liquide est retiré puis la grille est incubée face retournée sur une goutte d'acétate uranyl à 1% en eau durant 20 secondes. Une fois la grille séchée sur papier filtre, elle est insérée puis observée au microscope électronique à transmission (Tecnai G2, 200 kV). La seconde technique d'identification précoce des virus est une PCR directe en utilisant des amorces amplifiant des gènes cœurs des différentes familles virales. La polymérase Terra utilisée est tolérantes aux inhibiteurs de PCR et permet une amplification directe de l'ADN contenu dans des échantillons variés (tissus, broyats de plante...). Le protocole suivi est celui recommandé par le fabicant pour un volume final de 25 µL. Une fois les cycles terminés, 2 µL des produits PCR sont déposés sur un gel d'agarose 1,5% en tampon Tris-EDTA (TE) et un courant de 130 V durant 25 min permet la migration des fragments d'ADN dans le gel. Après révélation par incubation en BET, les échantillons amplifiés par PCR sont purifiés par sur les colonnes du kit Monarch PCR & DNA Cleanup Kit (New England Biolabs). Les amplicons purifiés sont ensuite envoyés pour séquençage via la plateforme en ligne Eurofins.

Comptage et Clonage

Etant donnée la taille des particules virales étudiées au laboratoire, la titration des suspensions virales peut être effectuée par microscope optique grâce à un comptage en cellule de Malassez. La préparation pour le comptage consiste à mélanger 5 µL de virus dilué en PBS (1/100e ou 1/1000e), 35 µL de glutaraldéhyde 2%, et 10 µL de bleu trypan. La présence de bleu Trypan permet une coloration de la capside virale et un comptage manuel. La concentration virale est calculée en virus par ml et déduit par le calcul suivant : moyenne des virus par carreau x facteur de dilution x 10 (dilution du bleu trypan) x 100 (100 carreaux = 1 µL) x 1 000.

Pour obtenir une production virale homogène sur le plan génétique, il est nécessaire de réaliser un clonage de ce dernier par production du virus à partir d'une unique cellule infectée. Le puit A1 d'une plaque 12 puits est ensemencé à quasi confluence, soit 70 000 cellules/cm^2 . Une fois les amibes adhérentes, elles sont infectées avec une quantité de virus 50 fois supérieures (MOI 50). Après 2h de contact entre les virus et les amibes, le tapis est lavé 17 fois avec 3 ml de PPYG puis 20 fois avec 5 ml de PPYG frais afin de retirer les particules virales non phagocytées. Dans les puits A2 à A4, 500 µL de PPYG frais est déposé pour permettre des dilutions sériées au 5/7ème du puit A1. Des gouttes de 0,5 µL du puit A4 sont déposés ensuite dans les puits restants à l'exception du puit C4 qui sert de témoin, et la présence d'une seule amibe par goutte est controlé par microscopie optique. Les puits contenant 1 cellule, ainsi que le puit témoin, sont ensuite ensemencés à faible densité, 500 cellules/cm^2, pour permettre la propagation de proche en proche du virus. Après 5 jours d'incubation à 32°C, le phénotype des cellules témoins est contrôlé et le surnageant des puits présentant une production virales est conservé.

Production et purification des virus

Quarante flasques contenant des amibes à 50 000 cellules/ cm^2 sont infectées à partir d'un clone. Après 4 jours d'incubation à 32°C et d'observations quotidiennes par microscopie optique du phénotype infectieux des amibes, les flasques sont grattées à l'aide d'un râteau stérile afin de décoller les cellules restantes non lysées, puis le contenu de ces flasques est centrifugé à 500 x g entre 5 à 10 minutes et le culots contenant les débris cellulaires est jeté. Le surnageant contenant les virus a ensuite été centrifugé à 6 800 x g de 45 minutes. Le culot viral est ensuite lavé par 3 re-suspensions en PBS puis centrifugation à 6 800 x g pendant 45 minutes afin de se débarrasser des résidus de PPYG et des débris les plus légers. Ensuite le culot viral est repris dans 5 mL de PBS pour conservation à 4°C.

Une fois le virus produit, il est purifié par dépôt sur gradient discontinu de chlorure de cesium. La couche la plus dense (3,5) est déposée au fond d'un tube d'ultracentrifugation suivi des couches moins denses : 3,4, 3,3. Le virus produit et conservé est centrifugé à 6 800 x g pendant 45 minutes puis le culot est repris dans la fraction la moins dense du gradient (3,2). Les gradients sont ensuite ultra-centrifugués à une vitesse de 103 000 g durant une nuit. L'anneau viral est ensuite prélevé à la seringue puis lavé 4 fois en PBS. Le virus purifié est ensuite conservé en PBS et des aliquots sont stockés à -80°C en présence d'un agent cryoprotecteur, le diméthylsulfoxyde (DMSO).

Etude du cycle infectieux

L'étude du cycle infectieux est réalisé par microscopie et chaque technique de microscopie

nécessite une préparation spécifique des échantillons.

L'étude du cycle infectieux par microscopie électronique à transmission permet l'étude fine des événements intra-cellulaires se déroulant au cours de l'infection virale : phase d'éclipse, devenir des virions, recyclage des membranes cellulaire, établissement de l'usine virale... Les échantillons de microscopie électronique sont préparés comme suit : douze flasques T25 sont ensemencées à 60 000 cellules/cm^2, dont 11 sont infectées par des virus à une MOI de 50, puis incubées à 32°C. Afin d'obtenir des infections synchrones, les tapis cellulaires sont lavés deux fois avec 5 ml de PPYG frais 50 minutes après la mise en contact des cellules avec les virus. Immédiatement après, une première flasque est contrôlée au microscope optique puis fixée par ajout d'une solution de glutaraldéhyde dilué en PBS à une concentration finale de 2,5%. Après un temps de fixation de 45 minutes le tapis cellulaire est décollé avec un râteau stérile et centrifugés à 5 000 x g pendant 45 minutes. Enfin, le culot est re-suspendu dans 1 mL de glutaraldéhyde à une concentration finale de 2,5%, lavé en PBS, puis stocké à 4°C. Cette opération est répétée pour une nouvelle flasque infectée toutes les heures jusqu'à 10h post-infection (PI). La onzième flasque est gardée pour un temps tardif (40h PI) et la flasque non-infectés sert de témoin et est fixée en même temps que la dernière flasque infectée.

L'usage de la microscopie optique permet d'étudier le devenir du noyau cellulaire au cours de l'infection virale par marquage de l'ADN au DAPI. Le principe est de réaliser les infections virales sur des lamelles de microscopie polylysinées. Pour cela, les lamelles de microscopie sont d'abord dégraissées dans un bain composé au 2/3 d'éthanol pur et d'un 1/3 d'acide chlorhydrique 1 M sur la nuit. Les lamelles dégraissées sont ensuite rincées à l'eau ultra-pure puis à nouveau trempées dans un bain d'éthanol pur puis séchées et disposées dans des puits de plaque 12 puits. La plaque est ensuite stérilisée 30 minutes sous UV et les lamelles sont incubées 10 minutes dans un bain de poly-L-lysine 0,01% (masse/volume). Après rinçage, les puits sont séchés sous le PSM II et à nouveau stérilisés sous UV pendant 30 minutes. Les puits sont ensuite ensemencés à 60 000 cellules/cm^2 et, comme pour la préparation des échantillons de microscopie électronique, une infection synchrone des puits (à MOI 50 puis lavage) est réalisée. Pour éviter toute fluorescence non spécifique, l'infection est stoppée par l'ajout de formaldéhyde à une concentration finale de 3,7% puis un lavage en PBS des lamelles est fait. Les lamelles sont ensuite séchées puis scellées sur des lames de microscopie avec une goutte d'environ 5 µ L de VECTASHIELD Antifade Mouting Medium contenant du DAPI à 1,5 µg/ml. Les lames sont ensuite stockées à 4°C et à l'obscurité.

Microscopie optique

Les montages sont observés sur un microscope à épifluorescence (Zeiss Axio Observer), en utilisant un objectif x 63 associé à un Optovar x 1,6.

Microscopie électronique

Deux protocoles d'inclusion sont réalisés pour permettre des observations fines du cycle infectieux par microscopie électronique à transmission. Deux types de traitement sont faits, le premier pour obtenir des sections présentant un contraste fort des bi-couches lipidiques et le second pour observer les structures membranaires fines par un marquage des lipides moins dense en métaux lourds. Pour faciliter la manipulation des échantillons au cours des différentes étapes de fixation et de marquage, ces derniers sont enrobés dans des boudins d'agarose comme suit : les culots de cellules infectées sont chauffés à 42°C dans un bain-marie ainsi que des Microvettes dont le tube externe est rempli d'eau chauffée (Sarstedt). Les culots de cellules chauds sont re-suspendus dans 250 µL d'agarose basse température de fusion en surfusion (2% en PBS), puis déposés dans des Microvettes et centrifugés à 10 000 x g pendant 3 minutes. Le culot enrobé ainsi formé dans le fond du tube Microvettes est incubés 30 minute dans de la glace pour permettre sa solidification. Une fois les culots enrobés extraits de la Microvettes, ils sont lavés 3 fois en PBS durant 5 minutes à température ambiante.

Le premier protocole, appelé OTO pour « osmium-thiocarbohydrazide-osmium » est un protocole lourd permettant un fort contraste des membranes et de s'affranchir d'étapes de post-marquage des grilles de microscopie électronique [81]. Ce protocole repose sur le tétroxyde d'osmium (OsO4), oxydant puissant, qui se fixe sur les lipides. Par une réaction d'oxydo-réduction, l'OsO4 vient se fixer aux lipides, le thiocarbohydrazide (TCH) sert d'amplificateur du signal par fixation à l'OsO4, libérant des radicaux libres pour lui-même fixer une seconde couche de molécules d'osmium. Le protocole est le suivant : chaque échantillon enrobé est ensuite incubé dans un mélange de tétroxyde d'osmium 2% final, 1,5% ferrocyanide en PBS, pendant 1 heure à 4°C. Après trois lavages en eau ultra-pure de 5 minutes, les échantillons sont plongés dans une solution filtrée (0,2 µm) de thiocarbohydrazide (TCH) préparée extemporanément par dissolution à 60°C de TCH solide dans de l'eau ultra-pure à une concentration finale de 10 mg/L. Après 3 lavages en eau ultra-pure de 5 minutes les échantillons sont plongés dans une seconde solution d'osmium 2% durant 30 minutes à 4°C. Une incubation sur la nuit dans une solution d'acetate uranyl à 1% en eau permet de renforcer encore le contraste. Après cette étape, un marquage supplémentaire est fait à partir d'aspartate de plomb. Une solution de nitrate de plomb à 6,5 mg/L solubilisée dans de l'acide

aspartique 0,003 M pH 5,5 est préparée extemporanément puis filtrée sur membrane 0,2 µm. Les échantillons sont incubés dans cette solution durant 30 min à 60°C. Après incubation, trois lavages en eau ultra-pure rapides sont réalisés pour éviter la précipitation du plomb, suivis de trois lavages de 5 minutes en eau ultra-pure. La déshydratation des échantillons se déroule en 6 étapes de 5 minutes sur glace dans des bains d'éthanol froid (4°C) à 50%, 75%, 85%, 95% et 100% deux fois. La dernière étape de déshydratation consiste en deux bains d'acétone sec froid durant 10 minutes à température ambiante.

Le second protocole est le suivant : les échantillons enrobés d'agarose sont post-fixés et marqués durant 1 heure sur glace par immersion dans un bain contenant 1% d'osmium et 1,5% de potassium ferrocyanide en tampon Hepes 0,1 M pH 7,1. Après 4 rinçages de 5 minutes à température ambiante avec du tampon Hepes 0,1 M pH 7,1, les échantillons sont contrastés par incubation dans une solution d'uranyl acetate 2% en eau durant 1h, à température ambiante et à l'obscurité. Après 4 lavages de 5 minutes en eau les échantillons sont déshydratés en 7 étapes : 5 minutes sur glace dans des bains d'éthanol froid (4°C) à 25%, 50%, 75%, 95% puis trois bains d'éthanol froid 100% durant 10 minutes.

Les échantillons sont ensuite imprégnés dans une résine Epon préparée comme suit : 50 mL de Embed 812 (EMS) avec 17 mL de DDSA (Dodecenyl Succinic Anhydride, EMS) et 34 mL de NMA (Methyl-5-Norbornene-2,3-Dicarboxylic Anhydride, EMS) : ce mélange est homogénéisé par agitation douce, pour éviter la formation de bulles. L'imprégnation est graduelle, tout d'abord le mélange Epon est dilué à 33% en acétone sec et les échantillons y sont incubés durant une nuit à 4°C sur une roue. Cette étape est répétée avec une concentration croissante de résine : 50%, 75%, 100% et à nouveau 100% à température ambiante. Un nouveau mélange de résine est alors préparé dans lequel est présent l'agent durcisseur permettant de catalyser la polymerisation de la résine : 16,7 mL de Embed 812, 5,7 mL de DDSA 11,3 mL de NMA et 0,525 ml de DMP30 (2,4,6-Tris-dimethylaminomthyl-phenol, EMS). Les échantillons sont imprégnés durant une journée à température ambiante puis mis en moule pour polymérisation durant 7 jours au four Pasteur à 60°C. Après durcissement des blocs, ces derniers sont découpés à l'aide d'un ultra-microtome (Leica) et d'un couteau de diamant. Les coupes ultra-fines (90 nm d'épaisseur) obtenues, sont déposées sur une grille de cuivre couverte de formvar et de carbone. Les échantillons traités selon le second protocole sont post-marqués pendant 45 minutes dans une solution d'acétate uranyl 2% en eau puis dans une solution de citrate de plomb durant 4 minutes dans le noir. Les observations sont ensuite réalisées au microscope Tecnai G2, 200 kV.

Structure des génomes

Pour confirmer expérimentalement la structure des génomes prédites *in silico* nous avons utilisé deux approches, la première, moléculaire, permet de confirmer les jonctions entre les différents fragments par amplification *in vitro* des jonctions ; la seconde approche consiste à faire migrer des longs fragments d'ADN par électrophorèse en champs pulsé.

Réaction par polymérisation en chaine

Pour amplifier des fragments a haute fidélité d'ADN purifié nous avons utilisé le kit PCR ThermoFisher Phusion selon les recommandations du producteur. Les génomes étudiés ici étant riches en GC nous avons utilisé le tampon adéquat. Les séquences d'un même couple d'amorce sont choisies de façon à ce que les températures d'appareillement soient similaires ±1°C et préférentiellement riches en GC en 3' de la séquence.

Électrophorèse en champ pulsé

L'électrophorèse en champ pulsé (PFGE) permet la migration d'ADN génomique entier et intègre. Pour cela il est impératif de ne pas pipeter l'ADN. Les virus sont d'abord enrobés dans un bloc d'agarose avant d'être lysés. Pour chaque puit du gel d'éléctrophorèse, 40 µL de virus à une concentration de $\sim 10^{10}$ particules virales par millilitres est mis à chauffer à 45°C. Chaque suspension virale est ensuite diluée au demi avec de l'agarose à basse température de fusion à 2% (m/v) en surfusion. Le mélange ainsi obtenu est ensuite placé dans un moule approprié et laissé à durcir 15 minutes à température ambiante puis 15 minutes à 4°C. Les blocs ainsi formés sont ensuite lysés dans un tampon de lyse contenant 50 mM de Tris-HCl pH 8, 50 mM 1% (v/v) N-laurylsarcosine, 1 mg/ml de protéinase K à 20 mg/ml durant 6 heures à 50°C sous agitation basse (500 rpm). Un second bain identique est ensuite réalisé en ajoutant 1 mM de DTT. Enfin, un troisième bain est réalisé, identique au premier mais incubé durant 12h. Les blocs sont ensuite rincés 4 fois durant 15 minutes à 50°C dans de l'eau ultra-pure au premier lavage puis dans du tampon TE (10 mM Tris-HCl pH 8 et 0,1 mM EDTA).

Dans le cas d'une digestion enzymatique, les blocs sont ensuite rincés dans de l'eau ultra-pure durant 15 minutes à 50°C puis deux fois durant 1h dans du tampon TE contenant un inhibiteur de protéase, le fluorure de phénylméthylsulfonyle (PSMF) et enfin 1 fois dans du tampon TE durant 30 minutes sur glace. Les blocs sont ensuite incubés dans le tampon de digestion de l'enzyme durant 30 minutes. A noter, en cas de digestions successives, il convient de réaliser la digestion dont le tampon est le moins salin en premier. Une fois l'incubation en tampon de digestion réalisée, ce

dernier est changé et est ajouté 20 unités enzymatiques de l'enzyme de restriction. Les blocs sont ainsi incubés durant 20 minutes à température ambiante puis 6h à la température recommandée. Les échantillons sont ensuite placés dans des réactifs frais pour à nouveau 12h de lyse. Une fois la lyse terminée, les blocs sont rincés 2 fois, une première fois 1 heure dans de l'eau ultra-pure et la seconde fois durant 2 heures à 50°C dans du tampon de lyse contenant 1 mg/ml de protéinase K. Après cette étape d'inactivation de l'enzyme de restriction, les blocs sont rincés 3 fois dans du tampon TE durant 15 minutes à 50°C.

Une fois les blocs prêts, ils sont chargés dans un gel d'agarose 1% (m/v) en TBE 0,5 X puis scellé avec l'excédent d'agarose. Le gel est placé dans la cuve de migration. Le générateur est ensuite programmé pour permettre une migration dans une gamme de taille de bande donnée avec une migration linéaire de 6 V/cm. La fréquence d'alternance du courant électrique ainsi que la durée de migration sont choisies de façon automatique par l'appareil en fonction des paramètres de taille de fragments attendus donnés.

Extraction et purification de l'ADN génomique

La nature des capsides virales n'etant pas homogène parmi les virus géants, il existe deux types de protocoles pour l'extraction de l'ADN génomique. Le premier protocole concerne les virus de la famille des *Mimiviridae*, *Marseilleviridae* et mollivirus. Le second protocole s'applique pour les *Pandoraviridae* et les *Pithoviridae.*

L'ADN des mollivirus est ainsi extrait à partir du premier protocole, à partir d'une quantité de 5 x 10^9 particules virale environ. Le protocole suivi est celui recommandé dans kit PureLink™ Genomic DNA mini kit (Invitrogen) en ajoutant au tampon de lyse présent dans le kit du DTT à une concentration finale de 1 mM.

L'ADN des pandoravirus est extrait en suivant le second protocole et fait intervenir un tensioactif fort, le Bromure de cetyltrimethylammonium (CTAB). Un culot de 2 x 10^{10} particules virales est re-suspendu dans 300 µL d'eau ultra-pure et 500 µL de tampon contenant du CTAB (20 g/L CTAB, 1,4 M NaCl, 100 mM Tris-HCl, 20 mM Na2EDTA, pH 8,0), 20 mg/mL de Protéinase K et 3 µL de DTT 1 M sont ajoutés. Après 1 h d'incubation à 65°C les ARN présents dans la suspension virale sont hydrolysés par ajout de RNase A (20 mg/mL final et incubation durant 10 minutes à 65°C). Pour solubiliser l'ADN extrait dans 500 µL de chloroforme l'échantillon est vortexé pendant 30 secondes, puis centrifugé à 16 000 g pendant 10 minutes à 4°C. La phase aqueuse située dans la partie supérieure du tube est récupérée puis centrifugée à 16 000 g pendant 5 minutes à 4°C. Le surnageant est ensuite mélangé avec un volume équivalent de chloroforme,

vortexé et centrifugé à 16 000 g pendant 5 minutes à 4°C. La phase aqueuse est de nouveau récupérée et l'ADN est précipité en ajoutant 2 volumes de tampon de précipitation contenant 5 g/L CTAB, 40 mM NaCl, pH 8,0. Par inversions délicates du tube on observe la formation d'un précipité blanc. Ensuite l'échantillon est incubé 1 heure à température ambiante avant d'être centrifugé à 16 000 g pendant 5 minutes. Une fois le précipité récupéré, il est resuspendu dans 350 µL de NaCl 1,2 M, puis l'ADN est extrait à nouveau par ajout de 350 µL de chloroforme. L'échantillon est vortexé très brièvement avant centrifugation à 16 000 g pendant 10 minutes à 4°C. La phase aqueuse est récupérée, et 0,6 volumes d'isopropanol sont ajoutés pour permettre une seconde précipitation de l'ADN. Le tube est inversé plusieurs fois jusqu'à ce que l'ADNg forme un précipité transparent et refringeant. Le culot est récupéré par centrifugation à 16 000 g pendant 10 minutes à température ambiante puis lavé avec 500 µL d'éthanol à 70% en retournant plusieurs fois le tube. Enfin la suspension d'ADN est à nouveau centrifugée à 16 000 g pendant 10 minutes, le culot d'ADN est séché à l'air libre puis re-solubilisé dans 60 µL d'eau ultra-pure.

Séquençage de génome

Le séquençage des génomes de virus géant, et en particulier les génomes présentant un grand nombre de séquences répétées et/ou inversées, nécessite la combinaison de deux technologies: l'une générant des fragments séquencés de petites tailles et avec un faible taux d'erreur, Illumina Hiseq 2500 rapid, et une seconde générant des fragments longs mais de qualité moindre, Oxford Nanopore R9-Long Read 1D. Les banques d'ADN ainsi que les séquençages sont réalisés par nos collaborateurs du Genoscope au C.E.A. Les lectures séquencées sont ensuite récupérées au format fastq.

Assemblage des génomes

La nécessité de l'assemblage découle du fait que les technologies actuelles de séquencage génèrent des fragments d'ADN séquencés de taille inférieure à l'ADN analysé. Le but de l'assemblage est donc de reconstruire la séquence d'ADN originale par l'obtention de chaines de nucléotides les plus longues possibles (*contig)*. L'ensemble de ces *contig* constitue une ébauche de la séquencé d'intérêt (*scaffold*). La combinaison des deux techniques de séquençage a pour but de faciliter l'assemblage par l'obtention de lectures longues tout en ayant un faible taux d'erreur grâce aux lectures courtes. L'usage de jeux de données Nanopore et Illumina impose des contraintes techniques, notamment dans le choix d'un assembleur capable de compiler les données *short-read* Illumina et *long-read* Nanopore. Plusieurs assembleurs *de novo* ont ainsi été testés dans le cadre d'un *benchmark* incluant deux logiciels supportant en données d'entrée les lectures Nanopore et

Illumina :

- Spades 3.13 [82] : l'algorithme recherche les chevauchements de longueur k-1 entre des mots de longueur k. Cette longueur k, ou *k*-mer, peut être choisie manuellement et est conditionnée par la taille des lectures, les petites tailles de *k*-mer permettent d'obtenir un graphe largement connecté tandis que les tailles de *k*-mer les plus grandes permettent de résoudre les chevauchements de zones répétées. Le choix des *k*-mer constitue un compromis entre temps de calcul et qualité de l'assemblage. Les chemins les plus longs ainsi constitués matérialisent donc les *contig*, fragments plus ou moins complets de la séquence génomique originale. Ici, les tailles de *k*-mer choisies ici sont, par itérations successives: 21, 41, 61, 81 et 99 et l'option – careful est utilisée pour réduire le nombre de décalage et d'insertions délétions.

- Unicycler v0.4.8 [83] : Unicycler reprenant le fonctionnement de Spades. Pour chaque longueur de kmer un score est donné au graphe produit en fonction du nombre de *contig* et du nombre de *dead end*. Ainsi, le logiciel choisit automatiquement les tailles optimales de *kmer*. Un échantillon des lectures Nanopore comprenant 90% des meilleures lectures de plus de 1000 pb est sélectionné grâce au logiciel filtlong (option –min_length 1000, --keep_percent 90).

Et deux logiciels dédiés à l'assemblage des lectures Nanopore uniquement, incluant donc une étape de correction du *scaffold* par alignement des lecture Illumina (*polishing) :*

- Canu 2.0 [84] : Ce logiciel utilise un logiciel d'alignement de séquence, MHAP, qui permet de détecter les zones de chevauchement entre les lectures. Ici l'option genomeSize=2m est utilisée en se basant sur la taille moyenne des génomes de pandoravirus. L'assemblage donné par Canu est ensuite corrigé par alignement des lectures Illumina contre le *scaffold* généré (*polishing*).

- Miniasm-0.3 [85] : La technologie déployée par Miniasm reprend le principe des graphes en chaîne de Canu. Le logiciel d'alignement minimap2 permet de détecter les zones chevauchantes entre les lectures longue.

Annotation du génome et annotation fonctionnelle

L'annotation d'un génome assemblé consiste à prédire les gènes codés. La stratégie globale d'annotation des génomes repose sur trois axes : la détection de gènes *ab initio*, l'utilisation de données externes (transcriptome, protéome...), et des données de conservation des gènes entre

espèces proches (Figure 34). Dans la suite de ce manuscript nous verrons que les différents logiciels suivants sont utilisés ; pour la prédiction *ab initio* :

- AUGUSTUS.2.5.5 [86] : programme qui utilise des données RNA-seq ou de protéomique ainsi que pour effectuer un entrainement automatique *ab initio*.
- GeneMark-ES : programme permettant la recherche de gènes dans les génomes eucaryotes qui effectue un entraînement automatique *ab initio* non supervisé.
- GeneMark-ET [87] : programme reprenant le fonctionnement de GeneMark-ES en y ajoutant la possibilité d'intégrer des données d'alignements de lectures RNA-Seq dans la procédure d'entrainement.
- GeneMark-S : programme permettant la recherche de gènes *ab initio* dans les génomes procaryotes sans entraînement automatique.
- GetORF : programme permettant de détecter tous les ORF d'une taille donnée.

Pour l'utilisation de données externes les logiciels suivants sont utilisés :

- BRAKER [88] : programme qui combine des données de transcriptomique et les logiciels AUGUSTUS et GeneMark-ES pour faire des prédictions de structures de gène *ab initio*.
- PASA [89] : programme d'annotation de génome eucaryote utilisant les transcrits produits par Trinity [90] pour la prédiction de gènes.

Pour l'utilisation de données de conservation le logiciel suivant est utilisé :

- Exonerate [91] : programme d'alignement permettant, entre autre, d'aligner sans saut des séquences protéiques contre des séquences nucléiques.

Une fois les différents types de données produites, ces dernières sont compilées. Chaque outil de prédiction est pondéré et le choix de la meilleure prédiction est permise grâce au logiciel EvidenceModeler-1.1.1 (EVM) [89].

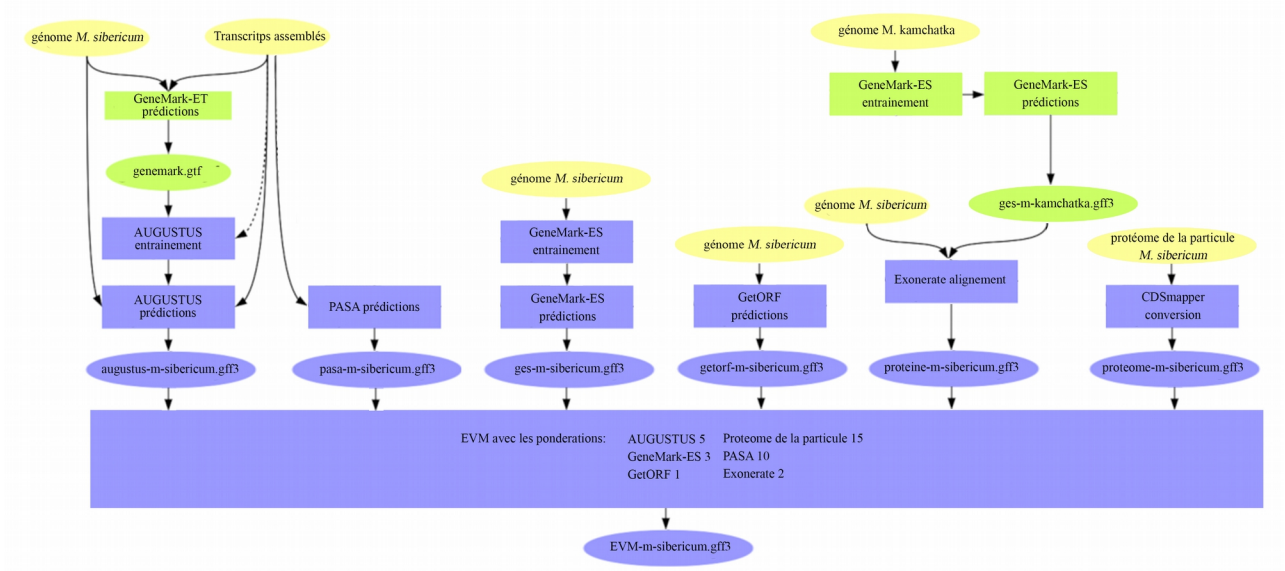


Figure 34 : Schéma de principe de la procédure d'annotation de M. sibericum.

Annotation du génome de *M. sibericum*

Le génome de *M. sibericum* avait été annoté en 2015 lors de sa découvert, l'utilisation de données de transcriptome avait permis une première annotation. Cependant, les données de transcriptomique n'avaient pas été assemblées en conservant l'orientation des transcrits. Il a donc été décidé de faire une nouvelle annotation de *M. sibericum* en mettant à profit les données de transcriptome orientées dans le processus d'annotation (Figure 35). Les lectures sont alignées contre le génome de *M. sibericum* grâce au logiciel Tophat-2.0.0 puis les lectures alignées sont assemblées en transcrits en prenant en compte l'orientation des ARN avec le logiciel Trinity-v2.11.0.

Une fois les données de transcriptome générées, la prédiction de gènes (données externes) est faite comme suit :

- PASA : les transcrits alignés sur plus de 75% de leur taille avec un pourcentage d'identité minimum de 95% sont utilisés. Les paramètres de PASA sont: --MAX_INTRON_LENGTH 1000 pour donner la taille maximale des introns, --transcribed_is_aligned_orient pour conserver l'orientation des transcrits et –stringent_alignement_overlap 30

Les trois logiciels de prédiction *ab initio* suivant sont utilisés pour l'annnotation du génome de *M. sibericum* avec les options suivantes :

- GeneMark-ES : --format GFF --fnn --faa et --virus

- GetORF : -filter et -min 300 pour récupérer l'ensemble des protéines de plus de 100 acides aminés
- AUGUSTUS.2.5.5 : paramètres par défaut

Les données de conservation sont produites comme suit :

- Exonerate : à partir d'un assemblage *de novo* de *M. kamchatka* les ORF predits sont alignés contre le génome de *M. sibericum*
- CDSmapper [92] : permet d'aligner le protéome de la particule de *M. sibericum* contre le génome.

Enfin, l'ensemble de ces données est compilé par le logiciel EVM avec les pondérations suivantes : AUGUSTUS 5, GeneMark-ES 3, GetORF 1, PASA 10, Protéome de la particule 15, Exonerate 2.

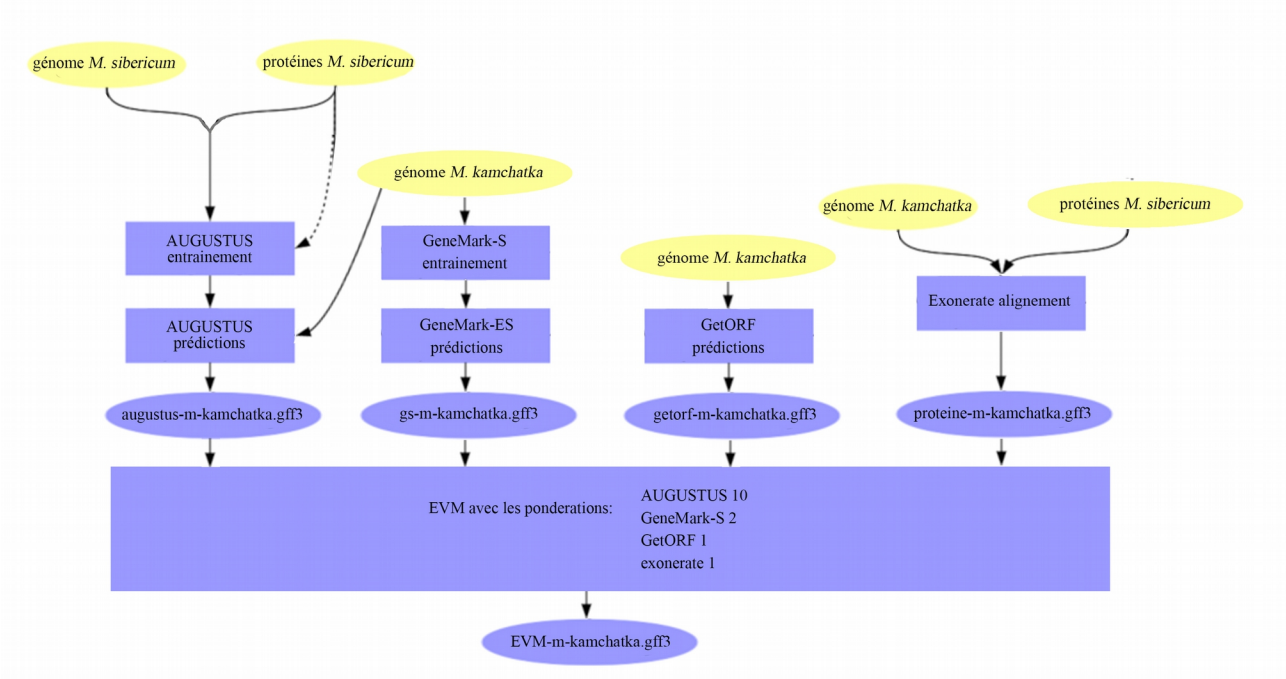


Figure 35 : Schéma de principe de la procédure d'annotation de M. kamchatka.

Annotation du génome de *M. kamchatka*

Le génome de *M. kamchatka* est annoté de façon similaire à celui de *M. sibericum*.

Trois logiciels de prédiction *ab initio* sont utilisés :

- GeneMark-S : avec les option par défaut
- GetORF : avec les options -filter et -min 300 pour récupérer l'ensemble des protéines de plus de 100 acides aminés

- AUGUSTUS.2.5.5 : paramètres par défaut

Les données de conservation sont produites ainsi :

- Exonerate : les ORF prédits de *M. sibericum* sont alignés contre le génome de *M. kamchatka*

L'ensemble de ces données sont compilées par le logiciel EVM avec les pondérations suivantes : AUGUSTUS 10, GeneMark-S 2, GetORF 1, PASA 10, Exonerate 1.

Annotation fonctionnelle des gènes prédits

L'annotation fonctionnelle permet de prédire les fonctions des gènes annotés sur le génome. Pour se faire deux approches sont combinées. La première consiste à chercher les domaines protéiques conservés, la seconde à rechercher des homologies des séquences avec des protéines connues.

La détection des domaines conservés se fait grâce à l'outil CD-Search [93]. Pour chaque protéine à analyser, l'algorithme de CD-Search appelle l'outil Reverse PSI-BLAST pour comparer la séquence à la base de données des domaines conservés (CDD). La base de données CDD compile les alignements de chaque protéines au domaine connu avec ses séquences homologues. Ces alignements permettent de construire des matrices où, à chaque position, est donné un score en fonction du niveau de conservation de l'acide aminé à cette position (matrice PSSM). Le résultat de l'alignement entre la protéine de fonction inconnue est la base de donnée CDD est une prédiction de domaine avec un score et une *E*-valeur, similaire aux résultats obtenus par Reverse PSI-BLAST. Dans notre cas, nous avons défini un seuil de *E*-valeur à 10^{-5} et les protéines prédites sont comparées à 55570 profils PSSMs.

La deuxième approche pour l'annotation fonctionnelle consiste à rechercher des protéines homologues par comparaison de profils HMM (Hidden Markov-Model) grâce au logiciel Hhblits [94]. A partir de la séquence d'entrée, Hhblits construit un alignement multiple de séquences (MSA). De façon itérative, un profil HMM de la séquence est construit à chaque tour d'alignement contre une base de donnée de profils HMM pré-construits. Les paramètres Hhblits utilisés sont les suivants : la base de donnée de profil HMM référence utilisée est uniclust 30 (version septembre 2018), -id 100 définissant le pourcentage maximal d'identité, -qid 50 le pourcentage d'identité minimal, -e 10^{-5} *E*-valeur de 10^{5}, Z 100 donnant le nombre de lignes de résultat maximal, -B 100 donnant le nombre d'alignements maximal.

Les résultats obtenus par ces deux outils sont ensuite compilés dans un fichier tabulé et manuellement inspectés. Quatre niveaux d'annotation sont faits : une prédiction concordante entre les deux logiciels entraine l'annotation de la protéine avec la fonction prédite, deux prédictions convergentes mais avec un domaine apparaissant incomplet au vu du résultat CD-Search entraine une annotation « partial domain » et en cas de discordance ou d'absence de prédiction la protéine est annotée « hypothétique ». Cette démarche combine deux approches et permet par conséquent de réduire les fausses prédictions. Cependant, une telle stringence entraine également une perte de sensibilité.

Génomique Comparative

« Clustering » de gènes homologues

Des protéines homologues sont des protéines dont les gènes qui les codent ont une origine commune. On distingue par conséquent deux cas d'homologie : deux gènes homologues dans deux génomes distincts sont dits orthologues et partagent une histoire évolutive commune : celle qui s'est écoulée avant la spéciation ; de même, deux gènes homologues dans un même génome sont dits paralogues et partagent une histoire évolutive commune : celle qui s'est écoulée avant la duplication. Ainsi, bien qu'il n'existe aucune corrélation stricte entre le niveau de conservation des gènes orthologues et le niveau de conservation des protéines codés par ces mêmes gènes, la présence de gènes communs dans deux espèces distinctes traduit une histoire évolutive commune entre ces deux organismes. Pour étudier l'histoire évolutive entre des virus d'une même famille ou entre deux familles virales, il convient donc de rechercher l'ensemble des groupes de gènes communs, appelés « orthogroupes », aux espèces considérées. Pour réaliser ces orthogroupes le logiciel OrthoFinder-2.2.7 compare l'ensemble des protéines une à une entre elles par BLASTp (seuil *E*-valeur 10^{-3}) [95]. Ce processus permet donc la formation d'un graphe dont les sommets sont l'ensemble des séquences considérées et les branches le *bit-score* de BLASTp entre deux séquences. La segmentation des sommets permet ensuite de définir les séquences d'un même orthogroupe. Pour ces travaux, le logiciel Orthofinder est utilisé avec les paramètres par défaut, en ajoutant l'option -M msa pour générer les alignements entre les séquences homologues. De même, pour l'ensemble des orthogroupes sans paralogues, le pourcentage d'identité est calculé par BLASTp réciproque entre les séquences.

Phylogénie

La détection de gènes homologues entre différentes espèces permet d'établir une histoire

évolutive commune entre ces organismes. La phylogénie est l'un des moyens permettant d'étudier cette histoire évolutive. La manière courante de figurer cette histoire est la construction d'un arbre phylogénétique. La construction d'un arbre phylogénétique se fait en trois étapes, il faut en premier lieu faire un alignement multiple global entre les séquences homologues d'intérêt. Puis, estimer le meilleur modèle d'évolution et enfin reconstruire l'arbre à proprement parlé.

Il existe deux principales approches pour évaluer la distance entre des séquences : la première est basée sur des mesures de distances (*i.e* nombre de substitutions par site) et la seconde est statistique, basée sur l'étude des états de caractères (nucléotide ou acide aminé présent à une position, présence ou absence d'une insertion/délétion). Parmi les méthodes usuelles de distance, on retrouve la méthode dite de *Neighbor Joining* (NJ), les méthodes statistiques usuelles sont le maximum de vraisemblance (ML) et Bayes.

Pour les études présentées dans ce manuscrit, nous avons choisi d'utiliser la méthode ML. Cette dernière est actuellement largement utilisée et repose sur l'étude indépendante de chaque position de l'alignement. La vraisemblance maximale des données est évaluée en calculant la vraisemblance sous différentes hypothèses évolutives d'un modèle d'évolution. Ainsi, chaque position est considérée indépendante et le logarithme de la vraisemblance est calculé pour une topologie donnée, en utilisant un modèle d'évolution particulier. Le logarithme de la vraisemblance est cumulé sur tous les sites et sa somme est maximisée pour estimer la longueur des branches de l'arbre. Une fois la reconstruction terminée, les arbres générés par méthode ML doivent être enracinés. Pour se faire il existe deux options, définir un groupe de séquences considérées comme groupe extérieur ou enraciner l'arbre au poids moyen (*mid-point rooting*). De même, la robustesse de l'arbre doit être évaluée. Pour se faire, un nombre donné de tirages avec remise est fait à partir du jeu de données initial. Pour chaque tirage on calcule la phylogénie correspondante par la même méthode et la présence de chaque nœud dans les différentes simulations est évaluée (*bootstrap*).

La procédure que nous avons choisie dans ce manuscrit est la suivante : les alignements globaux sont faits avec MAFFT v7.407 et l'option G-INS-i pour alignement global pair à pair avec raffinement itératif. Pour déduire les arbres phylogénétiques par maximum de vraisemblance nous avons choisi le logiciel IQ-TREE 1.6.7 avec l'option -m TEST qui permet une sélection rapide et précise du modèle dévolution le plus en adéquation avec les observations, la robustesse de l'arbre généré est évaluée en ajoutant l'option -bb 1000, permettant de réaliser 1000 ré-échantillonges avec remise du jeu de données. L'arbre généré est ensuite visualisé grâce au logiciel Archéoptéryx et est enraciné par méthode *mid-point*.

Pression de sélection

Lorsqu'on veut estimer la pression de sélection appliquée sur un couple de gènes orthologues, plusieurs paramètres sont calculés. Le premier est le taux de substituons synonymes par positions synonymes (dS) et le nombre de substitutions non synonymes par positions non synonymes (dN). Le ratio dN/dS calculé pour une paire de protéines orthologues, ou omega (ω), matérialise donc l'équilibre entre accumulation de mutation aboutissant à la modification de la séquence de la protéine et mutation conservant la structure primaire de la protéine. Il existe trois cas d'interprétation de la valeur ω :

- $\omega < 1$: il y a moins de substitutions non synonymes qu'attendu car les substitutions délétères ont été éliminées, le gène est donc sous pression de sélection négative
- $\omega > 1$: il y a plus de substitutions non synonymes qu'attendu, le gène est sous pression de sélection positive
- $\omega \approx 1$: le gène est sous pression de sélection neutre

On considère alors que les gènes sous pression de sélection négative, dont la fonction est strictement maintenue, sont essentiels au maintien et à la survie de l'organisme considéré. Il existe deux possibilités pour expliquer qu'un gène soit sous pression de sélection positive. Soit, il peut s'agir d'un gène dont l'accumulation de mutations non synonymes entraine une diversification permettant de maintenir sa fonction, c'est le cas par exemple des gènes codant des antigènes pathogènes [96] ; soit, le gène accumule des mutations non synonymes, conduisant potentiellement conférer un avantage sélectif a l'organisme étudié.

Il convient de préciser que cette méthode contient intrinsèquement des limites. Ainsi pour des espèces divergentes la saturation en mutation synonyme peut entrainer des valeurs d'ω aberrantes car il est virtuellement impossible de prédire le nombre de mutations accumulées à une position donnée. Ainsi il convient d'être prudent sur la valeur seuil du dS. De plus, des espèces trop proches, ou ayant divergés il y a très peu de temps, peuvent accumuler des mutations non-synonymes qui ne sont pas encore perdues, entrainant une sur-estimation des valeurs de dN. Par conséquent, il convient d'être prudent dans l'analyse des valeurs ω obtenues.

Pour calculer la pression de sélection appliquée sur une paire de gènes homologues il faut dans un premier temps aligner ces protéines, convertir cet alignement en un alignement des codons et ensuite calculer la valeur ω associée. Ici les protéines sont alignées avec MAFFT v7.407 et les alignements sont contrôlés manuellement [97]. L'alignement des codons correspondant est réalisé avec

pal2nal.v14 et l'option -nogap [98]. Les valeurs de ω sont calculées par le logiciel codeml avec les paramètres : runmode = -2 pour *pairwise alignment,* model = 2 pour le modèle de codon et CodonFreq = 2 qui suppose que la probabilité de substitution est égale pour chaque codon et la fréquence des codons est calculée en fonction de la fréquence de chaque nucléotide dans chaque position du codon [99]. Pour pallier aux éventuels biais, les valeurs de ω obtenues sont sélectionnées en considérant les seuils suivants : 0<dS <=2 et dN >0.

Calcul de l'adaptation des codons

L'index d'adaptation des codons (CAI) mesure l'écart entre les codons utilisés par un gène codant pour une protéine donnée par rapport à un ensemble de gènes de référence, plus la valeur de CAI est proche de 1 plus l'usage des codons entre la référence et l'organisme testé est proche. Le CAI permet de prédire le niveau d'expression d'un gène basé sur sa séquence nucléique. Le logiciel utilisé est cai de la suite EMBOSS et la table de référence est construite à partir d'un ensemble de 771 gènes constitutifs de *A. castellanii* [100].

Metagénomique

Extraction d'ADN total

L 'extraction d'ADN total des échantillons A, B et C, constitue un défi technique pour plusieurs raisons. En effet, il est nécessaire de maintenir les échantillons congelés et garantir leur stérilité lors de la manipulation, pour permettre l'extraction de l'ADN viral potentiellement contenu dans les virions présents dans l'échantillon, nécessite d'ajuster le protocole proposé par le fabricant du kit. De plus, la quantité minimale de matériel génétique pour le séquençage est de 1 µg d'ADN, pour atteindre cette quantité minimale d'ADN il faut donc multiplier les duplicats. Chaque série d'échantillon a été traitée en utilisant le kit Dneasy PowerSoil Kit de Qiagen. Les séries A et B sont traitées avec la version miniprep du kit Dneasy PowerSoil Kit de Qiagen, en suivant le protocole recommandé par le fabricant, en ajoutant au tampon de lyse C1 5 µL de DTT 1M. Les échantillons sont d'abord traités en double (2 tubes contenant 0,25 grammes d'échantillon par tube) puis en sextuple si la quantité d'ADN dosée était inférieure à 1 µg. La série C est traitée avec la version maxiprep du kit Dneasy PowerSoil Kit de Qiagen en duplicat (avec 20 grammes d'échantillon par colonne) selon les recommandations du fabricant en ajoutant 145 µL de DTT 1 M au tampon de lyse C1 et les étapes suivantes : Après ajout du tampon C1 contenant du DTT, les échantillons sont broyés mécaniquement durant 20 secondes dans un homogénéiseur MP FastPrep à une vitesse de 4

m/s, incubé 30 minutes à 65°C puis à nouveau broyé 20 secondes à 4 m/s, la suite du protocole est celle recommandée par le fabricant. Après élution dans 5 ml de tampon d'élution l'ADN extrait est concentré sur une colonne de silice issue du kit Monarch Génomic DNA purification Kit de NEB jusqu'à obtention d'une quantité d'ADN suffisante.

Assemblages, annotation et sélection des *contig*

L'ensemble des métagénomes a séquencé par méthode Illumina HiSeq 432 grâce à une collaboration avec le Génoscope. La qualité des séquences brutes a été contrôlée grâce au logiciel FASTQC. Les lectures contaminantes explicites sont enlevées et les extrémités 3' des lectures sont éliminées si leur score de qualité était inférieur à 30 grâce au logiciel BBTools (33). Les lectures sont ensuite assemblées en utilisant le logiciel MEGAHIT avec les options: "--k-list 33,55,77,99,127 436 --min-contig-len 1000". Les lectures sont ensuite alignées contre les *contigs* produits grâce au logiciel Bowtie2 avec l'option "--very-sensitive".

RESULTATS

Métagénome, diverses pistes à l'étude

Des 31 échantillons de cryosol, 16 ont pu fournir une quantité d'ADN suffisante de 1 µg pour le séquençage. Parmi les échantillons séquencés, 8 sont en provenance de la zone de Yukéchi (série C), 3 du Kamchatka (série A), deux des abords de la rivière Kolyma (série B), un de la rivière Omolon (série B) et l'échantillon en provenance de l'Antarctique (série B) (Tableau 1). Comme annoncé plus haut dans ce manuscrit, du fait de la réactivation et l'étude rapide de deux nouveaux virus, les travaux effectués sur les profils métagénomiques sont encore à l'état d'ébauche.

#	Collaboration	Collaborateur	Nom	Nomenclature	Origine
1	Kamtchatka	Alexander Morawitz	P2	C	shore of Kronotsky River
2	Kamtchatka	Alexander Morawitz	P4	D	near lake at Kizimen Volcano
3	Kamtchatka	Alexander Morawitz	P5	E	Shapina river bank
4	Pushchino	Stas Malavin	Lab1	F	River Bolshaya Chukochya
5	Pushchino	Stas Malavin	Lab8	G	Antarctic Modern Soil
6	Pushchino	Stas Malavin	Lab10	H	Kolyma River (old, permafrost nest)
7	Pushchino	Stas Malavin	Lab11	I	Modern soil Alazeya River
8	Pushchino	Stas Malavin	Lab14	J	Omolon River (old)
9	A. Wegener Inst	Jens Strauss	Yed6	K	Yedoma_Dry, -16m, frozen
10	A. Wegener Inst	Jens Strauss	Ala11	L	AlasCenter_Dry, -12m, frozen
11	A. Wegener Inst	Jens Strauss	Y2	M	Yedoma_Lake, -16m, frozen
12	A. Wegener Inst	Jens Strauss	Ala10	N	AlasCenter_Dry, -16m, frozen
13	A. Wegener Inst	Jens Strauss	Y1	O	Yedoma_Lake, -19m, frozen
14	A. Wegener Inst	Jens Strauss	Yed7	P	Yedoma_Dry, -11m, frozen
15	A. Wegener Inst	Jens Strauss	Y4	Q	Yedoma_Lake, -6m, thawed
16	A. Wegener Inst	Jens Strauss	Ala9	R	AlasCenter_Dry, -20m, frozen

Tableau 1 : Liste des métagénomes obtenus

Les travaux de Artemiya Goncharov du laboratoire de médecine expérimentale de l'Université de médécine Mechnikov à Saint-petersbourg ont démontré que le pergélisol arctique était un réservoir naturel d'éléments génétiques mobiles procaryotes porteurs de gènes anciens de résistance aux antibiotiques [101]. Il a été également démontré que ces éléments mobiles étaient capables de se conjuguer dans des génomes de bactéries modernes, ainsi le plasmide pKLH80 isolé dans du pergélisol du pleistocène et contenant le gène blaRTG-6 codant pour une beta-lactamase a été transmis à une souche nosocomiale d'*Acinetobacter baumanii* [102]. Les données de séquençage de pergélisol ont donc été exploitées dans l'objectif de découvrir les gènes de résistance aux antibiotiques présents dans les différents échantillons de cryosol. Les données de séquençage ont permis d'assembler un total de 869178 contig de plus de 5 kpb, desquels ont été prédits 10152851 ORF. Des membres du laboratoire sont actuellement chargés d'étudier la taxonomie des lectures assemblées dans le but de détecter de potentiels éléments génétiques mobiles porteurs de gènes de résistance aux antibiotiques.

Bien qu'à l'heure actuelle aucune taxonomie des lectures assemblées n'a été faite. Une première démarche pour détecter des micro-organismes dans les métagénomes a consisté à associer une taxonomie directement aux lectures filtrées par l'utilisation du logiciel Kraken [103] Les premiers pathogènes recherchés sont les pathogènes des catégories A et B tels que définis par le Centre de Contrôle des Maladies, aux Etat-Unis (CDC). Ce premier résultat indique que le pourcentage de lectures associées à ces pathogènes est très faible (Figure 36). La question qui se pose est donc de savoir si il est possible d'envisager que ces micro-organismes soient réellement présents dans les échantillons et, de surcroit encore infectieux. A l'heure actuelle, l'équipe en charge de la poursuite

de cette étude se concentre sur la mise en place d'un protocole permettant d'assembler les lectures ayant une taxonomie commune ensemble.

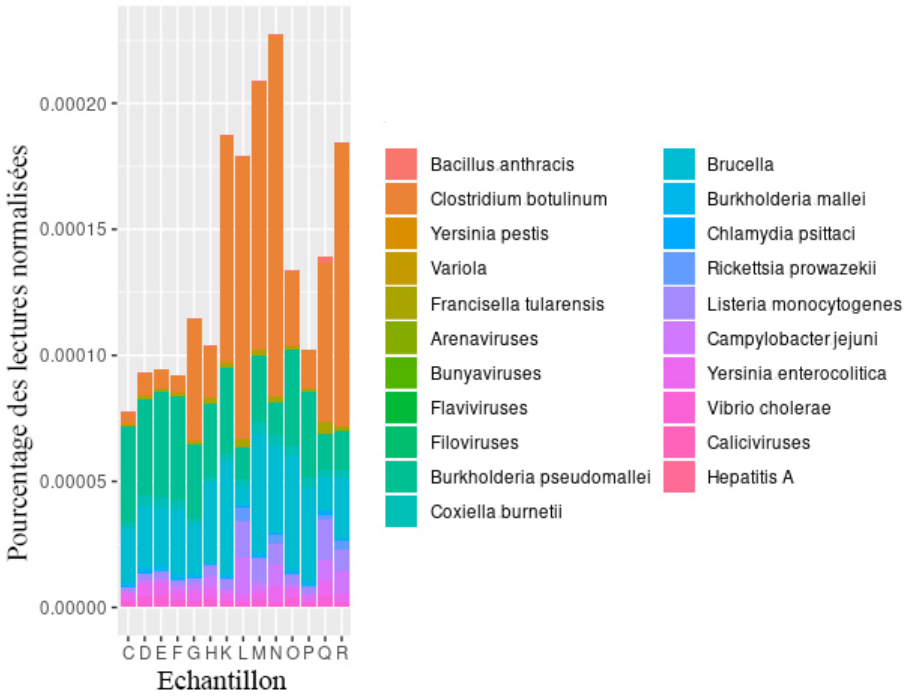


Figure 36 : Pourcentage des lectures associées à chacun des micro-organismes de classe A ou B dans les différents métagénomes.

Les pandoravirus, des virus géants au génome intriguant

Suivant la découverte historique de *P. dulcis* e t *P. salinus,* le laboratoire a continué d'explorer différents environnements à la recherche de nouveaux pandoravirus. Depuis lors, trois nouveaux virus ont été isolés et séquencés : *P. quercus* à partir d'un échantillon de sol en provenance de Marseille, *P. macleodensis* dans un échantillon d'eau douce en provenance de Melbourne et *P. neocaledonia* à partir d'eau saumâtre d'une mangrove aux abords de l'aéroport de Nouméa ; portant donc à 5 le nombre de virus séquencés et caractérisés au laboratoire. Il est donc apparu à ce moment que les *Pandoraviridae* constituaient une nouvelle famille virale dont les représentants étaient distribués sur tout le globe. L'obtention de ces 5 virus a rendu possible l'étude des caractéristiques intrinsèques et l'évolution de cette famille virale par la combinaison de plusieurs approches dites « omiques », incluant analyse transcriptomique et analyse protéomique de la particule, et expérimentales de microscopie, incluant observation par microscopie optique et électronique. Pour compléter l'analyse génomique des *Pandoraviridae,* la série de génomes viraux a été complétée par les données de séquençage d'un pandoravirus caractérisé en 2015 en Allemagne par un autre laboratoire et baptisé *P. inopinatum.* Paradoxalement, *P. inopinatum* avait été observé

sans que sa nature virale ait été découverte en 2008, suite à une étude des endosymbiontes procaryotes de *A. castellanii* publiée par le laboratoire allemand.

L'étude comparée des cycles infectieux de *P. quercus, P. macleodensis* et *P. neocaledonia* a permis de déceler des nouvelles particularités propres à la réplication des *Pandoraviridae.* Ainsi il est apparu que bien que les pandoravirus aient une dépendance stricte au noyau de l'hôte, des capsides virales en cours de synthèse ont été observées dans des cellules comportant encore une enveloppe nucléaire intègre. Ainsi, la rupture de l'enveloppe nucléaire n'est plus apparue comme systématique, comme avancé précédemment. De plus, dans le cadre de l'étude du cycle infectieux des *Pandoraviridae*, j'ai été amené à réaliser des observations par microscopie optique du cycle infectieux dans le but de caractériser le phénomène d'exocytose des particules virales. Pour ce faire, des lamelles couvertes de poly-lisine ont été ensemencées, infectées à MOI 1000, puis observées sans fixation sur un microscope (Zeiss Axio Observer) avec un grossissement total de 1008. Ce protocole a permis d'observer à 5h post infection une vésicule d'exocytose contenant des virions de pandoravirus fusionnant avec la membrane externe de l'amibe et relarguant les particules virales dans le milieux extérieur. Cela prouve donc que l'exocytose est le moyen de sortie des virions à ce stade du cycle infectieux. Ce résultat m'a permis d'être cité en tant qu'auteur dans l'article sous-cité.

Sur le plan génétique, la première caractéristique notable des pandoravirus est le gigantisme de leur génome, dont la taille varie de 1,8 Mpb pour *P. macleodensis* (1552 ORF) à 2,4 Mpb (2394 ORF) pour *P. salinus*. Faisant des *Pandoraviridae* les plus gros virus connus à ce jour. La recherche de protéines homologues aux protéines de pandoravirus dans les bases de données a fait apparaître que 70% des gènes prédits n'avaient pas d'homologues en dehors des *Pandoraviridae.* Ce nombre passe à 80% lorsqu'on s'intéresse uniquement aux gènes codants pour des protéines retrouvées dans la capside virale. Cette grande proportion de gènes inconnus, baptisés ORFan, fait donc des pandoravirus des réservoirs privilégiés de gènes codant pour des protéines aux fonctions nouvelles et dont la présence et l'origine restent à caractériser. Le séquençage de ses souches a également permis de réaliser une première phylogénie des *Pandoraviridae.* Il est ainsi apparu que les pandoravirus se regroupent en deux lignées distinctes, avec d'un côté la lignée A (*P. dulcis*, *P. salinus*, *P. quercus* et *P. celtis*) et de l'autre la lignée B (*P. macleodensis, P. neocaledonia*).

Le clustering des gènes homologues a fait apparaître 455 groupes de gènes partagés, délimitant ainsi le génome cœur des pandoravirus. Si on s'intéresse maintenant à l'ensemble de la diversité génétique des *Pandoraviridae,* il apparaît que l'allure de la courbe de saturation du pan-génome est caractéristique des pan-génomes dits ouverts ($\alpha = 0,83$). Ainsi, à chaque découverte d'un nouveau pandoravirus le pan-génome s'enrichit de 50 à 60 nouveaux gènes (Figure 37). Pour

déterminer l'origine de ces gènes uniques à chaque souche plusieurs hypothèses ont été testées.

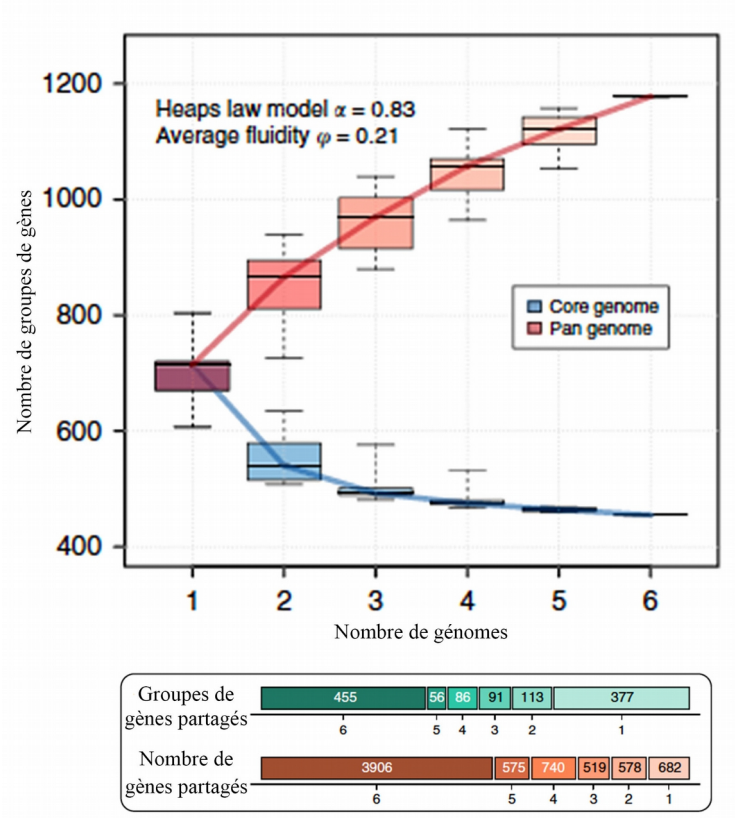


Figure 37 : Comparaison du contenu en gènes des Pandoraviridae. Le paramètre α, est caractéristique d'un pan-génome ouvert (α <1). Le nombre de groupes de gènes partagés défini le génome cœur des Pandoraviridae.

La présence de 70% d'ORFans laisse donc la place à ce que 30% des gènes soient des candidats potentiels à des transferts de gènes horizontaux. L'analyse phylogénétique de ces gènes a démontré que seuls 6 à 15% de ces derniers étaient issus de HGT potentiels. De plus, le sens de transfert est pour moitié seulement dans le sens cellule vers virus. Ce ratio de gènes acquis par transferts horizontaux, retrouvé chez d'autres NCLDV comme mimivirus dont le génome est plus petit, n'explique pas à lui seul le gigantisme des *Pandoraviridae*.

Le clustering des 11468 protéines prédites chez l'ensemble des 6 virus a mis en évidence que 50% des gènes étaient dupliqués (Figure 38). L'analyse de la position des paralogues détectés montre que ce phénomène de duplication a été ensuite suivi des réarrangements entraînant la perte de colinéarité entre les gènes paralogues. Enfin, le phénomène de duplication est marginal lorsqu'on considère uniquement les gènes uniques à chaque souche. Ainsi, bien que les *Pandoraviridae* soient les plus enclins au phénomène de duplication, cela n'explique pas entièrement le gigantisme de ces virus.

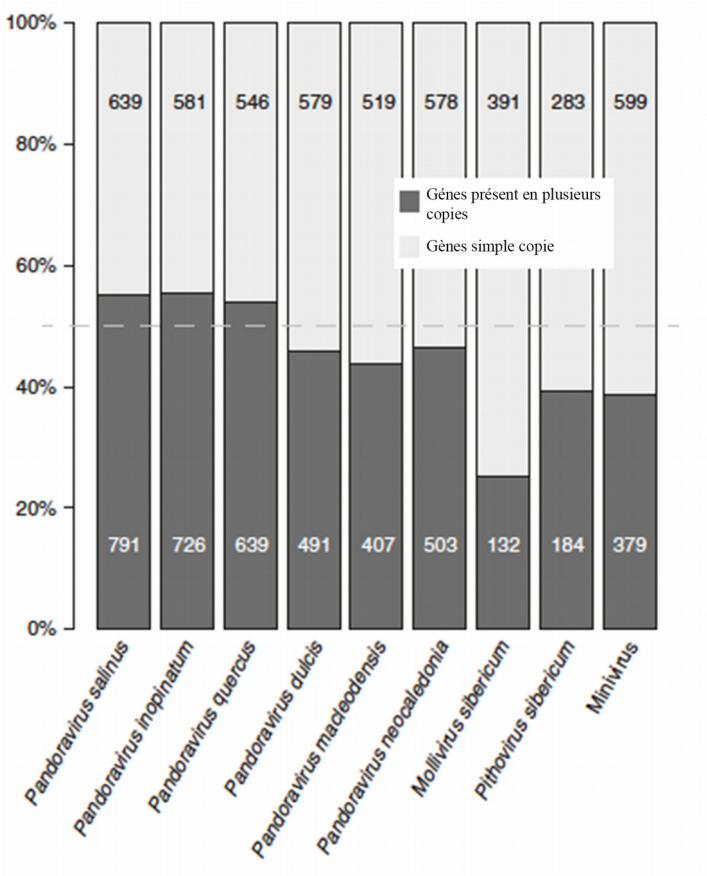


Figure 38 : Distribution des gènes présents en une seule copie VS les gènes présents en plusieurs copies dans plusieurs familles de virus géants.

En considérant que ni les HGT ni les duplications expliquent à eux seuls la complexité des génomes de pandoravirus, trois classes de gènes ont été étudiées séparément : les gènes du génome cœur, les ORFans spécifiques d'une lignée et les ORFans uniques à chaque virus (ORFans stricts). L'étude de ces trois classes de gènes a fait apparaître que ces derniers se distinguent selon trois critères : le CAI, la taille et le GC%. Ainsi, les ORFans spécifiques d'une lignée et les ORFans uniques à chaque virus ont des caractéristiques uniques, plus proches de caractéristiques des régions intergéniques : GC% statistiquement plus bas et CAI distinct des gènes coeurs. L'hypothèse la plus probable expliquant la présence de ces gènes est donc que ces derniers ont été crées *de novo* à partir des régions intergéniques. Cette hypothèse est consistante avec le fait que 38% du génome de pandoravirus est constitué de régions intergéniques, offrant donc la place à ce mécanisme pour s'établir. De même, la taille moyenne des ORFans stricts est supérieure à 100 acides aminés, augmentant la probabilité que les protéines produites forment des structures secondaires. Le processus de création de gènes *de novo* peut se dérouler de deux façons, soit la région intergénique acquiert la transcription pour ensuite devenir un ORF, soit l'inverse. La présence de deux ORF de *P. salinus* correspondant à un long ADN non codant de *P. néocaledonia* et à une région

intergénique de *P. quercus* suggère que l'apparition de gènes crées *de novo* est issue en premier lieu de l'acquisition par des régions intergéniques de la transcription puis d'une structuration en ORF.

Pour conclure, les trois mécanismes contribuent de façon différentes à l'acquisition de nouveaux gènes par les pandoravirus : HGT, duplication et création de gènes.

Papier 1 : Legendre, M., Fabre, E., Poirot, O., Jeudy, S., Lartigue, A., Alempic, J. M., ... & Labadie, K. (2018). Diversity and evolution of the emerging Pandoraviridae family. Nature communications, 9(1), 1-12.

nature COMMUNICATIONS

ARTICLE

DOI: 10.1038/s41467-018-04698-4 OPEN

Diversity and evolution of the emerging *Pandoraviridae* family

Matthieu Legendre[1], Elisabeth Fabre[1], Olivier Poirot[1], Sandra Jeudy[1], Audrey Lartigue[1], Jean-Marie Alempic[1], Laure Beucher[2], Nadège Philippe[1], Lionel Bertaux[1], Eugène Christo-Foroux[1], Karine Labadie[3], Yohann Couté[2], Chantal Abergel[1] & Jean-Michel Claverie[1]

With DNA genomes reaching 2.5 Mb packed in particles of bacterium-like shape and dimension, the first two Acanthamoeba-infecting pandoraviruses remained up to now the most complex viruses since their discovery in 2013. Our isolation of three new strains from distant locations and environments is now used to perform the first comparative genomics analysis of the emerging worldwide-distributed *Pandoraviridae* family. Thorough annotation of the genomes combining transcriptomic, proteomic, and bioinformatic analyses reveals many non-coding transcripts and significantly reduces the former set of predicted protein-coding genes. Here we show that the pandoraviruses exhibit an open pan-genome, the enormous size of which is not adequately explained by gene duplications or horizontal transfers. As most of the strain-specific genes have no extant homolog and exhibit statistical features comparable to intergenic regions, we suggest that de novo gene creation could contribute to the evolution of the giant pandoravirus genomes.

[1] Aix Marseille Univ, CNRS, Structural and Genomic Information Laboratory, UMR 7256 (IMM FR 3479), 163 Avenue de Luminy, Case 934, 13288 Marseille cedex 9, France. [2] Univ. Grenoble Alpes, CEA, Inserm, BIG-BGE, 38000 Grenoble, France. [3] CEA-Institut de Génomique, GENOSCOPE, Centre National de Séquençage, 2 rue Gaston Crémieux, CP5706, 91057 Evry Cedex, France. Correspondence and requests for materials should be addressed to M.L. (email: legendre@igs.cnrs-mrs.fr) or to C.A. (email: Chantal.Abergel@igs.cnrs-mrs.fr) or to J.-M.C. (email: Jean-Michel.Claverie@univ-amu.fr)

For 10 years after the serendipitous discovery of the first giant virus (i.e., easily visible by light microscopy) Acanthamoeba polyphaga Mimivirus[1, 2], environmental sampling in search of other Acanthamoeba-infecting viruses only succeeded in the isolation of additional members of the *Mimiviridae* family[3, 4]. Then, when we returned in 2013 to the Chilean coastal area from where we previously isolated *Megavirus chilensis*[3], we isolated the even bigger *Pandoravirus salinus*[5]. Its unique characteristics suggested the existence of a different family of giant viruses infecting Acanthamoeba. The worldwide distribution of this predicted virus family, the proposed *Pandoraviridae*, was quickly hinted by our subsequent isolation of *Pandoravirus dulcis* more than 15 000 km away, in a freshwater pond near Melbourne, Australia[5]. We also spotted pandoravirus-like particles in an article reporting micrographs of Acanthamoeba infected by an unidentified "endosymbiont"[6], the genome sequence of which has recently become available as that of the German isolate *Pandoravirus inopinatum*[7].

Here we describe three new members of the proposed *Pandoraviridae* family that were isolated from different environments and distant locations: *Pandoravirus quercus*, isolated from ground soil in Marseille (France); *Pandoravirus neocaledonia*, isolated from the brackish water of a mangrove near Noumea airport (New Caledonia); and *Pandoravirus macleodensis*, isolated from a freshwater pond near Melbourne (Australia), only 700 m away from where we previously isolated *P. dulcis*. Following the characterization of their replication cycles in *Acanthamoeba castellanii* by light and electron microscopy, we analyzed the five pandoravirus strains available in our laboratory through combined genomic, transcriptomic, and proteomic approaches. We then used these data (together with the genome sequence of *P. inopinatum*) in a comparative manner to build a global picture of the emerging family and refine the genome annotation of each individual strain. While the number of encoded proteins has been revised downward, we unraveled hundreds of previously unpredicted genes associated to non-coding transcripts. From the comparison of the six representatives at our disposal, the *Pandoraviridae* family appears quite diverse in terms of gene content, consistent with a family for which many members are still to be isolated. A large fraction of the pan-genome codes for proteins without homologs in cells or other viruses, raising the question of their origin. The purified virions are made of more than 200 different proteins, about half of which are shared by all tested strains in well-correlated relative abundances. This large core proteome is consistent with the highly similar early infection stages exhibited by the different isolates.

Results

Environmental sampling and isolation of pandoravirus strains. We used the same isolation protocol that led to the discovery of *P. salinus* and *P. dulcis*[5]. It consists in mixing the sampled material with cultures of Acanthamoeba adapted to antibiotic concentrations high enough to inhibit the growth of other environmental microorganisms (especially bacteria and fungi). Samples were taken randomly from humid environments susceptible to harbor Acanthamoeba cells. This led to the isolation of three new pandoravirus strains: *P. quercus*; *P. neocaledonia*; and *P. macleodensis* (Table 1, see Methods). They exhibit adequate divergence to start assessing the conserved features and the variability of the emerging *Pandoraviridae* family. When appropriate, our analyses also include data from *P. inopinatum*, isolated in a German laboratory from a patient with an Acanthamoeba keratitis[7].

Study of the replication cycles and virion ultrastructures. Starting from purified particles inoculated into *A. castellanii* cultures, we analyzed the infectious cycle of each isolate using both light and transmission electron microscopy (ultrathin section). As previously observed for *P. salinus* and *P. dulcis*, the replication cycles of these new pandoraviruses were found to last an average of 12 h[5] (8 h for the fastest *P. neocaledonia*). The infectious process is the same for all viruses, beginning with the internalization of individual particles by Acanthamoeba cells. Following the opening of their apical pore, the particles ("pandoravirions") transfer their translucent content to the cytoplasm through the fusion of the virion internal membrane with that of the phagosome. The early stage of the infection is remarkably similar for all isolates. While we previously reported that the cell nucleus was fully disrupted during the late stage of the infectious cycle[5], the thorough observation of the new strains revealed neosynthetized particles in the cytoplasm of cells still exhibiting nucleus-like compartments in which the nucleolus was no longer recognizable (Supplementary Fig. 1). Eight hours post infection, mature virions became visible in vacuoles and are released through exocytosis (Supplementary Movie). For all isolates, the

Table 1 Data on the pandoravirus isolates used in this work

Name	Origin	Isolate	RNA-seq	Virion proteome	Genome size (bp) (G + C)%	*N* ORFs[a] (standard)	*N* Genes (stringent)
P. salinus	Chile	Ref. [5]	This work	This work	2473870 62%	2394 (2541)[a]	1430 ORFs 214 lncRNAs 3 tRNAs
P. dulcis	Australia	Ref. [5]	This work	This work	1908524 64%	1428 (1487)[a]	1070 ORFs 268 lncRNAs 1 tRNA
P. quercus	France (Marseille)	This work	This work	This work	2077288 61%	1863	1185 ORFs 157 lncRNAs 1 tRNA
P. neocaledonia	New Caledonia	This work	This work	This work	2003191 61%	1834	1081 ORFs 249 lncRNAs 3 tRNA
P. macleodensis	Australia (Melbourne)	This work	—	—	1838258 58%	1552	926 ORFs 1 tRNA
P. inopinatum	Germany	Ref. [6]	—	—	2243109 61%	2397 (1839)[a]	1307 ORFs 1 tRNA

NC non-protein-coding transcripts
[a]The number of ORFs predicted in the original publications are indicated in parenthesis below the number obtained using the same standard reannotation protocol for all genomes. A more stringent estimate (next column) is taking into account protein sequence similarity as well as RNA-seq and proteomic data when available (see Methods)

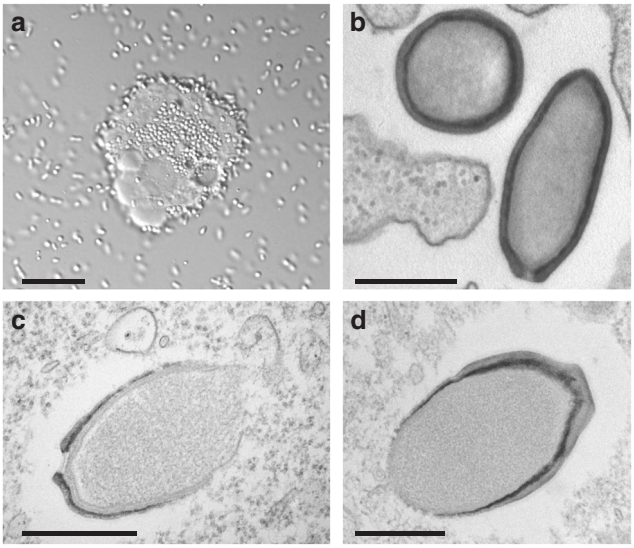


Fig. 1 The new pandoravirus isolates. **a** Overproduction by an *A. castellanii* cell of *Pandoravirus macleodensis* virions from the environmental sample prior cell lysis. Environmental bacteria can be seen in the culture medium together with *P. macleodensis* virions. (scale bar is 10 µm). **b** TEM image of an ultrathin section of *A. castellanii* cell during the early phase of infection by *P. neocaledonia*. The ameba pseudopods are ready to engulf the surrounding virions. Ten minutes pi, virions have been engulfed and are in vacuoles (scale bar is 500 nm). **c** TEM image of an ultrathin section of *A. castellanii* cell during the assembly process of a *P. salinus* virion (scale bar is 500 nm). **d** TEM image of an ultrathin section of a nascent *P. quercus* virion. (scale bar is 500 nm). The structures of the mature particles from the different strains do not exhibit any noticeable difference

replicative cycle ends with the cells lysis and the release of about a hundred particles (Fig. 1).

Genome sequencing and annotation. Genomic DNA of *P. neocaledonia*, *P. macleodensis*, and *P. quercus* were prepared from purified particles and sequenced using either the PacBio or Illumina platforms (see Methods). As for *P. salinus*, *P. dulcis*[5], and *P. inopinatum*[7], the three new genomes assembled as single linear double-stranded DNA (dsDNA) molecules (≈60% G + C) with sizes ranging from 1.84 to 2 Mb. In addition to their translucent amphora-shaped particles (Fig. 1), higher than average G + C content and genomic gigantism thus remain characteristic features shared by the *Pandoraviridae*[5, 8]. Given the high proportion of viral genes encoding proteins without database homolog, gene predictions based on purely ab initio computational approaches (i.e., "ORFing" and coding propensity estimates) are notoriously unreliable, leading to inconsistencies between teams using different values of arbitrary parameters (e.g., minimal open reading frame (ORF) size). For instance among families of large dsDNA viruses infecting eukaryotes, the average protein-coding gene density reportedly varies from one gene every 335 bp (*Phycodnaviridae*, NCBI: NC_008724) up to one gene every 2120 bp (*Herpesviridae*, NCBI: NC_003038), while the consensus is clearly around one gene every kb (such as for bacteria). As a result, one oscillates between situations where many genes are overpredicted and others where many real genes are probably overlooked. Such uncertainty about which genes are "real" introduces a significant noise in comparative genomic analyses and the subsequent testing of evolutionary hypotheses. In addition, computational methods are mostly blind to genes expressed as non-protein-coding transcripts.

To overcome the above limitations, we performed strand-specific RNA-seq experiments and particle proteome analyses, the results of which were mapped on the genome sequences. Only genes supported by experimental evidence (or protein similarity) were retained into this stringent reannotation protocol (see Methods, Supplementary Fig. 2). On one hand, this new procedure led to a reduced set of predicted proteins, on the other hand it allowed the discovery of an unexpected large number of non-coding transcripts (Table 1).

The new set of validated protein-coding genes exhibits a strongly diminished proportion of ORFs shorter than 100 residues, most of which are unique to each pandoravirus strain (Supplementary Fig. 3). The stringent annotation procedure also resulted in genes exhibiting a well-centered unimodal distribution of codon adaptation index (CAI) values (Supplementary Fig. 3).

For consistency, we extrapolated our stringent annotation protocol to *P. inopinatum* and *P. macleodensis*, reducing the number of predicted proteins taken into account in further comparisons (see Methods, Table 1). As expected, the discrepancies between the standard versus stringent gene predictions are merely due to the overprediction of small ORFs (length < 300 nucleotides). Such arbitrary ORFs are prone to arise randomly in G + C-rich sequences within which stop codons (TAA, TAG, and TGA) are less likely to occur by chance than in the non-coding regions of A + T-rich genomes. Indeed, the above standard and stringent annotation protocols applied to the A + T-rich (74.8%) *Megavirus chilensis* genome[3] resulted in two very similar sets of predicted versus validated protein-coding genes (1120 versus 1108). This control indicates that our stringent annotation is not simply discarding eventually correct gene predictions by arbitrary raising a confidence threshold, but specifically correcting errors induced by the G + C-rich composition. Purely computational gene annotation methods are thus markedly less reliable for G + C-rich genomes, especially when they encode a large proportion of ORFans (i.e., ORF without database homolog), as for pandoraviruses. However, it is worth noticing that even after our stringent reannotation, the fraction of predicted proteins without significant sequence similarity outside of the *Pandoraviridae* family remained quite high (from 67 to 73%, Supplementary Fig. 4).

An additional challenge for the accurate annotation of the pandoravirus genomes is the presence of introns (virtually undetectable by computational methods when they interrupt ORFans). The mapping of the assembled transcript sequences onto the genomes of *P. salinus*, *P. dulcis*, *P. quercus*, and *P. neocaledonia*, allowed the detection of spliceosomal introns in 7.5–13% of the validated protein-coding genes. These introns were found in the untranslated regions (UTRs) as well as in the coding sequences, including on average 14 genes among those encoding the 200 most abundant proteins detected in the particles (see below). Although spliceosomal introns are found in other viruses with a nuclear phase such as the chloroviruses[9], pandoraviruses are the only ones for which spliceosomal introns have been validated for more than 10% of their genes. These results support our previous suggestion that at least a portion of the pandoravirus transcripts are synthetized and processed by the host nuclear machinery[5]. Yet, the number of intron per viral gene remains much lower (around 1.2 in average) than for the host genes (6.2 in average[10]). Pandoravirus genes also exhibit UTRs twice as long (Supplementary Table 1) as those of *Mimiviridae*[11].

The mapping of the RNA-seq data led to the unexpected discovery of a large number (157–268) of long non-coding transcripts (LncRNAs) (Table 1, Supplementary Table 1 for detailed statistics). These LncRNAs exhibit a polyA tail and about 4% of them contain spliceosomal introns. LncRNAs are most often transcribed from the reverse strand of validated protein-

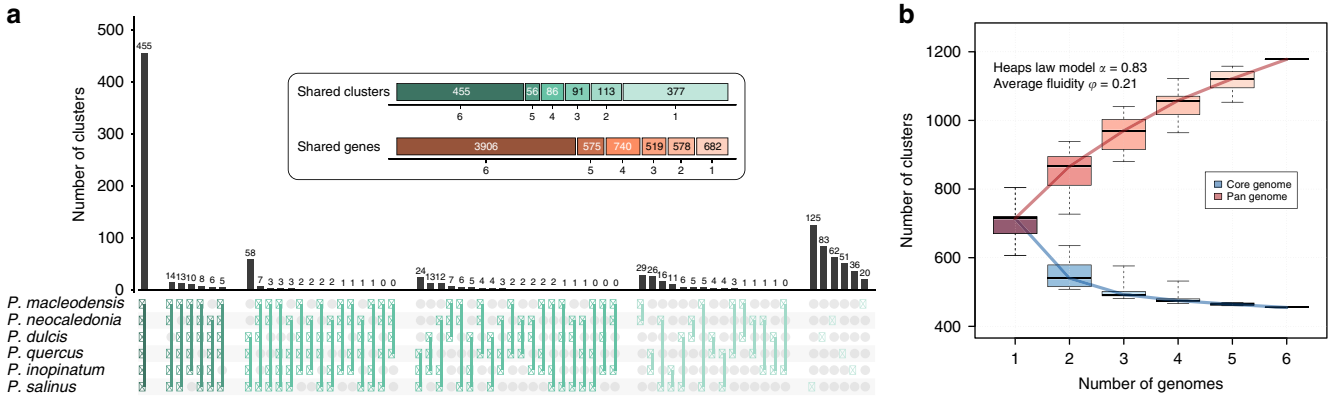


Fig. 2 Comparison of the pandoravirus gene contents. **a** The distribution of all the combinations of shared protein clusters is shown. The inset summarizes the number of clusters and genes shared by 6, 5, 4, 3, 2, and 1 pandoraviruses. **b** Core genome and pan-genome estimated from the six available pandoraviruses. The estimated heap law α parameter ($\alpha < 1$) is characteristic of an open pan-genome[50] and the fluidity parameter value characteristic of a large fraction of unique genes[51]. Box plots show the median, the 25th, and 75th percentiles. The whiskers correspond to the extreme data points

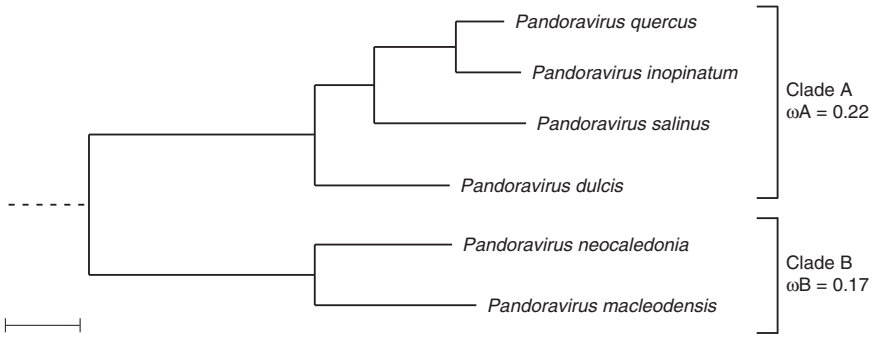


Fig. 3 Phylogenetic structure of the proposed *Pandoraviridae* family. Bootstrap values estimated from resampling are all equal to 1 and so were not reported. Synonymous to non-synonymous substitution rates ratios (ω) were calculated for the two separate clades and are significantly different (scale bar is 0.07 substitution/site)

coding genes while a smaller fraction are expressed in intergenic (i.e., inter-ORF) regions (Supplementary Fig. 5). These non-coding transcripts may play a role in the regulation of pandoravirus genes expression.

Overall, 82.7–87% of the pandoravirus genomes is transcribed (including ORFs, UTRs, and LncRNAs), but only 62–68.2% is translated into proteins. Such values are much lower than in giant viruses from other families (e.g., 90% of the Mimivirus[11] genome is translated), in part due to the larger UTRs flanking the pandoravirus genes.

Comparative genomics. The six protein-coding gene sets obtained from the above stringent annotation were then used as references for whole-genome comparisons aiming to identify specific features of the *Pandoraviridae* family. Following a sequence similarity-based clustering (see Methods), the relative overlaps of the gene contents of the various strains were computed (Fig. 2a), producing what we refer to as "protein clusters".

We then computed the number of shared (i.e., "core") and total genes as we incrementally incorporated the genomes of the various isolates into the above analysis, to estimate the size of the family core gene set and that of the accessory/flexible gene set. If the six available isolates appeared sufficient to delineate a core genome coding for 455 different protein clusters, the "saturation curve" leading to the total gene set is far from reaching a plateau, suggesting that the *Pandoraviridae* pan-genome is open, with each additional isolate predicted to contribute more than 50 additional genes (Fig. 2b). This remains to be confirmed by the analysis of additional *Pandoraviridae* isolates.

We then investigated the global similarity of the six pandoravirus isolates by analyzing their shared gene contents both in term of protein sequence similarity and genomic position. The pairwise similarity between the different pandoravirus isolates ranges from 54 to 88%, as computed from a super alignment of the protein products of the orthologous genes (Supplementary Table 2). A phylogenetic tree computed with the same data clusters the pandoraviruses into two separate clades (Fig. 3).

Interpreted in a geographical context, this clustering pattern conveys two important properties of the emerging family. On one hand, the most divergent strains are not those isolated from the most distant locations (e.g., the Chilean *P. salinus* versus the French *P. quercus*; the Neo-Caledonian *P. neocaledonia* versus the Australian *P. macleodensis*). On the other hand, two isolates (e.g., *P. dulcis* versus *P. macleodensis*) from identical environments (two ponds located 700 m apart and connected by a small water flow) are quite different. Pending a larger-scale inventory of the *Pandoraviridae*, these results already suggest that members of this family are distributed worldwide with similar local and global diversities.

Our analysis of the positions of the homologous genes in the various genomes revealed that despite their sequence divergence (Supplementary Table 2), 80% of the orthologous genes remain collinear. As shown in Fig. 4, the long-range architecture of the pandoravirus genomes (i.e., based on the positions of orthologous genes) is globally conserved, despite their differences in sizes (1.83–2.47 Mb). However, one-half of the pandoravirus chromosomes (the leftmost region in Fig. 4) curiously appears

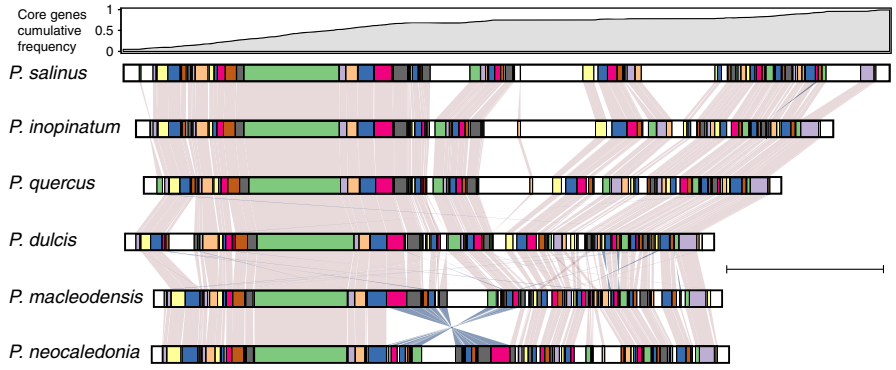


Fig. 4 Collinearity of the available pandoravirus genomes. Cumulative frequency of core genes is shown at the top. Conserved collinear blocks are colored in the same color in all viruses. White blocks correspond to non-conserved DNA segments (scale bar is 500 kb)

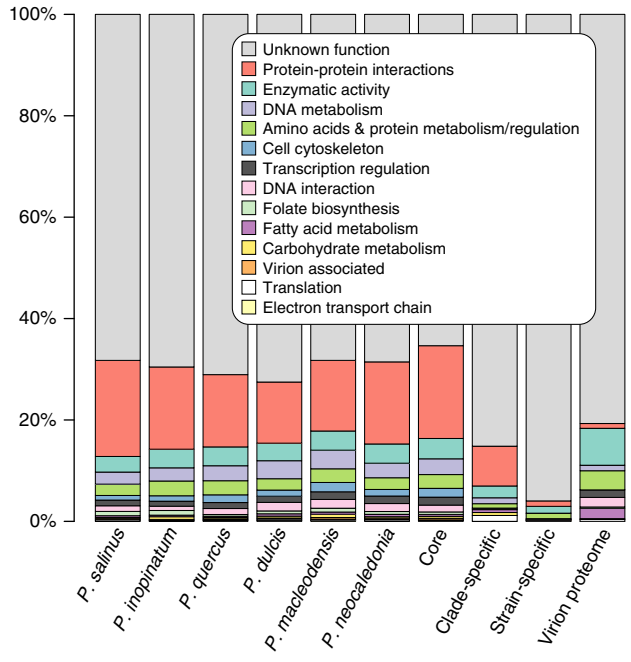


Fig. 5 Functional annotations

evolutionary more stable than the other half where most of the non-homologous segments occur. These segments contain strain-specific genes and are enriched in tandem duplications of non-orthologous ankyrin, MORN, and F-box motif-containing proteins. Conversely, the stable half of the genome concentrates most of the genes constituting the *Pandoraviridae* core genome (top of Fig. 4). Interestingly, the local inversion that distinguishes the chromosome of *P. neocaledonia* from the other strains is located near the boundary between the stable and unstable regions, and may be linked to this transition (although it may be coincidental). Finally, all genomes are also enriched in strain-specific genes (and/or duplications) at both extremities.

We then analyzed the distribution of the predicted proteins among the standard broad functional categories (Fig. 5). As it is now recurrent for large and giant eukaryotic DNA viruses, the dominant category is by far that of proteins lacking recognizable functional signatures. Across the six strains, an average of 70% of the predicted proteins correspond to "unknown functions". Such a high proportion is all the more remarkable as it applies to carefully validated gene sets, from which dubious ORFs have been eliminated. It is thus a biological reality that a large majority of these viral proteins cannot be linked to previously characterized pathways. Remarkably, the proportion of such anonymous proteins remains quite high (65%) among the products of the pandoravirus core genome, that is among the presumably essential genes shared by the six available strains (and probably all future family members, according to Fig. 2b). Interestingly, this proportion remains also very high (≈80%) among the proteins detected as constituting the viral particles. Furthermore, the proportion of anonymous proteins totally dominates the classification of genes unique to each strain, at more than 95%. The most generic functional category, "protein–protein interaction" is the next largest (from 11.7 to 18.9%), corresponding to the detection of highly frequent and uninformative motifs (e.g., ankyrin repeats). Overall, the proportion of pandoravirus proteins to which a truly informative function could be attributed is <20%, including a complete machinery for DNA replication and transcription.

We then investigated two evolutionary processes possibly at the origin of the extra-large size of the pandoravirus genomes: horizontal gene transfers (HGTs) and gene duplications. The acquisition of genes by HGT was frequently invoked to explain the genome size of ameba-infecting viruses compared to "regular" viruses[12, 13]. We computed that up to a third of the pandoravirus proteins exhibit sequence similarities (outside of the *Pandoraviridae* family) with proteins from the three cellular domains (Eukarya, Archaea, and Eubacteria) or other viruses (Supplementary Fig. 4). However, such similarities do not imply that these genes were horizontally acquired. They also could denote a common ancestral origin or a transfer from a pandoravirus to other microorganisms. We individually analyzed the phylogenetic position of each of these cases to infer their likely origin: ancestral —when found outside of clusters of cellular or viral homologs; horizontally acquired—when found deeply embedded in the above clusters; or horizontally transferred to cellular organisms or unrelated viruses in the converse situation (i.e., a cellular protein lying within a pandoravirus protein cluster). Supplementary Fig. 6 summarizes the results of this analysis.

We could make an unambiguous HGT diagnosis for 39% of the cases, the rest remaining undecidable or compatible with an ancestral origin. Among the likely HGT, 49% suggested a horizontal gain by pandoraviruses, and 51% the transfer of a gene from a pandoravirus. Interestingly, the acquisition of host genes, a process usually invoked as important in the evolution of viruses, only represent a small proportion (13%) of the diagnosed HGTs, thus less than from the viruses to the host (18%). Combining the above statistics with the proportion of genes (one-third) we started from, in the whole genome, suggests that at most 15% (and at least 6%) of the pandoravirus gene content could have been gained from cellular organisms (including 5–2% from their contemporary Acanthamoeba host) or other viruses. Such range of values is comparable to what was previously estimated

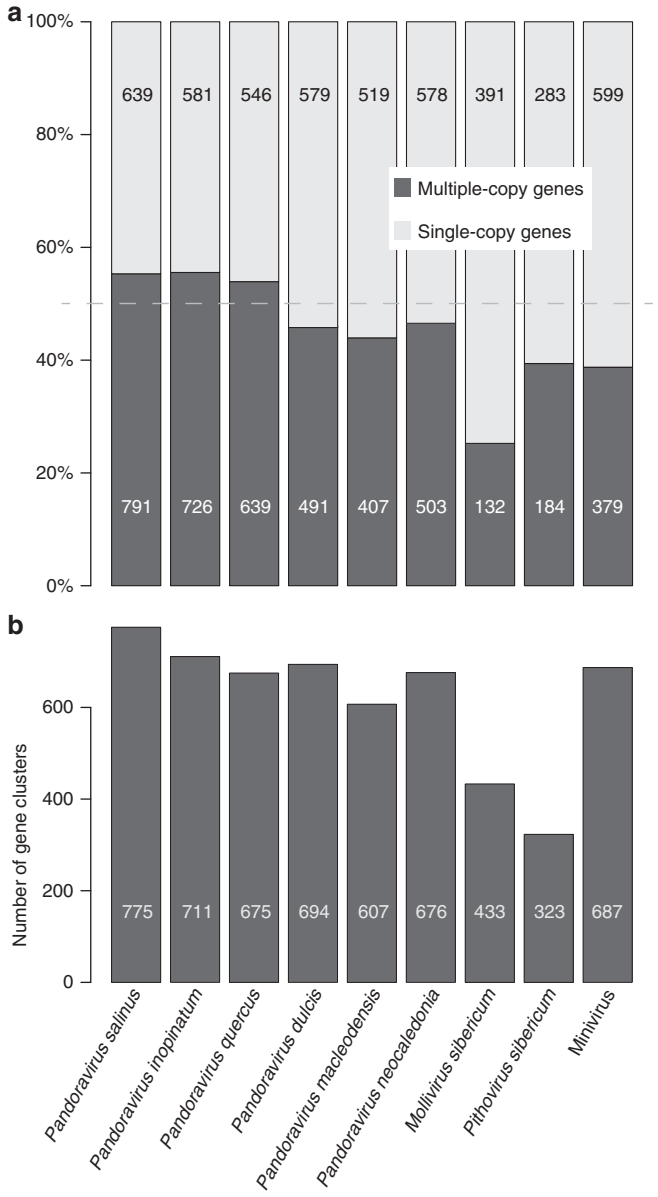


Fig. 6 Analysis of gene duplication in various giant virus families. **a** Distribution of single-copy versus multiple-copy genes in giant viruses. **b** Number of distinct gene clusters

for Mimivirus[14]. HGT is thus not the distinctive process at the origin of the giant pandoravirus genomes.

We then investigated the prevalence of duplications among pandoravirus genes. Figure 6a compares the proportions of single versus duplicated (or more) protein-coding genes of the six available pandoraviruses with that computed for representatives of the three other known families of giant DNA viruses infecting Acanthamoeba. It clearly shows that the proportion of multiple-copy genes (ranging from 55 to 44%) is higher in pandoraviruses, than for the other virus families, although it does not perfectly correlate with their respective genome sizes. The distributions of cluster sizes among the different pandoravirus strains are similar. Most multiple-copy genes are found in cluster of size 2 (duplication) or 3 (triplication). The number of bigger clusters then decreases with their size (Supplementary Fig. 7).

Fewer large clusters (size > 20) correspond to proteins sharing protein–protein interaction motifs, such as Ankyrin, MORN, and F-box repeats. Surprisingly, the absolute number of single-copy genes in pandoraviruses is similar to, and sometimes smaller (e.g., *P. neocaledonia*, 2 Mb) than that in Mimivirus, with a genome (1.18 Mb) half the size. Overall, the number of distinct gene clusters (Fig. 6b) overlaps between the *Pandoraviridae* (from 607 to 775) and Mimivirus (687), suggesting that despite their difference in genome and particle sizes, these viruses share comparable genetic complexities.

Gene duplication being such a prominent feature of the pandoravirus genomes, we investigated it further looking for more insight about its mechanism. First, we computed the genomic distances between pairs of closest paralogs, most likely resulting from the most recent duplication events. The distributions of these distances, similar for each pandoravirus, indicate that the closest paralogs are most often located next to each other (distance = 1) or separated by a single gene (distance = 2) (Supplementary Fig. 8).

We then attempted to correlate the physical distance separating duplicated genes with their sequence divergence as a (rough) estimate of their evolutionary distance. We obtained a significant correlation between the estimated "age" of the duplication event and the genomic distance of the two closest paralogs (Supplementary Fig. 9). These results suggest an evolutionary scenario whereby most duplications are first occurring in tandem, with subsequent genome alterations (insertions, inversions, and gene losses) progressively blurring this signal.

Comparative proteomic of pandoravirions. Our previous mass spectrometry proteomic analysis of *P. salinus* particles identified 210 viral gene products, most of which ORFans or without predictable function. In addition, we detected 56 host (Acanthamoeba) proteins. Importantly, none of the components of the virus-encoded transcription apparatus was detected in the particles[5]. In this work we performed the same analyses on *P. salinus*, *P. dulcis*, and two of the new isolates (*P. quercus* and *P. neocaledonia*) to determine to what extent the above features were conserved for members of the *Pandoraviridae* family with various levels of divergence, and identify the core versus the accessory components of a generic pandoravirion.

Due to the constant sensitivity improvement in mass spectrometry, our new analyses of purified virions led to the reliable identification of 424 proteins for *P. salinus*, 357 for *P. quercus*, 387 for *P. dulcis*, and 337 for *P. neocaledonia* (see Methods). However, this increased number of identifications corresponds to abundance values (intensity-based absolute quantification, iBAQ) spanning more than five orders of magnitude. Many of the proteins identified in the low abundance tail might thus not correspond to bona fide particle components, but to randomly loaded bystanders, "sticky" proteins, or residual contaminants from infected cells. This cautious interpretation is suggested by several observations:

- the low abundance tail is progressively enriched in viral proteins identified in the particles of a single pandoravirus strain (even though other strains possess the homologous genes),
- the proportion of host-encoded proteins putatively associated to the particles increases at the lowest abundances,
- many of these host proteins were previously detected in particles of virus unrelated to the pandoraviruses but infecting the same host,
- these proteins are abundant in the Acanthamoeba proteome (e.g., actin, peroxidase, etc) making them more likely to be retained as purification contaminants.

Unfortunately, the iBAQ value distributions associated to the pandoravirion proteomes did not exhibit a discontinuity that

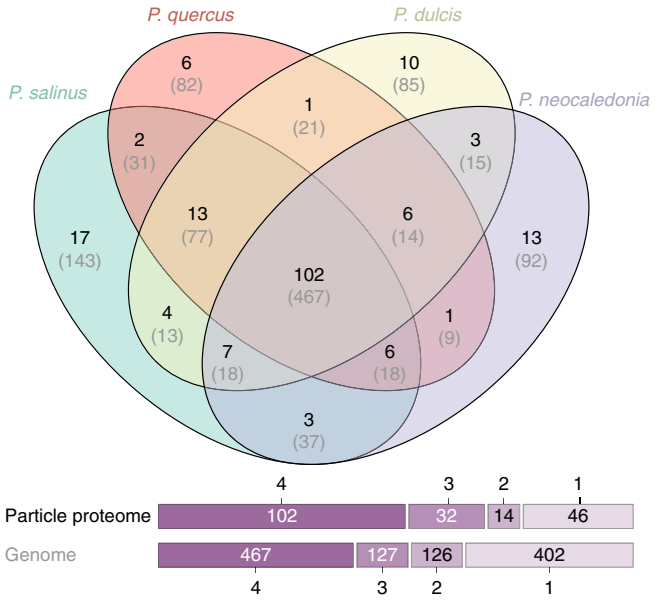


Fig. 7 Venn diagram of the particle proteomes of four different pandoravirus strains

could serve as an objective abundance threshold to distinguish bona fide particle components from dubious ones. However, the number of identified Acanthamoeba proteins sharply increases after rank ≈200 in the whole proteome (Supplementary Fig. 10). Following the same conservative attitude as for the genome reannotation, we decided to disregard the proteins identified below this rank as likely bystanders and only included the 200 most abundant proteins in our further analyses of the particle proteomes (Supplementary Data 1, Supplementary Table 3). Using this stringent proteome definition for each of the four different pandoravirions, we first investigated the diversity of their constituting proteins and their level of conservation compared to the global gene contents of the corresponding pandoravirus genomes.

Figure 7 shows that the particle proteomes include proteins belonging to 194 distinct clusters, 102 of which are shared by the four strains. The core proteome is thus structurally and functionally diverse. It corresponds to 52.6% of the total protein clusters globally identified in all pandoravirions. By comparison, the 467 protein clusters encoded by the core genome only represents 41.6% (i.e., 467/1122) of the overall number of pandoravirus-encoded protein clusters. The pandoravirus "box" used to propagate the genomes of the different strains is thus significantly more conserved than their gene contents ($p \ll 10^{-3}$, chi-square test). The genes encoding the core proteome also exhibit the strongest purifying selection among all pandoravirus genes (Supplementary Fig. 11a).

To evaluate the reliability of our proteome analyses we compared the abundance (iBAQ) values determined for each of the 200 most abundant proteins for two technical replicates and for two biological replicates performed on the same pandoravirus strain (Supplementary Fig. 12a & b). A very good correlation (Pearson's $R > 0.97$) was obtained in both cases for abundance values ranging over three orders of magnitude. We then compared the iBAQ values obtained for orthologous proteins shared by the virion proteomes of different isolates. Here again, a good correlation was observed ($R > 0.81$), as expected smaller than for the above replicates (Supplementary Fig. 12c & d). These results suggest that although the particles of the different strains appear morphologically identical (Supplementary Fig. 1), they admit a tangible flexibility both in terms of the protein sets they are made of (with 89% of pairwise orthologues in average), and in their precise stoichiometry.

We then examined the predicted functions of the proteins composing the particles, from the most to the least abundant, hoping to gain some insights about the early infectious process. Unfortunately, only 19 protein clusters could be associated to a functional/structural motif out of the 102 different clusters defining the core particle proteome (Supplementary Data 1, Supplementary Table 3). This proportion is less than for the whole genome (Fig. 5), confirming the alien nature of the pandoravirus particle as already suggested by its unique morphology and assembly process[5]. The pandoravirions are mostly made of proteins without homologs outside of the *Pandoraviridae* family. No protein even remotely similar to the usually abundant major capsid protein (MCP), a predicted DNA-binding core protein, or a DNA-packaging ATPase, hallmarks of most eukaryotic large DNA viruses, is detected. In particular, a *P. salinus* hypothetical protein (previously ps_862 now reannotated psal_cds_450) recently suggested by Sinclair et al.[15] to be a strong MCP candidate was not detected in the *P. salinus* virions, nor its homologs in the other strain proteomes. This negative result emphasizes the need for the experimental validation of computer predictions made from the "twilight zone" of sequence similarity. No trace of the pandoravirus-encoded RNA polymerase is detected either, confirming that the initial stage of infection requires the host transcription machinery located in the nucleus. Spliceosomal introns were validated for 56 pandoravirus genes the products of which were detected in the pandoravirions (Supplementary Data 1). This indicates the preservation of a functional spliceosome until the end of the infectious cycle, as expected from the observation of unbroken nuclei (Supplementary Fig. 1).

Among the 19 non-anonymous protein clusters, 4 exhibit generic motifs without specific functional clue: 2 collagen-like domains and 1 Pan/APPLE-like domain that are involved in protein–protein interactions, and 1 cupin-like domain corresponding to a generic barrel fold. Among the 10 most abundant core proteins, 9 have no predicted function, except for 1 exhibiting a C-terminal thioredoxin-like domain (psal_cds_383). It is worth noticing that the predicted membrane-spanning segment of 22 amino acids (85–107) is conserved in all pandoravirus strains. The 5′UTR of the corresponding genes exhibit 2 introns (in *P. salinus*, *P. dulcis*, and *P. quercus*) and 1 in *P. neocaledonia*. Thioredoxin catalyzes dithiol-disulfide exchange reactions through the reversible oxidation of its active center. This protein, with another one of the same family (psal_cds_411, predicted as soluble), might be involved in repairing/preventing phagosome-induced oxidative damages to viral proteins prior to the initial stage of infection. The particles also share another abundant redox enzyme, an ERV-like thiol oxidoreductase that may be involved in the maturation of Fe/S proteins. Another core protein (psal_cds_1260) with a remote similarity to a thioredoxin reductase may participate to the regeneration of the oxidized active sites of the above enzymes. Among the most abundant core proteins, psal_cds_232 is predicted as DNA-binding, and may be involved in genome packaging. One putative NAD-dependent amine oxidase (psal_cds_628), and one FAD-coupled dehydrogenase (psal_cds_1132) complete the panel of conserved putative redox enzymes. Other predicted core proteins include a Ser/thr kinase and phosphatase that are typical regulatory functions. One serine protease, one lipase, one patatin-like phospholipase, and one remote homolog of a nucleoporin might be part of the toolbox used to ferry the pandoravirus genomes to the cytoplasm and then to the nucleus (Supplementary Table 3). Finally, two core proteins (psal_cds_118 and psal_cds_874) share an

endoribonuclease motif and could function as transcriptional regulators targeting cellular mRNA.

At the opposite of defining the set of core proteins shared by all pandoravirions, we also investigated strain-specific components. Unfortunately, most of the virion proteins unique to a given strain (about 10 in average) are anonymous and in low abundance. No prediction could be made about the functional consequence of their presence in the particles.

Discussion

We isolated three new pandoraviruses (*P. neocaledonia*, *P. quercus*, and *P. macleodensis*) from distant locations (resp. New Caledonia, South of France, and Australia). As for the previously characterized members of this emerging family (*P. salinus* from Chile, *P. dulcis* from Australia, and *P. inopinatum* from Germany), their genomes consist in large linear G + C-rich dsDNA molecules around 2 Mb in size (Table 1). Using the four most divergent pandoravirus strains at our disposal, we combined RNA-seq, virion proteome, and sequence similarity analyses to design a stringent annotation procedure, eliminating most false positive gene predictions that could both inflate the proportion of ORFans and distort the results of comparative analyses. Our—probably over-cautious—gene-calling procedure (reducing the number of predicted proteins by up to 44%) (Table 1) nevertheless confirmed that an average of 70% of the experimentally validated genes encode proteins without detectable homolog outside of the *Pandoraviridae* family, and up to 80% for those detected in the particles.

Using the six strains known as of today, we determined that each new family member contributed genes not previously seen in the other genomes, at a rate suggesting that the *Pandoraviridae* pan-genome is open (Fig. 2). Moreover, this flexible (i.e., strain- and clade-specific) gene subset exhibits a much higher proportion of ORFans (respectively 96% and 90%) than the core genome (63%) (Supplementary Fig. 11d). According to the usual interpretation, the core genome corresponds to genes inherited from the last common ancestor of a group of viruses while the flexible genome corresponds to genes that appeared since their divergence, through various mechanisms. We then performed further comparative statistical analyses to investigate which mechanisms might be responsible for the large pandoravirus gene content and, possibly, of its continuous expansion.

We determined that gene duplication was a contributing factor in the genome size of pandoraviruses, with 50% of their genes present in multiple copies (Fig. 6, Supplementary Fig. 7). However, this value is not vastly different from the proportion (40%) computed for Mimivirus with a genome half the size (Fig. 6). Thus, duplication alone does not explain the much larger gene content of the pandoraviruses. The proportion of single-copy ORFans (from 50.7 to 62.7%) compared to those in multiple copies is significantly larger than for non-ORFans (from 30 to 44.5%) (Fischer exact test, p-value $< 2 \times 10^{-4}$). ORFan genes thus tend to be less frequently duplicated.

HGT is also frequently invoked as a mechanism for viral genome inflation[12,13,16–18]. We estimated that HGT might be responsible for 6 to 15% of the *P. salinus* gene content (Supplementary Fig. 6). Such a proportion is not exceptional compared to other large eukaryotic dsDNA viruses[14] with much smaller genomes, and thus does not explain the huge pandoravirus gene content. Furthermore, the large proportion of ORFans among the flexible genome (Supplementary Fig. 11) is arguing against recent acquisitions from HGTs, short of postulating that they originated from mysterious organisms none of which has yet been characterized. Alternatively, the phylogenetic signal from these newly acquired genes could have been erased due to accelerated evolution. However, this is not supported by our data showing that pandoravirus-specific ORFan genes are under strong purifying selection, just to a lesser extent than non-ORFans (Supplementary Fig. 11).

The proportion of ORFans (i.e., proteins without homologs in the databases) obviously depends on our limited knowledge of the virosphere. However, what characterizes the *Pandoraviridae* is the unprecedented number of family-specific ORFans they share, the increase of their proportion among the subsets of core genes (with orthologs in all strains), clade-specific and strain-specific genes (Supplementary Fig. 11d), as well as their distinctive statistical properties (Supplementary Fig. 13). Altogether, this suggests that the pandoravirus-specific ORFans are not just ancestral genes missing from the database, but genes with histories confined within the *Pandoraviridae*.

To further investigate the origin of the pandoravirus genes, we performed various statistical analyses in search of what would distinguish core genes from clade-specific genes, and from those unique to each strain. To ensure the assignation of each of the genes to their respective categories, we added a constraint on their genomic positions. For instance, we only considered strain-specific genes found interspersed within otherwise collinear sequences of clade-specific or core genes (Fig S13a). The genes from the three above categories appeared significantly different with respect to three independent properties (G + C content, ORF length, and CAI). Moreover, the clade-specific and strain-specific genes exhibited average values intermediate between that of the core genes and intergenic sequences (Supplementary Fig. 13b-d). Such a gradient unmistakably advocates what is referred to as the de novo protein creation (reviewed in refs. [19–22]). Our data support an evolutionary scenario whereby novel (hence strain-specific) protein-coding genes could randomly emerge from non-coding intergenic regions, then become alike protein-coding genes of older ancestry (i.e., clade-specific and core genes) in response to an adaptive selection pressure (Supplementary Fig. 11b). For a long time considered unrealistic on statistical ground[23], the notion that new protein-coding genes could emerge de novo from non-coding sequences[24] started to gain an increasing support following the discovery of many expressed ORFan genes in *Saccharomyces cerevisiae*[25], *Drosophila*[26], *Arabidopsis*[27], mammals[28], and primates[29]. This hypothesis was recently extended to giant viruses[30].

A different process, called overprinting, involves the use of alternative translation frame from preexisting coding regions. It appears mostly at work in small (mostly RNA) viruses and bacteria, the dense genomes of which lacks sufficient non-coding regions[31, 32]. However, overprinting would not generate the observed difference in G + C composition between strain-specific and core genes (Supplementary Fig. 13b).

The eukaryotic-like de novo gene creation hypothesis might apply to the pandoraviruses for several reasons. This process requires ORFs that are abundant (to compensate for its contingent nature) and large enough (e.g., >150 bp) to encode peptides capable of folding into minimal domains (40–50 residues). We previously pointed out that the high G + C content of the pandoraviruses, compared to the A + T richness of the other Acanthamoeba-infecting viruses[8], statistically increases the size of the random ORFs in non-coding regions. Moreover, these non-coding regions are also larger in average, representing up to 38% of the total genome (Supplementary Table 1). The pandoravirus genomes thus offer an ideal playground for de novo gene creation. However, a high G + C composition does not imply viral genome inflation and/or an open-ended flexible gene content, as shown by Herpesviruses, another family of dsDNA virus replicating in the nucleus[33]. Even though HSV-1 and HSV-2 exhibit a G + C content of 68% and 70% respectively, their genome

remained small (≈150 kb), and their genes coding for core proteins, non-core proteins, as well as their relatively large intergenic regions (≈250 ± 150 bp) do not display any significant difference in composition[34]. Accordingly, a single gene (*US12*) has been suggested to have emerged de novo[35]. Thus, pandoraviruses (and/or their amoebal host) must exhibit some specific features leading them to favor de novo gene creation. This might be the extensible genome space offered in their particle, the uncondensed state of their DNA genome, the absence of DNA repair enzymes packaged in the virion, or an unknown template-free machinery generating new DNA. The later mechanism, although highly speculative, would be easier to reconcile with the conserved collinearity of the pandoravirus genome than intense mutagenesis, duplication, or the shuffling of preexisting genes. This template-free generation process might be linked to the apparent instability of the right half of the pandoravirus chromosome, depleted in "core genes" (Fig. 4). We need more genomes to validate the bipartite heterogeneity of the pandoravirus chromosomes as a distinctive property of the family.

Conceptually, de novo gene creation can occur in two different ways: an intergenic sequence gains transcription before evolving an ORF, or the converse[21]. The numerous LncRNAs that we detected during the infection cycle of the various pandoraviruses would appear to favor the transcription first mechanisms. However, most of these non-coding transcripts are antisense of bona fide coding regions and would not generate the shift in G + C composition observed for strain-specific genes. Novel proteins might thus mostly emerge from the numerous intergenic (random) ORFs gaining transcription.

The best evidence of de novo gene creation, although rarely obtained, is the detection of a significant similarity between the sequence encoding a strain-specific ORFan protein and an intergenic sequence in a closely related strain[19]. Out of the 318 pandoravirus strain-specific genes that we tested, we found two of such occurrences. The *P. salinus* psal_cds_1065 (58 aa, 55% GC, CAI = 0.287) is similar to a non-coding RNA (pneo_ncRNA_241) in *P. neocaledonia*, and the *P. salinus* psal_cds_415 (96 aa, 54% GC, CAI = 0.173) is matching within an intergenic region in *P. quercus*. In both cases, the matches occur at homologous genomic location. Such low rate of success (yet a positive proof of principle) was expected given the sequence divergence of the available pandoravirus strains, especially in their intergenic regions.

If we now admit the hypothesis that de novo gene creation plays a significant role in the large proportion of strain-specific ORFans and in the open-ended nature of the *Pandoraviridae* gene content, it could also have contributed to the pool of family-specific ORFans genes (now shared by two to six strains) to an unknown extant. The nature of the ancestor of the *Pandoraviridae* thus remains an unresolved question. Invoking the de novo creation hypothesis greatly alleviates the problem encountered when attempting to explain the diversity of the *Pandoraviridae* gene contents by lineage-specific gene losses and reductive evolution[8]. Instead of postulating an increasingly complex ancestor as new isolates are exhibiting additional unique genes, we can now attribute them to de novo creation. Yet, lineage-specific losses can still account for the gene content partially shared among strains.

As seductive as it is, the de novo creation hypothesis is nevertheless plagued by its own difficulties. First, newly expressed (random) proteins have to fold in a compact manner, or at least in a way not interfering with established protein interactions. Although early theoretical studies suggested that stable folding of random amino-acid sequences might be improbable[36], several experimental studies have indicated success rate of up to 20%[37, 38]. It has also been suggested that proteins encoded by de novo-created genes might be enriched in disordered regions[39]. Accordingly, we observed a slightly albeit significant ($p < 10^{-15}$, Wilcoxon signed-rank test) higher fraction of predicted disordered residues[40] in ORFans (14%) versus non-ORFans (11%). Also challenging is the process by which a random protein would spontaneously acquire a function. For example, only four functional (ATP-binding) proteins resulted from the screening of 6×10^{12} random sequences followed by many iterations of in vitro selections and directed evolution[41]. At the same time, the spontaneous mutation rate of large dsDNA viruses is very low (estimated at $<10^{-7}$ substitution per position per infection cycle)[42]. In absence of a useful function on which to exert a purifying selection, it seems very unlikely that a newly created protein could remain in a genome long enough to acquire a selectable influence on the virus fitness. How the so-called protogene[25] manage to be retained through the intermediate steps eventually leading to a selectable function remains the dark part of any de novo gene creation scenario. Thus, if our comparative genomic studies suggest new hypotheses about the evolution of pandoraviruses and other giant amoebal viruses, it is far from closing the debate about the genetic complexity of their ancestor[8, 17, 18, 43, 44].

In the context of this debate, it was previously proposed[17] that the pandoraviruses were highly derived phycodnaviruses based on the phylogenetic analysis of a handful of genes while disregarding the amazingly unique structural and physiological features displayed by the first two pandoravirus isolates[5] as well as the huge number of genes unique to them. Now using the six available pandoravirus genomes, a cladistic clustering based on the presence/absence of homologous genes in the different virus groups robustly separates the proposed *Pandoraviridae* family from the previously established families of large eukaryote-infecting dsDNA viruses (Fig. 8). The only remaining uncertainty concerns the actual position of the yet unclassified *Mollivirus sibericum* virus that will eventually be the seed of a distinct viral family, or the prototype of an early diverging branch[45] of smaller pandoraviruses. More *Pandoravidae* members are needed to delineate the exact boundaries of this new family and resolve the many issues we raised about the origin and mode of evolution of its members.

Methods

Environmental sampling and virus isolation. *P. neocaledonia:* A sample from the muddy brackish water of a mangrove near Noumea airport (New Caledonia, Lat: 22°16′29.50″S, Long: 166°28′11.61″E) was collected. After mixing the mud and the water, 50 mL of the solution was supplemented with 4% of rice media (supernatant obtained after autoclaving 1 L of seawater with 40 grains of rice) and let to incubate in the dark. After 1 month, 1.5 mL were recovered and 150 µL of pure Fungizone (25 µg/mL final) was added to the sample, which was vortexed and incubated overnight at 4 °C on a stirring wheel. After sedimentation, 1 mL supernatant was recovered and centrifuged at 800 × *g* for 5 min. Acanthamoeba *A. castellanii* (Douglas) Neff (ATCC 30010TM) cells adapted to Fungizone (2.5 µg/mL) were inoculated with 100 µL of the supernatant as previously described[46] and monitored for cell death.

P. macleodensis: A muddy sample was recovered from a pond 700 m away but connected to the La Trobe University pond in which *P. dulcis* was isolated[5]. After mixing the mud and the water, 20 mL of sample were passed through a 20 µm sieve and the filtrate was centrifuged 15 min at 30 000 × *g*. The pellet was resuspended in 200 µL of phosphate-buffered saline (PBS) supplemented with antibiotics and 30 µL was added to six wells of culture of *A. castellanii* cells adapted to Fungizone (see SI Materials and Methods).

P. quercus: Soil under decomposing leaves was recovered under an oak tree in Marseille. Few grams were resuspended with 12 mL PBS supplemented with antibiotics. After vortexing 10 min, the tube was incubated during 3 days at 4 °C on a stirring wheel. The tube was than centrifuged 5 min at 200 × *g* and the supernatant was recovered, centrifuged 45 min at 6800 × *g*. The pellet was resuspended in 500 µL PBS supplemented with the antibiotics. A volume of 50 µL of supernatant and 20 µL of the resuspended pellet were used to infect *A. castellanii* cells adapted to Fungizone. As for *P. neocaledonia* and *P. macleodensis*, visible particles resembling pandoraviruses were visible in the culture media after cell lysis. All viruses were then cloned using a previously described procedure[45] prior to DNA extraction for sequencing and protein extraction for proteomic studies.

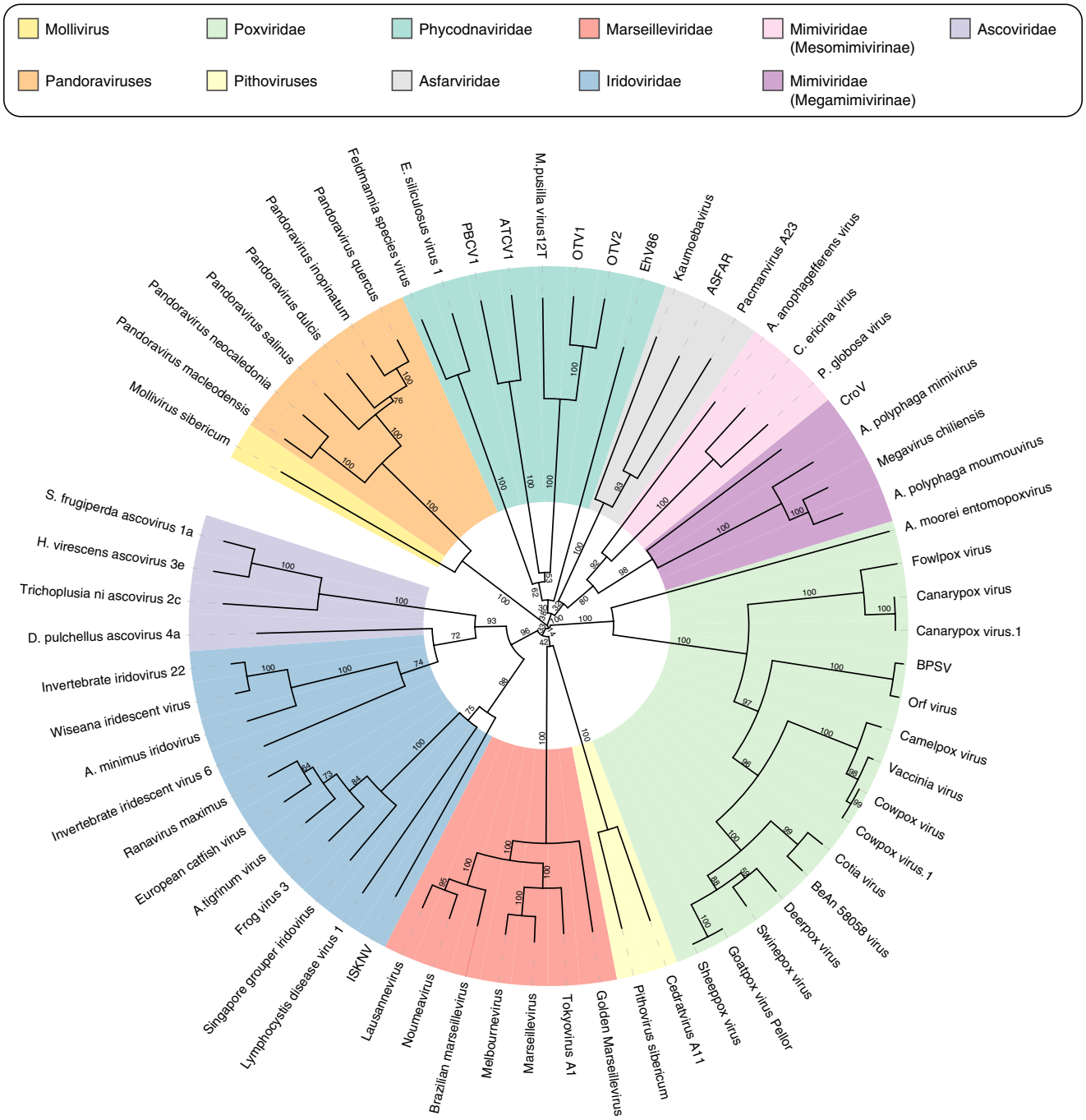


Fig. 8 Gene content-based cladistic tree of large DNA viruses. Long virus names have been replaced by acronyms (from top, clockwise). OTV1 Ostreococcus tauri virus 1, OTV2 Ostreococcus tauri virus 2, EhV86 Emiliania huxleyi virus 86, ASFAR African swine fever virus, CroV Cafeteria roenbergensis virus BV.PW1, BPSV Bovine papular stomatitis virus, ISKNV Infectious spleen and kidney necrosis virus. A maximum likelihood phylogenetic tree based on the DNA polymerase B protein sequence showing a globally similar topology was also computed (Supplementary Fig. 14)

Synchronous infections were performed for transmission electron microscopy observations of the infectious cycles. mRNA were extracted from the pooled infected cells prior to polyA+ enrichment and sent for library preparation and sequencing.

Genome sequencing and assembly. *P. neocaledonia* and *P. quercus* genomes were sequenced using the Pacbio sequencing technology. *P. macleodensis* genome was sequenced using the Illumina MiSeq technology with large insert (5–8 kb) mate pair sequences. Details on the strategy used for the genome assemblies are provided in the SI Materials and Methods.

Genome sequence stringent annotation. A stringent genome annotation was performed using a combination of ab initio gene prediction, strand-specific RNA-seq transcriptomic data, mass spectroscopy proteomic data as well as protein conservation data. The pipeline used is summarized in Supplementary Fig. 2 and described in the SI Materials and Methods.

Proteomic analyses. Virion proteomes were prepared as previously described in ref. [45] for mass spectrometry-based label-free quantitative proteomics. Briefly, extracted proteins from each preparation were stacked in the top of a 4–12% NuPAGE gel (Invitrogen) before R-250 Coomassie blue staining and in-gel digestion with trypsin (sequencing grade, Promega). Resulting peptides were analyzed by online nanoLC-MS/MS (Ultimate 3000 RSLCnano and Q-Exactive Plus, Thermo Scientific) using a 120-min gradient. Three independent preparations from the same clone were analyzed for each pandoravirus to characterize particle composition. Characterization of different clones and technical replicates were

performed for *P. dulcis*. Peptides and proteins were identified and quantified as previously described[45] (SI Materials and Methods).

Miscellaneous bioinformatic analyses. A detailed description of the bioinformatics analyses used for protein clustering and genome rearrangements is detailed in the SI Materials and Methods. CAI was measured using the cai tool from the EMBOSS package[47]. The reference codon usage was computed from the *A. castellanii* most expressed genes (Supplementary Data 2). DNA-binding prediction of pandoravirus proteins was computed using the DNABIND server[48].

Data availability. The annotated genomic sequence determined for this work as well as the reannotated genomic sequences have been deposited in the Genbank/EMBL/DDBJ database under the following accession numbers: *P. salinus*, KC977571; *P. dulcis*, KC977570; *P. quercus*, MG011689; *P. neocaledonia*, MG011690; and *P. macleodensis*, MG011691. The reannotated *P. inopinatum* genome used in our comparative analyses is provided as Supplementary Data 3.

The mass spectrometry proteomics data have been deposited to the ProteomeXchange Consortium via the PRIDE[49] partner repository with the dataset identifier PXD008167.

All the genomic data, gene annotations, and transcriptomic data can be visualized on an interactive genome browser at the following address: [http://www.igs.cnrs-mrs.fr/pandoraviruses/]. All data are available from the authors.

Received: 18 January 2018 Accepted: 17 May 2018
Published online: 11 June 2018

References

1. La Scola, B. et al. A giant virus in amoebae. *Science* **299**, 2033 (2003).
2. Raoult, D. et al. The 1.2-megabase genome sequence of Mimivirus. *Science* **306**, 1344–1350 (2004).
3. Arslan, D., Legendre, M., Seltzer, V., Abergel, C. & Claverie, J.-M. Distant Mimivirus relative with a larger genome highlights the fundamental features of Megaviridae. *Proc. Natl Acad. Sci. USA* **108**, 17486–17491 (2011).
4. Yoosuf, N. et al. 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 (2012).
5. Philippe, N. et al. Pandoraviruses: amoeba viruses with genomes up to 2.5 Mb reaching that of parasitic eukaryotes. *Science* **341**, 281–286 (2013).
6. Scheid, P., Zöller, L., Pressmar, S., Richard, G. & Michel, R. An extraordinary endocytobiont in Acanthamoeba sp. isolated from a patient with keratitis. *Parasitol. Res.* **102**, 945–950 (2008).
7. Antwerpen, M. H. et al. Whole-genome sequencing of a pandoravirus isolated from keratitis-inducing acanthamoeba. *Genome Announc.* **3**, e00136-15 (2015).
8. Abergel, C., Legendre, M. & Claverie, J.-M. The rapidly expanding universe of giant viruses: Mimivirus, Pandoravirus, Pithovirus and Mollivirus. *FEMS Microbiol. Rev.* **39**, 779–796 (2015).
9. Grabherr, R., Strasser, P. & Van Etten, J. L. The DNA polymerase gene from chlorella viruses PBCV-1 and NY-2A contains an intron with nuclear splicing sequences. *Virology* **188**, 721–731 (1992).
10. Clarke, M. et al. Genome of Acanthamoeba castellanii highlights extensive lateral gene transfer and early evolution of tyrosine kinase signaling. *Genome Biol.* **14**, R11 (2013).
11. Legendre, M. et al. mRNA deep sequencing reveals 75 new genes and a complex transcriptional landscape in Mimivirus. *Genome Res.* **20**, 664–674 (2010).
12. Filée, J., Pouget, N. & Chandler, M. Phylogenetic evidence for extensive lateral acquisition of cellular genes by nucleocytoplasmic large DNA viruses. *BMC Evol. Biol.* **8**, 320 (2008).
13. Yutin, N., Wolf, Y. I. & Koonin, E. V. Origin of giant viruses from smaller DNA viruses not from a fourth domain of cellular life. *Virology* **466-467**, 38–52 (2014).
14. Monier, A., Claverie, J.-M. & Ogata, H. Horizontal gene transfer and nucleotide compositional anomaly in large DNA viruses. *BMC Genomics* **8**, 456 (2007).
15. Sinclair, R. M., Ravantti, J. J. & Bamford, D. H. Nucleic and amino acid sequences support structure-based viral classification. *J. Virol.* **91**, e02275-16 (2017).
16. Filée, J. Genomic comparison of closely related Giant Viruses supports an accordion-like model of evolution. *Front. Microbiol.* **6**, 593 (2015).
17. Yutin, N. & Koonin, E. V. Pandoraviruses are highly derived phycodnaviruses. *Biol. Direct* **8**, 25 (2013).
18. Koonin, E. V., Krupovic, M. & Yutin, N. Evolution of double-stranded DNA viruses of eukaryotes: from bacteriophages to transposons to giant viruses. *Ann. N. Y. Acad. Sci.* **1341**, 10–24 (2015).
19. McLysaght, A. & Hurst, L. D. Open questions in the study of de novo genes: what, how and why. *Nat. Rev. Genet.* **17**, 567–578 (2016).
20. Schlötterer, C. Genes from scratch--the evolutionary fate of de novo genes. *Trends Genet.* **31**, 215–219 (2015).
21. Schmitz, J. F. & Bornberg-Bauer, E. Fact or fiction: updates on how protein-coding genes might emerge de novo from previously non-coding DNA. *F1000Res.* **6**, 57 (2017).
22. Light, S., Basile, W. & Elofsson, A. Orphans and new gene origination, a structural and evolutionary perspective. *Curr. Opin. Struct. Biol.* **26**, 73–83 (2014).
23. Jacob, F. Evolution and tinkering. *Science* **196**, 1161–1166 (1977).
24. Keese, P. K. & Gibbs, A. Origins of genes: 'big bang' or continuous creation? *Proc. Natl Acad. Sci. USA* **89**, 9489–9493 (1992).
25. Carvunis, A.-R. et al. Proto-genes and de novo gene birth. *Nature* **487**, 370–374 (2012).
26. Levine, M. T., Jones, C. D., Kern, A. D., Lindfors, H. A. & Begun, D. J. Novel genes derived from noncoding DNA in Drosophila melanogaster are frequently X-linked and exhibit testis-biased expression. *Proc. Natl. Acad. Sci. USA* **103**, 9935–9939 (2006).
27. Donoghue, M. T., Keshavaiah, C., Swamidatta, S. H. & Spillane, C. Evolutionary origins of Brassicaceae specific genes in Arabidopsis thaliana. *BMC Evol. Biol.* **11**, 47 (2011).
28. Heinen, T. J. A. J., Staubach, F., Häming, D. & Tautz, D. Emergence of a new gene from an intergenic region. *Curr. Biol.* **19**, 1527–1531 (2009).
29. Toll-Riera, M. et al. Origin of primate orphan genes: a comparative genomics approach. *Mol. Biol. Evol.* **26**, 603–612 (2009).
30. Forterre, P. & Gaïa, M. Giant viruses and the origin of modern eukaryotes. *Curr. Opin. Microbiol.* **31**, 44–49 (2016).
31. Sabath, N., Wagner, A. & Karlin, D. Evolution of viral proteins originated de novo by overprinting. *Mol. Biol. Evol.* **29**, 3767–3780 (2012).
32. Delaye, L., Deluna, A., Lazcano, A. & Becerra, A. The origin of a novel gene through overprinting in Escherichia coli. *BMC Evol. Biol.* **8**, 31 (2008).
33. Roizman, B. et al. in *Human Herpesviruses: Biology, Therapy, and Immunoprophylaxis* (eds. Arvin, A. et al.) (Cambridge Univ. Press, 2007, Cambridge, UK).
34. Brown, J. C. High G + C content of herpes simplex virus DNA: proposed role in protection against retrotransposon insertion. *Open Biochem. J.* **1**, 33–42 (2007).
35. Dolan, A., Jamieson, F. E., Cunningham, C., Barnett, B. C. & McGeoch, D. J. The genome sequence of herpes simplex virus type 2. *J. Virol.* **72**, 2010–2021 (1998).
36. Shakhnovich, E. I. & Gutin, A. M. Implications of thermodynamics of protein folding for evolution of primary sequences. *Nature* **346**, 773–775 (1990).
37. Davidson, A. R. & Sauer, R. T. Folded proteins occur frequently in libraries of random amino acid sequences. *Proc. Natl Acad. Sci. USA* **91**, 2146–2150 (1994).
38. Chiarabelli, C. et al. Investigation of de novo totally random biosequences, part II: on the folding frequency in a totally random library of de novo proteins obtained by phage display. *Chem. Biodivers.* **3**, 840–859 (2006).
39. Wilson, B. A., Foy, S. G., Neme, R. & Masel, J. Young genes are highly disordered as predicted by the preadaptation hypothesis of de novo gene birth. *Nat. Ecol. Evol.* **1**, 0146 (2017).
40. Ward, J. J., McGuffin, L. J., Bryson, K., Buxton, B. F. & Jones, D. T. The DISOPRED server for the prediction of protein disorder. *Bioinformatics* **20**, 2138–2139 (2004).
41. Keefe, A. D. & Szostak, J. W. Functional proteins from a random-sequence library. *Nature* **410**, 715–718 (2001).
42. Sanjuán, R., Nebot, M. R., Chirico, N., Mansky, L. M. & Belshaw, R. Viral mutation rates. *J. Virol.* **84**, 9733–9748 (2010).
43. Claverie, J.-M. & Abergel, C. Open questions about giant viruses. *Adv. Virus Res.* **85**, 25–56 (2013).
44. Claverie, J.-M. & Abergel, C. Giant viruses: the difficult breaking of multiple epistemological barriers. *Stud. Hist. Philos. Biol. Biomed. Sci.* **59**, 89–99 (2016).
45. Legendre, M. et al. In-depth study of Mollivirus sibericum, a new 30,000-y-old giant virus infecting Acanthamoeba. *Proc. Natl. Acad. Sci. USA* **112**, E5327–E5335 (2015).
46. Legendre, M. et al. Thirty-thousand-year-old distant relative of giant icosahedral DNA viruses with a pandoravirus morphology. *Proc. Natl. Acad. Sci. USA* **111**, 4274–4279 (2014).
47. Rice, P., Longden, I. & Bleasby, A. EMBOSS: the European Molecular Biology Open Software Suite. *Trends Genet.* **16**, 276–277 (2000).
48. Szilágyi, A. & Skolnick, J. Efficient prediction of nucleic acid binding function from low-resolution protein structures. *J. Mol. Biol.* **358**, 922–933 (2006).
49. Vizcaíno, J. A. et al. 2016 update of the PRIDE database and its related tools. *Nucleic Acids Res.* **44**, 11033 (2016).

50. Guimarães, L. C. et al. Inside the Pan-genome—methods and software overview. *Curr. Genomics* **16**, 245–252 (2015).
51. Kislyuk, A. O., Haegeman, B., Bergman, N. H. & Weitz, J. S. Genomic fluidity: an integrative view of gene diversity within microbial populations. *BMC Genomics* **12**, 32 (2011).

Acknowledgements

This work was partially supported by the French National Research Agency (ANR-14-CE14-0023-01), France Genomique (ANR-10-INSB-01-01), Institut Français de Bioinformatique (ANR–11–INSB–0013), the Fondation Bettencourt-Schueller (OTP51251), by a DGA-MRIS scholarship, and by the Provence-Alpes-Côte-d'Azur région (2010 12125). Proteomic experiments were partly supported by the Proteomics French Infrastructure (ANR-10-INBS-08-01) and Labex GRAL (ANR-10-LABX-49-01). We thank the support of the discovery platform and informatics group at EDyP. We thank Deborah for her thorough reading of the manuscript.

Author contributions

M.L., Y.C., C.A., and J.-M.C. conceived and designed the research; E.F., S.J., A.L., J.-M.A., L. Beucher, N.P., L. Bertaux, E.C.-F., and Y.C. performed experimental research; M.L., O. P., C.A., and J.-M. C. performed bioinformatic analyses; L. Beucher, K.L., and Y.C. contributed new reagents/analytic tools; M.L., C.A., and J.-M.C. analyzed data; M.L., C. A., and J.-M.C. wrote the paper.

Additional information

Supplementary Information accompanies this paper at https://doi.org/10.1038/s41467-018-04698-4.

Competing interests: The authors declare no competing interests.

Reprints and permission information is available online at http://npg.nature.com/reprintsandpermissions/

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

Pandoravirus Y4, vers une meilleure compréhension de la structure des génomes de Pandoravirus

Origine géographique et caractéristiques de l'échantillon

L'échantillon d'où a été activé *Pandoravirus Y4* (*P. Y4*) est issu de la série d'échantillon C, collectée dans le cadre d'une collaboration avec Jens Strauss du Wegner Institute. La zone de prélèvement, baptisée Yukechi, est une zone de yedoma continu. La carotte d'où provient l'échantillon a été prélevée dans un talik sous un alas à une profondeur de 6 mètres. La datation de cet échantillon fait remonter ce dépôt à 54 ans. *P. Y4* est donc le premier pandoravirus activé à partir d'une zone sub-arctique.

Activation, caractérisation moléculaire et phénotypique de Pandoravirus Y4

Une fraction de 10 g d'échantillon a été utilisée et mise en culture direct avec des cellules fraîches d'*Acanthamoeba castellanii*. Le phénotype des cellules a été contrôlé quotidiennement par microscopie optique. Après 48h de co-culture, les cellules étaient à 60% décollées et à 10% d'enkystement. Une observation plus approfondie au microscope optique à fort grossissement (objectif x63 avec Optovar x1,6) a permis de caractériser la présence de particules ovoïdes à l'intérieur des amibes ainsi qu'en suspension dans le milieu.

Nous avons caractérisé la nature de l'agent infectieux par des observations en microscopie électronique et par amplification directe d'ADN par PCR. La coloration négative de l'échantillon a révélé la présence de particules ovoïdes de 1 µm de longueur. Ici, les amorces choisies pour la PCR directe ont été conçues pour permettre l'amplification d'une zone de 1500 pb conservée chez les pandoravirus. Après migration sur gel d'agarose, la taille des amplicons générés atteint 1500 pb. La purification, le séquençage des amplicons puis l'analyse manuelle des chromatogrammes a permis de récupérer deux séquences d'ADN d'une taille de 667 pb, pour la séquence orientée dans le sens inverse, et 321 pb pour la séquence orientée dans le sens direct. L'alignement de ces deux séquences par BLASTn contre l'ensemble des génomes des virus géants étudiés au laboratoire a démontré qu'elles étaient similaires à une région centrale du génome de *Pandoravirus celtis* (comprise entre 1281151 pb et 1281638 pb), avec un pourcentage d'identité de 73% pour la séquence direct et 71% pour la séquence inverse (6% d'insertion/délétion).

Après amplification par 2 repiquages successifs, nous avons cloné le virus afin de le produire en quantité suffisante et de poursuivre la caractérisation de ce nouvel agent infectieux

d'*Acanthamoeba castellanii*. La purification sur gradient de chlorure de cesium a donné un anneau viral d'une densité comprise entre 1,4 et 1,5.

Extraction d'ADN, séquençage et assemblage

L'extraction d'ADN génomique de particules purifiées de P. *Y4* a permis de récupérer une quantité de 5,95 µg d'ADN, la purification sur colonne de silice a permis d'obtenir un ratio d'absorbance 260/280 de 1,89 et 230/260 de 2,23. La banque R9-Long Read 1D d'une taille souhaitée de 8 kpb a permis de générer 509 045 séquences d'une taille moyenne de 3,3 kpb. Le séquençage par méthode Illumina HiSeq 2500 rapide a permis de générer 4,6 millions de séquences en sens direct et inverse avec 79% des bases ayant au minimum 99,9% de chances d'être correctes (79% > Q30).

Les résultats des différents assembleurs ont permis de sélectionner l'outil le plus efficace, *i.e* l'assembleur donnant le nombre de contigs le plus faible. Avant de commencer le *benchmark* des assembleurs nous avons détecté des lectures Nanopores correspondant à un second virus de la famille des *Molliviridae*. L'absence de lectures contaminantes dans les lectures Illumina laisse à penser que la contamination est survenue durant la constitution des banques Nanopore, ou l'ADN génomique ou les deux virus ont été traités simultanément. Pour éliminer cette contamination nous avons utilisé l'outil carrierseq. Ce logiciel filtre les lectures en supprimant celles qui s'alignent contre un génome contaminant.

Nous avons donc sélectionné l'assemblage produit par Spades, comportant 2 contigs dont la taille totale est de 1846920 pb. Les deux contigs ont une taille respective de 1817691 pb pour le contig 1, et de 29229 pour le contig 2. La couverture du génome est calculée à partir de l'alignement des lectures Nanopore contre les deux contigs correspondant au génome de *P. Y4*. La couverture est homogène pour le contig 1 et double pour le contig 2. Pour vérifier que le contig 2 ne constitue pas un élément extra chromosomique sous forme d'épisome nous avons déposé *P. Y4* sur un gel de PFGE. Le résultat indique que l'ensemble du génome est présent sous forme d'un unique fragment d'ADN. Nous avons cherché à savoir comment s'assemblaient les contig 1 et 2 ensembles. Les différentes hypothèses sont les suivantes : le contig 2 flanque les deux extrémités du contig 1 où le contig 2 est présent deux fois à l'une des extrémités du contig 1. Pour tester ces différentes hypothèses nous avons construit ces modèles de génomes *in silico* puis regardé la couverture des lectures Nanopore contre ces différents modèles. Le modèle pour lequel la couverture était la plus homogène est le modèle : contig 2 suivi du contig 1 puis du contig 2. Après une observation base à base via l'outil de visualisation Ugene-1.32.0, une chute de couverture sur 100 pb a été observée à la

jonction entre les deux contig. Nous avons retiré 99 pb correspondant au motif répété de basse couverture, dont 75 pb à l'extrémité 3' du contig 2 et 24 pb en 3' du contig 1. Pour déterminer si l'assemblage proposé est consistant avec les autres génomes de pandoravirus séquencé, nous avons déterminé les plus proches parents de *P. Y4* et comparé la structure de leurs génomes. Nous avons récupéré par exonerate la séquence homologue à l'ADN polymerase des pandoravirus. La topologie de l'arbre place l'ADN polymerase de *P. Y4* au sein du clade A, et le génome de *P. quercus* a été séléctionné pour être comparé à celui de *P. Y4* (Figure 39).

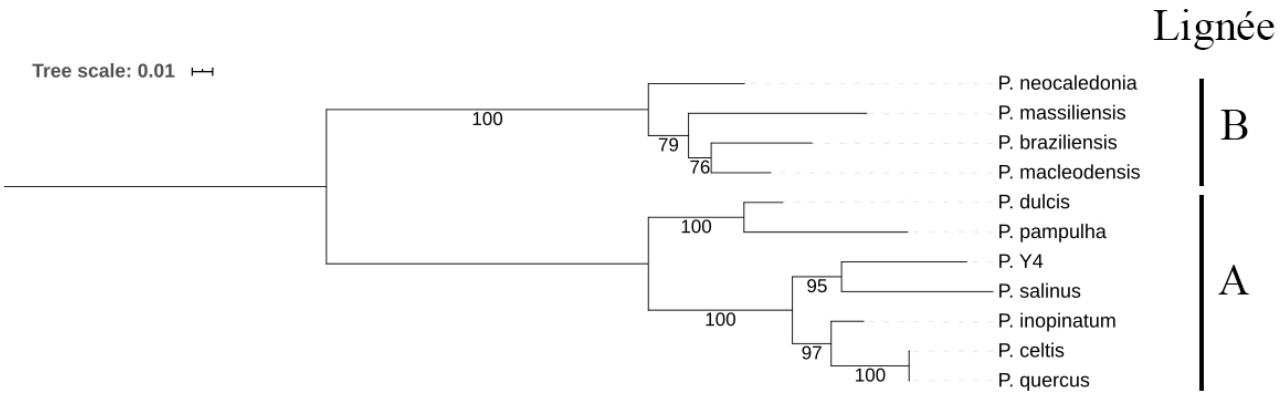


Figure 39 :Phylogénie des Pithoviridae. L'arbre phylogénétique est basé sur la séquence de l'ADN polymérase (modèle de substitution JTT+F+I+G4).

Un dotplot des génomes de *P. Y4* et *P. quercus* a montré la colinearité des deux génomes sur l'ensemble du contig 1 et l'ensemble du contig 2 à l'exception d'une zone de 9 kpb en 5' du contig 2 qui est inversée et repétée *P. quercus*. En conclusion, l'assemblage final du génome de *P. Y4* a donné un scaffold d'une taille totale de 1875975 pb de couverture homogène sur l'ensemble de sa longueur.

Structure du génome de Pandoravirus Y4

Pour valider l'architecture du génome de *P. Y4* prédite *in silico* deux approches expérimentales ont été mises en œuvre.

Tout d'abord, pour confirmer la présence et l'orientation des deux fragments du contig 2 aux extrémités du génome, nous avons réalisé des digestions du génome de *P. Y4* en blocs puis migration sur gel de PFGE. Les blocs de *P. Y4* ont été digérés par les enzymes HpaI et SwaI. La digestion par HpaI ne libère pas de fragment d'une taille avoisinant 30 kpb et la digestion par l'enzyme SwaI libère un fragment de 135,5 kpb (Figure 40). Ces résultats indiquent qu'il existe bien deux jonctions entre les contig 1 et 2 en 3' et en 5' du génome.

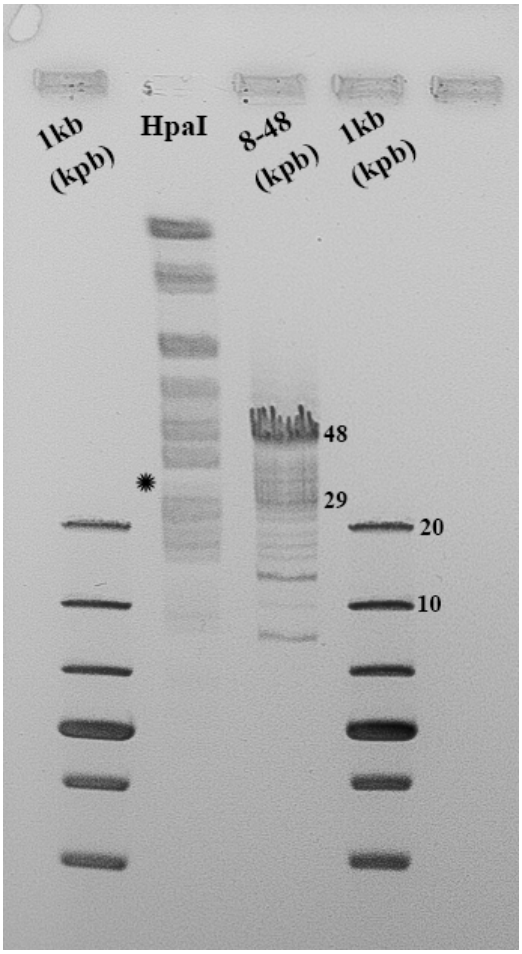


Figure 40 : Gel de PFGE. La digestion du génome de p. Y4 par HpaI ne libère pas de fragment d'une taille de 30 kpb (étoile). Cela suggère que l'extrémité 5' du génome de P. Y4 est d'une taille supérieur, suggérant la présence du contig 2 en amont de l'extrémité 5'. Le gel présentant la migration du génome de P. Y4 digéré par SwaI n'est pas encore optimal et n'est pas présenté ici.

Ensuite, pour déterminer les orientations du contig 2 aux extrémités du génome nous avons amplifié les jonctions par PCR en testant toutes les combinaisons d'orientations possibles. La présence d'amplicons dans les pistes correspondant aux amorces permettant d'amplifier la jonction en 5' du contig 2 dans le sens direct au contig 1 et la jonction en 3' du contig 1 au contig 2 orienté dans le sens reverse complémente permet donc de clarifier ces points (Figure 41). Ainsi, le génome de *P. Y4* est constitué d'un double brin d'ADN de 1875975 pb avec aux extrémités deux séquences répétées de 29154 pb.

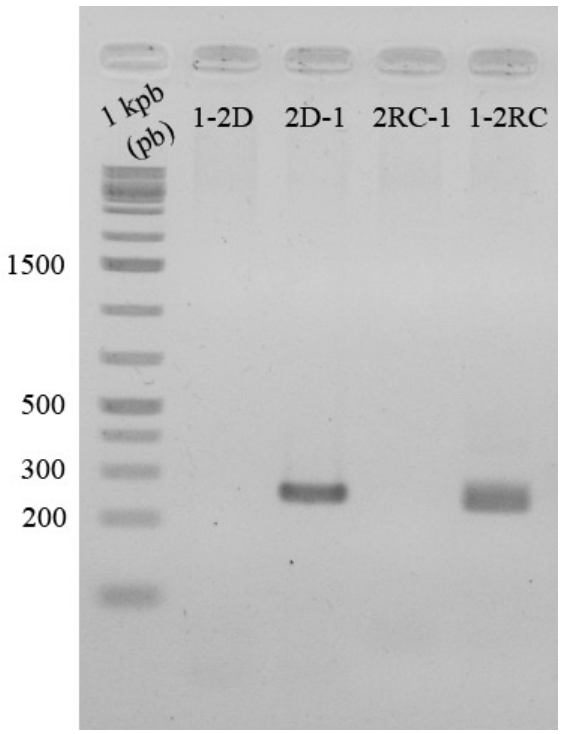


Figure 41 : Gel d'électrophorèse. Les différentes orientations du contig 2 avec le contig 1 ont été testées (D = sens direct, RC = reverse-complement).

Enfin, pour tenter d'expliquer la présence de ces séquences de 29 kpb identiques aux extrémités du génome nous avons recherché l'ensemble des motifs d'ADN suivant : répétitions A-phasées (répétition d'au moins 3 adénines à intervalle régulier), répétitions directes, répétitions miroirs (deux répétitions consécutives orientées en sens opposé sur un même brin) et répétitions inversées. Pour ce faire nous avons utilisé nBMST. Le résultat est le suivant : aux deux extrémités du génome on retrouve un motif inversé et répété de 60 pb en position 1, donnant une structure tige-boucle (*hairpin*) parfaitement appariée avec une boucle de 4 nucléotides ainsi qu'une longue séquence répétée dans le sens direct en position 17589 (Figure 42). Ce motif en *hairpin* a été recherché dans l'ensemble des génomes de pandoravirus séquencé. On le retrouve uniquement chez les pandoravirus de clade A, du même clade que *P. Y4*. On retrouve ce motif soit présent en une seule copie à une position inattendue : 9348 pour *P. celtis*, 9459 pour *P. quercus, s*oit à une unique position terminale : 1676109 pour *P. pampulha* et chez *P. salinus* en position 2473838.

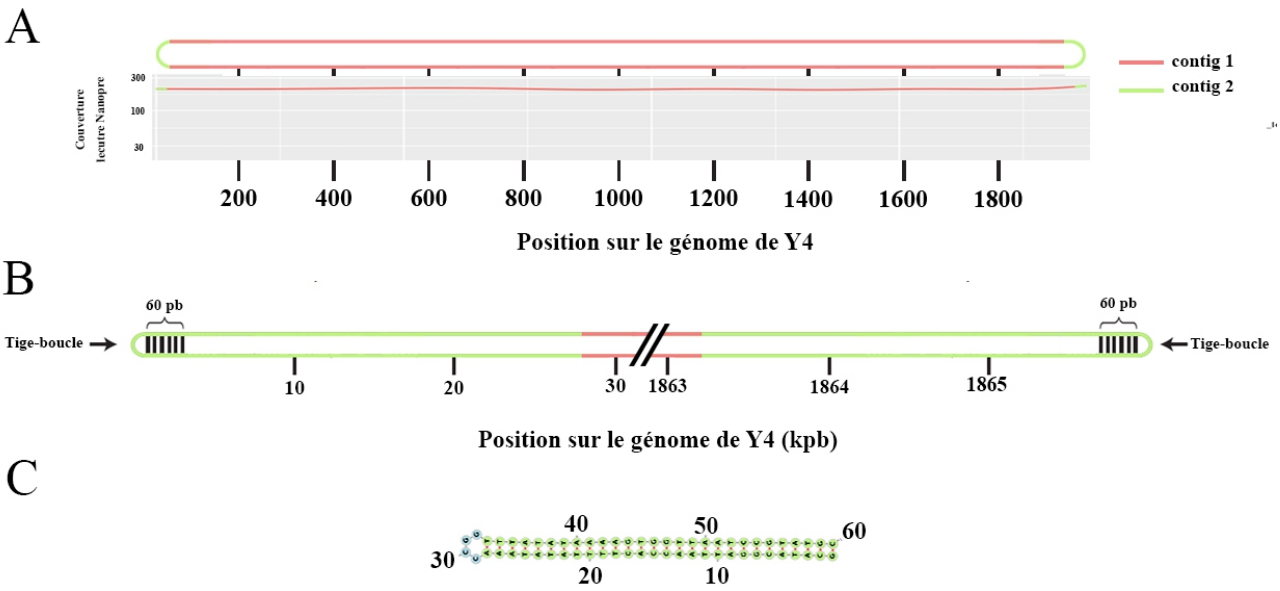


Figure 42 : Schéma présentant le modèle du génome de P. Y4. (A) Agencement des contig 1 et 2 et la couverture des lectures Nanopore associée (geom_smooth). (B) Détail du contig 2 aux extrémités du génome et la probable circularisation du génome grâce au motif tige-boucle. (C) Motif tige-boucle avec une boucle de 4 nucléotides.

Discussion

Il convient de noter d'ores et déjà qu'à l'endroit où a été détectée l'*hairpin* chez les pandoravirus de clade A, la couverture à cet endroit du génome est double. Les génomes de pandoravirus étant riches en GC (<60%) et de grande taille, il faudrait valider l'ensemble des assemblages par digestion et migration sur gel de PFGE pour tenter d'expliquer si l'architecture du génomes de *P. Y4* est commune à l'ensemble des *Pandoraviridae* de clade A . Cette zone, présente aux deux extrémités du génome de *P. Y4* et ayant un motif en *hairpin* pourrait conditionner le mécanisme de réplication du génome. Une telle organisation du génome a déjà été décrite pour une autre famille de NCDV. Ainsi, les génomes des *Poxviridae* ont une organisation similaire. D'une taille allant de 128 à 365 kpb, le génome des poxvirus est linéaire et ses extrémités sont liées de façons covalentes grâce à une structure en épingle à cheveux. Le génome possède à ses extrémités des répétitions terminales inversées (ITR) d'une dizaine de kpb (représentant en moyenne 8% de la taille totale du génome). Ces ITR comportent plusieurs structures spécifiques partagées par l'ensemble des *Poxviridae* : une *hairpin* terminale de 104 nucléotides comportant une boucle d'une dizaine de bases, une région conservée de 87 nucléotides en aval de la *hairpin* et des répétitions en tandem d'une taille allant de 54 à 125 pb distribuées sur 4 kpb. La structure des ITR semble conditionner le mode de réplication du génome des poxvirus. Ainsi, la réplication débuterait par une coupure simple brin dans l'un des TIR, libérant une extrémité 3' servant d'amorce à l'ADN

polymérase virale. La synthèse de ce nouveau brin va donc entraîner la synthèse d'un motif en *hairpin.* Ce brin néosynthétisé va donc lui même se replier et former à son tour une structure en épingle à cheveux, permettant la copie du second brin d'ADN génomique. Ce mode de réplication entraîne la formation de concatemer de génomes par appariement des *hairpin* et formation de structures cruciformes. Il a été démontré chez VACV que, lors de l'incorporation du génome viral dans les virions, la résolvase (A22) permettait la résolution des structures cruciformes et la maturation du génome viral. Bien que le mécanisme de réplication du génome des *Poxviridae* ne soit pas parfaitement connu, une NTPase ayant une activité primase et helicase a été indentifiée en plus de A22. Nous avons tenté de chercher des protéines homologues chez pandoravirus à ces deux protéines par comparaison de profils HMM, sans succès.

Mollivirus sibericum et Mollivirus kamchatka, proposition d'une nouvelle famille de virus géants

Isolement et caractérisation du deuxième mollivirus, *Mollivirus kamchatka*

Origine géographique et caractéristiques de l'échantillon

L'échantillon d'où à été activé *Mollivirus kamchatka* est issu de la série A, collectée par Alexander Morawitz. Le Kamtchatka est une péninsule volcanique de 1250 km de long. Le climat y est subarctique et tempéré à l'échelle locale. Ce climat relativement doux explique la présence de cryosols où le pergélisol y est « sporadique ». On note la présence de nombreux talik hydrothermaux. Le Kamtchatka est une zone riche en eau suite à la dernière glaciation et géologiquement active. On retrouve ainsi des formations volcaniques récentes, et à cet égard, le lac Kronotsky s'est formé au début de l'Holocène par obstruction du cours des fleuves locaux suite à une explosion du stratovolcan du même nom. L'échantillon en provenance des abords du lac Kronotsky, est caractéristique des beines d'érosion lacustre, et se présente sous forme de sable moyen. Les deux échantillons d'où ont été isolés respectivement *M. sibericum* et *M. kamchatka* correspondent à deux environnements différents, situés à plus de 1500 km de distance l'un de l'autre, permettant donc d'affirmer que les mollivirus ne sont pas uniquement retrouvés dans du pergélisol des régions subarctiques. De même, la découverte de ce mollivirus dans des sols de surface permet d'affirmer que ces derniers sont encore actifs.

Activation, caractérisation moléculaire et phénotypique de Mollivirus kamchatka

Afin d'enrichir la fraction virale, une part de l'échantillon a été placée en milieu riz et

incubée 1 mois à température ambiante et à l'obscurité. Après mise en culture d'un fragment de l'échantillon enrichi en présence de cellules d'*Acanthamoeba castellanii* le phénotype des cellules a été contrôlé quotidiennement par microscopie optique. Après 48h de co-culture, les cellules étaient à 80% décollées et à 30% d'enkystement. Une observation plus approfondie au microscope optique à fort grossissement (objectif x63 avec Optovar x1,6) a permis de caractériser la présence de petites particules sphérique d'environ 500 nm à l'intérieur des amibes ainsi qu'en suspension dans le milieu.

Pour déterminer si le changement phénotypique des amibes est lié à une infection virale et, si oui par quelle virus: nous avons combiné des observations directes, par microscopie électronique, d'une fraction du milieu de culture concentrée ainsi que par amplification et séquençage de fragments d'ADN amplifiés par PCR directe dans des conditions peu stringentes. Ici, les amorces choisies pour la PCR directe ont été conçues pour permettre l'amplification d'une zone de 1500 pb conservée entre *M. sibericum* et les pandoravirus. Après migration sur gel d'agarose, la taille des amplicons générés atteint 1000 pb soit 30% plus petit qu'attendu. La purification, le séquençage des amplicons, puis l'analyse manuelle des chromatogrammes a permis de récupérer deux séquences d'ADN d'une taille de 777 pb, pour la séquence orientée dans le sens inverse, et de 838 pb pour la séquence orientée dans le sens direct. L'alignement de ces deux séquences par BLASTn contre l'ensemble des génomes des virus géants étudiés au laboratoire a démontré qu'ils étaient similaires à une région terminale du génome de *M. sibericum* (comprise entre 538013 pb et 539769 pb). Le pourcentage d'identité de 94% pour la séquence direct et de 93% pour la séquence inverse (3% d'insertion/délétion) suggère la présence dans la flasque de culture d'un nouveau isolat de mollivirus. Pour confirmer cette hypothèse, une observation directe d'un aliquot concentré du surnageant de culture en coloration négative est nécessaire. L'observation de particules sphériques d'une taille légèrement supérieure à 600 nm renforce la suspicion d'une infection virale par un nouveau mollivirus.

Après amplification par 2 repiquages successifs nous avons cloné le virus afin de le produire en quantité suffisante et afin de poursuivre la caractérisation de ce nouvel agent infectieux d'*Acanthamoeba castellanii*. La purification sur gradient de chlorure de cesium a donné un anneau viral d'une densité voisine de 1,4.

Les résultats de séquençage en faveur de la découverte d'une nouvelle souche de mollivirus sont renforcés par analyse comparée par gel SDS-Page de l'ensemble du contenu protéique des particules purifiées de *M. sibericum* et du nouveau *M. kamchatka*. Après migration, les profils protéiques présentent trois différences significatives et un maintien d'une bande majoritaire.

Microscopie optique et électronique, vers une meilleure compréhension du cycle infectieux des mollivirus et de l'assemblage des particules virales

Le cycle infectieux de Mollivirus kamchatka et le devenir du noyau au cours du cycle infectieux

Dans le but d'obtenir un contraste fort des membranes', l'étude du cycle infectieux de *M. kamchatka* a été faite par MET en suivant le protocole OTO. En parallèle, pour permettre la mise en évidence du devenir du noyau au cours de l'infection virale, un suivi d'infection par microscopie optique à épi-fluorescence de cellules marquées au DAPI a été fait.

Au cours du cycle infectieux on observe des changements phénotypiques des amibes. A partir de 3h post-infection (PI) on observe une forme arrondie des cellules (5% des cellules à 2h PI). On observe ensuite une perte progressive des vacuoles (40% de cellules sans vacuoles à 6h PI) et la membrane externe des cellules apparaît de moins en moins réfringente. Les cellules apparaissent longues et étalées (60µm à 80µm de longueur). En phase tardive, après 10h, 40% à 50% des cellules ont lysées et les restantes semblent comme vidées de leur contenu cytoplasmique, prenant une forme arrondie plus petite (40 µm). Les changements de morphologies des amibes dessinent donc 3 grandes phases du cycle infectieux qui correspondent à des événements se déroulant dans le cytoplasme des amibes et observables par microscopie électronique.

Premièrement, les virus sont internalisés par les cellules à partir d'une heure PI. Les phagosomes contenant les particules virales fusionnent et on observe à 3h PI jusqu'à 7 particules virales dans un même vésicule. Durant cette phase précoce, on observe un nombre croissant de particules virales dont l'intérieur des capsides apparaît peu dense aux électrons, suggérant un échange entre le contenu de la capside et le cytoplasme de la cellule. Comme pour la majorité des virus géants, c'est une fusion de la membrane interne de la capside virale avec la membrane du phagosome qui permet cette interaction. autour de 4h PI, peu de virions sont visibles dans le cytoplasme. Cette disparition des virions suppose donc qu'après la phase précoce du cycle s'opère une phase d'éclipse. En parallèle, on observe une zone d'exclusion des organites intra-cellulaire, l'usine virale, se former dans le cytoplasme. L'usine virale de mollivirus semble accumuler en son sein de nombreux débris membranaires et de longues fibres denses aux électrons. C'est durant cette phase intermédiaire, A partir de 5h PI, que les premières particules se forment à la périphérie de l'usine virale. On observe des virions matures dans le cytoplasme ou dans des vacuoles. Il est difficile de conclure si les virus sont enveloppés suite à l'assemblage ou si ce phénomène est lié à une sur-infection. Après 7h PI, les cellules ont perdues l'ensemble de leurs vacuoles. La phase tardive du cycle cellulaire, se termine par l'exocytose des virions ou la lyse des cellules.

L’étude du noyau au cours de l'infection virale par marquage de l'ADN total au DAPI montre que les événements cellulaires vont de pairs avec les perturbations nucléaires suivantes : en phase précoce on observe une décompaction puis un mouvement du nucléole du centre vers la périphérie du noyau. Après la phase d’éclipse, on observe des structures fibrillaires s'accumuler dans le noyau, identiques à celles présentes dans l'usine virale. Enfin, après 6h PI on observe la rupture de l'enveloppe nucléaire dans 40% des cellules, libérant les structures fibrillaires accumulées au noyau vers le cytoplasme (Figure 43).

Pour conclure, la phase précoce permet l'entrée du virus dans les cellules, la sortie du matériel génétique et la transcription précoce au noyau des gènes viraux. Après une courte période d’éclipse, la phase intermédiaire permet la transcription des gènes viraux par l'appareil de transcription viral et l'installation d'une usine virale dans le cytoplasme. Durant la phase, les particules virales sont assemblées et les virions relâchés à l’extérieur de la cellule. Il apparaît donc que l’absence d'ARN polymérase dans la capside de *M. sibericum* est une donnée déterminante dans la dépendance au noyau de mollivirus au cours du cycle infectieux.

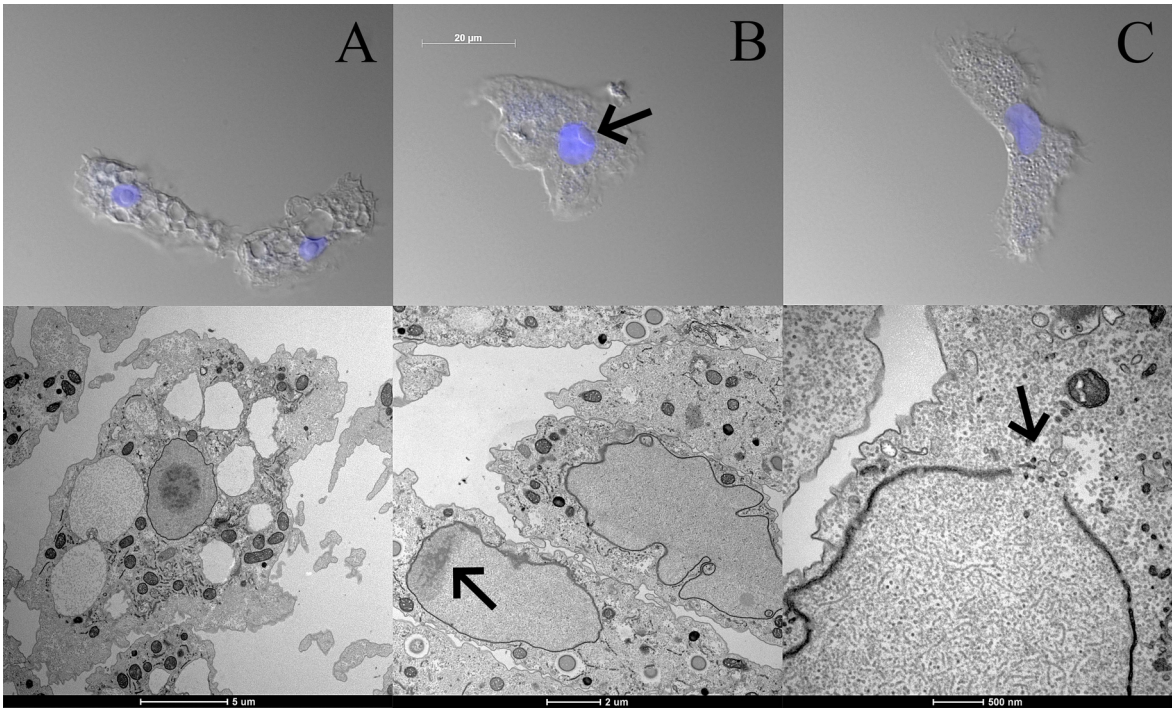


Figure 43:Suivi par microscopie optique et électronique du devenir du noyau cellulaire au cours de l'infection par M. kamchatka. (A) Cellule non infectée. On observe que le nucléole est circulaire et centré dans le noyau. Les amibes apparaissent vacuolées et le noyau semble est visible. (B) A 5h PI on constate un mouvement du nucléole vers la périphérie du noyau (flèche). Les amibes semblent plus étalées et contiennent peu de vacuoles. (C) A 7h PI on ne peut plus distinguer le noyau, les amibes sont pleines de particules virales. Par microscopie électronique on constate une rupture de la membrane nucléaire (flèche).

Microscopie électronique et assemblage des particules virales

Les observations du cycle infectieux de *M. kamchatka* sont venues enrichir une étude du cycle infectieux de *M. sibericum* par tomographie et sonde ionique focalisée (FIB) réalisée en collaboration avec l'Institut Pasteur de Paris. Les d'observations des cycles infectieux des deux mollivirus que j'ai réalisé ont permis à caractériser la structure membranaire à l'origine de l'assemblage des particules virales, me permettant ainsi d'être impliqué dans le processus de publication de l'article présenté à la fin de ce chapitre. Le modèle d'assemblage des particules virales proposé distingue quatre phases dont les particularités sont les suivantes :

1. Précurseur membranaire : l'initiation de l'assemblage se fait à partir d'une structure unique comportant une vésicule aplatie à laquelle est accolée en son centre une membrane linéarisée ressemblant à un *cisterna* ouvert et dont les extrémités sont recourbées. L'assemblage du virion commence par le recrutement de vésicules lipidiques (d'une taille moyenne de 66 nm) se linéarisant au niveau des extrémités recourbées du *cisterna,* entraînant l'assemblage de la membrane interne du virion. La synthèse de la couche dense de la capside se fait, quant à elle, de façon multidirectionnelle à partir du vésicule de la structure d'initiation (Figure 44).
2. Croissant viral : on observe, en parallèle de l'élongation de la capside, une accumulation du contenu interne du virion. Ce phénomène est permis par la présence d'une membrane circulaire à l'opposée de la structure d'initiation. Cette membrane circulaire semble couverte de grains denses aux électrons, laissant supposer qu'il s'agit de membranes issues du recyclage du réticulum endoplasmique granuleux de la cellule hôte (Figure 44).
3. Clôture du virion : c'est à la fin de l'assemblage que le génome viral est empaqueté dans la capside. Grâce à un immuno-marquage des coupes de cryomicroscopie, on observe que les structures fibrillaires, décrites plus haut et observables dès le début de la phase intermédiaire du cycle, sont liées à des molécules d'ADN. Ces structures nucléoproteiques denses aux électrons sont à priori embarquées dans les particules en fin d'assemblage car seuls les virions presque entièrement assemblés sont positifs à l'immuno-marquage (Figure 44).
4. Virion mature : la fin de l'assemblage se fait à l'opposé de la structure d'initiation, présente jusqu'à la fin de l'assemblage des particules. La structure externe du virion apparaît alors striée et entourée d'anneaux fins, suggérant que les virions s'entourent au cours de l'assemblage d'une couche de « cheveux » de nature inconnue (Figure 44).

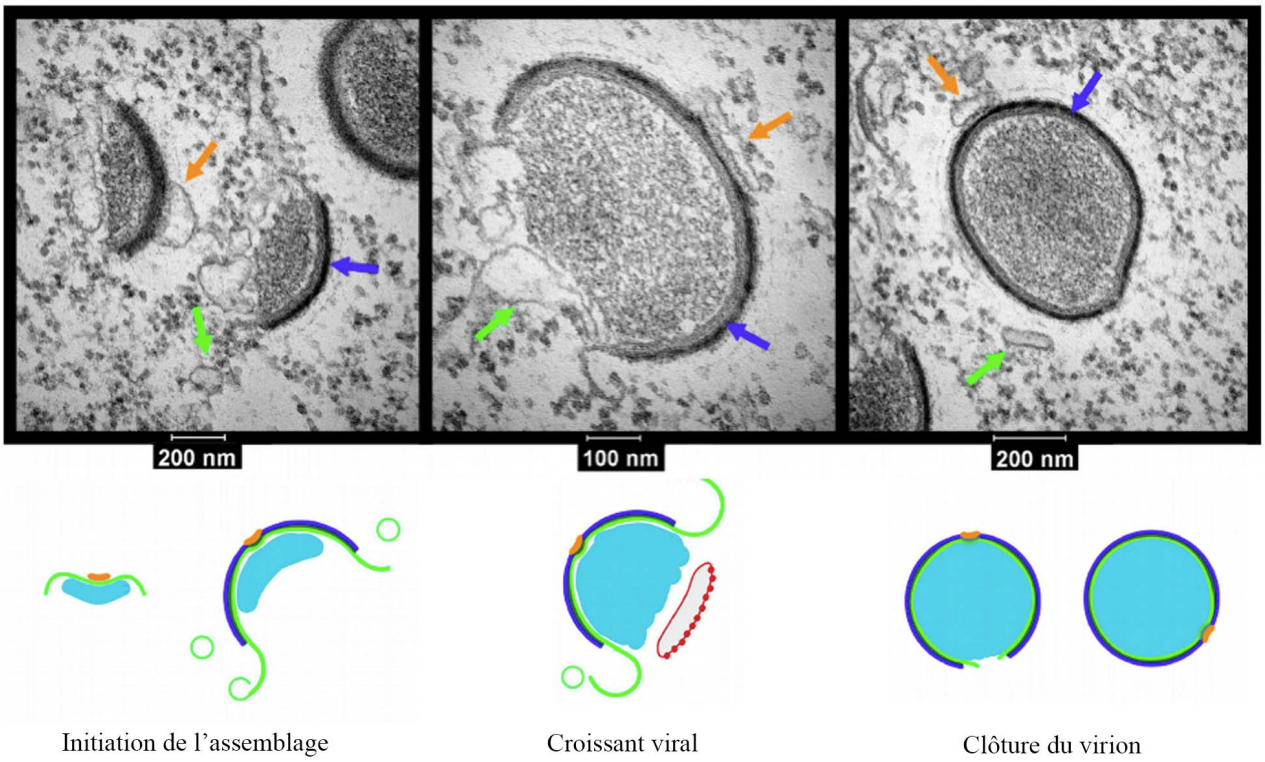


Figure 44 : Schéma des différentes phases de l'assemblage des virions de M. sibericum. En orange, le vésicule fermée du complexe d'initiation de l'assemblage. En vert, vésicules circulaires recrutés pour l'assemblage des structures membranaires internes du virion. On observe la linéarisation de ces vésicules au niveau de la zone d'assemblage. En bleu, extension de la structure lamellaire depuis le complexe d'initiation de l'assemblage.

Bien que les origines des membranes recrutées durant l'assemblage soient inconnues, il apparaît vraisemblable que plusieurs types de membranes interviennent au cours de la synthèse des virions. Ainsi par exemple, l'absence de réticulation par des ribosomes sur le précurseur membranaire au moment de l'initiation peut suggérer qu'il s'agit d'un fragment membranaire issu du recyclage de l'appareil de Golgi de la cellule hôte. De même, l'élongation de la membrane interne du virion, permise par le recrutement de vésicules s'ouvrant au contact de la zone d'assemblage rappelle le mouvement des saccules golgiens et suggère donc le probable recrutement d'éléments membranaires de l'appareil de Golgi. Enfin, la réticulation du croissant viral pourrait suggérer la présence de ribosomes à sa surface, indiquant le recyclage du réticulum endoplasmique granuleux de l'hôte.

Papier 2 : Quemin, E. R., Corroyer-Dulmont, S., Baskaran, A., Penard, E., Gazi, A. D., Christo-Foroux, E., ... & Krijnse-Locker, J. (2019). Complex membrane remodeling during virion assembly of the 30,000-Year-Old mollivirus Sibericum. Journal of virology, 93(13).

AMERICAN SOCIETY FOR MICROBIOLOGY | Journal of Virology®

VIRUS-CELL INTERACTIONS

Complex Membrane Remodeling during Virion Assembly of the 30,000-Year-Old Mollivirus Sibericum

E. R. Quemin,[a]* S. Corroyer-Dulmont,[a] A. Baskaran,[a]* E. Penard,[a] A. D. Gazi,[a] E. Christo-Foroux,[b] P. Walther,[c] C. Abergel,[b] J. Krijnse-Locker[a]

[a]Ultrastructural Bio-Imaging (UBI), Center for Ressources and Research in Technology (C2RT) and Department of Cell Biology and Infection, Institut Pasteur, Paris, France
[b]Aix-Marseille University, Centre National de la Recherche Scientifique, Information Génomique & Structurale, Unité Mixte de Recherche 7256, Institut de Microbiologie de la Méditerranée, FR3479, Marseille, France
[c]Electron Microscopy (EM) Core Facility, University of Ulm, Ulm, Germany

ABSTRACT Cellular membranes ensure functional compartmentalization by dynamic fusion-fission remodeling and are often targeted by viruses during entry, replication, assembly, and egress. Nucleocytoplasmic large DNA viruses (NCLDVs) can recruit host-derived open membrane precursors to form their inner viral membrane. Using complementary three-dimensional (3D)-electron microscopy techniques, including focused-ion beam scanning electron microscopy and electron tomography, we show that the giant Mollivirus sibericum utilizes the same strategy but also displays unique features. Indeed, assembly is specifically triggered by an open cisterna with a flat pole in its center and open curling ends that grow by recruitment of vesicles never reported for NCLDVs. These vesicles, abundant in the viral factory (VF), are initially closed but open once in close proximity to the open curling ends of the growing viral membrane. The flat pole appears to play a central role during the entire virus assembly process. While additional capsid layers are assembled from it, it also shapes the growing cisterna into immature crescent-like virions and is located opposite to the membrane elongation and closure sites, thereby providing virions with a polarity. In the VF, DNA-associated filaments are abundant, and DNA is packed within virions prior to particle closure. Altogether, our results highlight the complexity of the interaction between giant viruses and their host. Mollivirus assembly relies on the general strategy of vesicle recruitment, opening, and shaping by capsid layers similar to all NCLDVs studied until now. However, the specific features of its assembly suggest that the molecular mechanisms for cellular membrane remodeling and persistence are unique.

IMPORTANCE Since the first giant virus *Mimivirus* was identified, other giant representatives are isolated regularly around the world and appear to be unique in several aspects. They belong to at least four viral families, and the ways they interact with their hosts remain poorly understood. We focused on Mollivirus sibericum, the sole representative of "Molliviridae," which was isolated from a 30,000-year-old permafrost sample and exhibits spherical virions of complex composition. In particular, we show that (i) assembly is initiated by a unique structure containing a flat pole positioned at the center of an open cisterna, (ii) core packing involves another cisterna-like element seemingly pushing core proteins into particles being assembled, and (iii) specific filamentous structures contain the viral genome before packaging. Altogether, our findings increase our understanding of how complex giant viruses interact with their host and provide the foundation for future studies to elucidate the molecular mechanisms of Mollivirus assembly.

KEYWORDS electron tomography, focused-ion beam scanning electron microscopy, giant viruses, membrane remodeling, Mollivirus sibericum, nucleocytoplasmic large DNA viruses, viral factory, virus assembly

Citation Quemin ER, Corroyer-Dulmont S, Baskaran A, Penard E, Gazi AD, Christo-Foroux E, Walther P, Abergel C, Krijnse-Locker J. 2019. Complex membrane remodeling during virion assembly of the 30,000-year-old Mollivirus sibericum. J Virol 93:e00388-19. https://doi.org/10.1128/JVI.00388-19.

Editor Joanna L. Shisler, University of Illinois at Urbana Champaign

Address correspondence to C. Abergel, chantal.Abergel@igs.cnrs-mrs.fr, or J. Krijnse-Locker, jacomina.krijnse-locker@pasteur.fr.

* Present address: E. R. Quemin, Structural Cell Biology of Viruses, Center for Structural Systems Biology, Hamburg, Germany; A. Baskaran, Institut du Cerveau et de la Moelle Epinière, Hôpital Pitié Salpêtrière, Paris, France.

Received 5 March 2019
Accepted 11 April 2019
Accepted manuscript posted online 17 April 2019
Published 14 June 2019

Membrane dynamics is tightly controlled to perform cellular functions at specific locations. Fusion and fission reactions maintain organelle identity and integrity and ensure communication, protein secretion and degradation, lipid synthesis, uptake of nutrients, etc. (1, 2). As obligatory intracellular parasites, viruses hijack the host cell machinery and, in particular, target cellular membrane pathways for entry, i.e., fusion, endocytosis or phagocytosis (3), replication (4), assembly, and egress (5–7).

Nucleocytoplasmic large DNA viruses (NCLDV) originally encompassed an extended viral group with members of the families *Poxviridae*, *Asfarviridae*, *Iridoviridae*, and *Phycodnaviridae* (8, 9), characterized by large particle and genome sizes. Replication of NCLDVs is either exclusively carried out in the cytoplasm of the host cell or is partially dependent on the nucleus for early gene expression (10–17). NCLDVs form viral factories (VFs) in infected cells, which are dedicated sites for viral transcription, replication, and assembly, leading to a dramatic reorganization of the host cytoplasm (18–20). The presence of an inner membrane, surrounded and shaped by a viral capsid or scaffold protein, is also a common feature. The biogenesis and origin of this membrane have been the topic of intensive studies for vaccinia virus (VACV), the type species of the *Poxviridae* family. Our working model proposes that the assembling virion is composed of an open membrane sphere containing a single lipid bilayer shaped by a honeycomb-like scaffold. Open membrane intermediates, derived from the endoplasmic reticulum, contribute to the formation of the membrane sphere that closes once the DNA has been packed (21–26). Additional studies on African swine fever virus (27), *Paramecium bursaria* chlorella virus-1 (28), and *Acanthamoeba polyphaga* mimivirus (APMV) (29–32) also revealed the presence of open membrane intermediates and the formation of open membrane spheres shaped by the assembly of a capsid on their convex side, suggesting a novel and common strategy for viral inner membrane assembly.

APMV was the first giant virus discovered and the founding member of a new family of DNA viruses characterized by viral particle dimensions in the micrometer range and genomes encoding up to 1,500 proteins, the *Mimiviridae* of the recently proposed *Megavirinae* (10, 33, 34). This rapidly expanding family appeared to be phylogenetically related to the other families of NCLDVs. However, three recently proposed viral families, namely, *Pandoraviridae* (35), *Pithoviridae* (36), and *Molliviridae* (37), share only a few genes and question the likeliness of a common origin for all giant viruses (38). Notably, their particles exhibit an internal membrane lining the virion internal capsid shell, even though particle morphotypes range from spherical or ovoid to amphora shaped and are significantly different from the common icosahedral virions of APMV and most NCLDVs, with the exception of poxviruses (39, 40). The sole representative of the *Molliviridae*, Mollivirus sibericum, recently isolated from a 30,000-year-old permafrost sample, displays spherical virions (~600 nm) with a thick coat or tegument covered with fibers (37). Here, we use a combination of three-dimensional electron microscopy (3D-EM) techniques to provide a thorough characterization of membrane acquisition and virion assembly of the giant Mollivirus sibericum. Despite little homology with other large and giant viruses, our findings show that Mollivirus sibericum utilizes a strategy similar to NCLDVs for viral inner membrane assembly involving the recruitment and opening of cell-derived membrane intermediates. We also show that a cisterna with open curling ends and a central flat pole serves as a unique precursor for membrane assembly. The viral inner membrane is subsequently shaped by additional capsid layers assembled from the flat pole, and DNA is packed prior to closure of particles on the spherical counterpoint of this pole. Collectively, our analysis provides significant insights into the unique assembly of complex virions of the recently isolated Mollivirus sibericum and suggests that viral inner membrane assembly via open intermediates might be a common feature of several large and giant viruses infecting eukaryotes.

RESULTS

Reorganization of the whole cell during infection and formation of a prominent viral factory. At 3 h postinfection, viruses were found in large vacuoles, individ-

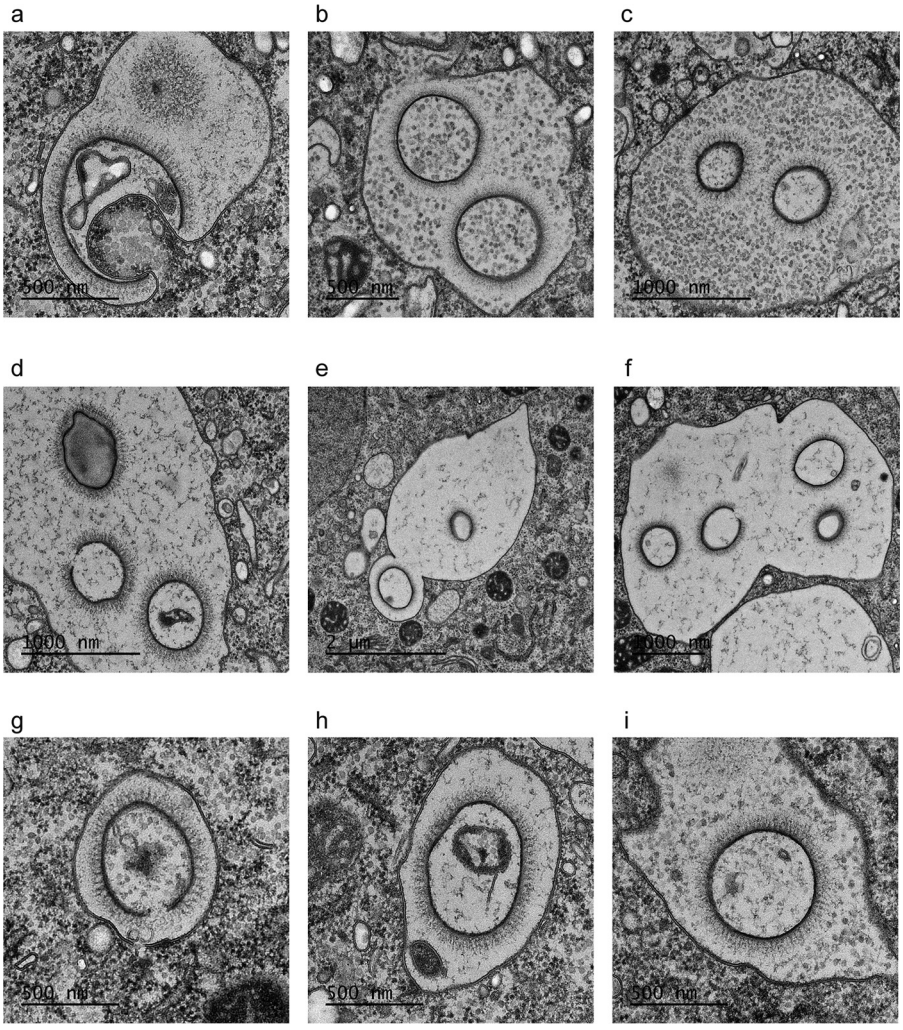


FIG 1 Transmission electron micrographs on the entry of Mollivirus sibericum. At 3 h postinfection, cells of *Acanthamoeba castellanii* were fixed and processed for transmission electron microscopy after cryo-fixation and freeze substitution. (a to i) Entry of Mollivirus sibericum leads to substitution of the viral inner content (a) and occurs via endocytosis with multiple particles present in one vacuole with different shape and content (b to f) and in individual particles with internal structures (g to i).

ually or in groups of 2 to 4 particles (Fig. 1), as previously observed (37). Particles of Mollivirus sibericum, like all other giant viruses characterized until now, are internalized by *Acanthamoeba castellanii* and undergo structural reorganization to fuse their inner membrane with the surrounding vacuolar membrane (40). While the particle interior is completely substituted with cytoplasmic content (Fig. 1), no electron-dense core that could contain the viral genome was detected within the host cytoplasm as opposed to APMV and VACV but reminiscent of *Pandoravirus* and *Pithovirus* (40). At 8 h postinfection, the host cell is completely rearranged with the formation of a prominent VF in the cytoplasm where viral replication and assembly of virions occur (Fig. 2). In the majority of infected cells, the nucleus appeared to be absent or greatly reorganized (Fig. 2), characterized by a loss of spherical appearance, a decrease in size, or a mislocalized nucleolus at the nuclear periphery rather than in the center. VFs occupy the majority of the cellular space and contain virions at different stages of assembly, and organelles, such as mitochondria, endoplasmic reticulum, and the Golgi complex, are excluded toward the cell periphery, close to the plasma membrane (Fig. 2) (20, 37, 41). Compared to uninfected cells (see Movie S1 in the supplemental material), mitochondria in infected cells are smaller and elongated rather than spherical, having a shorter diam-

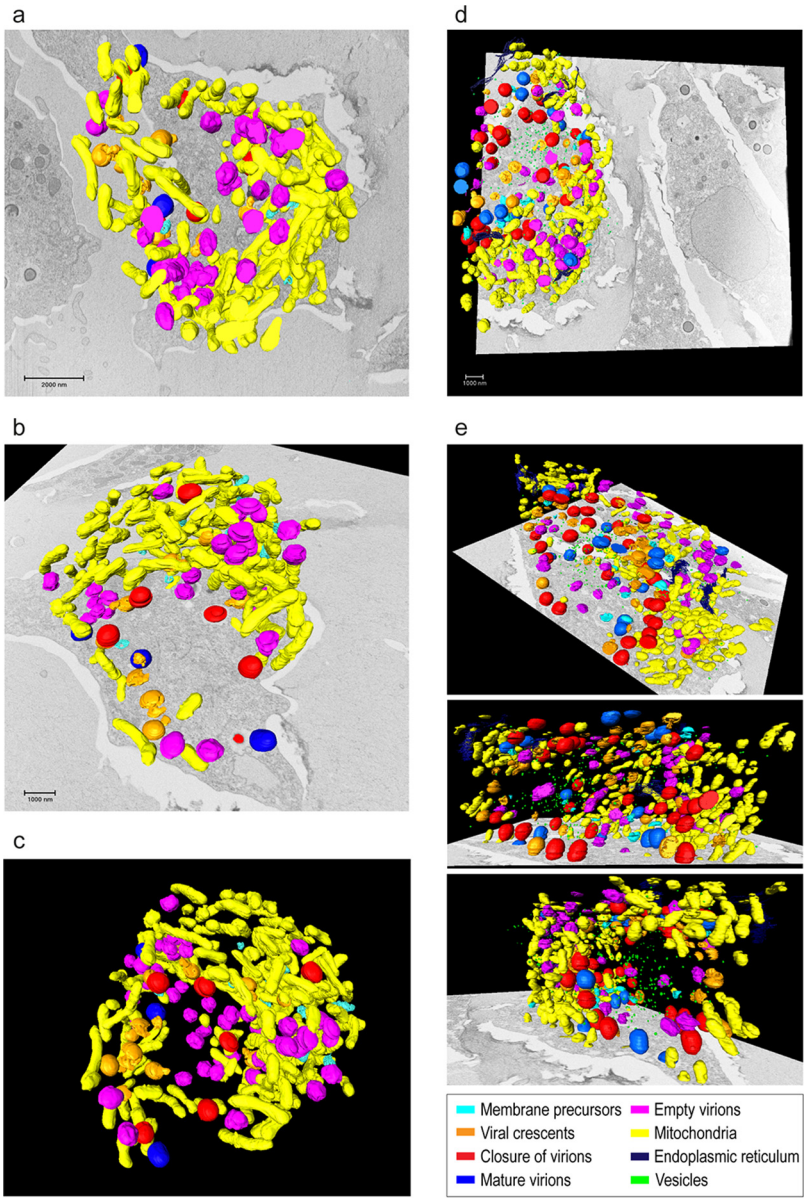


FIG 2 Analysis of whole infected cells by focused-ion beam scanning electron microscopy. At 8 h postinfection, cells of *Acanthamoeba castellanii* were fixed and processed for focused-ion beam scanning electron microscopy. (a) Slice 535 through the reconstructed volume (acquired at 10 by 10 by 10 nm) with 3D surface rendering of infected cells. Scale bar = 2 μm. (b) Slice 242 of the acquired volume and 3D surface rendering. (c) Projection with 3D surface rendering. (d) Slice 592 through the reconstructed volume (acquired at 10 by 10 by 10 nm) with 3D surface rendering of infected cells. Scale bar = 1 μm. (e) Slice 592 of the acquired volume and 3D surface rendering at different angles. (a to e) The video of volume reconstruction with segmentation is shown in Movie S2 (a to c) and Movie S3 (d to e). As indicated in the key, mitochondria are in yellow, endoplasmic reticulum is in black, vesicles are in green, empty virions are in magenta, mature virions are in dark blue, closure of virions is in red, viral crescents are in orange, and membrane precursors are in cyan.

eter (Table 1). The VFs appear less organized than do the assembly sites of APMV (13, 30, 31) and more resemble VACV-infected cells (20). The different stages of virus assembly, membrane precursors, viral crescents, and virions in the process of closure or maturation were found within the VF without any specific arrangement (Fig. 2 and 3). Inside the VFs, long filaments are observed (more visible in Fig. 4) which are specific to Mollivirus sibericum infection and most likely of viral origin (37). They might be involved in genome replication and packaging but do not seem to participate in particle assembly *per se* (discussed below).

TABLE 1 Shape and dimensions of mitochondria in uninfected and infected cells[a]

	Mitochondrial shape or dimension data					
Cell type	Length (nm)	Width (nm)	Breadth (nm)	Thickness (nm)	Vol (nm^3)	Sphericity
Uninfected	834.49 ± 230.16	437.19 ± 150.62	683.41 ± 130.41	420.83 ± 151.96	107,482,537.12 ± 82,758,689.89	1.21 ± 0.17
Infected	955.60 ± 343.93	387.54 ± 85.42	506.79 ± 85.34	357.34 ± 80.44	75,337,392.68 ± 38,567,468.94	1.92 ± 0.75

[a]Quantitative analysis of mitochondria observed in 3 uninfected ($n = 224$) and 3 infected cells ($n = 596$) acquired by focused-ion beam scanning electron microscopy in terms of length, width, breadth, thickness, volume, and sphericity. The average value and standard deviation are given for each condition (see Materials and Methods).

Membrane biogenesis guides the complex assembly of Mollivirus sibericum. **(i) Membrane precursors.** The assembly of virions of Mollivirus sibericum is initiated by a unique membrane precursor never reported for other NCLDVs (Fig. 3b and enlargement in Fig. 3c-1 and -3, white star). This membrane precursor resembles a cisterna of unknown origin which is specifically associated with a flat pole in its center (Fig. 5, black bracket) and displays open curling ends. In more mature particles with a thick electron-dense capsid layer, the flat pole appears as a membranous closed structure, appended on the convex side of the capsid, causing its flattening (Fig. 6, black bracket). Based on the fact that the flat pole is present on the membrane precursor we identified and is located at the center of virions being assembled (Fig. 3b and c), we propose that capsid assembly (Fig. 5 and 7) proceeds from this pole toward the open curling ends in a symmetrical manner.

(ii) Viral crescents. Typical crescent structures are formed upon elongation of the viral membrane precursor, shaped by a thick tegument or coat with at least two additional layers (Fig. 3b and enlargement in Fig. 3c-2). The open inner membrane ends are uncoated and curl toward the exterior of the particles (Fig. 7), similar to VACV crescents (21). Large amounts of coated (74 ± 7 nm in diameter) and uncoated (66 ± 24.7 nm in diameter) vesicles are present at the assembly sites that may actively participate in the assembly of the inner viral membrane (Fig. 7 and 8). Uncoated vesicles of roughly 66 nm in diameter were reproducibly found in close proximity to the open ends of the growing viral particles (Table 2). In contrast to VACV, for which open vesicular intermediates are recruited to the assembly sites (21), the vesicles found in the vicinity of the growing virions of Mollivirus are closed. However, vesicles in contact with the membrane ends are opened, suggesting that rupture occurs upon contact with the growing viral membrane (Fig. 7 and 8). We also observe a coassembly phenomenon, already reported for VACV, where particles at different stages of assembly can be connected via their inner viral membrane (Fig. 9). Packaging of the viral inner content is specifically associated with another cisterna-like element (Fig. 3c-2, white arrowheads), and this structure is present on the opposite side of the flat pole (Fig. 7 and Table 2). The presence of ribosomes segregated on the side facing the cytoplasm suggests that it originates from the endoplasmic reticulum (Fig. 10). Along the assembly process, the material found inside particles is densely packed and does not exhibit visible features. Interestingly, some material is already present at the concave side of the early crescents and becomes more obvious as the viral sphere expands (Fig. 3c-1 and -2).

(iii) Closure of virions. At the end of assembly, closure of the inner membrane occurs on the opposite side of the initiation pole (Fig. 3b and enlargement in Fig. 3c-3). At the site of closure, vesicles are often observed and might be involved in membrane sealing (Fig. 11). With DNA immunolabeling of thawed cryo-sections of infected cells, only particles that are mature or close to closure are labeled, indicating that genome packaging occurs at the final stages of assembly, similar to VACV (40) and APMV (21, 26, 29, 30, 42). The immunolabeling is also detected on long filamentous structures (Fig. 4) which are specific to Mollivirus sibericum infection and abundant in the VF. These DNA-associated filaments participate in DNA packaging and might therefore represent the viral nucleoproteins (Legendre et al. [37]) (Fig. 12). Notably, the flat pole is present

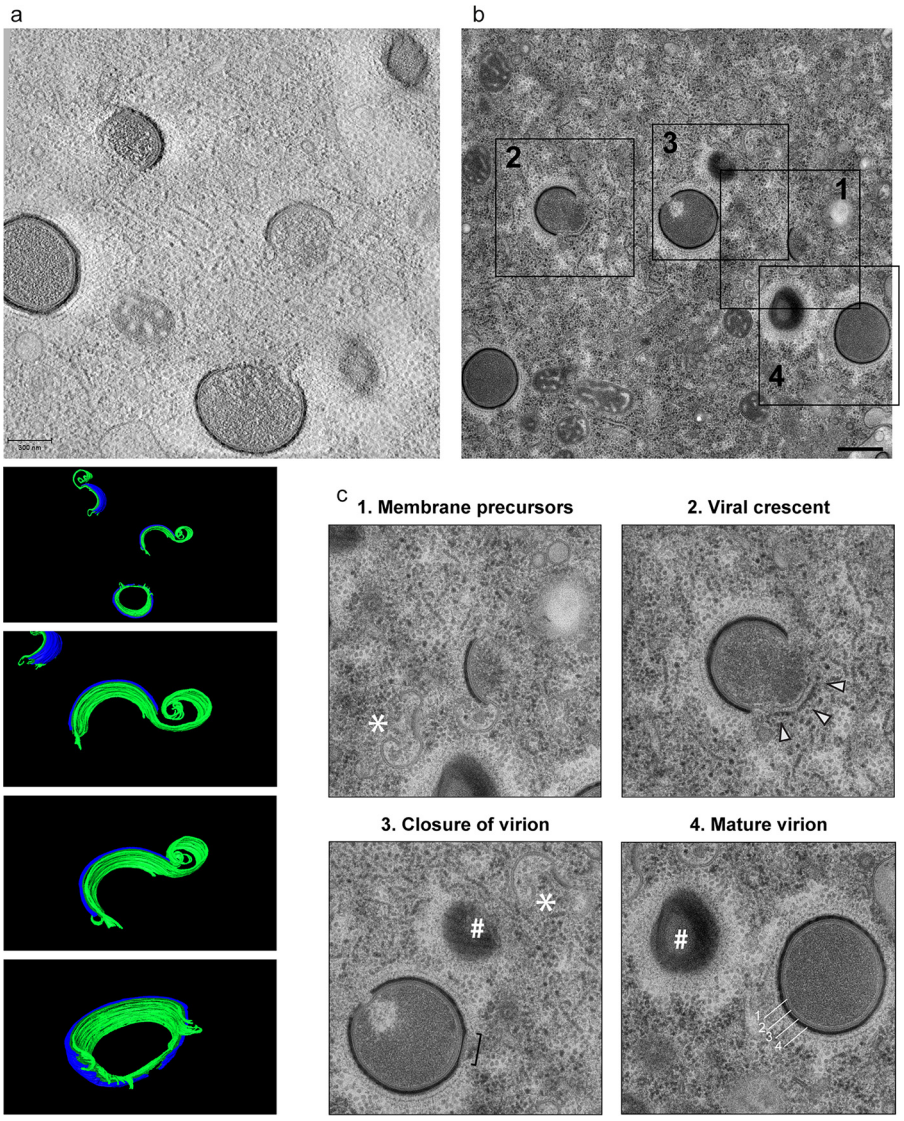


FIG 3 Observations of the different assembly stages of Mollivirus sibericum. At 8 h postinfection, cells of *Acanthamoeba castellanii* were fixed and processed for scanning transmission electron tomography and transmission electron microscopy after cryo-fixation and freeze substitution. (a) Slice 40 through a tomogram acquired by scanning transmission electron tomography; bottom, projections of 3D surface rendering of particles being assembled. A detailed overview of slices through the tomogram including 3D surface rendering is shown in Fig. 8. The video of the reconstructed volume with segmentation is shown in Movie S4. The tegument is in blue, additional capsid layers are in dark green, and the inner viral membrane is in light green. Scale bar = 300 nm. (b) Different stages of viral assembly can be observed by transmission electron microscopy, as follows: 1, curved membrane intermediates with open curling ends associated with a flat pole or covered with the tegument and additional capsid layers; 2, typical open crescents associated with cisternae with segregated ribosomes; 3, closure of particles beginning with the inner viral membrane on the opposite side of the flat pole; and 4, mature particles with thick tegument covered with fibers and encasing additional capsid layers, the inner viral membrane, and dense material. Scale bar = 500 nm. (c) Enlargements of the particles at different stages of assembly as depicted by black squares in panel b. Membrane precursors are indicated by white stars in squares 1 and 3, the cisternae associated with viral crescents and segregated ribosomes are indicated by white arrowheads in square 2, the flat pole is indicated by a black bracket in square 3, and the different capsid layers are numbered in square 4. Note the top views and tangential planes of particles in squares 3 and 4 which are indicated by a white pound sign and show the striations of the capsid outer surface.

at all stages of assembly as well as on mature particles and is always located on the spherical counterpart compared to the sites used for membrane elongation and closure or DNA uptake (Fig. 12, black bracket).

(iv) Mature virions. Mature viruses are composed of at least one inner membrane enclosing dense material, two additional capsid layers, and a thick tegument or coat

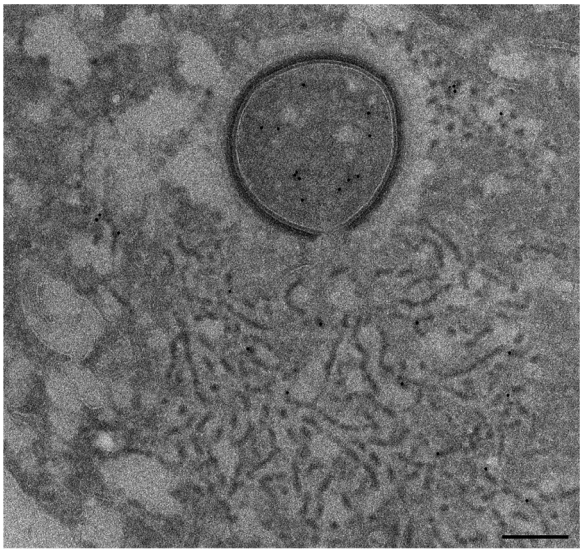


FIG 4 DNA-associated filaments involved in packaging during assembly of Mollivirus sibericum. At 8 h postinfection, cells of *Acanthamoeba castellanii* were fixed and processed for immunolabeling of thawed cryo-sections. DNA labeling is associated with viral filamentous structures and present in particles at the end of the assembly. Scale bar = 500 nm.

covered with two layers of fibers (Fig. 3b and enlargement in Fig. 3c-4; the different capsid layers are indicated and numbered). When observed from the top, on a tangential plane, the surface of the particle displays remarkable striations (Fig. 3c-3 and 2c-4, white pound sign), as previously reported (37). One of the two additional capsid layers present between the inner membrane and the tegument resembles a membrane

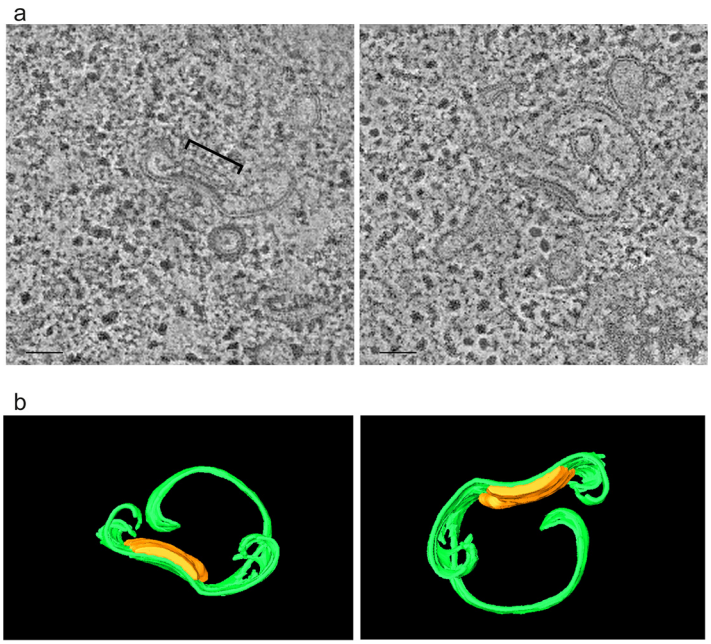


FIG 5 3D analysis of membrane precursor during assembly of Mollivirus sibericum particles. At 8 h postinfection, cells of *Acanthamoeba castellanii* were fixed and processed for transmission electron tomography. (a) Slices 12 (left) and 94 (right) through the tomogram showing curved membrane intermediates, with open curling ends associated with a flat pole and typical open crescents associated with cisternae. Scale bars = 200 nm. The flat pole is indicated by a black bracket. (b) Projection images of 3D surface rendering of the particle shown in panel a with different angles. The video of volume reconstruction with segmentation is shown in Movie S5. The tegument is in blue, the additional capsid layers are in dark green, the inner viral membrane is in light green, and the flat pole is in orange.

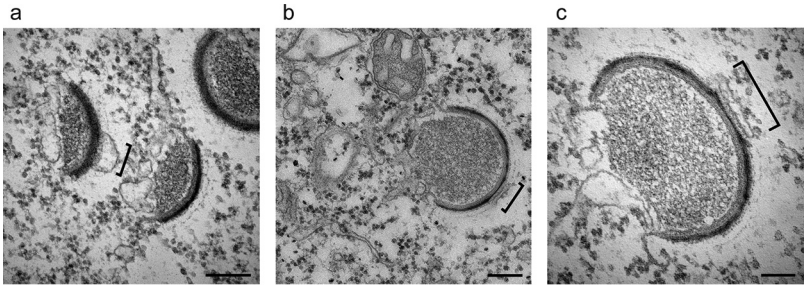


FIG 6 Transmission electron micrographs showing the flat pole on viral crescents. At 8 h postinfection, cells of *Acanthamoeba castellanii* were fixed and processed for transmission electron microscopy after substitution at room temperature. Different stages of virus assembly are shown. The flat pole, indicated by a black bracket, is always visible at the opposite of the site where assembly proceeds. Scale bars = 200 nm (a and b) or 100 nm (c).

when cells are embedded in resin (Fig. 13, left, white arrows). However, only the inner structure appears white (Fig. 3c-4, layer 3, and Fig. 13, right, white arrows) when using the Tokuyasu method, suggesting that there might be only one membrane in Mollivirus sibericum. Cryo-EM will be needed to solve virion structure at high resolution and under native conditions (43) in order to make conclusions.

Abundant membrane proteins in mollivirions. Using membrane prediction software programs (see Materials and Methods), we identified 27 proteins containing at least one transmembrane domain in the publicly available proteome of Mollivirus sibericum virions (37). Interestingly, structural and functional predictions of these membrane-containing proteins highlight a remote homologue of the fusion protein of flaviviruses (ml_417), lysosome-associated proteins (ml_309 and ml_331) and vesicle-associated proteins (ml_448 and ml_185), a possible cell adhesion protein (ml_499), and a homologue to a pore-forming protein (ml_330) (Table 3), which in the context of our findings are interesting targets for future studies. The presence of these numerous membrane proteins composing about 20% of the proteome of the purified viral

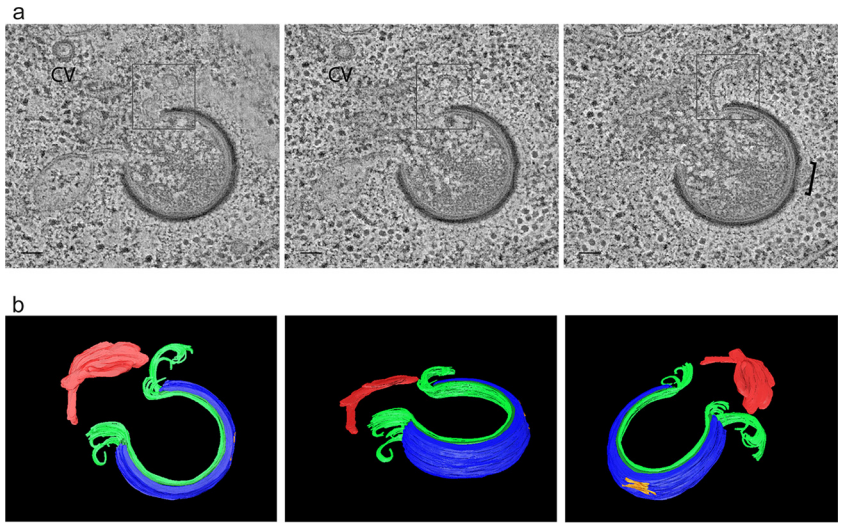


FIG 7 3D analysis of viral crescent during assembly of Mollivirus sibericum particles. At 8 h postinfection, cells of *Acanthamoeba castellanii* were fixed and processed for transmission electron tomography. (a) Slices 16, 24, and 56 (from left to right) through the tomogram showing a typical viral crescent associated with a cisterna and displaying open curling ends. On the open curling end highlighted with a black square, a vesicle comes into contact with the inner membrane and subsequently opens up. Scale bars = 200 nm. The flat pole is indicated by black bracket, and one coated vesicle is labeled CV. (b) Projection images of 3D surface rendering of the particle shown in panel a with different angles. The video of volume reconstruction with segmentation is shown in Movie S5. The tegument is in blue, the additional capsid layers are in dark green, the inner viral membrane is in light green, the flat pole is in orange, and the cisternae are in red.

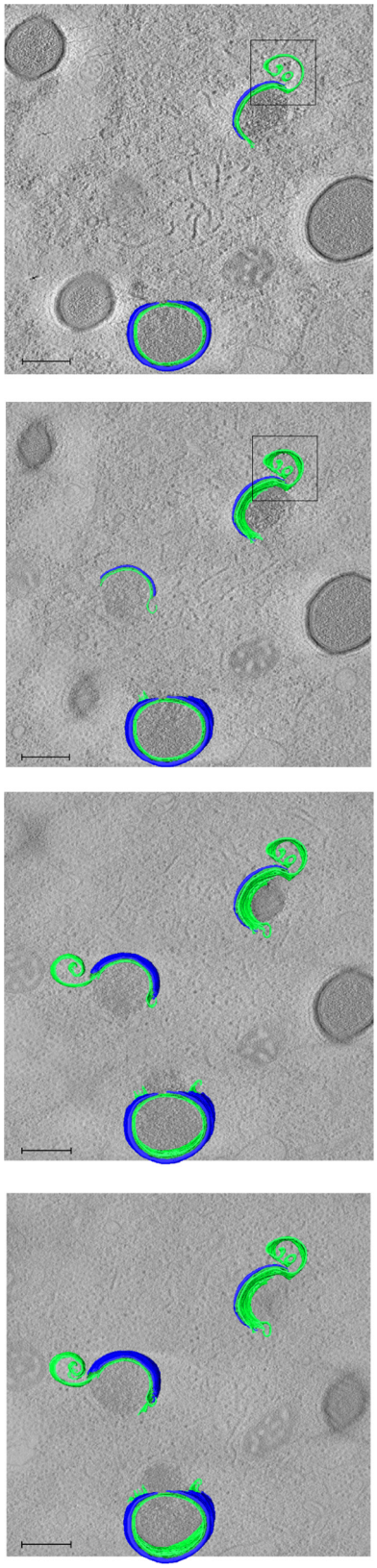


FIG 8 Elongation of the viral inner membrane is triggered by binding to and opening of vesicles. At 8 h postinfection, cells of *Acanthamoeba castellanii* were fixed and processed for transmission electron microscopy after cryo-fixation and freeze substitution. Slices 20, 40, 60, and 80 (from top to bottom) through the tomogram are shown. The video of the reconstructed volume with segmentation is shown in Movie S4. On the open curling end highlighted with a black square, a vesicle comes into contact with

(Continued on next page)

TABLE 2 Characteristics of virions observed by tomography at different stages of assembly[a]

Tomogram	Stage	End 1	End 2	Cisterna-like structure	Flat pole	No. of coated vesicles
1	Crescent	Open	Open	Present with ribosomes segregated	Observed	1
2	Crescent	Open	Open	Present with ribosomes segregated	Observed	2
3	Crescent	Open and connected to 2 vesicles	Open	Present with ribosomes segregated	Observed	0
4	Crescent	Open	Open	Present with ribosomes segregated	Observed	5
5	Crescent	Open	Open	Present with ribosomes segregated	Observed	5
6	Crescent	Open and connected to 2 vesicles	Open and connected to 2 vesicles	Present with ribosomes segregated	Observed	2
7	Crescent	Open and connected to 1 vesicle	Open	Present with ribosomes segregated	Not observed	0
8	Crescent	Open and connected to 3 vesicles	Open	Present with ribosomes segregated	Not observed	5
9	Crescent	Open and connected to 1 vesicle	Open	Present with ribosomes segregated	Not observed	10
10	Crescent	Open	Open	Present with ribosomes segregated	Not observed	5
11	Crescent	Open	Open	Present with ribosomes segregated	Not observed	0
12	Crescent	Open and connected to 1 vesicle	Open	Present with ribosomes segregated	Not observed	0
13	Crescent	Open	Open	Present with ribosomes segregated	Not observed	10
14	Crescent	Open and connected to 1 vesicle	Open	Present with ribosomes segregated	Not observed	10
15	Crescent	Open	Open	Present with ribosomes segregated	Not observed	5
16	Crescent	Open and connected to 1 vesicle	Open	Present with ribosomes segregated	Not observed	5
17	Closure of virions			Not observed	Not observed	1

[a]Characteristics of all virions imaged in 17 tomograms acquired through the course of our study are given depending on the stage of assembly of the virion. Crescents are defined as membrane structures that are composed of a membrane bended by the electron dense layer on its convex side and exclude (almost) a completed sphere. All viral crescents display two open ends that are sometimes found in close proximity and connected to vesicles; a cisterna-like structure is always associated with viral crescents but not with the one virion in the process of closure; and the flat pole is not always visible depending on the section plane. In the reconstructed volumes, coated vesicles are also often observed.

particles (37) suggests that they could play an important role in the virion assembly process and, in particular, in viral membrane acquisition, including recruitment of cell-derived vesicles, fusion, and rupture of membrane intermediates. Alternatively, they could also be involved in the initial fusion event necessary to initiate the infection, i.e., fusion of viral and cellular membranes prior to genome delivery. Furthermore, the major capsid protein (MCP; ml_347), which is only the 7th most abundant protein in the mature virions, might serve for scaffolding during virion assembly, as observed for VACV (44). Interestingly, maturation of VACV virions includes proteolysis of the MCP, a prerequisite to the formation of nonicosahedral infectious virions. A similar process could also apply to Mollivirus sibericum and explain the changes in particle morphology observed upon exit of the mollivirions in the extracellular medium. The spherical neosynthesized virions become more flexible and elongated after being released from the host cells and undergo a dramatic change at the flat pole to form the large depression that will open and fuse with the vacuole membrane during the next round of infection (37). Such changes in terms of particle dimensions, shape, and size of the flat pole could be the result of proteolysis among other means and might correspond to the last stage of mollivirion maturation.

DISCUSSION

Through the course of our study, protocols for sample preparation for 3D-EM analysis of *Acanthamoeba* spp., infected or not, have been optimized and are now

FIG 8 Legend (Continued)
the inner membrane. Scale bars = 200 nm. The tegument is in blue, additional capsid layers are in dark green, and the inner viral membrane is in light green. On the open curling end of a viral crescent highlighted with a black square, a vesicle comes into contact with the inner membrane.

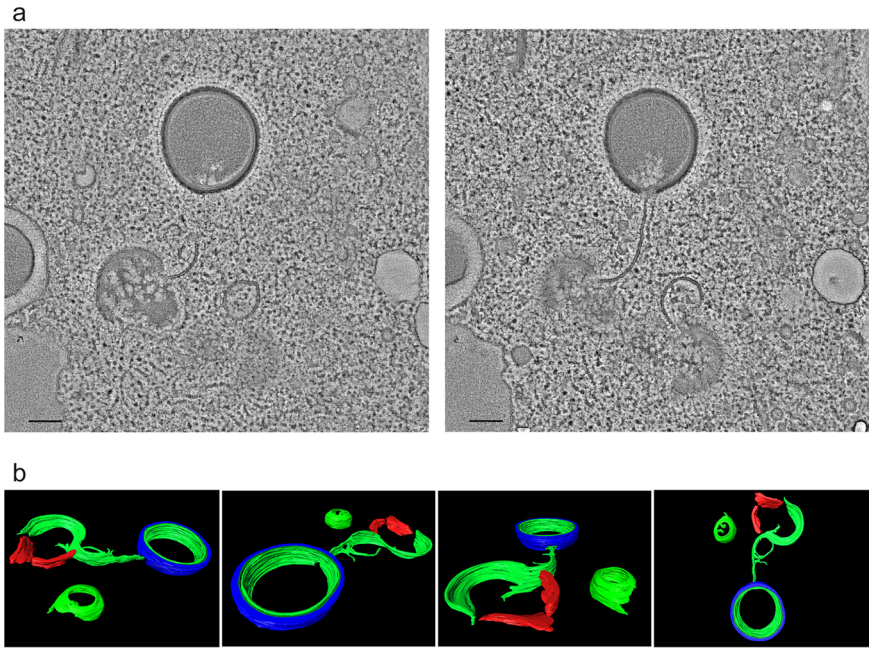


FIG 9 Coassembly of multiple particles of Mollivirus sibericum. At 8 h postinfection, cells of *Acanthamoeba castellanii* were fixed and processed for transmission electron tomography. (a) Slices 17 (left) and 80 (right) through the tomogram showing three particles being coassembled, including two connected through the inner viral membrane. Scale bars = 200 nm. (b) Projection images of 3D surface rendering of the particle shown in panel a with different angles. The video of volume reconstruction with segmentation is shown in Movie S7. The tegument is in blue, the additional capsid layers are in dark green, the inner viral membrane is in light green, and the cisternae are in red.

available. Our approach combines complementary EM techniques, i.e., focused-ion beam-scanning electron microscopy (FIB-SEM), electron tomography (ET), and immunolabeling, to provide an in-depth characterization of the infection cycle of Mollivirus sibericum. We believe that our experimental design, which focuses mainly on 3D analysis of virion assembly, will serve as a basis for future studies on other giant and large DNA viruses in general.

We confirm previous findings on the internalization of the virions into vacuoles (37). Infection proceeds through the formation of a large cytoplasmic structure dedicated to genome replication and virus assembly, the viral factory (VF) (20, 37). In contrast to APMV, the VF appears to contain viruses at different stages of assembly with no specific arrangement similar to VACV (18–20). We used FIB-SEM to provide insights into the overall rearrangement of the cellular cytoplasm upon virus infection (45). VFs occupy most of the cell interior, pushing organelles toward the cell periphery, and nuclei are also rarely observed at late stages of infection and may appear smaller, with mislocalized nucleoli at the periphery. In general, infection with giant and large DNA viruses has different impacts on the host nucleus, such as persistence for members of the *Mimiviridae* (19, 31, 46), *Pithoviridae* (36), and *Poxviridae* (47), disruption, e.g., *Pandoraviridae* (35), or even transient recruitment of nuclear factors as recently reported for *Marseilleviridae* (48). In the case of Mollivirus sibericum, the viral genome is initially delivered into the nucleus, and nuclear proteins are essential for early replication. At late stages, however, the function and fate of the nucleus remain ambiguous, and it can either be maintained in some cells or completely destroyed in others. During infection, mitochondria become smaller and elongated compared to control cells. Interestingly, it has been shown previously that several viruses, including poxviruses and asfarviruses, modulate mitochondrial metabolism toward fission and mitophagy to increase replication, attenuate apoptosis, and promote viral persistence (49).

With both transmission EM and scanning transmission EM, we observed that virion assembly of Mollivirus sibericum relies on the general strategy of vesicle recruitment,

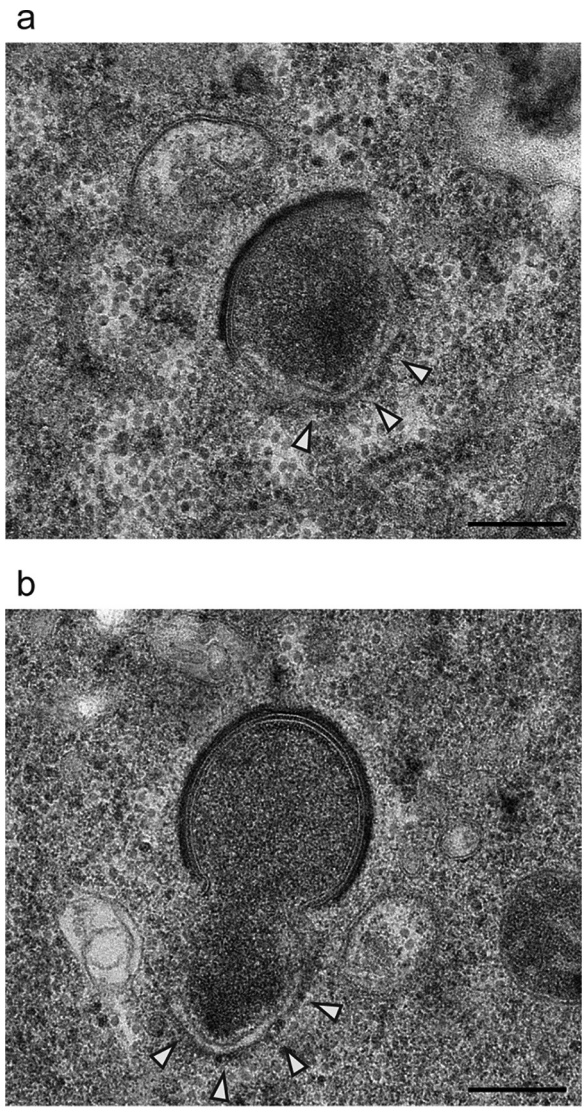


FIG 10 Viral crescents are associated with cisterna-like structures. At 8 h postinfection, cells of *Acanthamoeba castellanii* were fixed and processed for transmission electron microscopy after cryo-fixation and freeze substitution. The assembly of viral crescents early (a) or late (b) is associated with cisterna-like structures which are seemingly pushing the viroplasm inside the particle being assembled. Scale bars = 250 nm. The cisternae associated with viral crescents with segregated ribosomes are indicated by white arrowheads.

opening and shaping by capsid layers similar to all NCLDVs studied until now. However, membrane assembly in particular exhibits specific features, as follows: (i) assembly is initiated by a unique structure containing a flat pole positioned at the center of an open cisterna of unknown origin; (ii) the final packing of core proteins involves another cisterna-like element seemingly pushing core material into the particle being assembled; (iii) using DNA immunolabeling, we also reveal the genomic nature of the filamentous structures unique to Mollivirus sibericum (Fig. 14) (37). These DNA-associated filaments are abundant inside the VF; since virions are only labeled at the latest stages of assembly, we propose that DNA uptake takes place at the very final stage, similar to VACV and APMV (26, 29, 30, 42). Finally, for all the studied NCLDVs, the recruitment of open membrane intermediates in the VF was shown. For Mollivirus sibericum, however, the vesicles abundant in the VF were closed and seemingly opened only by contacting the curling ends of the growing viral membrane (27, 29, 30), suggesting that these membrane ends mediated their rupture.

The viral membrane is a key component of enveloped viruses and is necessary for

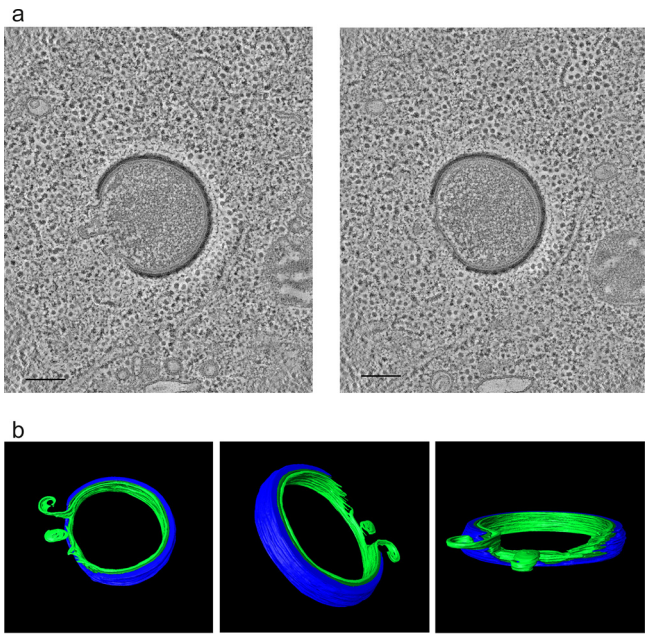


FIG 11 3D analysis of closure of particles during assembly of Mollivirus sibericum particles. At 8 h postinfection, cells of *Acanthamoeba castellanii* were fixed and processed for transmission electron tomography. (a) Slices 53 (left) and 85 (right) through the tomogram showing closure of the inner membrane of particles. Scale bars = 200 nm. (b) Projection images of 3D surface rendering of the particle shown in panel a with different angles. The video of volume reconstruction with segmentation is shown in Movie S6. The tegument is in blue, the additional capsid layers are in dark green, and the inner viral membrane is in light green. Due to the orientation of the particle in the tomogram, the flat-shaped pole is not visible and therefore not represented.

cytoplasmic delivery of the viral core upon fusion. Cellular membrane modified by viruses may also serve as an anchor for both nonstructural and structural viral proteins, driving virion replication and assembly, respectively (50). Although NCLDVs and giant viruses exhibit different virion morphotypes and infect a wide range of hosts, they all contain an internal membrane. This internal membrane can work as a support for additional structural components, guiding the assembly process of virions and playing a key role during disassembly to release the virion content including the viral genome into the cellular cytoplasm upon fusion. In light of our findings, we propose that the unconventional assembly of the viral inner membrane is a shared feature of NCLDVs and giant viruses, which involves the recruitment and opening of cell-derived vesicles. This general strategy could testify to a common origin or rather be the sign of a convergent evolution.

In the case of VACV and African swine fever virus, collective data show that the viral membrane is derived from the endoplasmic reticulum (22, 27, 51, 52), and it has also been proposed that the other giant APMV assembles an inner viral membrane from vesicles originating from the endoplasmic reticulum (30). The origin of the membrane of Mollivirus sibericum remains to be investigated. The lack of ribosomes on the cisterna initiator may suggest a role for the Golgi complex or the endoplasmic reticulum-Golgi interphase, which is consistent with the extensive recruitment of coated vesicles in the VF. Membrane elongation *per se* is mediated by uncoated vesicles of unknown origin, whereas at the final stage of core packaging, we observed a role for a closed cisternal element containing segregated ribosomes that could also be derived from the endoplasmic reticulum.

Taken as a whole, our analysis improves our understanding of the replication cycle of the complex virions of Mollivirus sibericum and in particular how it assembled in comparison with other virus-host systems. We could show that the particles contain at least one inner membrane covered by a thick coat and two additional capsid layers with a specific pole. However, the detailed organization of the internal structure and packed

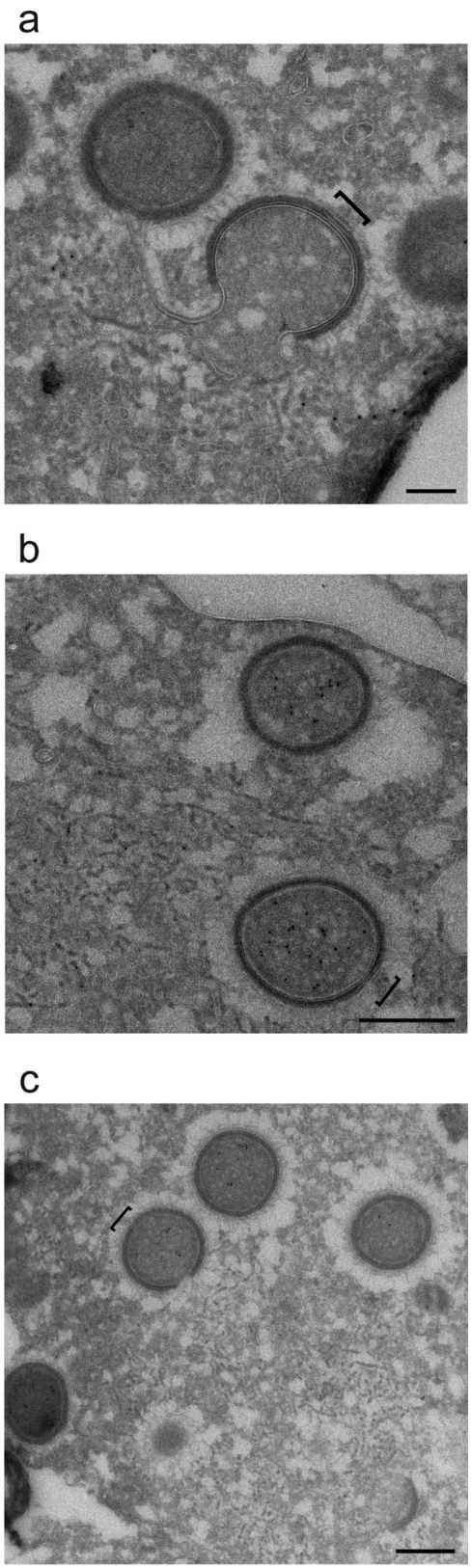


FIG 12 DNA packaging occurs at the end of the assembly of Mollivirus sibericum. At 8 h postinfection, cells of *Acanthamoeba castellanii* were fixed and processed for immunolabeling of thawed cryo-sections. DNA labeling is absent from half-assembled particles (a) and observed inside mature and associated with viral filamentous structures (b and c). Scale bars = 200 nm (a) and 500 nm (b and c). The flat pole is indicated by a black bracket.

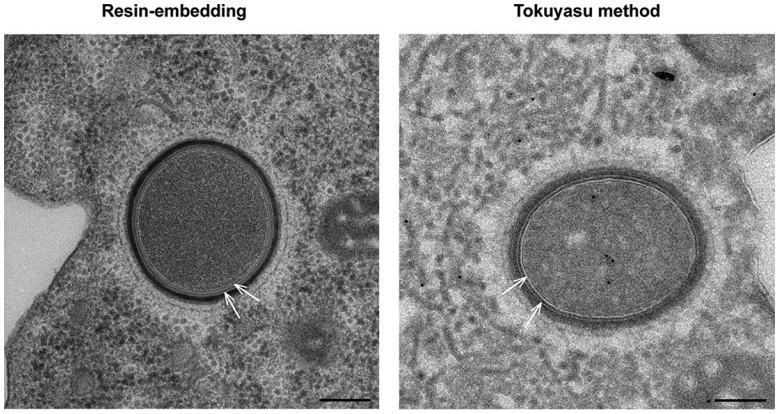


FIG 13 Comparison of virions of Mollivirus sibericum prepared by classical embedding or the Tokuyasu method. At 8 h postinfection, cells of *Acanthamoeba castellanii* were processed for transmission electron microscopy after either cryo-fixation and freeze-substitution (left) or fixation and immunolabeling of thawed cryo-sections (right). The white arrows indicate the appearance of the two capsid layers by the classical embedding method and in the Tokuyasu-prepared sample, which resemble membranes. Scale bars = 200 nm (A) and 250 nm (B).

material remains to be determined, and future studies should aim at resolving the structure of virions at high resolution, as was previously done for APMV (32, 43, 53, 54). Notably, we also report on cisternae as unique precursors of the viral inner membrane with open curling ends triggering fusion and rupture of vesicles for elongation. The flat

TABLE 3 Hypothetical membrane proteins and major capsid proteins in Mollivirus sibericum[a]

Protein no.	Rank in Mollivirus proteome	Prediction and/or no. of transmembrane domains (TM)	NCBI accession no.
ml_436	3	1	YP_009165402.1
ml_347	7	Major capsid protein related to NCLDVs	YP_009165313.1
ml_492	8	3	YP_009165458.1
ml_222	9	1	YP_009165188.1
ml_309	17	Lysosome-associated membrane glycoprotein, 1	YP_009165275.1
ml_361	20	1	YP_009165327.1
ml_499	21	Hydrolase, 1	YP_009165465.1
ml_355	26	2	YP_009165321.1
ml_331	33	Cell adhesion, 1	YP_009165297.1
ml_330	34	Perforin-related protein related pleurotolysin, 1	YP_009165296.1
ml_393	35	Antifreeze protein, 3	YP_009165359.1
ml_285	37	Integrin, 1	YP_009165251.1
ml_196	40	Hydrolase, 1	YP_009165162.1
ml_367	41	ABC transporter, 1	YP_009165333.1
ml_212	44	1	YP_009165178.1
ml_402	49	1	YP_009165368.1
ml_442	50	Receptor, 1	YP_009165408.1
ml_417	53	Fusion protein, 2	YP_009165383.1
ml_435	56	Sulfhydryl oxidase, 1	YP_009165401.1
ml_353	65	Toxin, 1	YP_009165319.1
ml_448	66	Vesicle-associated membrane protein, 2	YP_009165414.1
ml_333	67	Channel protein, 1	YP_009165299.1
ml_185	68	Vesicle associated protein, 1	YP_009165151.1
ml_201	72	Carboxylesterase, 1	YP_009165167.1
ml_399	78	Capsid protein related to poliovirus, 1	YP_009165365.1
ml_321	94	1	YP_009165287.1
ml_341	101	2	YP_009165307.1
ml_307	104	Chemosensor, 2	YP_009165273.1

[a]A homologue to the major capsid protein of NCLDV containing hydrophobic segments and predicted transmembrane proteins identified in the proteome of Mollivirus sibericum are listed with protein number, rank in the proteome, predictions along with the number of transmembrane domains, and accession numbers. Possible transmembrane segments were searched using the TMHMM v2.0 and Phobius servers. Functional and structural predictions were performed using the HHpred server.

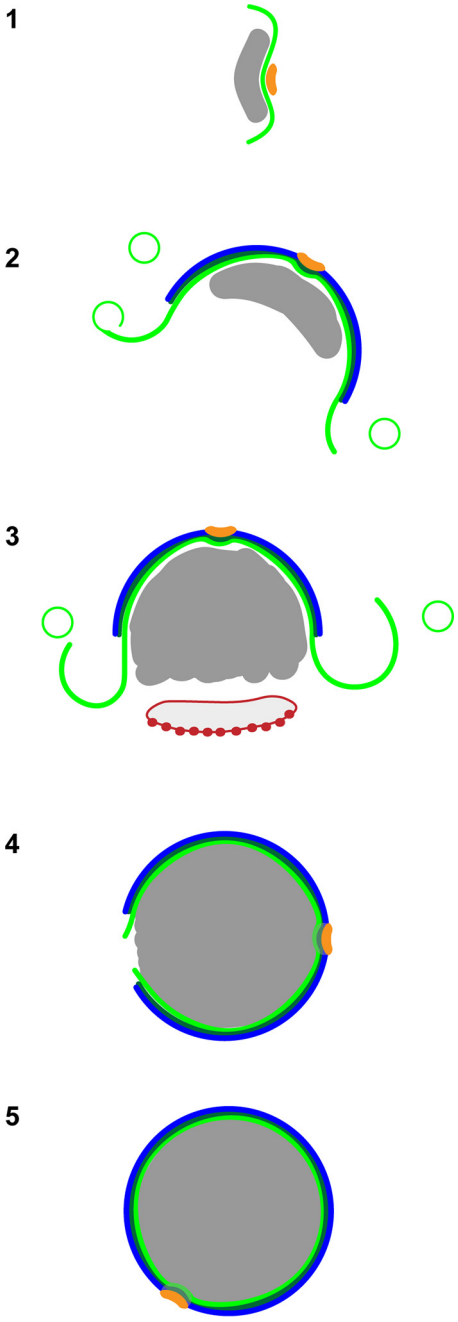


FIG 14 Schematic model of the assembly of Mollivirus sibericum. (1) Membrane assembly is initiated by an open cisternal element (light green) containing a flat pole (orange) at its center, and core proteins (gray) are recruited at the opposite side of the membrane. (2) From this pole, additional protein layers, the coat (light blue) and inner capsid layer (dark green), are recruited and shape the growing membrane into an open sphere or crescent. (3) The membrane grows by recruitment of vesicles that open upon contacting the open ends and final packaging of viral core proteins in crescents by a cisterna (red) located at the exact opposite of the flat pole; note that ribosomes are segregated on the side facing the cytoplasm. (4) Elongation of the membrane, packing of core proteins, and recruitment of the tegument layer result in an almost completed particle with a remaining opening located at the opposite side of the flat pole, likely required to take up the viral genome (not shown). (5) Closure of the internal membrane with removal of material in excess.

pole found at the center of the opened cisternae mediates the assembly of additional capsid layers and remains visible on mature particles, supporting an important role for virus entry and genome delivery as well. Such mechanism of membrane acquisition and virion assembly is, so far, unique to Mollivirus sibericum and appears as a new variation

of the general strategy of recruitment and opening of cell-derived vesicles for the inner membrane assembly of NCLDVs (22, 29, 31).

MATERIALS AND METHODS

Cell culture and virus infection. Conditions for cell culture and virus purification were as described previously (37). Briefly, *Acanthamoeba castellanii* (Douglas) Neff (ATCC 30010) was grown in PPYG medium [20 g of proteose peptone and 1 g of yeast extract per liter supplemented with 4 mM $MgSO_4$, 0.4 mM $CaCl_2$, 50 μM $Fe(NH_4)_2(SO_4)_2$, 2.5 mM Na_2HPO_4, 2.5 mM KH_2PO_4, and 100 μM glucose] at 32°C and infected with Mollivirus sibericum at a multiplicity of infection of 10 particles/cell. Titration of particles was carried out with a Malassez counting chamber after staining with toluidine blue.

Sample preparation for electron microscopy. Control and infected cells at 3 and 8 h postinfection were chemically fixed in 2 steps. First, cells were incubated for 1 h with 4% paraformaldehyde (EMS, USA) and 0.1% glutaraldehyde (Merck, Germany) in PHEM buffer [60 mM piperazine-*N*,*N*′-bis(2-ethanesulfonic acid) (PIPES), 25 mM HEPES, 10 mM EGTA, and 4 mM $MgSO_4$ (pH 6.9)]. Second, postfixation was done overnight at 4°C with 2.5% glutaraldehyde in PHEM buffer.

Cells were washed with PHEM buffer, postfixed for 1 h with 1% OsO_4 (Merck, Germany) and 1.5% potassium ferricyanide (Merck, Germany), incubated with 2% uranyl acetate (Merck) in H_2O for 1 h, and dehydrated in successive steps of 15 min in 25, 50, 75, 95, and 100% ethanol. The samples were subsequently infiltrated with Agar 100 epoxy resin (Agar Scientific, United Kingdom). Alternatively, cells were washed four times in phosphate-buffered saline (PBS), incubated with 50 mM ammonium chloride for 10 min, pelleted (3,000 $\times$ g for 30 min), and resuspended in a minimal amount of 20% bovine serum albumin (BSA) in PPYG. Cell pastes were transferred into a lecithin-coated sample holder type A (EMS, USA) and cryo-fixed with a high-pressure freezing machine (HPM 010; Bal-Tec Products, USA). Following cryo-fixation, the samples were freeze substituted with 0.2% OsO_4, 0.1% uranyl acetate, 2% H_2O, and 4% methanol in acetone (EMS), according to the following schedule: −90°C for 1 h, 1.25°C/h for 8 h, 3.8°C/h for 8 h, 12.5°C/h for 4 h, and 0°C for 1 h. The samples were rinsed four times in acetone and slowly infiltrated with Agar 100 epoxy resin.

After polymerization, samples were first processed on an Ultracut R microtome (Leica, Austria) to prepare 70 nm-thin sections that were collected on Formvar-coated copper grids (EMS, Washington, PA, USA) and poststained with 4% uranyl acetate for 45 min and Reynold's lead citrate (Delta Microscopies, France) for 5 min. The grids were analyzed in a Tecnai T12 transmission electron microscope (Thermo Fisher Scientific, USA) operated at 120 kV and equipped with a Gatan BM Ultra scan (Gatan, Inc., USA).

Focused-ion beam scanning electron microscopy. For focused-ion beam scanning electron microscopy, embedded samples after cryo-fixation and freeze substitution were cut with an Ultra R microtome to obtain a flat surface. The sample was subsequently placed on a pin stub (Agar Scientific, United Kingdom) and recovered with silver paint (Agar Scientific) except on the flat surface, which was coated with a 20 nm-thick layer of gold-palladium in an Ion Beam Coater (Gatan, Inc., USA) to avoid a charge effect. The 3D imaging was done with an Auriga FIB-SEM (Zeiss, Germany). Another protective layer of platinum of 1 μm was deposited on the surface of the volume of interest using ion beam-assisted deposition with 30 kV acceleration potential. The cross-section was milled using a 10-nA ion beam current. The surface obtained was then polished using this time a 2-nA ion beam current. Sections during acquisition were made using a 500-μA ion beam current. SEM images were acquired at 10-nm pixel size for a picture size of 2,048 with 2 keV acceleration voltage and an aperture of 30 μm with the backscattered electron detector. Alignment of the acquired stack of images was done using ImageJ and all the segmentations and measurements with Amira 6.3 (Thermo Fisher Scientific, USA).

Electron tomography. For transmission electron tomography, embedded samples after cryo-fixation and freeze substitution were cut into serial 200-nm-thick sections with an Ultracut R microtome and collected on Formvar-coated copper parallel bar grids (EMS, USA). Protein A-gold particles of 15 nm (UMC-Utrecht, The Netherlands) were applied on both sides of the sections, and poststaining was performed as described above, if necessary. Dual-axis electron tomograms were collected on a F20 transmission electron microscope (Thermo Fisher Scientific, USA) operated at 200 kV and equipped with a Gatan US4000. Tomographic tilt ranges were collected using the SerialEM program (55), typically from −60° to +60°, with an angular increment of 1°. For scanning transmission electron tomography, embedded samples were cut into serial 750-nm-thick sections and collected on Formvar-coated copper bar grids. Protein A-gold particles of 15 nm were applied on both sides of the sections, and poststaining was performed as described above, if necessary. Single-axis tomograms were collected on a Jeol JEM-2100F microscope operated at 200 kV and a bright-field detector (Jeol, Japan) at a size of 1,024 by 1,024 pixels. Continuous tilt series were collected typically from −72° to 72° with an angular increment of 1.5°.

Image analysis. For ET, tilted images were aligned using gold fiducials as a reference, and weighted back projection (WBP) reconstruction was carried out in the IMOD 4.9 software package. Manual segmentation, 3D surface rendering, and videos were prepared with Amira 6.3. Measurements on FIB-SEM data were done on the segmented objects with the module "Label analysis" from Amira 6.3. For each object, the following measurements were done: (i) length, maximum of the Feret diameters; (ii) width, minimum of the Feret diameters; (iii) breadth, largest distance between two parallel lines touching the object without intersecting it and lying in a plane orthogonal to the maximum Feret diameter; (iv) thickness, the largest segment that touches the object by its endpoints and lying in a plane orthogonal to the maximum Feret diameter and orthogonal to the breadth diameter; and (v) sphericity, the ratio between the length and breadth of an object with a ratio of 1 corresponding to a sphere.

Immunolabeling of thawed cryo-sections. Control and infected cells at 8 h postinfection were chemically fixed with 4% paraformaldehyde and 0.1% glutaraldehyde in PHEM buffer for 1 h and kept in

4% paraformaldehyde in PHEM until processed for Tokuyasu cryo-sectioning, as previously described (56). In brief, thawed cryo-sections were blocked with 1% BSA in PBS, immunolabeled with anti-DNA antibody clone 3519 (Abcam, United Kingdom) diluted to 1/1,000 and 10 nm protein A-gold prior to contrasting with uranyl acetate in methylcellulose (1:9). Once dried, sections were observed on a Tecnai T12 transmission electron microscope operated at 120 kV and equipped with a Gatan BM UltraScan. At least three independent experiments were performed for each condition.

Virion protein sequence analyses. Possible transmembrane segments were searched using the TMHMM v2.0 (http://www.cbs.dtu.dk/services/TMHMM/) (57) and Phobius (http://phobius.sbc.su.se/) servers (58). Functional and structural predictions were performed using the HHpred server (https://toolkit.tuebingen.mpg.de/#/tools/hhpred) (59).

SUPPLEMENTAL MATERIAL

Supplemental material for this article may be found at https://doi.org/10.1128/JVI.00388-19.

SUPPLEMENTAL FILE 1, PDF file, 0.1 MB.
MOVIE S1, MP4 file, 9.8 MB.
MOVIE S2, MP4 file, 19.5 MB.
MOVIE S3, MP4 file, 11.9 MB.
MOVIE S4, MP4 file, 11.1 MB.
MOVIE S5, MP4 file, 15.5 MB.
MOVIE S6, MP4 file, 16.2 MB.
MOVIE S7, MP4 file, 17.3 MB.

ACKNOWLEDGMENTS

The Ultrastructural Bio-Imaging Facility of the Institut Pasteur, Paris, France, directed by J. Krijnse-Locker, is a member of the national infrastructure France-BioImaging and supported by the French National Research Agency (ANR-10-INBS-04). The laboratory of C. Abergel is supported by the Fondation Bettencourt-Schueller (OTP51251) and by a DGA-MRIS scholarship (201760003). E. R. Quemin is currently funded by the Alexander von Humboldt Foundation (FRA 1200789 HFST-P).

We thank all the members of UBI, in particular, Martin Sachse for useful discussions, technical assistance, and critical readings of the manuscript. We also thank Elsa Garcin for useful discussions and critical editing of the manuscript.

REFERENCES

1. Schmick M, Bastiaens P. 2014. The interdependence of membrane shape and cellular signal processing. Cell 156:1132–1138. https://doi.org/10.1016/j.cell.2014.02.007.
2. Harayama T, Riezman H. 2018. Understanding the diversity of membrane lipid composition. Nat Rev Mol Cell Biol 19:281–296. https://doi.org/10.1038/nrm.2017.138.
3. Yamauchi Y, Helenius A. 2013. Virus entry at a glance. J Cell Sci 126:1289–1295. https://doi.org/10.1242/jcs.119685.
4. Reid CR, Airo AM, Hobman TC. 2015. The virus-host interplay: biogenesis of +RNA replication complexes. Viruses 7:4385–4413. https://doi.org/10.3390/v7082825.
5. Rossman JS, Lamb RA. 2011. Influenza virus assembly and budding. Virology 411:229–236. https://doi.org/10.1016/j.virol.2010.12.003.
6. Votteler J, Sundquist WI. 2013. Virus budding and the ESCRT pathway. Cell Host Microbe 14:232–241. https://doi.org/10.1016/j.chom.2013.08.012.
7. Münz C. 2017. Autophagy proteins in viral exocytosis and anti-viral immune responses. Viruses 9:E288. https://doi.org/10.3390/v9100288.
8. Iyer LM, Aravind L, Koonin EV. 2001. Common origin of four diverse families of large eukaryotic DNA viruses. J Virol 75:11720–11734. https://doi.org/10.1128/JVI.75.23.11720-11734.2001.
9. Iyer LM, Balaji S, Koonin EV, Aravind L. 2006. Evolutionary genomics of nucleo-cytoplasmic large DNA viruses. Virus Res 117:156–184. https://doi.org/10.1016/j.virusres.2006.01.009.
10. Gallot-Lavallée L, Blanc G, Claverie JM. 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.
11. Arslan D, Legendre M, Seltzer V, Abergel C, Claverie JM. 2011. Distant Mimivirus relative with a larger genome highlights the fundamental features of Megaviridae. Proc Natl Acad Sci U S A 108:17486–17491. https://doi.org/10.1073/pnas.1110889108.
12. Legendre M, Audic S, Poirot O, Hingamp P, Seltzer V, Byrne D, Lartigue A, Lescot M, Bernadac A, Poulain J, Abergel C, Claverie JM. 2010. mRNA deep sequencing reveals 75 new genes and a complex transcriptional landscape in Mimivirus. Genome Res 20:664–674. https://doi.org/10.1101/gr.102582.109.
13. Suzan-Monti M, La Scola B, Barrassi L, Espinosa L, Raoult D. 2007. Ultrastructural characterization of the giant volcano-like virus factory of Acanthamoeba polyphaga Mimivirus. PLoS One 2:e328. https://doi.org/10.1371/journal.pone.0000328.
14. Renesto P, Abergel C, Decloquement P, Moinier D, Azza S, Ogata H, Fourquet P, Gorvel JP, Claverie JM. 2006. Mimivirus giant particles incorporate a large fraction of anonymous and unique gene products. J Virol 80:11678–11685. https://doi.org/10.1128/JVI.00940-06.
15. Rojo G, Garcia-Beato R, Vinuela E, Salas ML, Salas J. 1999. Replication of African swine fever virus DNA in infected cells. Virology 257:524–536. https://doi.org/10.1006/viro.1999.9704.
16. Van Etten JL, Burbank DE, Joshi J, Meints RH. 1984. DNA synthesis in a Chlorella-like alga following infection with the virus PBCV-1. Virology 134:443–449. https://doi.org/10.1016/0042-6822(84)90311-8.
17. Rohrmann G, Moss B. 1985. Transcription of vaccinia virus early genes by a template-dependent soluble extract of purified virions. J Virol 56:349–355.
18. Risco C, de Castro IF, Sanz-Sanchez L, Narayan K, Grandinetti G, Subramaniam S. 2014. Three-dimensional imaging of viral infections. Annu Rev Virol 1:453–473. https://doi.org/10.1146/annurev-virology-031413-085351.
19. Mutsafi Y, Fridmann-Sirkis Y, Milrot E, Hevroni L, Minsky A. 2014. Infec-

tion cycles of large DNA viruses: emerging themes and underlying questions. Virology 466-467:3–14. https://doi.org/10.1016/j.virol.2014.05.037.
20. de Castro IF, Volonte L, Risco C. 2013. Virus factories: biogenesis and structural design. Cell Microbiol 15:24–34. https://doi.org/10.1111/cmi.12029.
21. Chlanda P, Carbajal MA, Cyrklaff M, Griffiths G, Krijnse-Locker J. 2009. Membrane rupture generates single open membrane sheets during vaccinia virus assembly. Cell Host Microbe 6:81–90. https://doi.org/10.1016/j.chom.2009.05.021.
22. Krijnse Locker J, Chlanda P, Sachsenheimer T, Brügger B. 2013. Poxvirus membrane biogenesis: rupture not disruption. Cell Microbiol 15:190–199. https://doi.org/10.1111/cmi.12072.
23. Moss B. 2018. Origin of the poxviral membrane: a 50-year-old riddle. PLoS Pathog 14:e1007002. https://doi.org/10.1371/journal.ppat.1007002.
24. Moss B. 2015. Poxvirus membrane biogenesis. Virology 479–480:619–626. https://doi.org/10.1016/j.virol.2015.02.003.
25. Liu L, Cooper T, Howley PM, Hayball JD. 2014. From crescent to mature virion: vaccinia virus assembly and maturation. Viruses 6:3787–3808. https://doi.org/10.3390/v6103787.
26. Condit RC, Moussatche N, Traktman P. 2006. In a nutshell: structure and assembly of the vaccinia virion. Adv Virus Res 66:31–124. https://doi.org/10.1016/S0065-3527(06)66002-8.
27. Suárez C, Andres G, Kolovou A, Hoppe S, Salas ML, Walther P, Krijnse Locker J. 2015. African swine fever virus assembles a single membrane derived from rupture of the endoplasmic reticulum. Cell Microbiol 17:1683–1698. https://doi.org/10.1111/cmi.12468.
28. Milrot E, Mutsafi Y, Fridmann-Sirkis Y, Shimoni E, Rechav K, Gurnon JR, Van Etten JL, Minsky A. 2016. Virus-host interactions: insights from the replication cycle of the large Paramecium bursaria chlorella virus. Cell Microbiol 18:3–16. https://doi.org/10.1111/cmi.12486.
29. Suárez C, Welsch S, Chlanda P, Hagen W, Hoppe S, Kolovou A, Pagnier I, Raoult D, Krijnse Locker J. 2013. Open membranes are the precursors for assembly of large DNA viruses. Cell Microbiol 15:1883–1895. https://doi.org/10.1111/cmi.12156.
30. Mutsafi Y, Shimoni E, Shimon A, Minsky A. 2013. Membrane assembly during the infection cycle of the giant Mimivirus. PLoS Pathog 9:e1003367. https://doi.org/10.1371/journal.ppat.1003367.
31. Mutsafi Y, Zauberman N, Sabanay I, Minsky A. 2010. Vaccinia-like cytoplasmic replication of the giant Mimivirus. Proc Natl Acad Sci U S A 107:5978–5982. https://doi.org/10.1073/pnas.0912737107.
32. Kuznetsov YG, Xiao C, Sun S, Raoult D, Rossmann M, McPherson A. 2010. Atomic force microscopy investigation of the giant mimivirus. Virology 404:127–137. https://doi.org/10.1016/j.virol.2010.05.007.
33. Abrahão J, Silva L, Silva LS, Khalil JYB, Rodrigues R, Arantes T, Assis F, Boratto P, Andrade M, Kroon EG, Ribeiro B, Bergier I, Seligmann H, Ghigo E, Colson P, Levasseur A, Kroemer G, Raoult D, La Scola B. 2018. Tailed giant Tupanvirus possesses the most complete translational apparatus of the known virosphere. Nat Commun 9:749. https://doi.org/10.1038/s41467-018-03168-1.
34. Raoult D, Audic S, Robert C, Abergel C, Renesto P, Ogata H, La Scola B, Suzan M, Claverie JM. 2004. The 1.2-megabase genome sequence of Mimivirus. Science 306:1344–1350. https://doi.org/10.1126/science.1101485.
35. Philippe N, Legendre M, Doutre G, Coute Y, Poirot O, Lescot M, Arslan D, Seltzer V, Bertaux L, Bruley C, Garin J, Claverie JM, 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.
36. Legendre M, Bartoli J, Shmakova L, Jeudy S, Labadie K, Adrait A, Lescot M, Poirot O, Bertaux L, Bruley C, Coute Y, Rivkina E, Abergel C, Claverie JM. 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. https://doi.org/10.1073/pnas.1320670111.
37. Legendre M, Lartigue A, Bertaux L, Jeudy S, Bartoli J, Lescot M, Alempic JM, Ramus C, Bruley C, Labadie K, Shmakova L, Rivkina E, Coute Y, Abergel C, Claverie JM. 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–E5335. https://doi.org/10.1073/pnas.1510795112.
38. Koonin EV, Yutin N. 2018. Multiple evolutionary origins of giant viruses. F1000Res 7:F1000 Faculty Rev-1840. https://doi.org/10.12688/f1000research.16248.1.
39. Colson P, La Scola B, Levasseur A, Caetano-Anollés G, Raoult D. 2017. Mimivirus: leading the way in the discovery of giant viruses of amoebae. Nat Rev Microbiol 15:243–254. https://doi.org/10.1038/nrmicro.2016.197.
40. Abergel C, Legendre M, Claverie JM. 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.
41. Miller S, Krijnse-Locker J. 2008. Modification of intracellular membrane structures for virus replication. Nat Rev Microbiol 6:363–374. https://doi.org/10.1038/nrmicro1890.
42. Roberts KL, Smith GL. 2008. Vaccinia virus morphogenesis and dissemination. Trends Microbiol 16:472–479. https://doi.org/10.1016/j.tim.2008.07.009.
43. Kaelber JT, Hryc CF, Chiu W. 2017. Electron cryomicroscopy of viruses at near-atomic resolutions. Annu Rev Virol 4:287–308. https://doi.org/10.1146/annurev-virology-101416-041921.
44. Hyun JK, Accurso C, Hijnen M, Schult P, Pettikiriarachchi A, Mitra AK, Coulibaly F. 2011. Membrane remodeling by the double-barrel scaffolding protein of poxvirus. PLoS Pathog 7:e1002239. https://doi.org/10.1371/journal.ppat.1002239.
45. Narayan K, Subramaniam S. 2015. Focused ion beams in biology. Nat Methods 12:1021–1031. https://doi.org/10.1038/nmeth.3623.
46. Claverie JM, Abergel C. 2009. Mimivirus and its virophage. Annu Rev Genet 43:49–66. https://doi.org/10.1146/annurev-genet-102108-134255.
47. Carter GC, Law M, Hollinshead M, Smith GL. 2005. Entry of the vaccinia virus intracellular mature virion and its interactions with glycosaminoglycans. J Gen Virol 86:1279–1290. https://doi.org/10.1099/vir.0.80831-0.
48. Fabre E, Jeudy S, Santini S, Legendre M, Trauchessec M, Coute Y, Claverie JM, Abergel C. 2017. Noumeavirus replication relies on a transient remote control of the host nucleus. Nat Commun 8:15087. https://doi.org/10.1038/ncomms15087.
49. Khan M, Syed GH, Kim SJ, Siddiqui A. 2015. Mitochondrial dynamics and viral infections: a close nexus. Biochim Biophys Acta 1853:2822–2833. https://doi.org/10.1016/j.bbamcr.2014.12.040.
50. Perlmutter JD, Hagan MF. 2015. Mechanisms of virus assembly. Annu Rev Phys Chem 66:217–239. https://doi.org/10.1146/annurev-physchem-040214-121637.
51. Weisberg AS, Maruri-Avidal L, Bisht H, Hansen BT, Schwartz CL, Fischer ER, Meng X, Xiang Y, Moss B. 2017. Enigmatic origin of the poxvirus membrane from the endoplasmic reticulum shown by 3D imaging of vaccinia virus assembly mutants. Proc Natl Acad Sci U S A 114:E11001–E11009. https://doi.org/10.1073/pnas.1716255114.
52. Sodeik B, Krijnse-Locker J. 2002. Assembly of vaccinia virus revisited: de novo membrane synthesis or acquisition from the host? Trends Microbiol 10:15–24. https://doi.org/10.1016/S0966-842X(01)02256-9.
53. Xiao C, Rossmann MG. 2011. Structures of giant icosahedral eukaryotic dsDNA viruses. Curr Opin Virol 1:101–109. https://doi.org/10.1016/j.coviro.2011.06.005.
54. Ekeberg T, Svenda M, Abergel C, Maia FR, Seltzer V, Claverie JM, Hantke M, Jonsson O, Nettelblad C, van der Schot G, Liang M, DePonte DP, Barty A, Seibert MM, Iwan B, Andersson I, Loh ND, Martin AV, Chapman H, Bostedt C, Bozek JD, Ferguson KR, Krzywinski J, Epp SW, Rolles D, Rudenko A, Hartmann R, Kimmel N, Hajdu J. 2015. Three-dimensional reconstruction of the giant mimivirus particle with an x-ray free-electron laser. Phys Rev Lett 114:098102. https://doi.org/10.1103/PhysRevLett.114.098102.
55. Mastronarde DN. 2005. Automated electron microscope tomography using robust prediction of specimen movements. J Struct Biol 152:36–51. https://doi.org/10.1016/j.jsb.2005.07.007.
56. de Castro Martin IF, Fournier G, Sachse M, Pizarro-Cerda J, Risco C, Naffakh N. 2017. Influenza virus genome reaches the plasma membrane via a modified endoplasmic reticulum and Rab11-dependent vesicles. Nat Commun 8:1396. https://doi.org/10.1038/s41467-017-01557-6.
57. Krogh A, Larsson B, von Heijne G, Sonnhammer EL. 2001. Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. J Mol Biol 305:567–580. https://doi.org/10.1006/jmbi.2000.4315.
58. Käll L, Krogh A, Sonnhammer EL. 2004. A combined transmembrane topology and signal peptide prediction method. J Mol Biol 338:1027–1036. https://doi.org/10.1016/j.jmb.2004.03.016.
59. Zimmermann L, Stephens A, Nam SZ, Rau D, Kubler J, Lozajic M, Gabler F, Soding J, Lupas AN, Alva V. 2018. A completely reimplemented MPI Bioinformatics toolkit with a new HHpred server at its core. J Mol Biol 430:2237–2243. https://doi.org/10.1016/j.jmb.2017.12.007.

Mollivirus kamchatka, représentant moderne des *Molliviridae,* vers une meilleure compréhension de l'histoire évolutive des *Molliviridae*

Extraction d'ADN, séquençage et assemblage

L'extraction d'ADN génomique de particules purifiées de *M. kamchatka* a permis de récupérer une quantité de 6,79 µg d'ADN, la purification sur colonne de silice a permis d'obtenir un ratio d'absorbance 260/280 de 1,86 et 230/260 de 1,62. Le séquençage de type Nanopore a généré plus de 1,1 millions de séquences d'une taille moyenne de 6,8 kpb. Le séquençage par méthode Illumina HiSeq 2500 rapide a permis de générer 4,5 millions de séquences en sens direct et inverse avec 84% des bases ayant au minimum 99,9% de chances d'être correctes (84% > Q30). L'assemblage des lectures Nanopore et Illumina par Spades a permis de générer 951 contigs dont 10 d'une taille supérieure à 10 kpb. Pour discriminer ces contigs nous avons calculé leur couverture et leur pourcentage en GC (GC%). Nous avons déterminé que le plus grand contig, d'une taille de 648864 pb, d'une couverture réelle moyenne de 7000X était le génome de *M. kamchatka.* Les contigs 2, 3, 5, 6, 7, 8, 9, 10 d'une taille totale de 351811 pb, d'une couverture de 2X et d'un GC% de 43 correspondent à un génome quasi-complet de *Marseilleviridae.* L'alignement par BLASTn du contig 2 (133366 pb) donne une similarité de 96% avec le génome de *Insectomime virus*. Le contig 5, d'une taille de 41330 pb, d'une couverture de 5X et d'un GC% de 29 correspond quant à lui au génome mitochondrial complet d'*Acanthamoeba castellanii.* La présence du génome mitochondriale complet d'amibe (99% d'identité par BLASTn) n'a pas fait l'objet d'études approfondies mais soulève des questions quant à la capacité de la capside à embarquer du matériel génétique exogène, comme déjà constaté chez les *Poxviridae* [104] (Figure 45).

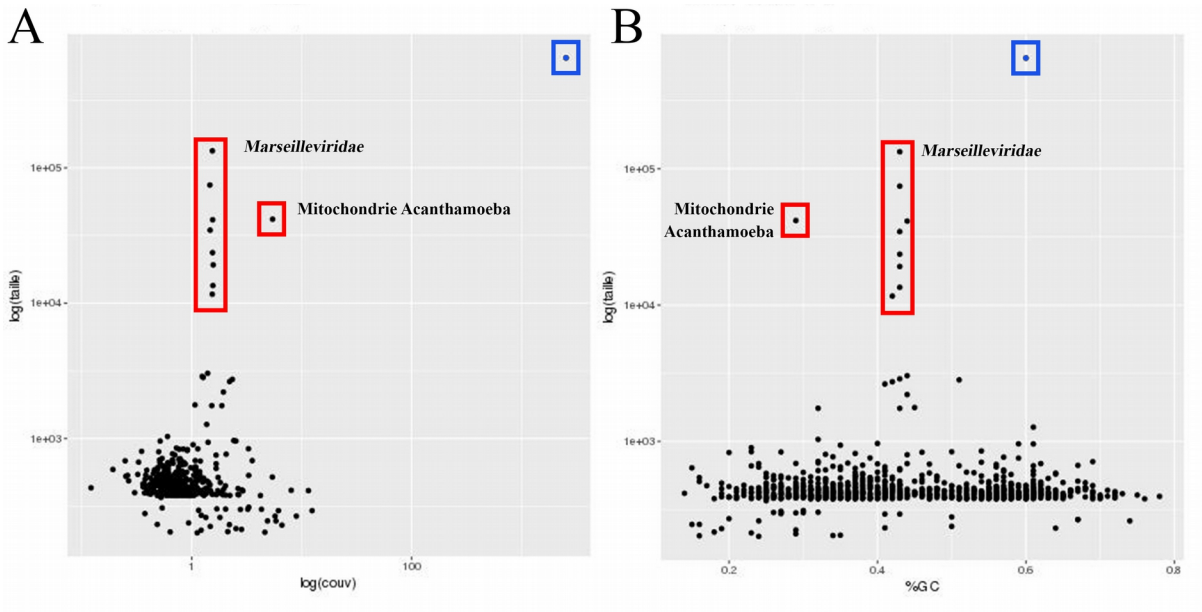


Figure 45 : Graphiques présentant les contig de plus de 200 pb. (A) La distribution de la couverture en fonction de la taille des contig montre la présence de 3 groupes de contig distincts. Par BLASTp M. kamchatka a été identifié (bleu) ainsi que deux autres organismes. (B) La distribution du GC% en fonction de la taille des contig confirme la présence de 3 groupes de contig distincts. En bleu le contig associé à M. kamchatka.

La couverture des lectures calculée à partir de l'alignement des lectures Nanopore contre le contig correspondant au génome de *M. kamchatka* est homogène à l'exception d'une région de 10 kpb en position 3' du génome pour laquelle la couverture est double. Pour savoir si ce doublement de couverture est lié à une circularisation du génome, une digestion de l'ADN génomique par les enzymes AseI et ApaI suivie d'une migration par électrophorèse en champ pulsé a été faite. Les résultats indiquent qu'il y a autant de fragments que de sites de coupure, démontrant la linéarité des génomes de mollivirus et excluant, de fait, que la présence de séquences répétées soit liée à une circularisation du génome. Il a été démontré chez *M. sibericum* que les extrémités du génome étaient flanquées de régions inversées et répétées d'une longueur de 10 kpb. On peut donc supposer que la structure du génome de *M. kamchatka* est similaire. La taille du génome est donc de 648864 ± 10000 pb en comptant les zones répétées avec un contenu en GC (GC%) de 60, similaire à *M. sibericum* (Figure 46).

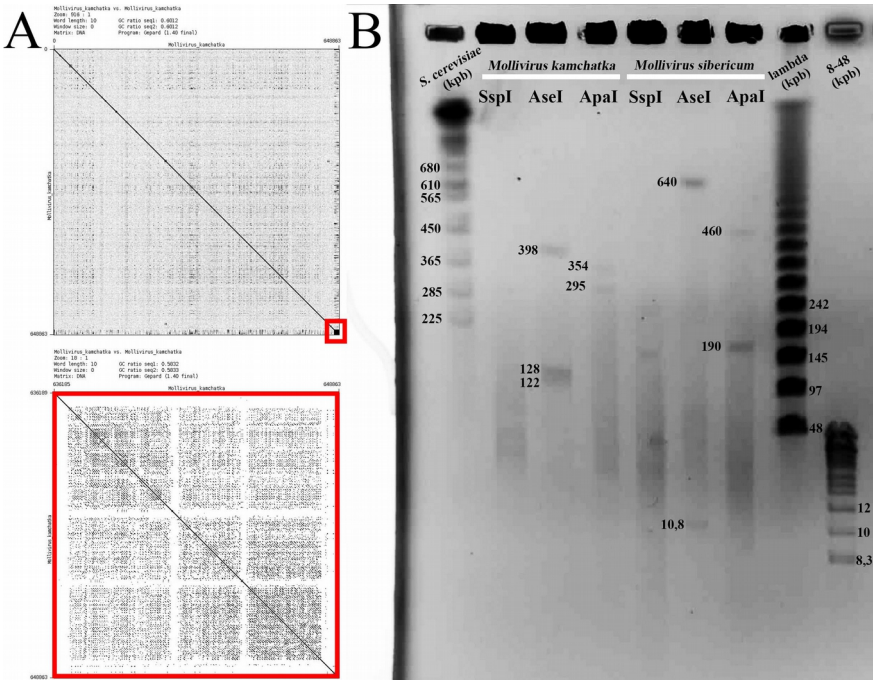


Figure 46 : Structure du génome de M. kamchatka. (A) Dotplot entre M. kamchatka (en abscisse) et M. sibericum (en ordonné), on observe une colinéarité entre les deux génomes à l'exception des régions terminales (extrémité 3' en rouge). Ces zones répétée flanquent à priori le génome de M. kamchatka. (B) Résultat de migration d'ADN génomique digéré de M. kamchatka sur gel par PFGE. Le nombre de fragments étant supérieur de 1 au nombre de sites de coupure, les génomes de mollivirus sont linéaires.

Annotation des génomes de Mollivirus sibericum et Mollivirus kamchatka

Une nouvelle annotation du génome de *Mollivirus sibericum* a été faite en utilisant des données de transcriptomique récemment acquises pour permettre de mieux supporter les prédictions de gènes et de détecter la présence d'introns. Les résultats d'annotation *ab inito* sont les suivants : 489 prédictions avec Augustus, 197 avec exonerate et 483 avec GenMark-ES. La conciliation de ces prédictions par EVM a permis d'obtenir 437 prédictions de gènes. De plus, l'outil Trinity d'assemblage *de novo* de données de transcriptome a permis l'obtention de 1899 transcrits. Ces transcripts ont été alignés contre les prédictions *ab initio* d'EVM grâce au logiciel PASA. Une dernière étape de curation manuelle, via l'outil de visualisation de génome WebAppollo, compilant données *ab initio* et données de transcriptome, a permis d'annoter 495 gènes d'une taille allant de 51 à 2171 acides aminés ainsi que leurs UTR.

Une stratégie similaire a été adoptée pour annoter le génome de *M. kamchatka*. La compilation de ces données et la curation manuelle sur WebAppollo ont permis l'identification de 485 protéines d'une taille allant de 57 à 2176 acides aminés. La suite de ce manuscrit tentera

d'explorer l'histoire évolutive de ces deux virus.

Caractéristiques protéomiques de Mollivirus kamchatka et transfert de gènes horizontaux

Parmi les 485 protéines prédites de *M. kamchatka*, 291 (60.4 % des gènes) n'ont pas d'homologues dans les bases de données en dehors de *M. sibericum*. Cette proportion d'ORFan est caractéristique des différentes familles de virus géants. Des 190 protéines prédites comme ayant au moins un homologue par BLASTp (*E*-valeur <10^{-5}) on retrouve 96 protéines homologues à d'autres protéines virales, et en particulier 78 protéines ayant leur homologue le plus proche chez pandoravirus (16% des gènes). On trouve aussi 75 protéines ayant un homologue chez un organisme eucaryote dont 51 (11% des gènes) chez *A. castellanii,* 17 protéines homologues à des protéines bactériennes et 2 protéines homologues à des protéines d'*Archaea* (Figure 47). Nous nous sommes dans un premier temps intéressés à l'ensemble des fonctions communes aux deux mollivirus. Dans un second temps nous avons investigué les fonctions uniques à *M. kamchatka,* absentes chez *M. sibericum*. Enfin, nous avons cherché à déterminer l'origine de ces gènes aux fonctions connues, candidats potentiels au transfert horizontal de gène.

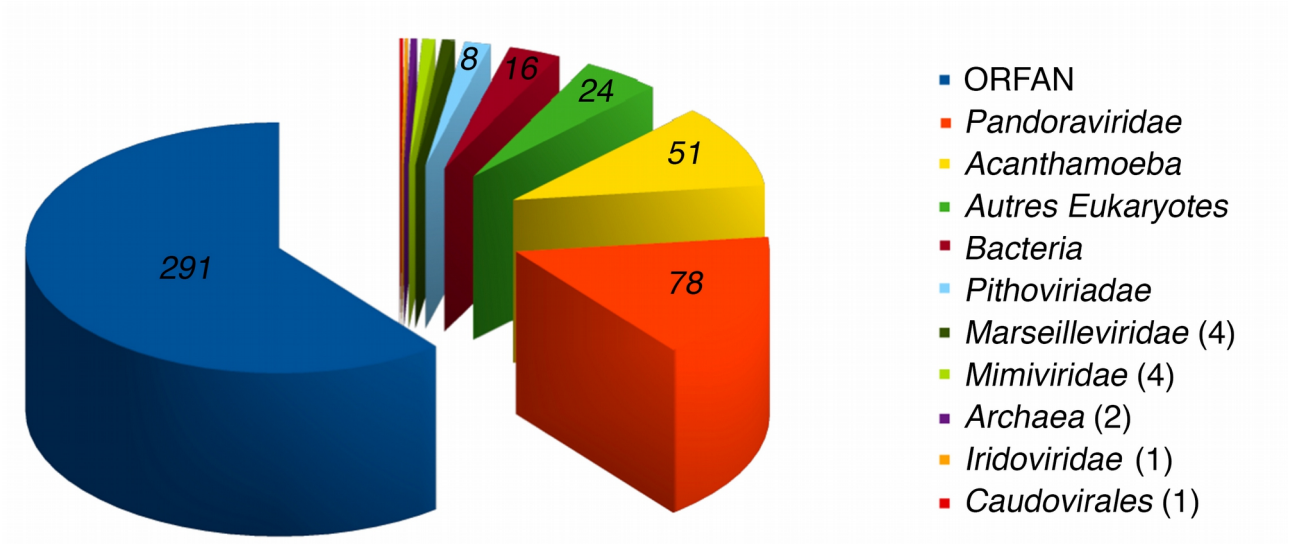


Figure 47 : Distribution des protéines homologues aux protéines prédites de M. kamchatka. La détection d'homologue a été faite par BLASTp (E-valeur < 10^{-5}) contre la base de données non redondante (NR) et les protéines associées à M. sibericum ont été exclues.

Sur le plan fonctionnel, 64 protéines prédites ont une fonction connue (13%). Parmi ces protéines portant une fonction, 59 sont communes aux deux mollivirus. On retrouve chez *M. kamchatka* les enzymes clefs du traitement de l'information génétique : une ADN polymérase de type B, un facteur d'initiation de la traduction, trois ARN polymérases. En plus des protéines liées au traitement de l'information génétique, on retrouve des enzymes du métabolisme : une serine-

threonine kinase, une NAD/NADP deshydrogenase, et une deoxycytidylate deaminase. On peut noter la présence d'une MCP chez *M. kamchatka* strictement identique à celle de *M. sibericum*. En construisant un arbre phylogénétique à partir de la MCP présente chez mollivirus, il apparaît qu'il existe une protéine homologue chez *A. castellanii* et aucune chez pandoravirus. La topologie de l'arbre, fait apparaître ces protéines amibiennes au milieu des protéines virales, suggérant un transfert latéral de gène du virus vers l'hôte. L'absence de ribonucleoside-diphosphate reductase, habituellement retrouvé chez les virus géants, fait apparaître que les mollivirus ont un appareil d'expression des gènes plus restreints que les autres familles de virus géants. De plus, plusieurs enzymes clefs de la biosynthèse de l'ADN telles que la thymidylate synthase, la thymidylate kinase et la thymidine kinase sont absentes de chez *M. kamchatka* et *M. sibericum*.

Si on s'intéresse maintenant aux 20 gènes uniques à *M. kamchatka,* 12 sont ORFans*,* 6 ont une séquence homologue chez pandoravirus et 2 chez *A. castellanii.* Parmi ces 6 protéines, 5 ont une fonction qui leur est associée. Sur le plan fonctionnel on peut donc noter l'acquisition par *M. kamchatka* d'une peptidase supplémentaire (mk_166), d'une protéine à domaine BI-1 (mk_159), de deux protéines à doigts de zinc (mk_93 et mk_104), et d'une methyltransferase de l'ADN (mk_92).

Ces gènes, ayant un homologue dans les bases de données et codant pour une protéine ayant une fonction probable, peuvent avoir été acquis par *M. kamchatka via* des transferts horizontaux de gènes. Pour déterminer le sens de ce transfert, s'il existe, nous avons réalisé des arbres phylogénétiques des gènes partagés par *M. kamchatka* et pandoravirus (mk_92, mk_93, mk_104, mk_159, mk_165 et mk_166). Seul l'arbre retraçant l'histoire évolutive de la protéine mk_165 a donné un signal clair de transfert de gènes horizontal, des pandoravirus de clade B vers mollivirus. Le cas de la methyltransferase mk_92 est quant à lui plus ambigu. La comparaison de la séquence à la base de données Rebase fait apparaître que mk_92 est une protéine homologue à une methyltransferase de *P. quercus* (*E*-valeur > 10^{-5}) reconnaissant le site CCTNAGG, cependant l'arbre phylogénetique présente une longue branche associée à *P. dulcis.* Ce phénomène peut s'expliquer par une rapide divergence de ces deux gènes ou par l'acquisition de ces gènes suite à un remplacement du gène original par recombinaison. Plus étrange encore, aucune endonucléase associée au site de restriction CCTNAGG n'a été trouvée chez *M. kamchatka*, en particulier mk_93 la protéine contigue à mk_92 et portant un motif doigt de zinc. La raison pour laquelle il existe des échanges de methyltransferase entre virus est encore mal connue. Le peu de séquences homologues trouvées pour les gènes mk_93, mk_104, mk_159 et mk_166 suggère que ces gènes étaient vraisemblablement partagés par l'ancêtre commun aux mollivirus et pandoravirus puis perdus par *M. sibericum.* En suivant le même raisonnement, nous avons étudié l'origine des deux gènes

uniques à *M. kamchatka,* mk_467 et mk_469, ayant des homologues chez *Acanthamoeba castellanii*. Il n'a pas été possible de déterminer le sens du transfert pour mk_467. L'absence de séquences homologues à mk_469 en dehors de celle retrouvée chez *Acanthamoeba castellanii*, suggère que ce gène a été acquis par l'hôte au cours d'un transfert d'un ancêtre commun aux mollivirus vers l'amibe (Figure 48).

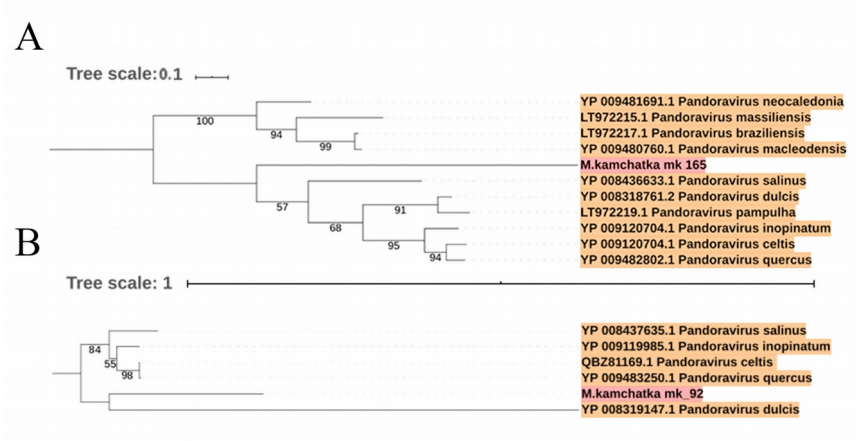


Figure 48 : Transferts éventuels de gènes d'un pandoravirus vers M. kamchatka. (A) Protéine mk_165 (pas de fonction prédite). La longue branche menant à mk_ 165 suggère une divergence rapide après acquisition à partir d'un pandoravirus. (B) Protéine mk_92 prédite comme étant une méthyltransférase. La longue branche menant à l'homologue de P. dulcis pourrait être interprétée comme un HGT ou un remplacement non orthologue de la version ancestrale du gène.

Pour conclure, comparés aux *Pandoraviridae,* leurs plus proches voisins, les mollivirus ont une machinerie de réplication plus partielle (en plus d'une absence de DNA ligase et de topoisomerase constatée chez les pandoravirus), laissant penser que les mollivirus ont évolués vers une dépendance croissante vis-à-vis de la cellule hôte par réduction du nombre de leurs gènes. Ces observations font des mollivirus les virus géants les plus dépendants de leur hôte à ce jour, renforçant donc l 'idée que le contenu en gène d'un virus ainsi que le protéome de sa particule définissent le niveau de parasitisme de la famille virale. A contrario, il apparaît surprenant que la MCP soit conservée chez mollivirus et perdue par les pandoravirus alors même que leurs capsides virales ne sont pas icosahedriques. La caractérisation de deux transferts horizontaux de gènes du virus vers l'amibe, et la présence de 51 protéines ayant un homologue chez *A. castellanii* (11% des gènes) démontre la capacité des mollivirus à échanger leurs gènes avec leur hôte.

Génomique comparative et création de gènes chez les Molliviridae

La découverte d'un second mollivirus a permis de réaliser une étude de génomique

comparative entre *M. sibericum* et *M. kamchatka*, dans la suite des questions soulevées par l'étude des pandoravirus présentée plus tôt.

Les génomes sont largement colinéaires. Avec un pourcentage d'identité moyen en nucléotide de 93% sur l'ensemble du génome, il apparaît que le génome de *M. kamchatka* n'a pas connu d'événements de réarrangements majeurs et aucun élément transposable n'a été détecté comme chez les pandoravirus.

Sur le plan génétique, 463 des 480 gènes de *M. kamchatka* ont leur homologue le plus similaire chez *M. sibericum,* desquels 60% sont des ORFans. Les résultats OrthoFinder font apparaître 434 groupes de gènes homologues entre les deux virus dont 411 sans paralogues. Ce résultat indique un faible taux de duplication du génome et permet de faire une estimation haute du génome cœur des mollivirus.

Sur le plan génétique, nous avons étudié la différence entre *M. kamchatka* et *M. sibericum* au travers de l'étude des gènes uniques à ces deux virus. Ainsi, on retrouve 12 ORFans stricts chez *M. kamchatka* et 26 ORFans stricts de *M. sibericum*. Il faut noter que la présence d'un nombre d'ORFan strict plus faible chez *M. kamchatka* peut s'expliquer par l'absence de données de transcriptomique, par conséquent il convient de préciser que cette valeur est sans doute une estimation minimale. En comparant l'ensemble des ORFans stricts aux gènes partagés du génome cœur avec, comme critères, les paramètres suivants : le CAI, la taille et le GC%, il apparaît que les ORFans stricts sont statistiquement différents des gènes cœurs. Les ORFans stricts ont des caractéristiques proches des régions intergéniques : GC% plus faible que les régions géniques et taille moyenne plus faible que les gènes. Les codons utilisés pour coder les ORFans stricts sont différents que ceux utilisés préférentiellement pour le reste du génome (Figure 49).

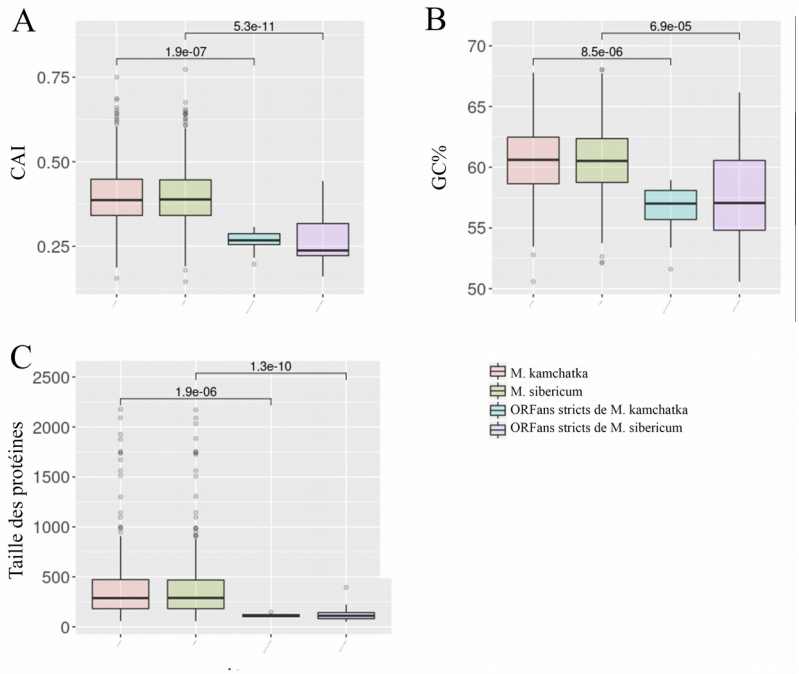


Figure 49 : Caractéristiques génomiques de differentes catégories de gènes : gènes de M. kamchatka (orange), gènes de M. sibericum (vert), ORFans spécifiques à M. kamchatka (bleu) et les ORFans spécifiques à M. sibericum (violet). (A) Indice d'adaptation des codons (CAI). (B) GC%. (C) Longueur des protéines. Les boîtes à moustaches montrent la médiane, les 25e et 75e centiles. La significativité est calculées en utilisant le test de Wilcoxon.

En conclusion, bien que le contenu en gènes de *M. kamchatka* et *M. sibericum* soit à plus de 80% identique, les différences observées sont en partie liées à la capacité des *Molliviridae* à créer leurs propres gènes *de novo*. Il semble, comme décrit pour les pandoravirus, que ce mécanisme a lieu dans les régions intergéniques. Pour confirmer ces observations, des expériences de transcriptomiques sont nécessaires. Bien que la création de gènes apparaît comme un mécanisme de « créativité » génétique commun au *Pandoraviridae* et aux *Molliviridae*, la duplication de gènes est un processus plus fréquent chez les *Pandoraviridae.*

Dynamique et histoire évolutive des Molliviridae

Si on considère que *M. kamchatka* est un parent contemporain de *M. sibericum* et que 30 000 ans séparent ces deux virus, il s'offre à nous trois modèles expliquant leur lien de parenté : une descendance directe, deux lignées sœurs ayant divergées juste avant que *M. sibericum* se retrouve gelé dans le pergélisol ou bien deux lignées sœurs ayant un ancêtre commun. Du fait de l'impossibilité d'évaluer le nombre de cycles infectieux de *M. kamchatka* ces 30 000 dernières années, ces modèles restent théoriques. En ce sens, la relation phylogénétique la plus parcimonieuse

à envisager consiste à considérer *M. kamchatka* et *M. sibericum* comme deux virus issus d'un même ancêtre commun et ayant divergé il y a plus de 30 000 ans (Figure 50). C'est à la lecture de ce constat que l'évocation des 30 000 ans séparant les deux virus doit être interprétée dans la suite du manuscrit.

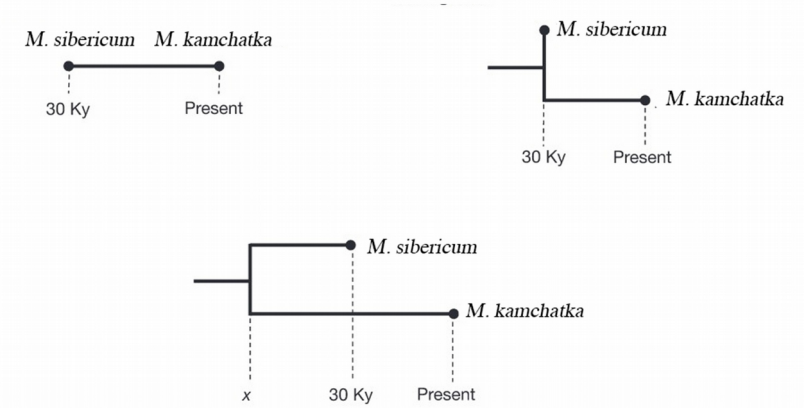


Figure 50 : Différents modèles évolutif établissant les liens de parenté entre M. kamchatka et M. sibericum. De haut en bas, descendance directe (gauche), lignées sœurs ayant divergées juste avant que M. sibericum se retrouve gelé dans le pergélisol (gauche) ou deux lignées sœurs ayant un ancêtre commun ayant divergé avant que M. sibericum se retrouve gelé dans le pergélisol (bas).

Avec un pourcentage d'identité en acide aminé moyen de 92% entre les protéines du génome cœur, les conditions sont optimales pour réaliser des alignements de qualité permettant un calcul optimal de la pression de sélection appliquée sur ces gènes. Ainsi, 397 valeurs de ω ont pu être calculées à partir des 411 paires de gènes homologues. Onze valeurs ont été retirées du fait d'une identité trop forte entre les séquences, et 3 du fait d'alignements mauvais probablement liés à une pseudogénisation de l'une des deux protéines. La valeur moyenne d' ω pour les 397 paires retenues est de 0,24 ± 0.14. Ce résultat indique donc une pression de sélection négative forte sur la majorité du génome, indiquant donc que *M. kamchatka* a accumulé des mutations conservatives durant les 30 000 dernières années et que l'intégrité du génome de *M. sibericum* a été conservée durant cette même période de temps dans le pergélisol. Si on restreint maintenant cette étude aux 244 paires d'ORFan, on obtient une valeur d'ω de 0,29 ± 0,15. Cela indique que, bien que n'ayant pas d'homologues dans des bases de données, ces gènes codent pour des protéines qui apparaissent nécessaires à la survie du virus, renforçant donc l'idée qu'il s'agit de « vrais » gènes à la fonction encore inconnue. Les cinq paires de gènes homologues ayant une valeur de ω strictement supérieure à 1 : ms_160/mk_141; ms_280/mk_262; ms_171/mk_151; ms_430/mk_411 et 223 ms_60/mk_48 sont tous des ORFan (Figure 51). Cette pression de sélection positive sur ces ORFan peut

s'expliquer de deux façons, soit ces gènes accumulent des mutations non synonymes pour maintenir leur fonction dans un contexte évolutif impliquant une forte pression de sélection (comme les gènes codants pour les parties variables des IgG), soit ces ORFan sont des gènes nouvellement créés en cours de raffinement depuis 30 000 ans.

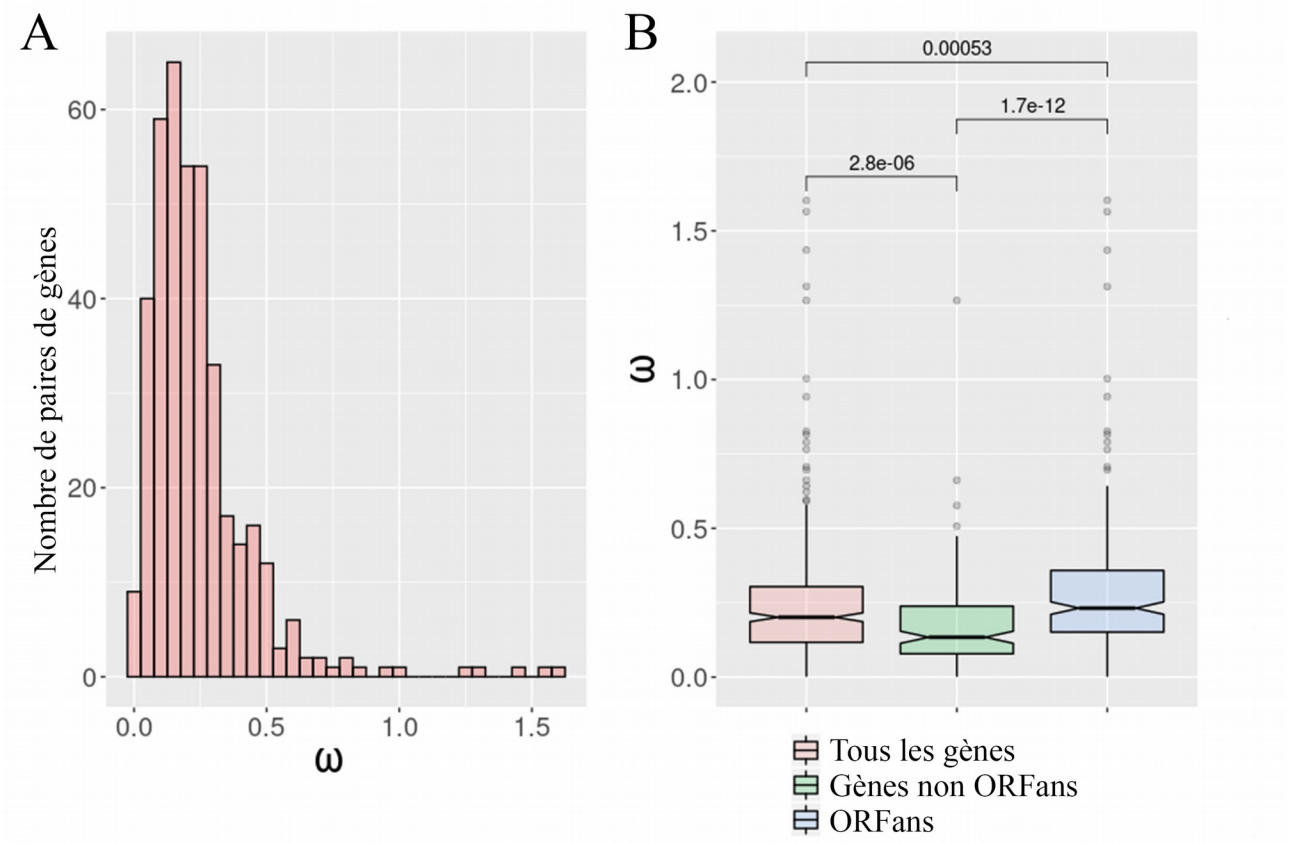


Figure 51 : Pression de sélection (ω) appliquée sur différentes catégories de gènes partagés par M. kamchatka et M. sibericum : tous les gènes (orange, n=397), gènes ayant un homologue en dehors des Molliviridae (vert, n=154) et gènes n'ayant aucun homologue connu dans les bases de données en dehors des Molliviridae (bleu, n=243).(A) Distribution des valeurs de ω. (B) Boîtes à moustaches montrant la médiane, les 25e et 75e centiles. La significativité est calculées en utilisant le test de Wilcoxon.

En considérant que 81% des séquences virales homologues aux gènes de *M. kamchatka* sont retrouvées chez des pandoravirus et que deux transferts horizontaux probables entre ces deux familles virales ont été trouvés, il existe probablement une histoire évolutive commune entre les *Pandoraviridae* et les mollivirus. Nous avons donc évalué la proximité phylogénétique des pandoravirus et des mollivirus par calcul du génome cœur partagé entre ces deux familles virales. Pour ce faire, nous avons récupéré les séquences des 10 génomes de pandoravirus disponibles sur les bases de données : *P. braziliensis, P. celtis, P. dulcis, P. salinus, P. inopinatum, P. macleodensis, P. massiliensis, P.neocaledonia, P. quercus* et *P. pampulha*. Les résultats OrthoFinder font apparaître 90 groupes de gènes partagés, desquels 64 sont composés uniquement

de gènes sans paralogues. On retrouve 465 groupes de gènes partagés uniquement par tous les pandoravirus et 385 groupes partagés uniquement par *M. kamchatka* et *M. sibericum* (Figure 52). Ces résultats sont légèrement différents des estimations initialement faites du génome cœur de ces mêmes familles : 455 groupes de gènes pour pandoravirus et 411 pour mollivirus. Par conséquent, mollivirus partage 23% de ses groupes gènes cœurs avec pandoravirus (90/385). La présence de 64 gènes communs à l'ensemble des génomes comparés constitue donc un « super génome coeur » regroupant 13% des gènes de *M. kamchatka* (64/485). Avec un ω moyen de 0,17 ± 0,1, la pression de sélection appliquée sur le super génome cœur chez mollivirus indique que ces gènes sont sous pression de sélection négative forte, confirmant que le génome cœur partagé par ces deux familles virales est strictement nécessaire au maintien de ces virus.

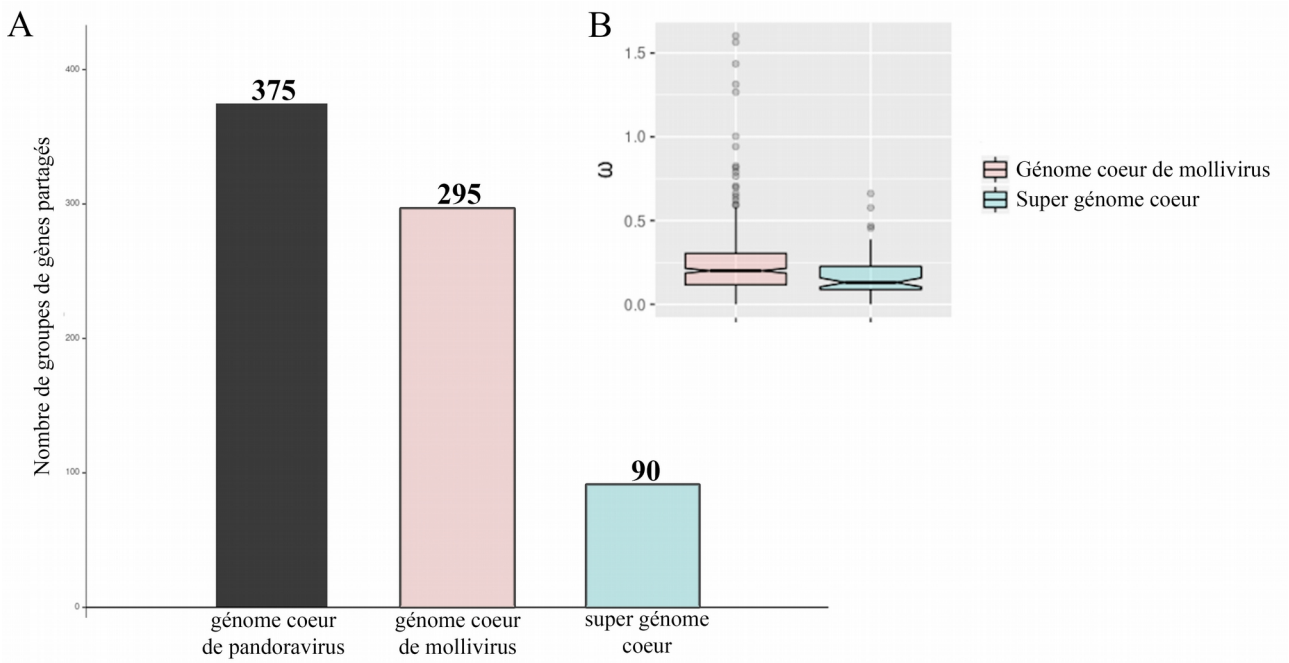


Figure 52 : Comparaison du contenu en gènes des Pandoraviridae et des Molliviridae.(A) Distribution des groupes de gènes communs à l'ensemble des Pandoraviridae (gris), distribution des groupes de gènes communs aux Molliviridae (rose) et des groupes de gènes partagés par les deux familles virales (bleu), ce super génome cœur représente donc 23% de la première estimation du génome cœur des mollivirus. (B) Boîtes à moustaches montrant la médiane, les 25e et 75e centiles des valeurs de ω pour le super génome cœur et le génome cœur des mollivirus.

De même, pour confirmer cette observation chez mollivirus, nous avons sélectionné deux pandoravirus au génome à 96% identiques : *P. celtis* et *P. quercus*. Ce niveau de conservation a permis de générer des alignements de même qualité que ceux générés pour mollivirus et la pression de sélection a été calculée à partir des 279 alignements ayant passés les critères de sélection. Avec une valeur ω moyenne de 0,15 ce résultat concorde avec les observations faite pour mollivirus et confirme que ce sous *set* de gènes est donc essentiel aux deux familles virales (Figure 53).

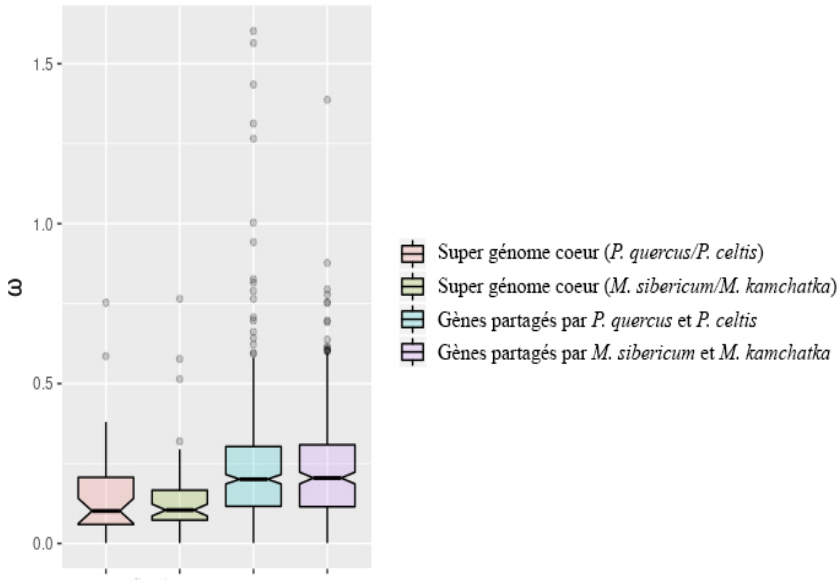


Figure 53 : Pression de sélection appliquée sur les gènes partagés entre deux pandoravirus proches : P. quercus et P. celtis. On observe que ces valeurs sont similaires à la pression de selection appliquée sur le génome cœur des Molliviridae. Boîtes à moustaches montrant la médiane, les 25e et 75e centiles des valeurs de ω.

En conclusion, la majorité du génome des *Molliviridae* est sous pression de sélection négative forte indiquant donc que la majorité des gènes est strictement nécessaire au virus. De plus, la présence d'un super génome cœur commun à la famille des *Pandoraviridae* et des *Molliviridae* confirme l'histoire évolutive commune de ces virus. Afin de déterminer si nous avions deux familles distinctes ou deux sous-familles d'une même famille, nous nous sommes intéressés à une autre famille de NCLDV, les poxvirus. Un échantillon de 21 génomes de *Poxviridae* a été sélectionné au hasard parmi les génomes complets déposés sur GenBank Virus. Ainsi, 6 génomes complets d' *Entomopoxviridae* et 15 de *Chordopoxviridae* ont été utilisés. Il existe 4 gènes strictement conservés entre poxvirus, mollivirus et pandoravirus : une ADN polymérase, une hélicase et deux ARN polymérases. Nous avons construit un arbre à partir de chaque orthogroupe. Une façon d'estimer si la divergence de deux sous-familles de *Poxviridae* est postérieure ou antérieure à la divergence des *Molliviridae* et des *Panddoraviridae* consiste à mesurer la taille des branches séparant chaque sous-famille de leur ancêtre commun potentiel. Il apparaît que pour ces 4 gènes la distance évolutive entre *Entomopoxviridae* e *Chordopoxviridae* semble plus faible qu'entre mollivirus et pandoravirus, suggérant que mollivirus et pandoravirus ont divergé avant la divergence entre *Entomopoxviridae* et *Chordopoxviridae*. Cet argument supplémentaire rend plus parcimonieux la du classification des *Molliviridae* comme nouvelle famille virale. Enfin, en

considérant la relation de parenté entre *M. sibericum* et *M. kamchatka* ainsi que le niveau de conservation des protéines, on évalue le taux de substitution à 1,7 changement d'acide aminé par position par an. Ce résultat constitue une surestimation car les deux virus ont probablement commencé à diverger l'un de l'autre il y a plus de 30 000 ans. Cette valeur est comparable aux estimations calculées pour poxvirus [105]. Ainsi, en l'absence de nouveaux isolats de mollivirus ou de virus intermédiaires, il apparaît plus parcimonieux de définir les *Molliviridae* comme nouvelle famille virale.

Distribution des gènes le long des génomes des Molliviridae

Il a été démontré que les gènes coeurs des *Pandoraviridae* étaient préférentiellement retrouvés en 3' des génomes. Du fait de la proximité génétique entre les *Pandoraviridae* et la nouvelle famille des *Molliviridae* nous avons décidé d'étudier la répartition de plusieurs catégories de gènes le long du génome de *M. sibericum* et *M. kamchatka.* Dans un premier temps nous avons regardé la distribution des gènes du super génome cœur (n= 64), les gènes issus de transfert horizontaux entre mollivirus et son hôte A. *castellanii* (n= 55 pour *M. sibericum* et n= 51 pour *M. kamchatka*), et enfin les gènes uniques à chacun des virus (n= 26 pour *M. sibericum* et n= 12 pour *M. kamchatka*) (Figure 54).

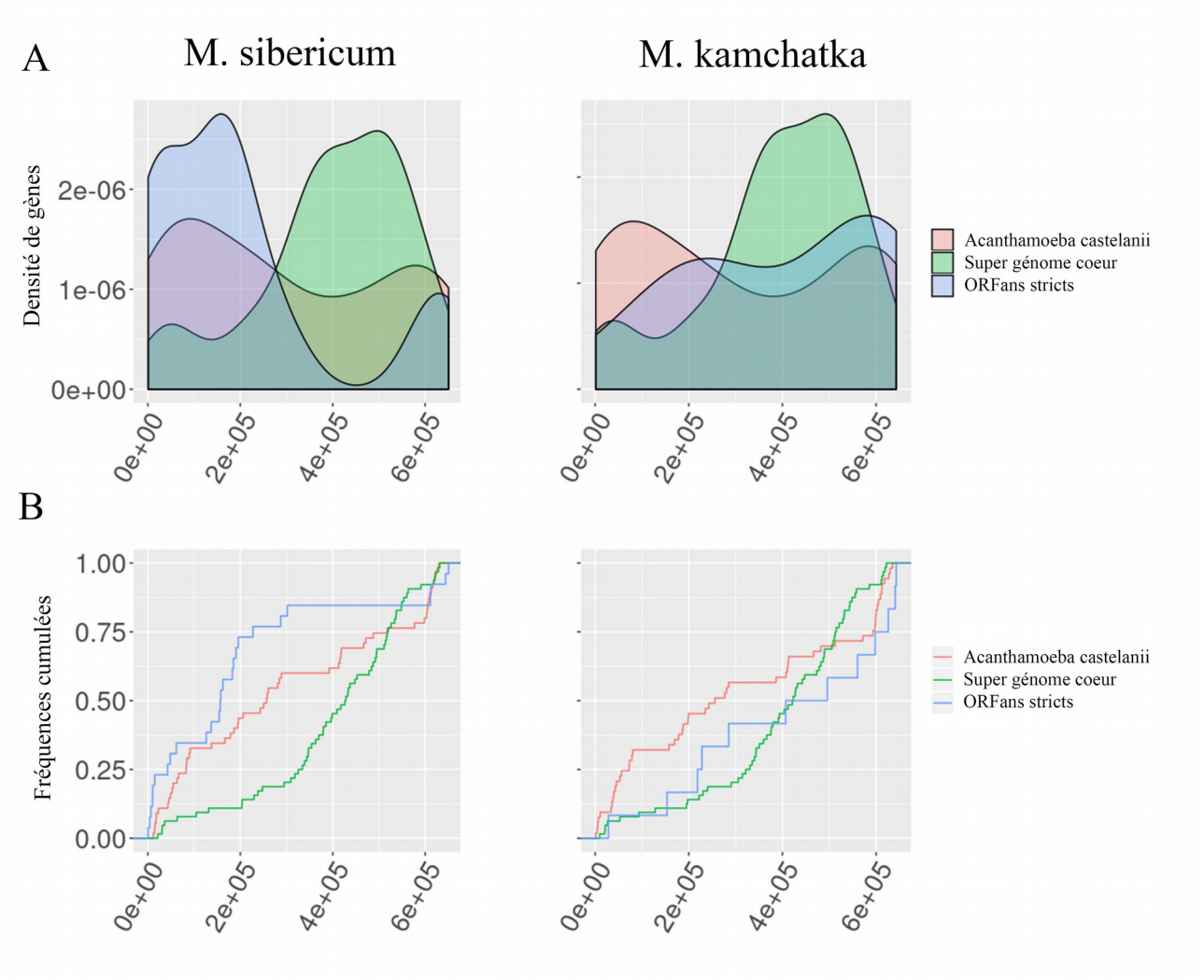


Figure 54 : Distribution de différentes catégories de gènes le long des génomes de M. kamchatka et M. sibericum : gènes ayant l'homologue le plus similaire chez A. castellanii (rouge), gènes du super génome cœur (vert) et ORFans strict à chaque virus (bleu). (A) Graphique montrant la densité de ces différentes catégories de gènes. (B) Fréquences cumulées de ces différentes catégories de gènes

On constate que le génome des mollivirus est polarisé. D'un coté s'accumulent les gènes du super génome cœur et de l'autre les ORFans stricts. De même, nous avons étendu cette étude aux gènes issus de duplications et en simple copie. Les gènes ayant des paralogues semblent s'accumuler à une extrémité du génome tandis-que la distribution des gènes en simple copie est uniforme (Figure 55).

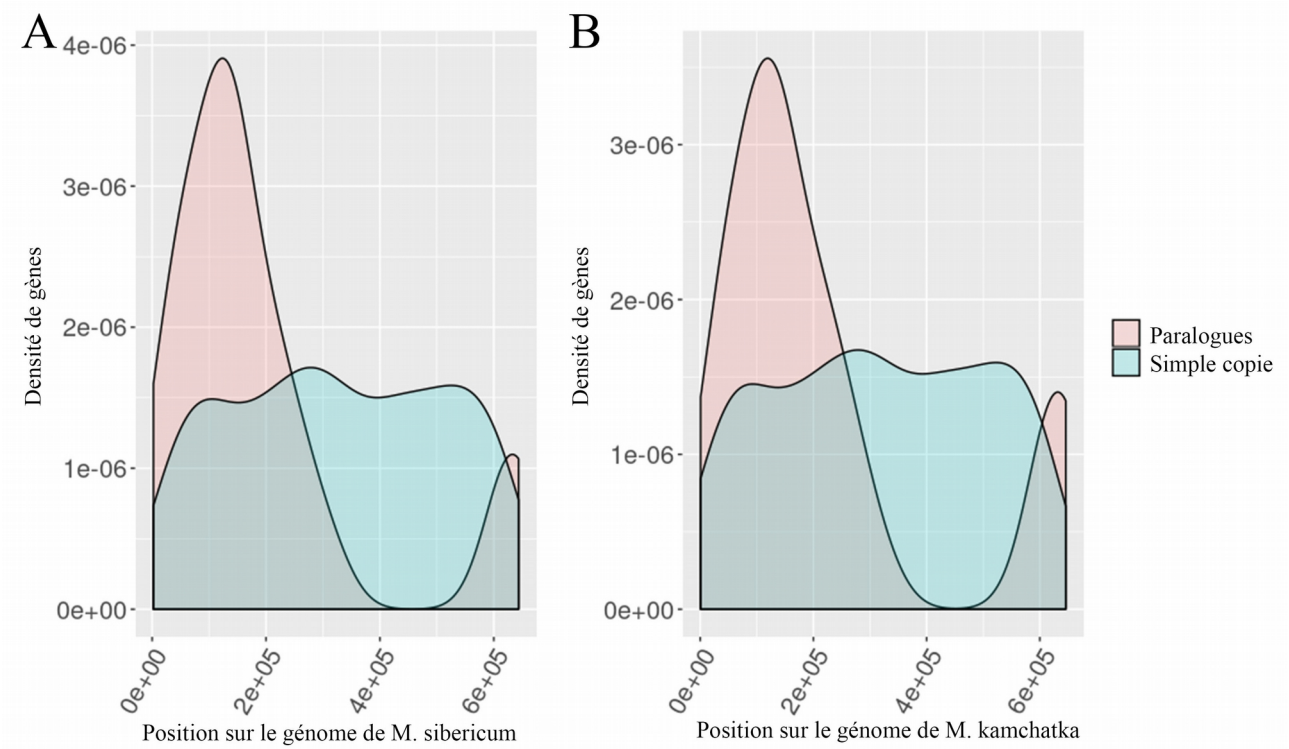


Figure 55 : Densité en gènes ayant au moins un paralogue (rose) et en gènes présents en une seule copie (bleu) le long des génomes de M. sibericum et M. kamchatka. (A) Densité le long du génome de M. sibericum. (B) Densité le long du génome de M. kamchatka.

En conclusion nous pouvons donc dire que le génome des mollivirus concentre sa « créativité génomique » : duplication, création de gènes.... à une extrémité du génome. A l'opposé de cette zone concentrant cette créativité se trouvent les gènes du super génome cœur. Nous avons tenté de vérifier si ces observations étaient corrélées à une distribution des valeurs de ω, sans succès.

Discussion

Alors que les familles des *Pithoviridae* s'est largement enrichie de nouveaux isolats de virus, *M. sibericum* est resté, depuis sa découverte, l'unique prototype de cette nouvelle famille virale. La découverte de *M. kamchatka,* à plus de 1500 km du lieu d'échantillonnage de *M. sibericum,* réfute donc l'hypothèse que mollivirus constitue l'unique représentant d'une famille virale éteinte depuis la dernière période glaciaire. A ce jour, le peu de lectures retrouvées dans les métagénomes des échantillons ayant permis la réactivation des *mollivirus* (0,03 particule par million (ppm) pour *M. kamchatka* et 1 ppm pour *M. sibericum*), ainsi que l'absence de *contig* associé à mollivirus dans les données de métagénome publiques assemblées, laisse supposer que la famille des *Molliviridae* est bien moins abondante que cellle des *Pandoraviridae* ou des *Mimiviridae,* deux familles virales considérées comme ubiquitaires.

L'étude comparée des capsides et des cycles infectieux de *M. sibericum* et *M. kamchatka* soulève des questions quant à la place des mollivirus dans le monde viral. Pour commencer, la couche externe fibrillaire de la capside des *Molliviridae* ressemble en tout point à la couche externe des capsides de *Pandoraviridae.* Des études préliminaires ont permis de déceler de la cellobiose et du glucose, sous produits de la dégradation de la cellulose, dans les capsides de *Pandoraviridae* et de *Molliviridae*. Ces résultats laissent penser que les Pandoravirus et Mollivirus détournent les celluloses synthases de leur hôte pour la synthèse de leurs capsides virales, s'affranchissant donc de la nécessité d'une MCP. Etrangement, bien qu'ayant une forme sphérique, les *Molliviridae* ont gardé une MCP qui est conservée à l'identique chez *M. sibericum* et *M. kamchatka* et dont une protéine homologue (à 64% identique) chez *A. castellanii.* Sur le plan évolutif ce résultat est remarquable, si il apparaît que cette protéine au motif *double-jelly roll* est encore nécessaire à l'intégrité des particules virales de mollivirus et constitue la 5ème protéine majoritaire du protéome de la particule virale de *M. sibericum*, la raison pour laquelle les *Pandoraviridae* ont perdu cette protéine reste encore inconnue.

Comme pour l'ensemble des particules de NCLDV, l'assemblage des membranes de la capside de mollivirus se fait par le recrutement de membranes cellulaires circulaires qui, au contact de la zone d'assemblage, permettent l'élongation du virion. De plus, l'embarquement du génome viral de mollivirus se fait en phase tardive de l'assemblage comme déjà caractérisé pour certains autres NCLDV comme *Vaccinia virus* ou les *Mimiviridae*. L'assemblage des virions de *Molliviridae* a également deux spécificités : les vésicules recrutées sont présentes en abondance dans la VF puis sont circularisés au contact de la zone d'assemblage et la synthèse de la couche externe du virion est permise par la présence d'un précurseur membranaire faisant office d'initiateur de la synthèse du virion. Ainsi, bien que les mécanismes restent à comprendre et la nature des membranes recyclées à établir, il est clair que la stratégie d'assemblage des virions de mollivirus partage des similarités avec celle des autres NCLDV tout en ayant des spécificités. Le cycle infectieux des *Molliviridae* repose sur une expression précoce des gènes viraux dans le noyau cellulaire. Ainsi, les *Molliviridae* constituent une deuxième famille de virus géants, avec les *Pandoraviridae*, ayant une dépendance stricte au noyau. Cette dépendance se manifeste par un appareil d'expression viral (transcription, réplication et traduction) extrêmement restreint comparé à l'ensemble des autres virus géants. Dans ce contexte, deux visions s'opposent. Dans un sens, on peut considérer que les *Pandoraviridae* ont diminué leur dépendance vis-à-vis de l'hôte par acquisition des gènes manquants à mollivirus, ainsi que certains gènes de la transcription. Dans l'autre sens, on peut également envisager une réduction du génome de mollivirus.

La découverte et la caractérisation d'un second mollivirus a permis d'élargir à la fois nos connaissances sur les *Molliviridae* en tant que nouvelle famille virale ainsi que de questionner leur origine et les liens de parenté avec les autres familles de NCLDV. En premier lieu, la grande similarité génétique entre *M. sibericum* et *M. kamchatka* n'a permis que de proposer une sur-estimation du génome cœur propre à cette famille virale. Ce haut niveau de conservation entre les paires de protéines homologues (92%) correspond à un taux de substitution faible de 1,7 acides aminés par position et par an, valeur similaire à celle déjà observée pour les *Poxviridae*. Bien que ce résultat reste une estimation biaisée par le nombre de cycles infectieux ayant réellement eu lieu durant les 30000 ans séparant les deux isolats, le calcul de la pression de sélection appliquée sur les gènes communs aux molliviridae démontre que la quasi totalité du génome est strictement nécessaire au maintien du virus. Ce résultat confirme également que les ORFans de mollvirus, eux-mêmes sous pression de sélection négative, sont de vrais gènes strictement nécessaires.

Il est apparu que le phénomène de duplication est marginal comparé à d'autres familles virales comme les *Pandoraviridae*. Nous avons également cherché à caractériser la présence d'éléments transposables, types HaT, largement présents chez les *Pandoraviridae*, sans succès. La présence de 51 gènes partagés par *M. sibericum* et *M. kamchatka* ayant comme uniques homologues des protéines de *A. castellanii* suggère des transferts de gènes horizontaux anciens du virus vers l'hôte. De plus, la découverte de deux transferts horizontaux de gènes entre *Pandoraviridae* et *Molliviridae,* dont une methyltransferase suggère que l'hôte *A. castellanii* constitue, en plus d'un *melting pot* d'ADN procaryote, un lieu de transfert de gènes horizontaux entre les virus et la cellule dans laquelle les *Molliviridae* jouent un rôle actif. En regardant les caractéristiques des ORFans uniques à chaque isolat nous avons pu émettre l'hypothèse que la création de gènes *de novo*, mécanisme commun aux mollivirus et aux pandoravirus, est en partie responsable du gigantisme de cette nouvelle famille virale.

Lors du calcul du génome cœur nous avons identifié un ensemble de gènes partagés par les *Molliviridae* et les *Pandoraviridae* sous pression de sélection négative plus forte que le reste du génome. En l'absence de nouveaux isolats de mollivirus il est difficile d'affirmer que cette relation phylogénétique est liée à une histoire évolutive commune ou à une accumulation de transferts horizontaux de gènes. La présence d'une structure de capside similaire et d'un génome cœur partagé 10 fois plus important que le génome cœur des NCLDV constituent néanmoins deux arguments forts en faveur d'un ancêtre commun à ces deux familles virales. Enfin, nous avons pu constater que ces gènes cœur étaient distribués de façon asymétrique le long du génome, en 3' du génome, comme déjà observé chez les *Pandoraviridae*. A l'opposé de cette zone extrêmement conservée se

concentre, en 5', l'ensemble de la « créativité » génétique des *Molliviridae :* duplication, création de gène. La présence de zones répétées aux extrémités des génomes de mollivirus et pandoravirus, couplée à ces observations pourrait suggérer que ces deux familles virales partagent certains mécanismes de réplication.

Détection des virus activés

Pour confirmer la présence de *M. kamchatka* dans l'échantillon à partir duquel il a été activé, nous avons aligné les lectures illumina contre le génome assemblé.

Ainsi, l'alignement des lectures contre le génome de *M. kamchatka* a permis de récupérer 16 lectures parfaitement alignées (100 d'identité) d'une taille de 151 pb. Parmi ces 16 lectures, on retrouve 4 paires appareillées et 8 non appareillées. La distribution des lectures alignées est homogène le long du génome. Pour vérifier la significativité des alignements nous avons aligné ces lectures contre le génome de *M. sibericum.* Le résultat a permis de récupérer 15 lectures ayant un pourcentage d'identité allant de 97% à 100% et chevauchant les ORF : ms_275, ms_274, ms_353, ms_433, ms_452, ms_459, ms_466 de *M. sibericum*.

En conclusion, la présence de 0,03 ppm de *M. kamchatka* dans l'échantillon rend la probabilité de réactivation très faible.

Papier 3 : Christo-Foroux, E., Alempic, J. M., Lartigue, A., Santini, S., Labadie, K., Legendre, M., ... & Claverie, J. M. (2020). Characterization of Mollivirus kamchatka, the first modern representative of the proposed molliviridae family of giant viruses. *Journal of Virology*, *94*(8).

Characterization of *Mollivirus kamchatka*, the first modern representative of the proposed *Molliviridae* family of giant viruses.

Running title: characterization of the first modern mollivirus

Eugene Christo-Foroux[#a], Jean-Marie Alempic[a], Audrey Lartigue[a], Sebastien Santini[a], Karine Labadie[b], Matthieu Legendre[a] , Chantal Abergel[a] , Jean-Michel Claverie[#a]

a Aix Marseille Univ., CNRS, IGS, Information Génomique & Structurale (UMR7256), Institut de Microbiologie de la Méditerranée (FR 3489), Marseille, France

b Genoscope, Institut François Jacob, CEA, Université Paris-Saclay, Évry, France

Correspondance to: Eugene.christo-foroux@igs.cnrs-mrs.fr, jean-michel.claverie@univ-amu.fr

Keywords: Paleovirology; Kamchatka; Comparative Genomics; Nucleo-cytoplasmic Large DNA viruses; C.

Abstract

Microbes trapped in permanently frozen paleosoils (permafrost) are the focus of increasing researches in the context of global warming. Our previous investigations led to the discovery and reactivation of two Acanthamoeba-infecting giant viruses, *Mollivirus sibericum* and *Pithovirus sibericum* from a 30,000-year old permafrost layer. While several modern pithovirus strains have since been isolated, no contemporary mollivirus relative was found. We now describe *Mollivirus kamchatka*, a close relative to *M. sibericum*, isolated from surface soil sampled on the bank of the Kronotsky river in Kamchatka. This discovery confirms that molliviruses have not gone extinct and are at least present in a distant subarctic continental location. This modern isolate exhibits a nucleo-cytoplasmic replication cycle identical to that of *M. sibericum*. Its spherical particle (0.6-µm in diameter) encloses a 648-kb GC-rich double stranded DNA genome coding for 480 proteins of which 61 % are unique to these two molliviruses. The 461 homologous proteins are highly conserved (92 % identical residues in average) despite the presumed stasis of *M. sibericum* for the last 30,000 years. Selection pressure analyses show that most of these proteins contribute to the virus fitness. The comparison of these first two molliviruses clarify their evolutionary relationship with the pandoraviruses, supporting their provisional classification in a distinct family, the *Molliviridae*, pending the eventual discovery of intermediary missing links better demonstrating their common ancestry.

Importance

Virology has long been viewed through the prism of human, cattle or plant diseases leading to a largely incomplete picture of the viral world. The serendipitous discovery of the first giant virus visible under light microscopy (i.e., >0.3µm in diameter), mimivirus, opened a new era of environmental virology, now incorporating protozoan-infecting viruses. Planet-wide isolation studies and metagenomes analyses have shown the presence of giant viruses in most terrestrial and aquatic environments including upper Pleistocene frozen soils. Those systematic surveys have led authors to propose several new distinct families, including the *Mimiviridae*, *Marseilleviridae*, *Faustoviridae*, *Pandoraviridae*, and *Pithoviridae*. We now propose to introduce one additional family, the *Molliviridae*, following the description of *M. kamchatka*, the first modern relative of *M. sibericum*, previously isolated from 30,000-year old arctic permafrost.

Introduction

The serendipitous discovery of the Acanthamoeba-infecting Mimivirus (1) and its detailed characterization (2, 3) more than 15 years ago, started a new era in virology that has now revealed the existence of several families of so-called “giant” viruses exhibiting particles and particles rivaling in size and gene contents with the cellular world. As of today, the use of Acanthamoeba (or related amoebozoa) as laboratory hosts (and environmental baits) has allowed the discovery and isolation of previously overlooked giant and large DNA viruses, exhibiting very diverse distinct virion morphologies and sizes, gene contents, and intracellular modes of replication (4). Few of them have yet received a formal ICTV taxonomy (5), with the exception of the *Mimiviridae* (i.e. Mimivirus relatives) (6) and the *Marseilleviridae* (7), that were chronologically first described and appears the most abundant in the environment. More recent discoveries include the Pandoraviruses in 2013 (i.e. the proposed *Pandoraviridae* family) (8), the number of which also increases rapidly (9), Pithovirus (prototype of the proposed *Pithoviridae* family) in 2014 (10, 11), Faustovirus (12)(related to the Asfarviruses) in 2015, and the more elusive Mollivirus (13), without known relative until this work. The latest addition to the list of large DNA virus infecting amoebas is Medusavirus (14). The main giant virus groups are positioned relative to each other in Fig. 1, using a phylogenetic tree of the DNA polymerase (the most informative of the only 3 core proteins strictly conserved across all large eukaryotic DNA viruses)(15). As expected from such a small number of common markers (especially relative to the hundreds of proteins they encode), the evolutionary relationships and origins of these amoeba-infecting giant viruses, remain a controversial subject (4, 15-17).

Studies started about 30 years ago, have provided multiple evidence that soils frozen since the late Pleistocene and predominantly located in arctic and subarctic Siberia, do contain a wide diversity of microbes that can be revived upon thawing (18-20) after tens of thousands of years.

These studies culminated by the regeneration of a plant from 30,000 year-old fruit tissue (21). Inspired by those studies, we then isolated from a similar sample two different Acanthamoeba-infecting large DNA viruses, named *Pithovirus sibericum* (10) and *Mollivirus sibericum* (13), demonstrating the ability for these viruses, and maybe many others, to remain infectious after similarly long periods of stasis in permafrost. In contrast to *P. sibericum* of which several modern relatives have since been characterized (11, 22-24), no other relative of *M. sibericum* was found despite the increasing sampling efforts deployed by several laboratories. Without additional isolates, its classification as a prototype of a new family, or as a distant relative of the pandoraviruses (with which it shared several morphological features and 16% of its gene content) (13) remained an open question. Here we report the discovery and detailed characterization of the first modern *M. sibericum* relative, named *Mollivirus kamchatka*, after the location of the Kronotsky river bank where it was retrieved. The comparative analysis of these first two molliviruses highlights their evolutionary processes and suggests their provisional classification into their own family, the *Molliviridae*, distinct from the *Pandoraviridae*, pending the eventual discovery of intermediary missing links clearly establishing their common ancestry.

Journal of Virology

Results

Virus isolation

The original sample consisted of about 50 ml of vegetation-free superficial soil scooped (in sterile tubes) from the bank of the Kronotsky river (coordinates: N 54°32'59'' E 160°34'55'') on july 6th, 2017. Before being stored at 4°C, the sample was transported in a backpack for a week at ambient temperature (5° C up to 24° C). This area corresponds to a continental subarctic climate: very cold winters, and short, cool to mild summers, low humidity and little precipitation. Back in the laboratory, few grams of the sample were used in the Acanthamoeba co-cultivation procedure previously described (13). After a succession of passages and enrichment on *Acanthamoeba castellanii* cultures, viral particles were produced in sufficient quantity to be recovered and purified.

Virion morphology and ultrastructure

As for *M. sibericum*, light microscopy of infected cultures showed the multiplication of particles. Using transmission electron microscopy (TEM), these particles - undistinguishable from that of *M. sibericum* - appear approximately spherical, 600 nm in diameter, lined by an internal lipid membrane and enclosed in a 20 nm-thick electron-dense thick tegument covered with a mesh of fibers (Fig. 2).

Analysis of the replication cycle

The replication cycle in *A. castellanii* cells was monitored using TEM and light microscopy of DAPI-stained infected cultures as previously described (13). The suite of events previously described for host cells infected by *M. sibericum*, was similarly observed upon *M. kamchatka* replication (13, 25). After entering the amoeba cell through phagocytosis, *M. kamchatka* virions are found gathered in large vacuoles individually or in groups of 2-6 particles. Multiple nuclear events occur during the infection starting with the drift of the host cell nucleolus to the periphery of the nucleus 4-5h post infection (p.i.). 7h p.i., the nucleus appears filled with numerous fibrils that may correspond to viral

genomes tightly packed in DNA-protein complexes (Fig. 3A). About 30% of the nuclei observed at that time exhibit a ruptured nuclear membrane (Fig. 3B). Besides those internal nuclear events, we observed a loss of vacuolization within the host cell from 4h p.i. to the end of the cycle (in average 9h p.i.). Large viral factories are formed in the cytoplasm at the periphery of the disorganized nucleus. These viral factories display the same characteristics as those formed during *M. sibericum* infections, involving an active recycling of membrane fragments (25).

Comparative genomics

DNA prepared from purified *M. kamchatka* particles was sequenced using both Illumina and Oxford Nanopore platforms. *M. kamchatka* genome, a linear double stranded DNA molecule (dsDNA), was readily assembled as a unique sequence of 648.864 bp. The read coverage was uniform throughout the entire genome except for a 10 kb terminal segment presumably repeated at both ends and exhibiting twice the average value. The *M. kamchatka* genome is thus topologically identical to that of a *M. sibericum*, slightly larger in size (when including both terminal repeats), and with the same global nucleotide composition (G+C= 60%). This similarity was confirmed by the detailed comparison of their genome sequences exhibiting a global collinearity only interrupted by a few insertions and deletions.

Prior to the comparison of their gene contents, *M. sibericum* and *M. kamchatka* were both annotated using the same stringent procedure that we previously developed to correct for gene overpredictions suspected to occur in G+C rich sequences such as those of pandoraviruses (8, 26, 27). A total of 495 and 480 genes were predicted for *M. sibericum* and *M. kamchatka*, with the encoded proteins ranging from 51 to 2171 residues and from 57 to 2176 residues, respectively. *M. kamchatka* predicted protein sequences were used in similarity search against the non-redundant protein sequence database (28) and the re-annotated *M. sibericum* predicted proteome. Out of the 480 proteins predicted to be encoded by *M. kamchatka* 463 had their closest homologs in *M. sibericum*, with 92% identical residues in average. After clustering the paralogs,

these proteins corresponded to 434 distinct genes clusters delineating a first estimate of the mollivirus core gene set. Four hundred and eleven of these clusters contained a single copy (singletons) gene for each strain. 290 of the 480 (60.4 %) *M. kamchatka*-encoded proteins did not exhibit a detectable homolog among cellular organisms or previously sequenced viruses (excluding *M. sibericum*). Those will be referred to as "ORFans". Among the 190 proteins exhibiting significant ($E<10^{-5}$) matches in addition to their *M. sibericum* counterparts, 78 (16% of the total gene content) were most similar to Pandoravirus predicted proteins, 18 (3.7%) to proteins of other virus families, 51 (10.6%) to *A. castellanii* proteins, 24 (5%) to proteins of other eukaryotes, 17 (3.5%) to bacterial proteins and 2 (0.4%) to proteins of *Archaea* (Fig. 4).

The interpretation of these statistics are ambiguous as, on one hand, the large proportion of "ORFans" (>60%) is characteristic of what is usually found for the prototypes of novel giant virus families (4). On the other hand, the closest viral homologs are not scattered in diverse previously defined virus families, but mostly belongs to the Pandoraviridae (78/96=81%) (Fig. 4). The two molliviruses thus constitute a new group of viruses with their own specificity but with a phylogenetic affinity with the pandoraviruses, as previously noticed (4). The proportion of *M. kamchatka* proteins with best matching counterparts in Acanthamoeba confirms a gene exchange propensity with the host, already noticed for *M. sibericum* (13).

Recent evolutionary events since the *M. sibericum*/*M. kamchatka* divergence

We investigated the evolutionary events specific for each of the molliviruses by focusing on proteins lacking reciprocal best matches between the two strains. We found 63 such cases of which 10 corresponded to unilateral strain-specific duplications of genes, and 53 were unique to a given strain. These unique genes (Table 1 & 2) result from gains or losses in either of mollivirus strains (20 in *M. Kamchatka*, 33 in *M. sibericum*). The likely origins of these strain-specific genes (horizontal acquisition, *de novo* creation (26,27), or differential loss) are listed in Table 1 and Table 2.

Six *M. kamchatka* proteins, absent from *M. sibericum*, have homologs in pandoraviruses suggesting common gene ancestors (and loss in *M. sibericum*) or horizontal acquisitions. According to its embedded position within the pandoravirus phylogenetic tree, only one anonymous protein (mk_165) (sharing 57% identical residues with its homolog in *P. salinus*) could be interpreted as an ancient horizontal transfer from pandoraviruses (Fig. 5A). Another candidate, mk_92, shares 75% of identical residues with a pandoravirus DNA methyltransferase (pqer_cds_559). However the very long branch associated with the P. dulcis homolog (eventually due to a non-orthologous replacement) raises some doubts as for the origin of the *M. kamchatka* gene (Fig. 5B).

Two *M. kamchatka*-specific proteins, encoded by adjacent genes (mk_466, mk_467), have homologs in Acanthamoeba, suggesting potential host-virus exchanges. Phylogenetic reconstruction did not suggest a direction for the transfer of mk_466. However, since the unique homolog of mk_467 is found in Acanthamoeba (and not in other eukaryotes), the corresponding gene could have been recently acquired from a close relative of *M. kamchatka*.

Three proteins unique to *M. sibericum* have homologs in pandoraviruses (Table 2), suggesting common gene ancestors (and loss in *M. kamchatka*) or horizontal acquisitions. One protein (ms_14) has a unique homolog in Acanthamoeba, suggesting a virus to host exchange. The homolog of ms_312 in *Cavenderia fasciculata* (29) might be the testimony of past interactions between molliviruses and a deeply rooted ancestor of the Amoebozoa clade.

The above analyses of the genes unique to each mollivirus indicate that if horizontal transfer may contribute to their presence, it is not the predominant mechanisms for their acquisition. We then further investigated the 12 genes unique to *M. kamchatka* and 26 genes unique to *M. sibericum* (i.e. “strain ORFans” without homolog in the databases) by computing three independent sequence properties (Fig. 6): the codon adaptation usage index (CAI), the G+C composition, and the ORF length.

With an average CAI value of 0.26, the strain-specific ORFans appear significantly different from the rest of the mollivirus genes (mean = 0.40, Wilcoxon test $p< 2\ 10^{-7}$). These genes also exhibit a significantly lower G+C content (56% for *M. kamchatka* and 57% for *M. sibericum*) than the rest of the genes (60.5% for both viruses), also closer to the value computed for intergenic regions (54% in average for both viruses). Moreover, the strain-specific ORFans are smaller in average compared to the rest of the genes (115bp/378bp for *M. kamchatka* and 122bp/369bp for *M. sibericum*). Altogether, those results suggest that *de novo* gene creation might occur in the intergenic regions of molliviruses as already postulated for pandoraviruses (26,27).

New predicted protein functions in *M. kamchatka*

Sixty four of the *M. kamchatka* predicted proteins exhibit sequence motifs associated with known functions. Fifty nine of them are orthologous to previously annotated genes in *M. sibericum* (13). This common subset confirms the limited complement of DNA processing and repair enzymes found in molliviruses: mainly a DNA polymerase: mk_287, a primase: mk_236, and 3 helicases: mk_291, mk_293, mk_351. *M. kamchatka* confirms the absence of key deoxynucleotide synthesis pathways (such as thymidylate synthase, thymidine kinase and thymidylate kinase), and of a ribonucleoside-diphosphate reductase (present in pandoraviruses). The five *M. kamchatka* specific proteins (Table 1) exhibiting functional motifs or domain signatures correspond to:

- two proteins (mk_93 and mk_104) containing a type of zinc finger (Ring domain) mediating protein interactions,
- one protein (mk_469) with similarity to the (BI)-1 like family of small transmembrane proteins,
- one predicted LexA-related signal peptidase (mk_166),
- one DNA methyltransferase (mk_92).

Evaluation of the selection pressure exerted on mollivirus genes

The availability of two distinct strains of mollivirus allows the first estimation of the selection pressure exerted on their shared genes during their evolution. This was done by computing the

222 ratio ω=dN/dS of the rate of non-synonymous mutations (dN) over the rate of synonymous
223 mutations (dS) for pairs of orthologous genes. ω values much less than one are associated with
224 genes where mutations have the strongest negative impact on the virus fitness. The high sequence
225 similarity of proteins shared by *M. sibericum* and *M. kamchatka* allowed the generation of flawless
226 pairwise alignments and the computation of highly reliable ω values for most (*i.e.* 397/411) of their
227 orthologous singletons.

228 Fourteen singleton pairs were not taken into account in the selection pressure analysis
229 because of their either identical (11 of them) or quasi-identical (one, >98% identical nucleotides)
230 sequences, or unreliable pairwise alignments (2 of them). For the 397 gene pairs retained in the
231 analysis, the mean ω value was 0.24 ± 0.14 (Fig. 7). This result corresponds to a strong negative
232 selection pressure indicating that most of the encoded proteins greatly contribute to the
233 molliviruses' fitness. Together with the high level of pairwise similarity (92%) of their proteins, this
234 also indicates that *M. kamchatka* evolved very little during the last 30,000 years and that the *M.*
235 *sibericum* genome was not prominently damaged during its cryostasis in permafrost.

236 The analysis restricted to the 244 pairs of ORFan-coding genes resulted in a very similar ω
237 value of 0.29± 0.15 (Fig. 7). This indicates that although homologs of these proteins are only found
238 in molliviruses, they have the same impact on the virus fitness than more ubiquitous proteins. This
239 confirms that they do encode actual proteins, albeit with unknown functions. In contrast, four
240 orthologous pairs (ms_160/mk_141; ms_280/mk_262; ms_171/mk_151; ms_430/mk_411;
241 ms_60/mk_48) exhibit ω value larger than one. Those ORFans under positive selection might
242 correspond to newly created gene products undergoing refinement or pseudogenization.

243 We further examined the selection pressure of protein-coding genes with homologs in
244 pandoraviruses. We used their 10 sequenced genomes to generate the corresponding gene
245 clusters (Fig. 8). The 90 clusters shared by both virus groups included 64 singletons (single copy
246 gene present in all viruses), among which 55 were suitable for dN/dS computations. The mean ω

value (0.17 ± 0.1) was very low, indicating that these genes, forming a "super core" gene set common to the molliviruses and pandoraviruses, are under an even stronger negative selection pressure than those constituting the provisional (most likely overestimated) mollivirus core gene set .

Genomic inhomogeneity

The original genome analysis of Lausannevirus (a member of the Marseilleviridae family) (30) revealed an unexpected non-uniform distribution of genes according to their annotation. "Hypothetical" genes (i.e. mostly ORFans) were segregated from "annotated" (i.e. mostly non-ORFans) in two different halves of the genome. In a more recent work, we noticed a similar bias in the distribution of Pandoravirus core genes (26). The availability of a second mollivirus isolate gave us the opportunity to investigate this puzzling feature for yet another group of Acantamoeba-infecting virus. In Fig. 9, we plotted the distribution of three types of genes: 1- those with homologs in *A. castellanii* (n= 55 for *M. sibericum* and n= 51 for *M. kamchatka*), 2- those belonging to the super core set shared by both molliviruses and pandoraviruses (n=64), 3- those unique to either mollivirus strains (n=26 for *M. sibericum*, n=12 for *M. kamchatka*). These plots reveal a strong bias in the distribution of the super core vs. ORFan genes (Fig. 9). The first half of the *M. sibericum* genome exhibits 90% of its ORFans while the second half contains most of the members of the super core gene set. In contrast, genes eventually exchanged with the host display a more uniform distribution. The lack of an apparent segregation in the distribution of ORFans in the *M. kamchatka* genome might be due to their unreliable prediction as no transcriptome information is available for this strain. Fig. 10 shows that there is also a strong bias in the distribution of single-copy genes *vs.* those with paralogs in either *M. sibericum* and/or *M. kamchatka*. Altogether, these analyses suggest that the right and left genome halves follow different evolutionary scenario, the first half concentrating the genomic plasticity (*de novo* gene creation, gene duplication), the other half concentrating the most conserved, eventually essential, gene content.

Discussion

Following the discovery of their first representatives, each families of giant (e.g. Mimiviridae, Pithoviridae, Pandoraviridae) and large (e.g. Marseilleviridae) viruses infecting acanthamoeba have expanded steadily, suggesting they were relatively abundant and present in a large variety of environments. One noticeable exception has been the molliviruses, the prototype of which remained unique after its isolation from 30,000-year old permafrost. The absence of *M. sibericum* relatives from the large number of samples processed by others and us since 2014, raised the possibility that they might have gone extinct, or might be restricted to the Siberian arctic. Our isolation of a second representative of the proposed Molliviridae family, *M. kamtchatka*, at a location more than 1,500 km from the first isolate and enjoying a milder climate, is now refuting these hypotheses. Yet, the planet-wide ubiquity of these viruses remains to be established, in contrast to other acanthamoeba-infecting giant viruses (4). Even when present, mollivirus-like viruses appear to be in very low abundance, as judged from the very small fraction of metagenomic reads they represent in total sample DNA for *M. kamchatka* (about 0.02 part per million) as well as for *M. sibericum* (about one part per million)(13). Another possibility would be that the preferred environmental host is not Acanthamoeba, the model host used in our laboratory, making the reactivation less effective. However, evidences of specific gene exchanges with acanthamoeba (including a highly conserved homolog major capsid protein) (13, 31, 32) make this explanation unlikely. We speculate that members of the proposed Molliviridae family are simply less abundant than other acanthamoeba-infecting viruses, a conclusion further supported by the paucity of Mollivirus-related sequences in the publicly available metagenomics data (data not shown).

As always the case, the characterization of a second representative of a new virus representative opened new opportunities of analysis. Unfortunately, the closeness of *M.*

297 *kamchatka* with *M. sibericum* limited the amount of information that could be drawn from their
298 comparison. For instance, the number of genes shared by the two isolates is probably a large
299 overestimate of the "core" gene set characterizing the whole family. On the other hand, the
300 closeness of the two isolates allowed an accurate determination of the selection pressure
301 (ω=dN/dS) exerted on many genes, showing that most of them, including mollivirus ORFans,
302 encode actual proteins under strong negative selection contributing to the virus fitness. Given the
303 partial phylogenetic affinity (i.e. 90 shared gene clusters) of the mollivirus with the pandoraviruses,
304 we also assessed the selection pressure exerted on 55 of these "super core" genes, and found
305 them under even stronger negative selection (Fig. 8). This suggests that this super core gene set
306 might have been present in a common ancestor to both proposed families.

307 If we postulate that *M. sibericum* underwent into a complete stasis when it became frozen in
308 permafrost while *M. kamchatka* remained in contact with living acanthamoeba, we could consider
309 the two viral genomes to be separated by at least 30,000 years of evolution (eventually more if
310 they are not in a direct ancestry relationship)(33). The high percentage of identical residues (92%)
311 in their proteins corresponds to a low substitution rate of 1.7 10^{-6} amino acid change/position
312 /year. This is an overestimate since the two viruses probably started to diverge from each other
313 longer than 30,000-year ago. This value is nevertheless comparable with estimates computed for
314 poxviruses (34) given the uncertainty on the number of replicative cycles occurring per year. The
315 high level of sequence similarity of *M. kamchatka* with *M. sibericum* also indicates that the later
316 did not suffer much DNA damage during its frozen stasis, even in absence of detectable virus-
317 encoded DNA repair functions.

318 Horizontal gene transfers with the host were suggested by the fact that 51 proteins shared by
319 the two mollivirus strains exhibited a second best match in acanthamoeba. Because no homolog is
320 detected in other eukaryotes for most of them, these transfers probably occurred in the mollivirus-
321 to-host direction.

322 The clearest case is that of a major capsid protein homolog (mk_314, ml_347) sharing 64%
323 identical residues with a predicted acanthamoeba protein (locus: XP_004333827). Two other genes
324 encoding proteins that have also homologs in molliviruses flank the corresponding host gene.
325 However, the corresponding viral genes are not collinear in *M. sibericum* or *M. kamchatka* and
326 were probably transferred from a different, yet unknown mollivirus strain. The presence of a 100%
327 conserved major capsid protein homolog in the genome of *M. kamchatka* and *M. sibericum* is itself
328 puzzling. Such protein (with a double-jelly roll fold) is central to the structure of icosahedral
329 particles (35). Consistent with its detection in *M. sibericum* virions (13), its conservation in *M.*
330 *kamchatka*, suggests that it still plays a role in the formation of the spherical mollivirus particles,
331 while it has no homolog in the pandoraviruses. Inspired by previous observations made on the
332 unrelated Lausannevirus genome (30), we unveiled a marked asymmetry in the distribution of
333 different types of protein-coding genes in the Mollivirus genomes. As shown in Fig. 9 the left half
334 of the genome concentrates most of the genes coding for strain-specific ORFans while the right
335 half concentrates most of super core genes shared with pandoraviruses. This asymmetry is even
336 stronger for the multiple copy genes while single-copy genes are uniformly distributed along the
337 genome (Fig. 10). The molliviruses thus appear to confine their genomic "creativity" (*de novo*
338 creation and gene duplication) in one-half of their genome, leaving the other half more stable. An
339 asymmetry in the distribution of the core genes was previously noticed in the pandoravirus
340 genomes (26). Such features might be linked to the mechanism of replication that is probably
341 similar for the two virus families. Further studies are needed to investigate this process. The
342 asymmetrical genomic distribution of pandoravirus core genes and mollivirus super core genes
343 might be a testimony of their past common ancestry.
344 Despite their differences in morphology, as well as in virion and genome sizes, the comparative
345 analysis of the prototype *M. sibericum* and of the new isolate *M. kamchatka* confirms their
346 phylogenetic affinity with the Pandoraviruses (Fig. 1, Fig. 4). However, it remains unclear whether

this is due to a truly ancestral relationship between them, or if it is only the consequence of numerous past gene exchanges favored by the use of the same cellular host. From the perspective of the sole DNA polymerase sequence, the two known molliviruses do cluster with the pandoraviruses, albeit at a larger evolutionary distance than usually observed between members of the same virus family (Fig. 1). In absence of an objective threshold, and pending the characterization of eventual "missing links", we thus propose to classify *M. sibericum* and *M. kamchatka* as members of the proposed Molliviridae family, distinct from the Pandoraviridae.

Materials and Methods

Virus isolation

We isolated *M. kamchatka* from muddy grit collected near Kronotski Lake, Kamchatka (Russian Federation N :54 32 59, E :160 34 55). The sample was stored for twenty days in pure rice medium (36) at room temperature.

An aliquot of the pelleted sample triggered an infected phenotype on a culture of *Acanthamoeba castellanii Neff* (ATCC30010TM) cells adapted to 2.5µg/mL of Amphotericin B (Fungizone), Ampicillin (100µg/ml), Chloramphrenicol (30µg/ml) and Kanamycin (25µg/ml) in protease-peptone–yeast-extract–glucose (PPYG) medium after two days of incubation at 32°C. A final volume of 6 mL of supernatant from two T25 flasks exhibiting infectious phenotypes was centrifuged for 1 hour at 16,000xg at room temperature. Two T75 flasks were seeded with 60,000 cells/cm² and infected with the resuspended viral pellet. Infected cells were cultured in the same conditions as described below. We confirmed the presence of viral particles by light microscopy.

Validation of the presence of *M. kamchatka* in the original sample

To confirm the origin of the *M. kamchatka* isolate from the soil of the Kronotsky river bank, DNA was extracted from the sample and sequenced on an Illumina platform, leading to 340,320,265 pair-ended reads (mean length 150bp). These metagenomic reads were then mapped

onto the genome sequence of *M. kamchatka*. Seven matching (100 % Identity) pair-ended reads (hence 14 distinct reads) were detected, indicating the presence of virus particles in the original sample, although at very low concentration. However, the very low probability of such matches by chance ($p<10^{-63}$) together with the scattered distribution of these matches along the viral genome, further demonstrate the presence of *M. kamchatka* in the original sample.

Virus Cloning

Fresh *A. castellanii* cells were seeded on a 12-well culture plate at a final concentration of 70,000cells/cm². Cell adherence was controlled under light microscopy after 45 minute and about 50 viral particles per host cell were added (multiplicity of infection (MOI) = 50). After 1 h, the well was washed 15 times with 3 mL of PPYG to remove any viral particle in suspension. Cells were then recovered by gently scrapping the well, and a serial dilution was performed in the next three wells by mixing 200µL of the previous well with 500µL of fresh medium. Drops of 0.5µL of the last dilution were recovered and observed by light microscopy to confirm the presence of a unique *A. castellanii* cell. The 0.5µL droplets were then distributed in each well of three 24-well culture plate. Thousand uninfected *A. castellanii* cells in 500µL of PPYG were added to the wells seeded with a single cell and incubated at 32°C until witnessing the evidence of a viral production from the unique clone. The corresponding viral clones were recovered and amplified prior purification, DNA extraction and cell cycle characterization by electron microscopy.

Virus mass production and purification

A total number of 40 T75 flasks were seeded with fresh *A. castellanii* cells at a final concentration of 60,000 cells/cm². We controlled cell adherence using light microscopy after 45 minute and flasks were infected with a single clone of *M. kamchatka* at MOI=1. After 48h hours of incubation at 32°C, we recovered cells exhibiting infectious phenotypes by gently scrapping the flasks. We centrifuged for 10 min at 500×g to remove any cellular debris and viruses were pelleted

by a 1-hour centrifugation at 6,800xg. The viral pellet was then layered on a discontinuous cesium chloride gradient (1.2g/cm²/ 1.3g/cm²/ 1.4g/cm²/ 1.5g/cm²) and centrifuged for 20h at 103,000xg. The viral fraction produced a white disk, which was recovered and washed twice in PBS and stored at 4 °C or −80 °C with 7.5% DMSO.

Infectious cycle observations using TEM

Twelve T25 flasks were seeded with a final concentration of 80,000 cells/cm² in PPYG medium containing antibiotics. In order to get a synchronous infectious cycle eleven flasks were infected by freshly produced *M. kamchatka* at substantial MOI=40. The *A. castellanii* infected flasks were fixed by adding an equal volume of PBS buffer with 5 % glutaraldehyde at different time points after the infection : 1h pi, 2h pi, 3h pi, 4h pi, 5h pi, 6h pi, 7h pi, 8h pi, 9h pi, 10h pi and 25h pi. After 45min of fixation at room temperature, cells were scrapped and pelleted for 5min at 500xg. Then cells were resuspended in 1ml of PBS buffer with 2.5 % glutaraldehyde and stored at 4°C. Each sample was coated in 1mm³ of 2 % low melting agarose and embedded in Epon-812 resin. Optimized osmium-thiocarbohydrazide-osmium (OTO) protocol was used for staining the samples: 1h fixation in PBS with 2 % osmium tetroxide and 1.5 % potassium ferrocyanide, 20 min in water with 1 % thiocarbohydrazide, 30 min in water with 2 % osmium tetroxide, overnight incubation in water with 1 % uranyl acetate and finally 30min in lead aspartate. Dehydration was made using an increasing concentration of ethanol (50 %, 75 %, 85 %, 95 %, 100 %) and cold dry acetone. Samples were progressively impregnated with an increasing mix of acetone and Epon-812 resin mixed with DDSA 0.34v/v and NMA 0.68v/v (33 %, 50 %, 75 % and 100%). Final molding was made using a hard Epon-812 mix with DDSA 0.34v/v NMA 0.68v/v and 0.031v/v of DMP30 accelerator and hardened in the oven at 60°C for 5 days. Ultrathin sections (90nm thick) were observed using a FEI Tecnai G2 operating at 200kV.

DNA Extraction

M. kamchatka genomic DNA was extracted from approximately 5 x 10^9 purified virus particles using Purelink Genomic extraction mini kit according to the manufacturer's recommendation. Lysis was performed with in a buffer provided with the kit and extra DTT at a final concentration of 1mM.

Genome sequencing and assembly

The genome of *M. kamchatka* was sequenced using both Oxford Nanopore Technologies (ONT) and HiSeq 2500 plateforms. The DNA-seq paired-end protocol respectively produced 4,515,973 and 4,577,450 reads with an average quality score of 34.9 (84% of the reads > Q30). The sequencing of 1,026 ng using the ONT plateform allowed us to retrieve 1,104,003 long reads (average size 6,768 bp and N50= 9,513). Reads were assembled using Spades (37)(version SPAdes-3.12.0) with a stringent k-mer parameter using various iteration steps (k= 21, 41, 61, 81, 99, 127), and options "—nanopore" and "—careful" to minimize the number of mismatches in the final contig.

Annotation of *Mollivirus sibericum* and *Mollivirus kamchatka*

A stringent gene annotation of *M. sibericum* was performed as previously described (26) using RNA-seq transcriptomic data (13). Stranded RNA-seq reads were used to accurately annotate protein-coding genes. Stringent gene annotation of *M. kamchatka* was performed w/o RNA seq data but taking into account protein similarity with *M. sibericum*. Gene predictions were manually curated using the web-based genomic annotation editing platform Web Apollo (38). Functional annotations of protein-coding genes of both genomes were performed using a two-sided approach as already previously described (26). Briefly, protein domains were searched with the CD-search tool (39) and protein sequence searching based on the pairwise alignment of hidden Markov models (HMM) was performed against the Uniclust30 database using HHblits tool (40). Gene

clustering was done using Orthofinder's default parameters (41) adding the "-M msa -oa" option. Strict orthology between pairs of proteins was confirmed using best reciprocal blastp matches.

Selection Pressure Analysis

Ratios of non-synonymous (dN) over synonymous (dS) mutation rates for pairs of orthologous genes were computed from MAFFT global alignment (42) using the PAML package and codeml with the « model = 2 » (43). A strict filter was applyied to the dN/dS ratio: dN > 0, dS > 0, dS ≤ 2 and dN/dS ≤ 10. The computation of the Codon Adaptation Index (CAI) of both Mollivirus was performed using the cai tool from the Emboss package (44).

Metagenome sequencing, assembly and annotation

Total DNA was extracted from 0.25 and 0.26 g of soil sample using the PowerSoil DNA isolation kit (QIAGEN) following the manufacturer's protocol except for the addition of 80 mM DTT to the lysis buffer (C1) to increase the particle lysis effectiveness. 250 ng of purified DNA was sequenced on the Illumina HiSeq platform (French National Sequencing Center, Genoscope, Paris) using the DNA-seq paired-end protocol producing a dataset of 340,350,938 read pairs (2x150bp in length) with an average quality score of 37.87 (90.45% > Q30). Raw reads quality was evaluated with FASTQC (45). Identified contaminants (primers and chimeric reads) were discarded. Valid reads were trimmed on the right end using 30 as quality threshold with BBTools (46). All filtered reads were then mapped to the generated contigs using Bowtie2 (47) with the "--very-sensitive" option.

Availability of data

The *M. kamtchatka* annotated genome sequence is freely available from the public through the Genbank repository (URL://www.ncbi.nlm.nih.gov/genbank/) under accession number MN812837.

Acknowledgements

We are deeply indebted to our volunteer collaborator Alexander Morawitz for collecting the Kamchatka soil sample. We thank N. Brouilly, F. Richard and A. Aouane (imagery platform, Institut de Biologie du Développement de Marseille Luminy) for their expert assistance, and the PACA Bioinfo platform for computing support. Eugène Christo-Foroux is the recipient of a DGA-MRIS scholarship (201760003) and this project was funded by the French National Research Center (PRC1484-2018). The funding bodies had no role in the design of the study, analysis, and interpretation of data and in writing the manuscript.

Competing interests

The authors declare that they have no competing interests

Figure legends

Fig 1. Phylogeny of DNA polymerase B of large and giant dsDNA viruses. This neighbor-joining tree was computed (JTT substitution model, 100 resampling) on 397 amino acid positions from an alignment of 42 sequences computed by MAFFT (29). Branches with boostrap values <60% were collapsed.

Fig 2. Ultrathin section TEM image of a neo-synthetized *M. kamchatka* particle in the cell cytoplasm 7h post infection. The structure of the mature particles appear identical to that of *M. sibericum.*

Fig 3. Ultrathin section TEM image of *A. castellanii* cell 7 to 10 hours post infection by *M. kamchatka***.** (A) Viral factory exhibiting fibrils (F), a nascent viral particle (V), and surrounding mitochondria (M). Fragments of the ruptured nuclear membrane are visible as dark bead strings. (B) Details of a nuclear membrane rupture through which fibrils synthetized in the nucleus (N) are shed into the cytoplasm (C).

Fig 4. Distribution of the best-matching NR homologs of *M. kamchatka* predicted proteins. Best-matching homologous proteins were identified using BLASTP (E value $<10^{-5}$) against the non-redundant (NR) database (15) (after excluding *M. sibericum*). Green shades are used for eukaryotes, red shades for viruses.

Fig 5. Eventual gene transfers from a pandoravirus to *M. kamchatka*. Both phylogenetic trees were computed from the global alignments of orthologous protein sequences using MAFFT (29). IQtree (48) was used to determine the optimal substitution model (options: « -m TEST » and « –bb 1000 »). (A) Protein mk_165 (no predicted function). The corresponding long branch suggests its accelerated divergence since an ancient acquisition from a pandoravirus. (B) Predicted

methyltransferase mk_92. The long branch leading to the *P. dulcis* homolog might alternatively be interpreted as a non-orthologous replacement of the ancestral pandoravirus version of the gene.

Fig 6. Genomic features of strain-specific ORFans. (A) Codon adaption index (CAI). (B) G+C content. (C) Protein length. Box plots show the median, the 25th and 75th percentiles. P-values are calculated using the Wilcoxon test.

Fig 7. Selection pressure among different classes of genes. Values of ω (i.e. dN/dS) were computed from the alignments of homologous coding regions in *M. kamchatka* and *M. sibericum*. (A) Distribution of calculated ω values (n=397). (B) Box plots of the ω ratio among ORFan genes (n=243) and non ORFan genes (n=154). Box plots show the median, the 25th, and 75th percentiles. All p-values are calculated using the Wilcoxon test.

Fig 8. Comparison of the mollivirus and pandoravirus core gene contents. (A) The distribution of the protein clusters shared by all pandoraviruses (black), the two Molliviruses (pink), and by both virus groups (super core genes) (blue). (B) Box plot of ω values calculated from the alignment of molliviruses core genes (pink), and super core genes (blue). Box plots show the median, the 25th, and 75th percentiles.

529 **Fig 9.** Distribution of different classes of genes along mollivirus genomes. (A) Variation of the gene
530 density as computed by the ggpplot2 "geom_density" function (49). The distribution of super core
531 genes (n=64, in green) is strongly biased toward the right half of the genome, in contrast to the
532 genes with best-matching homologs in *A. castellanii* (in the NR database excluding mollivirus) more
533 evenly distributed (in pink)(n=55 and n=51 for *M. sibericum* and *M. kamchatka*, respectively). *M.*
534 *sibericum*-specific ORFans (in blue) also exhibit a non-uniform distribution toward the left half of
535 the genome. (B) Cumulative distribution of the above classes of genes using the same color code.
536
537 **Fig 10.** Distribution of single-copy *vs.* multiple copy genes along mollivirus genomes. Single copy
538 genes (in blue) in both strains are evenly distributed in contrast to genes with paralogs (pink) that
539 cluster in the left half of the genomes. (A) *M. sibericum* (n=48). (B) *M. kamchatka* (n= 46).
540 .

541 **Tables**

542

543 **Table 1.** Status of the protein-coding genes unique to *M. kamchatka*

ORF ID	Predicted function	Putative evolutionary scenario
mk_25	None (ORFan)	*de novo* creation or loss in *M. sibericum*
mk_92	DNA methyltransferase	loss in *M. sibericum* (present in some pandoraviruses)
mk_93	Ring domain	loss in *M. sibericum* (present in some pandoraviruses)
mk_104	Ring domain	loss in *M. sibericum* (present in some pandoraviruses)
mk_127	None (ORFan)	*de novo* creation or loss in *M. sibericum*
mk_159	None	loss in *M. sibericum* (present in some pandoraviruses)
mk_165	None	HGT from Pandoravirus
mk_166	Peptidase	loss in *M. sibericum* (present in some pandoraviruses)
mk_172	None (ORFan)	*de novo* creation or loss in *M. sibericum*
mk_182	None (ORFan)	*de novo* creation or loss in *M. sibericum*
mk_231	None (ORFan)	*de novo* creation or loss in *M. sibericum*
mk_313	None (ORFan)	*de novo* creation or loss in *M. sibericum*
mk_369	None (ORFan)	*de novo* creation or loss in *M. sibericum*
mk_415	None (ORFan)	*de novo* creation or loss in *M. sibericum*
mk_441	None (ORFan)	*de novo* creation or loss in *M. sibericum*
mk_466	None (ORFan)	*de novo* creation or loss in *M. sibericum*
mk_467	None	loss in *M. sibericum* (HGT to Acanthamoeba)
mk_469	B1-1 like	loss in *M. sibericum* (present in Acanthamoeba)
mk_476	None (ORFan)	*de novo* creation or loss in *M. sibericum*
mk_478	None (ORFan)	*de novo* creation or loss in *M. sibericum*

544

545 **Table 2.** Status of the protein-coding genes unique to *M. sibericum*

ORF ID	Predicted function	**Putative evolutionary scenario**
ms_1	None (ORFan)	*De novo* creation or loss in *M. kamchatka*
ms_3	None (ORFan)	*De novo* creation or loss in *M. kamchatka*
ms_5	None (ORFan)	*De novo* creation or loss in *M. kamchatka*
ms_7	None (ORFan)	*De novo* creation or loss in *M. kamchatka*
ms_8	None (ORFan)	*De novo* creation or loss in *M. kamchatka*
ms_13	None (ORFan)	*De novo* creation or loss in *M. kamchatka*
ms_14	None	HGT to Acanthamoeba
ms_38	None (ORFan)	*De novo* creation or loss in *M. kamchatka*
ms_42	None (ORFan)	*De novo* creation or loss in *M. kamchatka*
ms_53	None (ORFan)	*De novo* creation or loss in *M. kamchatka*
ms_64	None	loss in *M. Kamchatka* (present in Noumeavirus)
ms_109	None (ORFan)	*De novo* creation or loss in *M. kamchatka*
ms_120	None (ORFan)	*De novo* creation or loss in *M. kamchatka*
ms_136	None (ORFan)	*De novo* creation or loss in *M. kamchatka*
ms_138	None (ORFan)	*De novo* creation or loss in *M. kamchatka*
ms_139	None (ORFan)	*De novo* creation or loss in *M. kamchatka*
ms_144	None (ORFan)	*De novo* creation or loss in *M. kamchatka*
ms_157	None (ORFan)	*De novo* creation or loss in *M. kamchatka*
ms_159	None (ORFan)	*De novo* creation or loss in *M. kamchatka*
ms_166	None (ORFan)	*De novo* creation or loss in *M. kamchatka*
ms_172	None (ORFan)	*De novo* creation or loss in *M. kamchatka*
ms_190	None	loss in *M. kamchatka* (present in some pandoraviruses)
ms_193	None (ORFan)	*De novo* creation or loss in *M. kamchatka*
ms_246	None (ORFan)	*De novo* creation or loss in *M. kamchatka*
ms_258	None (ORFan)	*De novo* creation or loss in *M. kamchatka*
ms_311	Zinc-finger domain	loss in *M. kamchatka (*present in *Gossypium hirsutum)*
ms_312	Zinc-finger domain	loss in *M. kamchatka* (present in *Cavenderia fasciculata)*
ms_313	None	loss in *M. kamchatka* (present in some pandoraviruses)
ms_464	None (ORFan)	*De novo* creation or loss in *M. kamchatka*
ms_465	None (ORFan)	*De novo* creation or loss in *M. kamchatka*
ms_479	DNA methyltransferase	loss in *M. kamchatka* (present in some pandoraviruses)
ms_494	None (ORFan)	*De novo* creation or loss in *M. kamchatka*
ms_495	None (ORFan)	*De novo* creation or loss in *M. kamchatka*

546

547

548 **References**

549 1. La Scola B, Audic S, Robert C, Jungang L, de Lamballerie X, Drancourt M, Birtles R, Claverie JM,
550 Raoult D. 2003. A giant virus in amoebae. Science 299:2033.
551 https://doi.org/10.1126/science.1081867.
552 2. Raoult D, Audic S, Robert C, Abergel C, Renesto P, Ogata H, La Scola B, Suzan M, Claverie JM.
553 2004. The 1.2-megabase genome sequence of Mimivirus. Science 2004 306:1344-1350.
554 https://doi.org/10.1126/science.1101485.
555 3. Claverie JM, Grzela R, Lartigue A, Bernadac A, Nitsche S, Vacelet J, Ogata H, Abergel C. 2009.
556 Mimivirus and Mimiviridae: giant viruses with an increasing number of potential hosts, including
557 corals and sponges. J Invertebr Pathol 101:172-180. https://doi.org/10.1016/j.jip.2009.03.011.
558 4. Abergel C, Legendre M, Claverie JM. The rapidly expanding universe of giant viruses: Mimivirus,
559 Pandoravirus, Pithovirus and Mollivirus. 2015. FEMS Microbiol Rev 39:779-796.
560 https://doi.org/10.1093/femsre/fuv037.
561 5. Lefkowitz EJ, Dempsey DM, Hendrickson RC, Orton RJ, Siddell SG, Smith DB. 2018. Virus
562 taxonomy: the database of the International Committee on Taxonomy of Viruses (ICTV). Nucleic
563 Acids Res 46(D1):D708-D717. https://doi.org/10.1093/nar/gkx932.
564 6. Claverie JM, Abergel C. 2018. Mimiviridae: An expanding family of highly diverse large dsDNA
565 viruses infecting a wide phylogenetic range of quatic eukaryotes. Viruses 10(9). pii: E506.
566 https://doi.org/10.3390/v10090506.
567 7. Colson P, Pagnier I, Yoosuf N, Fournous G, La Scola B, Raoult D. 2013. "Marseilleviridae", a new
568 family of giant viruses infecting amoebae. Arch Virol 2013 158:915-920.
569 doi: 10.1007/s00705-012-1537-y.
570 8. Philippe N, Legendre M, Doutre G, Couté Y, Poirot O, Lescot M, Arslan D, Seltzer V, Bertaux L,
571 Bruley C, Garin J, Claverie JM, Abergel C. 2013. Pandoraviruses: amoeba viruses with genomes up
572 to 2.5 Mb reaching that of parasitic eukaryotes. Science 341:281-286.

https://doi.org/10.1126/science.1239181.

9. Poirot O, Jeudy S, Abergel C, Claverie JM. 2019. A puzzling anomaly in the 4-Mer composition of the giant pandoravirus genomes reveals a stringent new evolutionary selection process. J Virol 93(23). pii: e01206-19. https://doi.org/10.1128/JVI.01206-19.

10. 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 JM. 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. https://doi.org/ 10.1073/pnas.1320670111.

11. Bertelli C, Mueller L, Thomas V, Pillonel T, Jacquier N, Greub G. 2017. Cedratvirus lausannensis - digging into Pithoviridae diversity. Environ Microbiol 19:4022-4034. https://doi.org/10.1111/1462-2920.13813.

12. 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.

13. Legendre M, Lartigue A, Bertaux L, Jeudy S, Bartoli J, Lescot, M, Alempic JM, Ramus C, Bruley C, Labadie K, Shmakova L, Rivkina E, Couté Y, Abergel C, Claverie JM. 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-E5335. https://doi.org/10.1073/pnas.1510795112.

14. Yoshikawa G, Blanc-Mathieu R, Song C, Kayama Y, Mochizuki T, Murata K, Ogata H, Takemura M. 2019. Medusavirus, a novel large DNA virus discovered from hot spring water. J Virol 93(8). pii: e02130-18. https://doi.org/10.1128/JVI.02130-18.

15. Guglielmini J, Woo AC, Krupovic M, Forterre P, Gaia M. 2019. Diversification of giant and large eukaryotic dsDNA viruses predated the origin of modern eukaryotes. Proc Natl Acad Sci U S A. 116:19585-19592. https://doi.org/10.1073/pnas.1912006116.

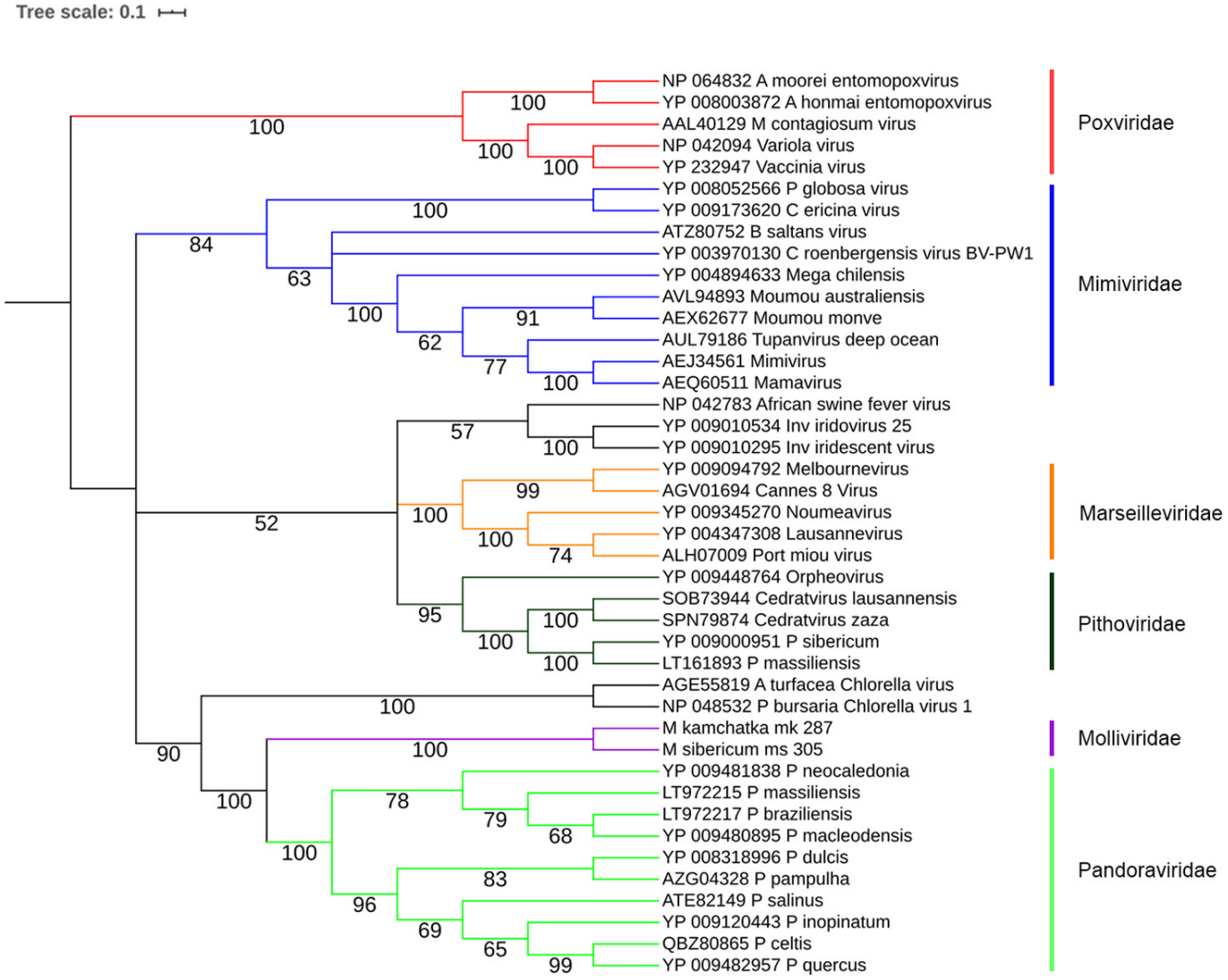
598 16. Nasir A, Caetano-Anollés G. A phylogenomic data-driven exploration of viral
599 origins and evolution. 2015. Sci Adv. 1(8):e1500527.
600 https://doi.org/10.1126/sciadv.1500527.
601 17. Koonin EV, Yutin N. 2019. Evolution of the large nucleocytoplasmic DNA viruses of
602 eukaryotes and convergent origins of viral gigantism. Adv Virus Res 103:167-202.
603 https://doi.org/10.1016/bs.aivir.2018.09.002.
604 18. Shi T, Reeves RH, Gilichinsky DA, Friedmann EI, 1997. Characterization of Viable Bacteria from
605 Siberian Permafrost by 16S rDNA Sequencing. Microb Ecol 33:169-179.
606 https://doi.org/10.1007/s002489900019.
607 19. Vishnivetskaya T, Kathariou S, McGrath J, Gilichinsky D, Tiedje JM. 2000. Low-temperature
608 recovery strategies for the isolation of bacteria from ancient permafrost sediments. Extremophiles
609 4:165-173. https://doi.org/10.1007/s007920070031.
610 20. Graham DE, Wallenstein MD, Vishnivetskaya TA, Waldrop MP, Phelps TJ, Pfiffner SM, Onstott
611 TC, Whyte LG, Rivkina EM, Gilichinsky DA, Elias DA, Mackelprang R, VerBerkmoes NC, Hettich RL,
612 Wagner D, Wullschleger SD, Jansson JK. 2012. Microbes in thawing permafrost: the unknown
613 variable in the climate change equation. ISME J 6:709-712.
614 https://doi.org/10.1038/ismej.2011.163.
615 21. Yashina S, Gubin S, Maksimovich S, Yashina A, Gakhova E, Gilichinsky D. 2012. Regeneration of
616 whole fertile plants from 30,000-y-old fruit tissue buried in Siberian permafrost. Proc Natl Acad Sci
617 U S A 109:4008-4013. https://doi.org/10.1073/pnas.1118386109.
618 22. Levasseur A, Andreani J, Delerce J, Bou Khalil J, Robert C, La Scola B, Raoult D. 2016.
619 Comparison of a Modern and Fossil Pithovirus Reveals Its Genetic Conservation and Evolution.
620 Genome Biol Evol 8:2333-2339. https://doi.org/10.1093/gbe/evw153.

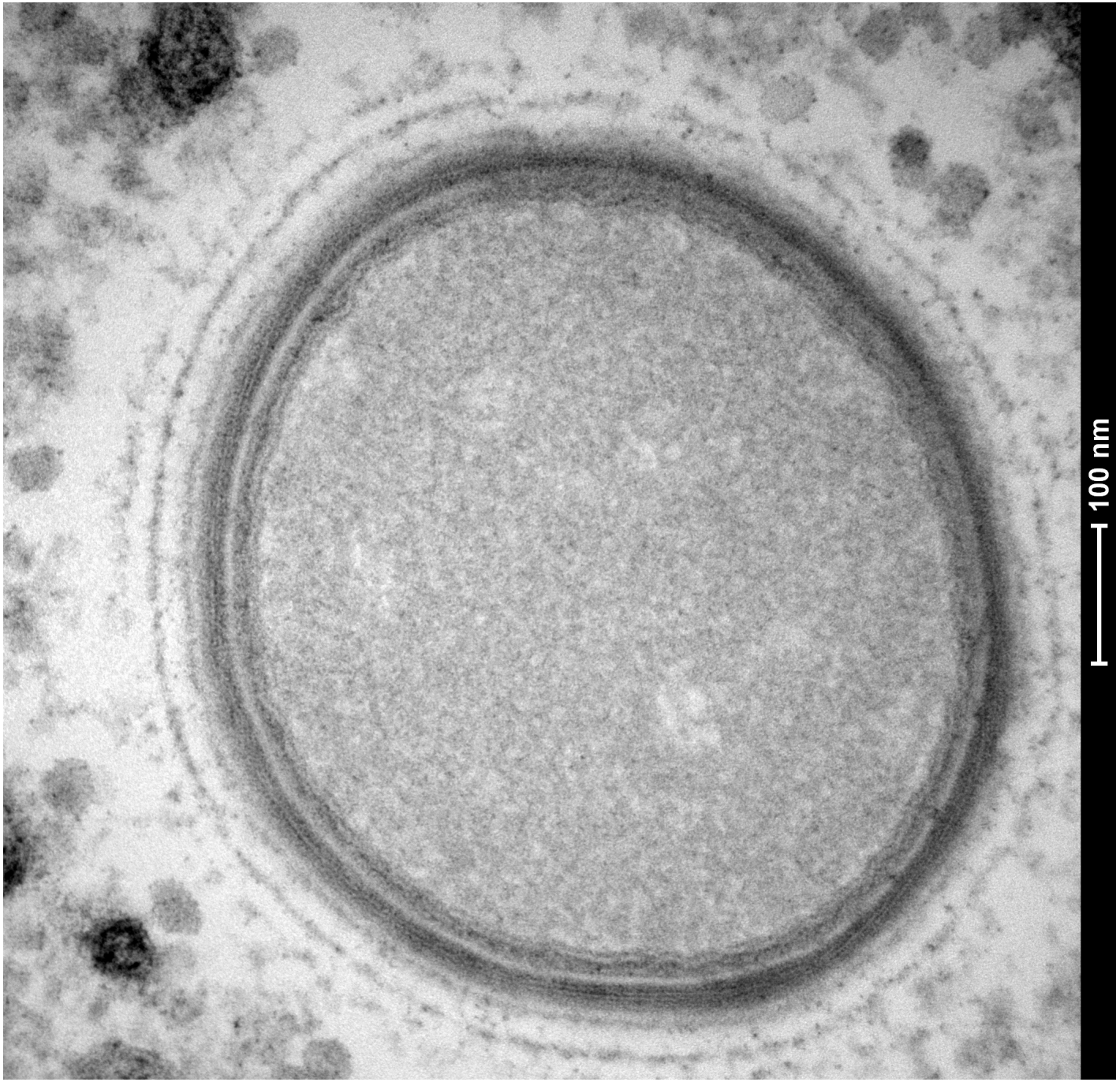
621 23. Andreani J, Aherfi S, Bou Khalil JY, Di Pinto F, Bitam I, Raoult D, Colson P, La Scola B. 2016.
622 Cedratvirus, a Double-Cork Structured Giant Virus, is a Distant Relative of Pithoviruses. Viruses
623 8:300. https://doi.org/10.3390/v8110300.
624 24. Andreani J, Khalil JYB, Baptiste E, Hasni I, Michelle C, Raoult D, Levasseur A, La Scola B.
625 Orpheovirus IHUMI-LCC2: A new virus among the giant viruses. 2018. Front Microbiol 8:2643.
626 https://doi.org/10.3389/fmicb.2017.02643.
627 25. Quemin ER, Corroyer-Dulmont S, Baskaran A, Penard E, Gazi AD, Christo-Foroux E, Walther P,
628 Abergel C, Krijnse-Locker J. 2019. Complex membrane remodeling during virion assembly of the
629 30,000-year-old Mollivirus sibericum. J Virol 93(13). pii: e00388-19.
630 https://doi.org/10.1128/JVI.00388-19.
631 26. Legendre M, Fabre E, Poirot O, Jeudy S, Lartigue A, Alempic JM, Beucher L, Philippe N, Bertaux
632 L, Christo-Foroux E, Labadie K, Couté Y, Abergel C, Claverie JM. 2018. Diversity and evolution of the
633 emerging Pandoraviridae family. Nat Commun 9:2285.
634 https://doi.org/10.1038/s41467-018-04698-4.
635 27. Legendre M, Alempic JM, Philippe N, Lartigue A, Jeudy S, Poirot O, Ta NT, Nin S, Couté Y,
636 Abergel C, Claverie JM. 2019. Pandoravirus celtis illustrates the microevolution processes at work
637 in the giant Pandoraviridae genomes. Front Microbiol 10:430.
638 https://doi.org/ 10.3389/fmicb.2019.00430.
639 28. NCBI Resource Coordinators. 2018. Database resources of the National Center for
640 Biotechnology Information. Nucleic Acids Res 46:D8-D13. https://doi.org/10.1093/nar/gkx1095.
641 29. Heidel AJ, Lawal HM, Felder M, Schilde C, Helps NR, Tunggal B, Rivero F, John U, Schleicher M,
642 Eichinger L, Platzer M, Noegel AA, Schaap P, Glöckner G. 2011. Phylogeny-wide analysis of social
643 amoeba genomes highlights ancient origins for complex intercellular communication. Genome Res
644 21:1882-1891. https://doi.org/10.1101/gr.121137.111.

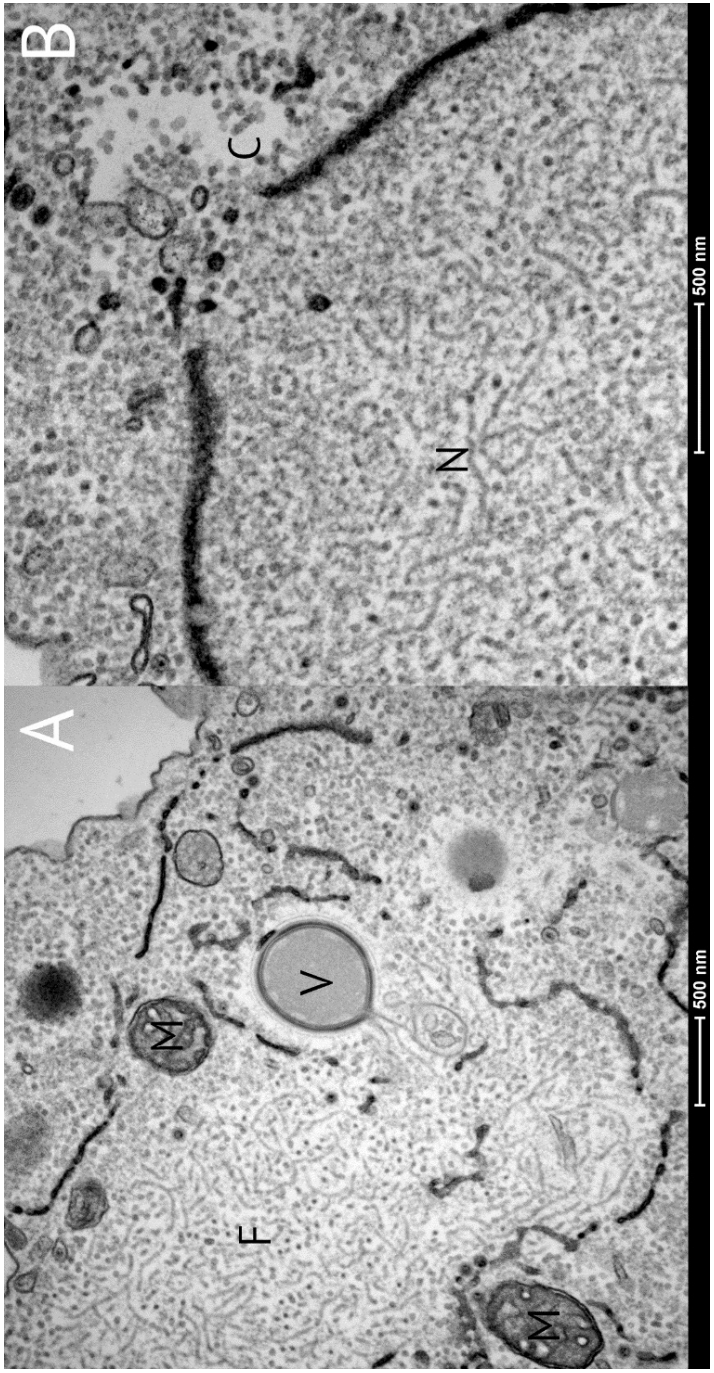
645 30. Thomas V, Bertelli C, Collyn F, Casson N, Telenti A, Goesmann A, Croxatto A, Greub G. 2011.
646 Lausannevirus, a giant amoebal virus encoding histone doublets. Environ Microbiol 13:1454-1466.
647 https://doi.org/10.1111/j.1462-2920.2011.02446.x.
648 31. Maumus F, Blanc G. 2016. Study of gene trafficking between acanthamoeba and giant viruses
649 suggests an undiscovered fof amoeba-infecting viruses. Genome Biol
650 Evol 8:3351-3363. https://doi.org/10.1093/gbe/evw260.
651 32. Chelkha N, Levasseur A, Pontarotti P, Raoult D, Scola B, Colson P. 2018. A phylogenomic study of
652 *Acanthamoeba polyphaga* draft genome sequences suggests genetic exchanges with giant viruses.
653 Front Microbiol 9:2098. https://doi.org/10.3389/fmicb.2018.02098.
654 33. Duchêne S, Holmes EC. 2018. Estimating evolutionary rates in giant viruses using
655 ancient genomes. Virus Evol 4:vey006. https://doi.org/10.1093/ve/vey006.
656 34. Hughes AL, Irausquin S, Friedman R. 2010. The evolutionary biology of poxviruses. Infect Genet
657 Evol 10:50-59. https://doi.org/10.1016/j.meegid.2009.10.001.
658 35. San Martín C, van Raaij MJ. 2018. The so far farthest reaches of the double jelly roll capsid
659 protein fold. Virol J 15:181. https://doi.org/10.1186/s12985-018-1097-1.
660 36. Arslan D, Legendre M, Seltzer V, Abergel C, Claverie JM. 2011. Distant Mimivirus relative with a
661 larger genome highlights the fundamental features of Megaviridae. Proc Natl Acad Sci U S A
662 108:17486-17491. https://doi.org/10.1073/pnas.1110889108.
663 37. Bankevich A, Nurk S, Antipov D, Gurevich A, Dvorkin M, Kulikov AS, Lesin V, Nikolenko S, Pham
664 S, Prjibelski A, Pyshkin A, Sirotkin A, Vyahhi N, Tesler G, Alekseyev MA, Pevzner P. 2012. SPAdes: A
665 new genome assembly algorithm and its applications to single-cell sequencing. J Comput Biol
666 19:455-477. https://doi.org/10.1089/cmb.2012.0021.
667 38. Dunn NA, Unni DR, Diesh C, Munoz-Torres M, Harris NL, Yao E, Rasche H, Holmes I, Elsik C,
668 Lewis S. 2019. Apollo: Democratizing genome annotation. PLoS Comput Biol 15: e1006790.
669 https://doi.org/10.1371/journal.pcbi.1006790.

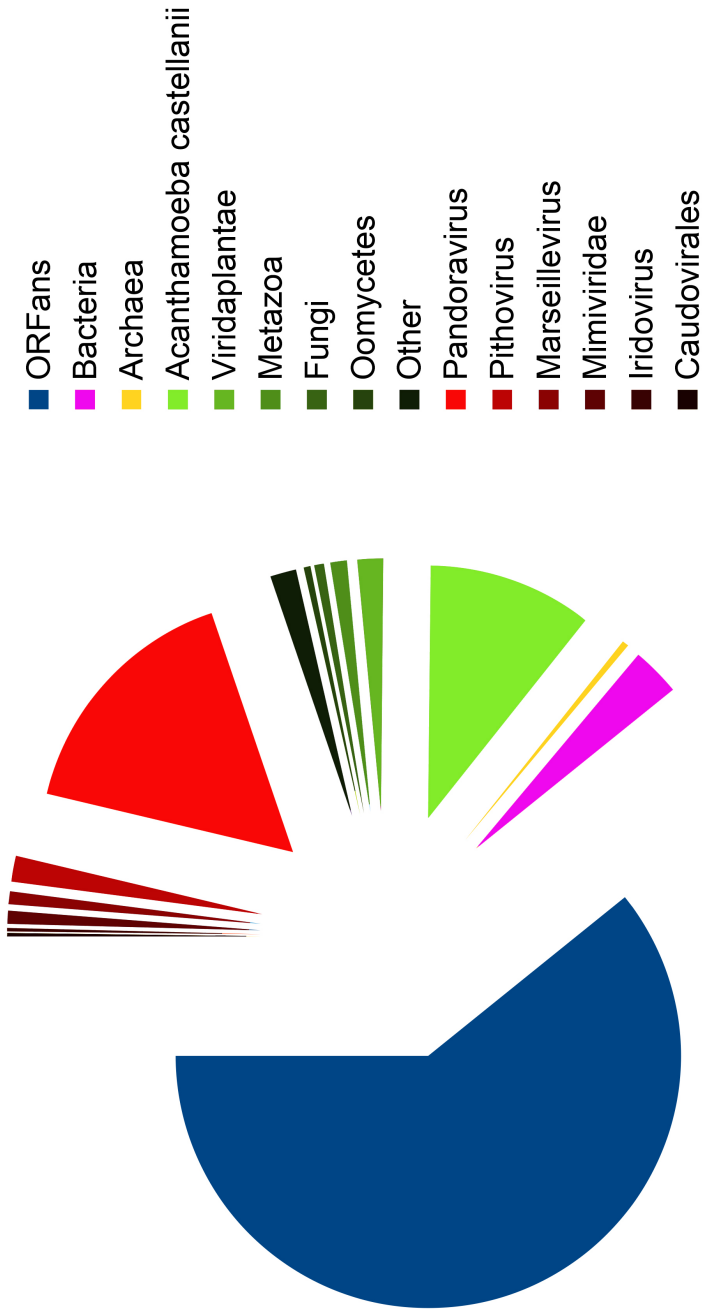
670 39. Marchler-Bauer A, Derbyshire MK, Gonzales NR, Lu S, Chitsaz F, Geer LY, Geer RC, He J, Gwadz
671 M, Hurwitz DI, Lanczycki CJ, Lu F, Marchler GH, Song JS, Thanki N, Wang Z, Yamashita RA, Zhang D,
672 Zheng C, Bryant SH. 2015. CDD: NCBI's conserved domain database. Nucleic Acids Res 43:D222-
673 226. https://doi.org/10.1093/nar/gku1221.
674 40. Remmert M, Biegert A, Hauser A, Söding J. 2011. HHblits: lightning-fast iterative
675 protein sequence searching by HMM-HMM alignment. Nat Methods. 9:173-175.
676 https://doi.org/10.1038/nmeth.1818.
677 41. Emms DM, Kelly S. 2015. OrthoFinder: solving fundamental biases in whole genome
678 comparisons dramatically improves orthogroup inference accuracy. Genome Biol 16:157.
679 https://doi.org/10.1186/s13059-015-0721-2.
680 42. Katoh K, Misawa K, Kuma K-I, Miyata T. 2002. MAFFT: a novel method for rapid multiple
681 sequence alignment based on fast Fourier transform. Nucleic Acids Res 30:3059-3066.
682 https://doi.org/10.1093/nar/gkf436.
683 43. Yang Z. 2007. PAML 4: phylogenetic analysis by maximum likelihood. Mol Biol Evol 24:1586-
684 1591. https://doi.org/10.1093/molbev/msm088.
685 44 . Rice P, Longden I, Bleasby A. 2000. EMBOSS: the European molecular biology open software
686 suite. Trends Genet 16:276-277. https://doi.org/10.1016/s0168-9525(00)02024-2.
687 45. Andrews S. 2010. FastQC: a quality control tool for high throughput sequence data. Available
688 online at: http://www.bioinformatics.babraham.ac.uk/projects/fastqc .
689 46. Bushnell B, Rood J, Singer E. 2017. BBMerge – Accurate paired shotgun read merging via
690 overlap. PLoS ONE 12: e0185056. https://doi.org/10.1371/journal.pone.0185056.
691 47. Langmead B, Salzberg SL. 2012. Fast gapped-read alignment with Bowtie 2. Nat Methods
692 9:357-359. https://doi.org/10.1038/nmeth.1923.

693 48. Nguyen LT, Schmidt HA, von Haeseler A, Minh BQ. 2015. IQ-TREE: A fast and effective stochastic
694 algorithm for estimating maximum likelihood phylogenies. Mol Biol Evol 32:268-274.
695 https://doi.org/10.1093/molbev/msu300.
696 49. Wickham H. ggplot2: Elegant Graphics for Data Analysis, pp 33-74. 2016. Springer-Verlag New
697 York.

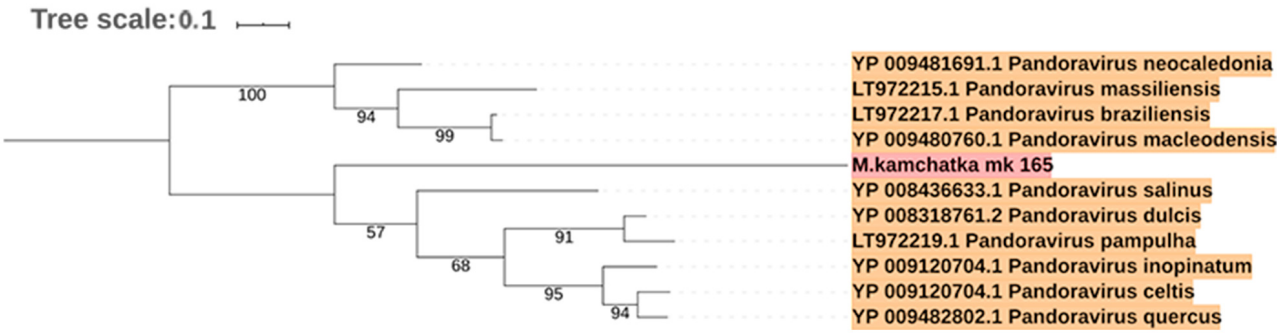




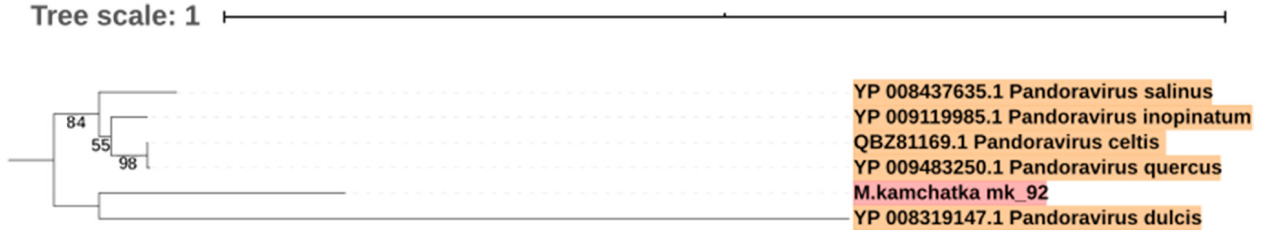


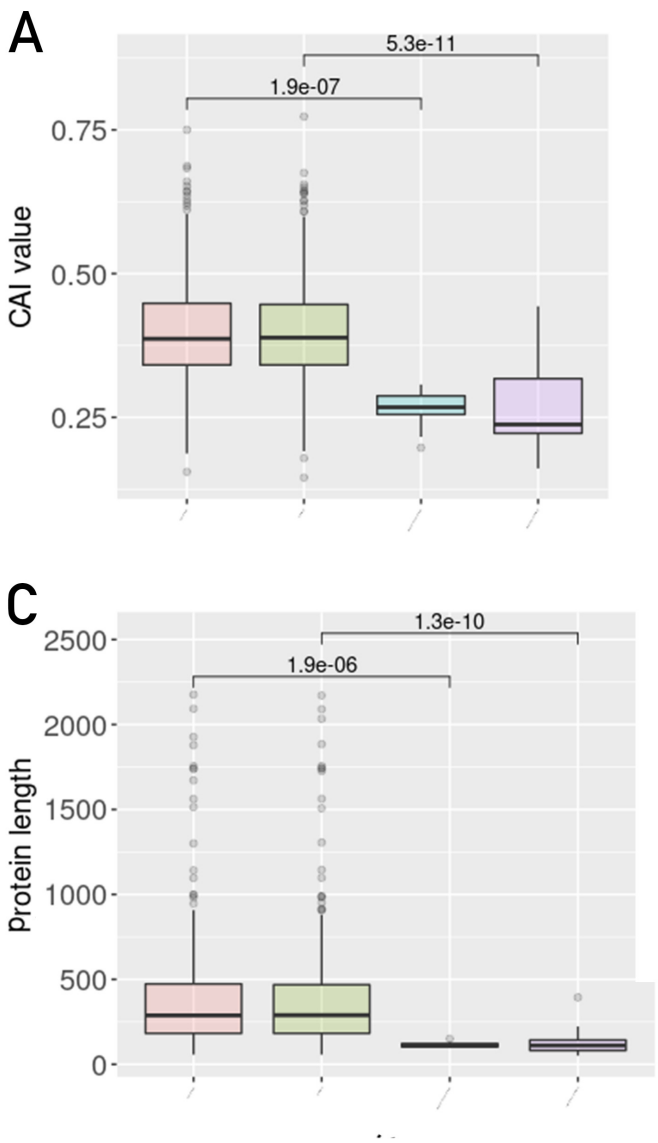


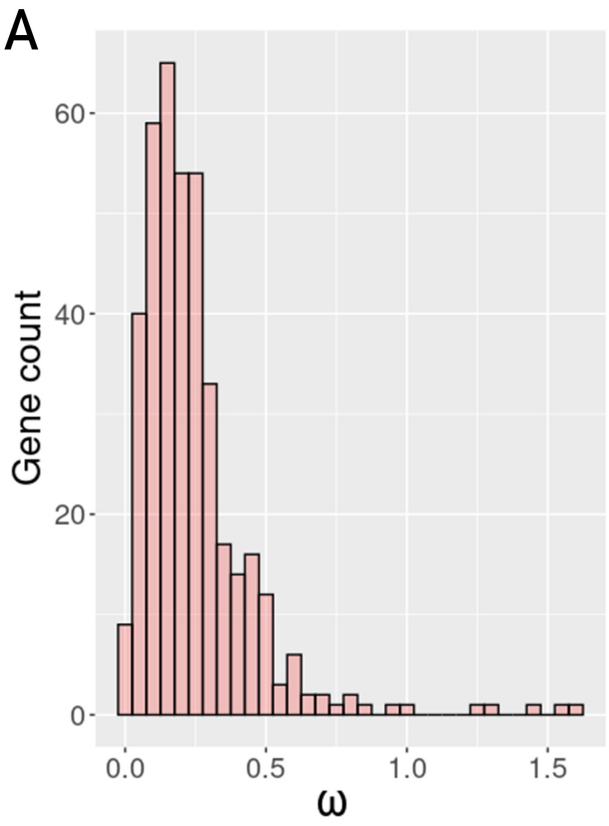
A

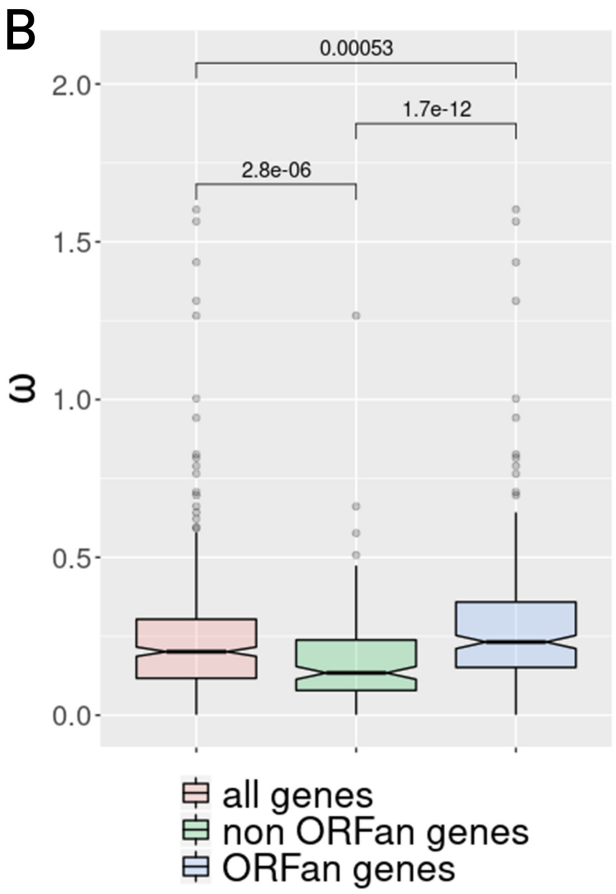


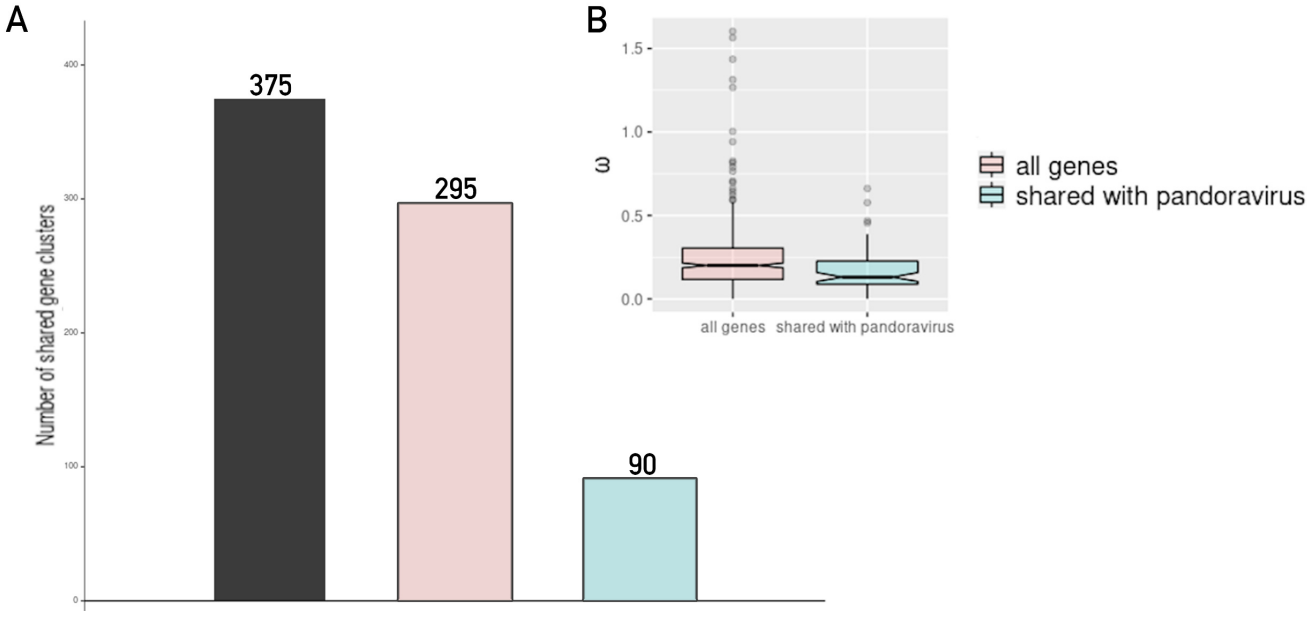
B

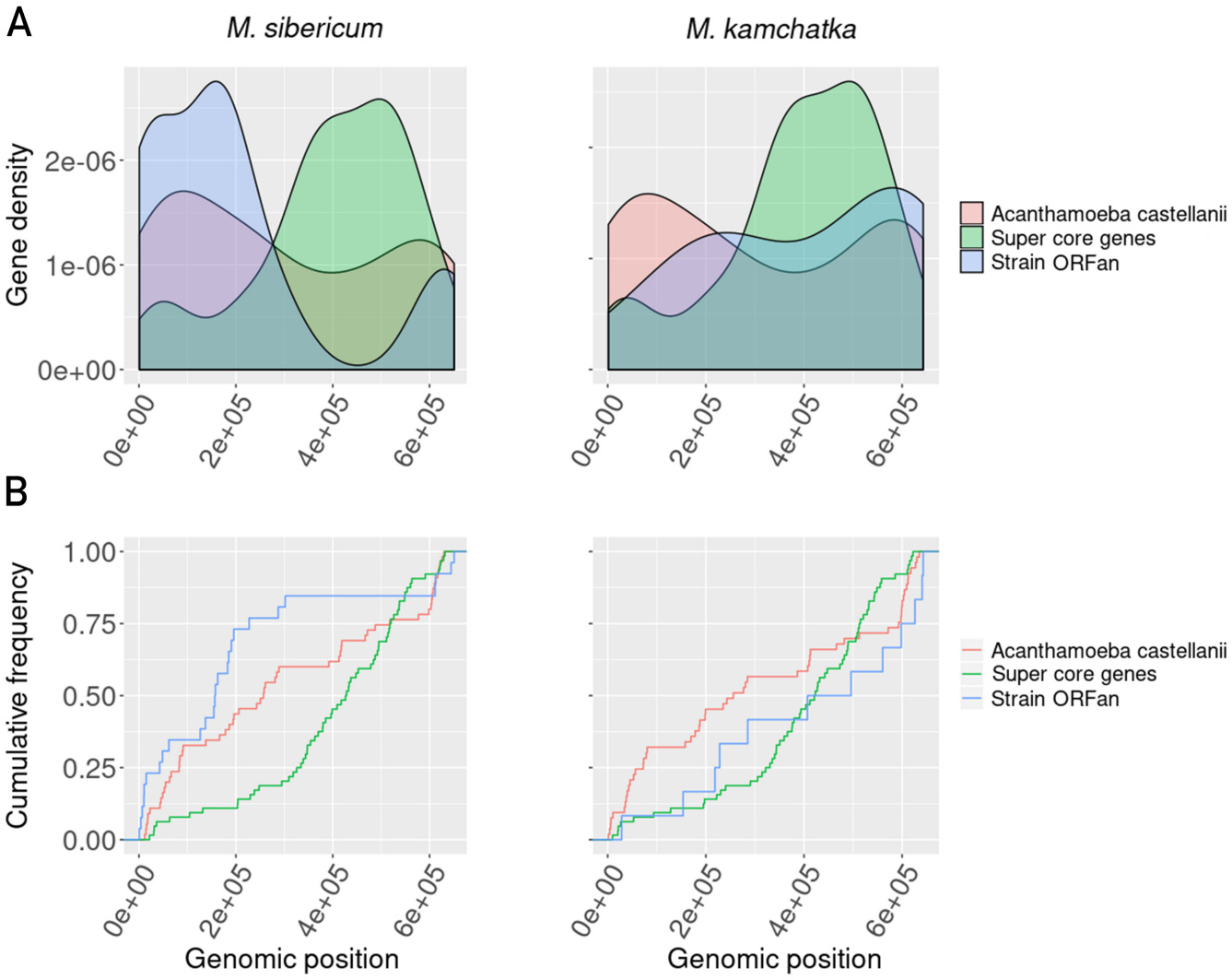


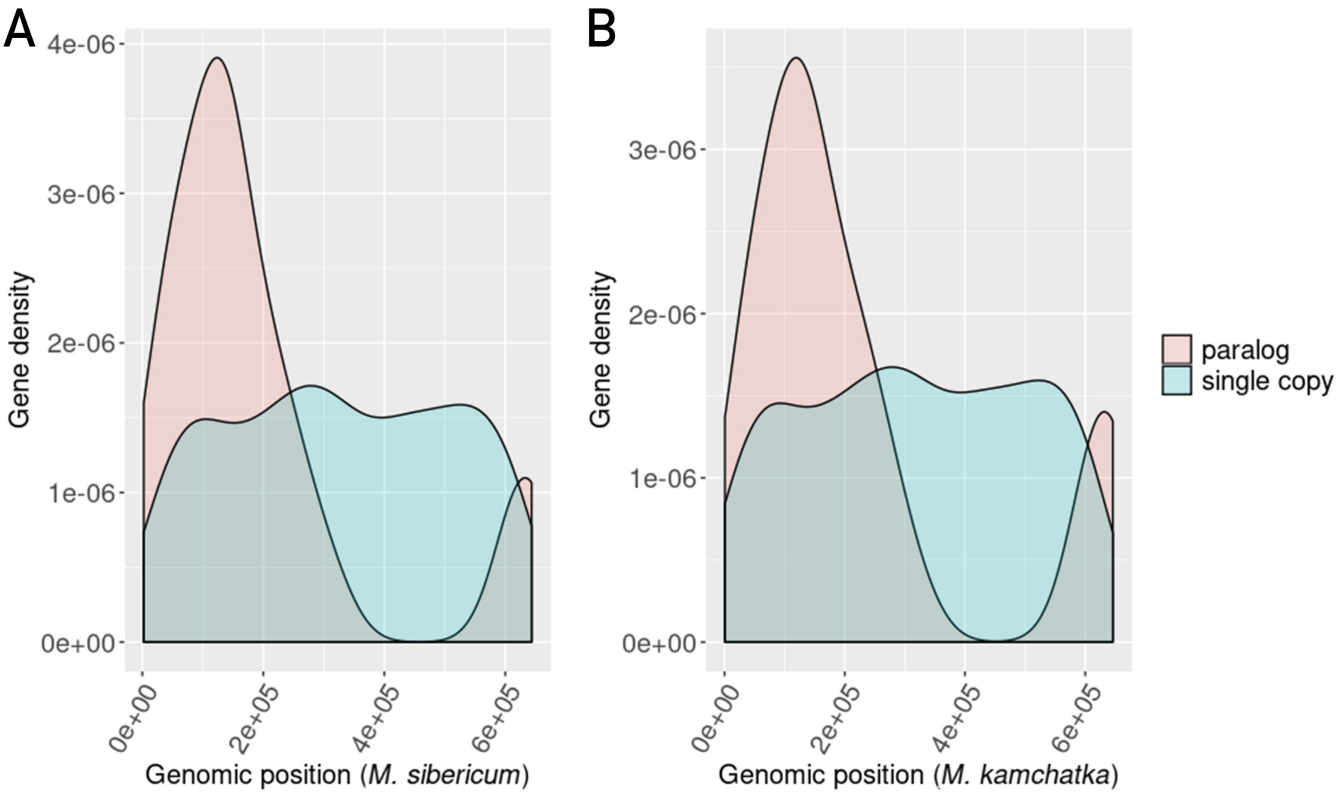












CONCLUSION

Le bref travail historiographique présenté en introduction montre que l'avènement de la virologie comme domaine d'étude est intrinsèquement lié à l'évolution de la microbiologie, et plus globalement de la Biologie dans son ensemble. Cette lecture non linéaire de l'Histoire de la virologie illustre, à mon sens, qu'à chaque instant donné, les connaissances scientifiques permettent d'établir des cadres théoriques, conditionnant la perception de la réalité. Le dépassement perpétuel de ces dogmes scientifiques amène donc à considérer chaque théorie, dans le pire des cas, comme « jamais vraie », et dans le meilleur des cas comme « pas encore fausse ». Les raisons qui font de la Science un processus de ruptures cyclique avec des dogmes peuvent s'expliquer dans notre cas par plusieurs facteurs. Premièrement, les découvertes n'émergent pas d'un seul individu visionnaire mais d'un large débat contradictoire mêlant un travail collectif non concerté, additionnant ajustements de théories passées et validation d'hypothèses avant-gardistes. En pleine hégémonie de la théorie des miasmes on constate par exemple que l'intuition de Girolamo Fracastoro appelait à remettre en cause ce dogme, sans permettre son dépassement à ce moment là. De même, la découverte des agents filtrants juste après les travaux de Pasteur sur la génération spontanée n'a pas entrainé un retour en arrière sur le plan conceptuel. Deuxièmement, il faut voir dans l'évolution des outils techniques (microscope, optique puis électronique, la centrifugation...) une ouverture permanente vers de nouvelles pratiques expérimentales. Plus généralement, le développement des techniques et des outils d'observations façonnent de nouvelles interactions avec le phénomène à caractériser, amenant donc une extension du champ des résultats possibles. Troisièmement, comment la communauté scientifique aurait-elle pu systématiser la recherche de virus dans des fractions inférieures à 0,2mm si Chamberland n'avait pas répondu aux politiques hygiénistes de son époque? Il semble ainsi difficile d'ignorer la pratique sociale et son importance dans l'émergence de nouveaux concepts. Si Ivanovsky n'a pas mesuré l'importance de ses travaux, il convient également de supposer que la pratique scientifique relève autant de l'expérience individuelle que collective. La découv*erte de Mim*ivirus illustre ce processus à plusieurs égards :

- Les virus ne sont pas des « agents filtrants » et il n'y a, à priori, pas de limite à la taille de leur génome et de leur capside.
- Les virus peuvent coder pour une partie de la machinerie de traduction
- Les capsides virales peuvent empaqueter de l'information génétique sous forme d'ADN et d'ARN

Ces constats, permis entre autre par les avancées des techniques de séquençage, ont donc permis de repousser les limites de la définition des virus faite par Lwoff il y a 70 ans.

Le paradoxe de ce manuscrit est d'avoir été partiellement rédigé durant la pandémie de SARS-CoV-2. Bien que ce virus soit très loin de ceux qui ont été décrits dans ce manuscrit, cette situation de pandémie m'a conduit à mener sur les réseaux sociaux, cadres privilégiés d'appréhension du monde en tant que réalité sociale, une brève expérience. Sous la forme d'un questionnaire à choix multiples anonyme, les 152 sondés tirés au hasard ont majoritairement associé les virus à des entités infectieuses (96%) pathogènes pour l'être humain (87%). A la question « les virus sont-ils vivants ? » 33% des sondés, à part égale avec ceux qui estiment que les virus sont « vivants », estiment que les virus sont des entités vivantes sans pour autant les associer à des « organismes vivants », une minorité (11%) estime que les virus sont des éléments inertes. Pour finir, 62% des sondés estiment que répondre à la question « les virus sont-ils vivants ? » nécessite un consensus scientifique sur la définition de « vivant ». La première conclusion de cette rapide enquête est que l'épidémie de SARS-CoV-2 objective la « virologie » pour le grand public de façon parcellaire et passéiste, loin des récentes découvertes faites dans le champ de la virologie environnementale. Dans ce cadre d'appréhension de la réalité, le « virologue » est réduit à son rôle historique d'investigateur des entités infectieuses et pathogènes dont l'objectif vise à soigner, contrôler, voire éradiquer les maladies virales. La deuxième conclusion est que l'étude des virus soulève des questions à la fois conceptuelles et philosophiques. D'une part, il apparaît nécessaire que les débats scientifiques autour des questions « Que sont les virus, sont-il vivants ? » et « Quelle est la place des virus dans le monde biologique ? » restent vivaces. C'est l'avancée des réflexions, des expériences et des débats contradictoires autour de ces questionnements qui permettront de dépasser les barrières épistémologiques entourant les problématiques de définition des virus et de leur rôle biologique.

Modestement, l'étude comparée de deux virus, l'un ancien et l'autre récent, *M. sibericum* et *M. kamchatka,* a contribué à alimenter le débat contradictoire portant sur l'origine évolutive des virus géants. La théorie dominante, directement issue de la vision historique des virus comme entités extérieures au reste du monde biologique, pourrait se résumer de la façon suivante : les virus géants sont originairement des petits virus dont le génome aurait « gonflé » par acquisition de gènes via des HGT avec leurs hôtes cellulaires. L'observation de 60% d'ORFans chez mollivirus vient donc en contradiction directe avec cette approche expansionniste [106]. Si il est vrai que les *Molliviridae* constituent l'une des famille virale ayant le plus large répertoire de gènes partagés avec son hôte, le faible nombre de gènes cellulaires homologues en dehors *d'A. castellanii* ainsi que la

caractérisation de transferts de gènes dans le sens « virus vers hôte » (comme la MCP) repousse encore d'avantage les limites de cette hypothèse évolutive orthodoxe. La comparaison entre *M. sibericum* e t *M. kamchatka* a également permis d'estimer que cette famille virale mutait probablement trop lentement pour envisager que la signature des gènes cellulaires potentiels ait pu disparaître. Enfin, il apparaît que les liens entre virus et hôtes ne semblent pas se résumer à un ensemble de HGT entre les deux entités. En effet, la caractérisation d'un échange de methyltransférase entre *M. kamchatka* et les pandoravirus, et une étude récente du méthylome des virus géants menée au laboratoire et ayant conclu que les méthyltransferases virales ont toutes une origine bactérienne, suggère que l'hôte *A. castellanii* constitue, en plus d'un *melting pot* d'ADN procaryote, un lieu privilégié d'échanges de gènes viraux [107].

L'étude de l'histoire évolutive des *Pandoraviridae* et des *Molliviridae* soulève également des questions intéressantes relatives à l'origine des virus géants. La présence d'un génome cœur partagé souligne une origine commune entre les deux familles virales. Force est de constater que le génome des *Molliviridae* est plus petit que celui de leurs cousins, les *Pandoraviridae.* Ainsi, certains gènes fondamentaux au cycle viral des pandoravirus ont disparu chez mollivirus, entrainant *de facto* une dépendance plus grande de ces derniers à leur cellule hôte. Ainsi, il est probable d'envisager que les *Molliviridae* et les *Pandoraviridae* aient pu avoir un ancêtre commun plus complexe que les représentants actuels de ces deux familles virales. Cette idée renverse donc l'hypothèse purement expansionniste à l'origine des virus géants et suggère qu'au contraire, ces derniers aient pu dériver d'un ancêtre plus complexe par réduction évolutive. Ce constat est renforcé par le maintien du gène de la MCP chez les *Molliviridae.* Ainsi, le scénario proposé par l'étude de ce gène pourrait suggérer qu'il s'agit là d'une trace d'un ancêtre commun aux deux familles virales et possédant une capside virale icoasédrique. Bien que le mécanisme évolutif ayant conduit à l'acquisition d'une capside virale composée d'un tégument lamellaire tel qu'observé aujourd'hui pour les capsides de mollivirus et de pandoravirus soit largement incompris, il est envisageable que la structure sphérique des mollivirus soit un état transitoire vers la structure en amphore des pandoravirus. Au-delà des virus géants, l'exemple des bracovirus de la famille des *Polydnaviridae*, symbiontes obligatoires de guêpes parasites de Lépidoptères, illustre parfaitement la théorie réductionniste. Il a été démontré qu'un ancêtre des nudivirus, virus à ADN double brin codant pour 98 à 154 protéines, avait probablement intégré son génome dans une guêpe il y a environ 100 millions d'années. Le génome des bracovirus étant donc le résultat d'une longue co-évolution entre la guêpe et le virus ayant conduit à une organisation bi-partite du génome de bracovirus, ou les gènes viraux homologues aux gènes de nudivirus sont portés par la guêpe tandis que le génome embarqué dans les particules virales porte des gènes de guêpe [108]. Ce constat, d'un passage d'un génome complexe à un état de

dépendance total à l'hôte n'est pas unique au monde viral et peut être étendu à l'ensemble des entités biologiques parasitant une cellule hôte. Ainsi, dans le monde bactérien il existe un grand nombre d'exemple de bactéries intra-cellulaires ayant subi au cours de l'évolution une réduction du nombre de leurs gènes. L'exemple le plus probant est le génome du parasite obligatoire *Mycoplasma genitallium* [109]. Il a été démontré que parmi les 485 gènes codés par ce génome de 0,6 Mpb (plus petite bacterie cultivable connue alors), 100 étaient non essentiels [110].

Lorsqu'on considère que les virus peuvent avoir dérivés du monde cellulaire par pertes successives d'ensembles de gènes associés à divers fonctions, il convient d'observer qu'une partie de la réponse à la question « Que sont les virus, sont-ils vivants ? » réside dans la notion de « parasitisme obligatoire ». Envisager cette question sous ce prisme permet de sortir l'habitude intellectuelle qui pousserait à chercher une origine aux virus dans ce qu'ils ont « en commun » avec le monde cellulaire, mais plutôt dans ce qu'ils « n'ont pas » en commun [106]. Tout en ne niant pas les fonctions communes aux NCLDV, cet angle d'approche permet de justifier la co-existence de diverses stratégies de réplication virale ainsi que de la diversité de capside observée au sein des NCLDV. De même, envisager l'évolution des virus sous ce prisme est consistant avec le faible nombre de gènes partagés par l'ensemble des familles de NCLDV. L'étude des virus géants au travers de ce prisme peu orthodoxe illustre parfaitement la nécessité d'une gradation dans la notion de parasitisme : d'un côté certains virus se répliquent intégralement dans le cytoplasme (*Mimiviridae* e t *Pithoviridae*), de l'autre, certains virus sont entièrement dépendants de la machinerie nucléaire d'*A. castellanii* (*Pandoraviridae* et mollivirus). A l'intermédiaire entre ces deux modes de réplication on retrouve les *Marseilleviridae* dont le cycle infectieux mobilise de façon transitoire la machinerie nucléaire de l'amibe. Cette gradation dans le parasitisme, se traduit par des répertoires de gènes spécifiques à chaque famille virale. Ainsi, certaines familles codent pour une partie de la machinerie de traduction, comme les *Mimiviridae* tandis-que certaines ne possédant pas de protéines clefs de la biosynthèse des nucléotides (*Molliviridae*). Il semble donc que ces différentes familles de virus géants sont le résultat de trajectoires évolutives différentes au cours desquelles différentes fonctions ont pu être perdues. Malgré ce constat, il convient cependant de rappeler que l'absence de voies de production d'ATP est un dénominateur commun à l'ensemble du monde viral. Par contre, si on considère que cette caractéristique n'est qu'une des spécificités possibles à la condition de parasite strict, à nouveau, en se penchant sur le monde « vivant », on observe des cas de bactéries ne produisant pas d'ATP, comme *Carsonella ruddii*. Cet endosymbionte de cellules d'insectes phlomophage, est à peu de chose près en passe de remplir le rôle d'un organite membranaire, faisant de l'absence de protéines du métabolisme énergétique une des caractéristiques des parasites les plus dépendants de leur hôte et non *stricto sensu* des virus [111].

Si on peut admettre, à ce stade de la conclusion, que les virus sont des entités biologiques dont la condition de parasite les a amenés à dépendre plus ou moins de leur hôte, se pose désormais la question de peuvent-ils être qualifiés de « vivant » ? Claude Bernard, dans la première des *Lecons sur les phénomènes de la vie communs aux animaux et aux végétaux*, répond « Il suffit que l'on s'entende sur le mot vie pour l'employer (…) il est illusoire et chimérique, contraire à l'esprit même de la science, d'en chercher une définition absolue ». Dès lors, si on considére que la vie est un processus, au même titre que les gamètes portent l'information génétique nécessaire à la constitution d'un organisme complexe « vivant » alors, à plusieurs égards le cycle infectieux des virus est un processus vivant. L'étude des virus géants a largement contribué à faire évoluer cette vision du monde « vivant ». Ainsi, la capside virale de *Mimivirus* apparaît comme une boîte inerte permettant la dissémination de l'information génétique, tandis que l'usine virale, regroupant l'ensemble du matériel nécessaire à l'expression du génome viral (largement codé par le virus), constitue la phase vivante du virus. Cette analogie qui de prime abord apparaît grossière, n'est pas sans fondement matériel. De récentes études sur les usines virales de *Mononegavirales* ont permis de démontrer que ces dernières se formaient par séparation de phase entre le cytoplasme des cellules infectées et la suspension de protéines virales [112]. Ces caractéristiques physico-chimiques propres aux usines virales permettent de concentrer les protéines virales nécessaires à la production des virions ainsi qu'à échapper aux mécanismes de défense de la cellule hôte. En ce sens, émerge donc l'idée que les usines virales peuvent être considérées comme des organites liquides autonomes.

L'étude de l'origine des virus géants ne peut cependant pas se résumer à une opposition entre théorie expansionniste et réductionniste. En effet, bien que les *Pandoraviridae* et les *Molliviridae* puissent être considérés comme les virus les plus dépendants de leur hôte, l'importante taille du génome des pandoravirus est en contradiction avec une théorie purement réductionniste. Comme exposé dans ce manuscrit, nous avons vu que le mécanisme de création de gènes *de novo* pouvait probablement expliquer la complexité de ces génomes. Cette découverte questionne donc le rôle des virus géants dans le monde biologique. La création de gènes viraux *de novo*, phénomène pour l'instant unique aux virus géants, ouvre la possibilité que les virus pourraient constituer un réservoir de diversité génétique à l'échelle de l'ensemble du monde vivant [113]. Ainsi, la découverte de *M. kamchatka,* la définition des *Molliviridae* comme nouvelle famille virale et l'étude des liens évolutifs avec les pandoravirus contribue à alimenter le débat autour de la question de l'origine des virus géants et plus généralement de l'origine et du rôle des virus au sein du monde vivant.

Comme nous l'avons vu en introduction en abordant les différents points de vue sur la nature des virus durant la période « héroïque » de la virologie, c'est la pratique du débat scientifique contradictoire qui permet de franchir des barrières épistémologiques. Les questions soulevées par la

découverte des virus géants, et la re-définition nécessaire des virus, n'échappe pas à ce processus. Ainsi à la fin de l'écriture de ce manuscrit, en juin 2020, a été publiée une proposition de classification du monde viral en partant de la classification de Baltimore [114]. Les virus de la classe I de Baltimore, qui regroupe les virus à ADN double brin et donc les virus géants, ont été regroupés pour cette étude. Bien que cette classe de virus ne partage aucun gène en commun, les différentes familles virales ont été séparées grâce à un réseau de gènes partagés. Les virus géants tombent, dans ce cadre, dans le « super-groupe » des virus à ADN codant pour une protéine de capside ayant un motif *double jelly roll*. La reconstruction de la phylogénie de ce « super-groupe » donne les virus géants comme des descendants d'éléments génétiques mobiles (polintons, polintons-*like*). Cette phylogénie sous-entend donc que les virus géants sont issus de virus plus petits par accumulation de gènes par HGT. Cette proposition peut paraître paradoxale dans la mesure où les *Pandoraviridae* ni les *Pithoviridae* semblent avoir perdu la MCP. De façon surprenante, cette démarche et cette classification ont été validées par l'ICTV, menant donc à une réorganisation de la classification approuvée par l'ICTV en septembre.

BIBLIOGRAPHIE

1. Grmek, M. D. Histoires de la virologie, des viroses et des virologues. *History and Philosophy of the Life Sciences* **16**, 339–354 (1994).
2. Byl, S. Mirko D. Grmek (Dir.), Histoire de la pensée médicale en Occident. 1. Antiquité et Moyen-Age. *L'Antiquité Classique* **65**, 411–412 (1996).
3. Pasteur, L. (1822-1895) A. du texte, Joubert, J. (1834-1910) A. du texte & Chamberland, C. (1851-1908) A. du texte. *La théorie des germes et ses applications à la médecine et à la chirurgie / lecture faite à l'Académie de médecine par M. Pasteur en son nom et au nom de MM. Joubert et Chamberland, le 30 avril 1878*. (1878).
4. Harden, V. A. Koch's postulates and the etiology of AIDS: an historical perspective. *Hist Philos Life Sci* **14**, 249–269 (1992).
5. 1898 -The Beginning of Virology...time marches on. *1898 -The Beginning of Virology...time marches on.* https://www.apsnet.org/edcenter/apsnetfeatures/Pages/BeginningofVirology.aspx.
6. Bos, L. Beijerinck's work on tobacco mosaic virus: historical context and legacy. *Philos Trans R Soc Lond B Biol Sci* **354**, 675–685 (1999).
7. Stanley, W. M. ISOLATION OF A CRYSTALLINE PROTEIN POSSESSING THE PROPERTIES OF TOBACCO-MOSAIC VIRUS. *Science* **81**, 644–645 (1935).
8. Kay, L. E. W. M. Stanley's crystallization of the tobacco mosaic virus, 1930-1940. *Isis* **77**, 450–472 (1986).
9. Virus as Organism. *Nature* **157**, 174–174 (1946).
10. Hirst, G. K. THE QUANTITATIVE DETERMINATION OF INFLUENZA VIRUS AND ANTIBODIES BY MEANS OF RED CELL AGGLUTINATION. *J Exp Med* **75**, 49–64 (1942).
11. Morange, M. What history tells us III. André Lwoff: From protozoology to molecular definition of viruses. *J. Biosci.* **30**, 591–594 (2005).
12. Lwoff, A. Principles of Classification and Nomenclature of Viruses. *Nature* **215**, 13–14 (1967).
13. International Committee on Taxonomy of Viruses (ICTV). https://talk.ictvonline.org/taxonomy/.
14. Williams, T., Barbosa- Solomieu, V. & Chinchar, V. G. A Decade of Advances in Iridovirus

Research. in *Advances in Virus Research* vol. 65 173–248 (Academic Press, 2005).

15. Oliveira, G. P., Rodrigues, R. A. L., Lima, M. T., Drumond, B. P. & Abrahão, J. S. Poxvirus Host Range Genes and Virus–Host Spectrum: A Critical Review. *Viruses* **9**, 331 (2017).
16. Kieser, Q., Noyce, R. S., Shenouda, M., Lin, Y.-C. J. & Evans, D. H. Cytoplasmic factories, virus assembly, and DNA replication kinetics collectively constrain the formation of poxvirus recombinants. *PLoS ONE* **15**, e0228028 (2020).
17. Burrell, C. J., Howard, C. R. & Murphy, F. A. Chapter 16 - Poxviruses. in *Fenner and White's Medical Virology (Fifth Edition)* (eds. Burrell, C. J., Howard, C. R. & Murphy, F. A.) 229–236 (Academic Press, 2017). doi:10.1016/B978-0-12-375156-0.00016-3.
18. Moss, B. Poxvirus DNA Replication. *Cold Spring Harb Perspect Biol* **5**, (2013).
19. Van Etten, J. L., Graves, M. V., Müller, D. G., Boland, W. & Delaroque, N. Phycodnaviridae–large DNA algal viruses. *Arch. Virol.* **147**, 1479–1516 (2002).
20. Milrot, E. *et al.* Virus–host interactions: insights from the replication cycle of the large Paramecium bursaria chlorella virus. *Cellular Microbiology* **18**, 3–16 (2016).
21. McKeown, D. A. *et al.* Phaeoviruses discovered in kelp (Laminariales). *The ISME Journal* **11**, 2869–2873 (2017).
22. Wilson, W. H., Van Etten, J. L. & Allen, M. J. The Phycodnaviridae: The Story of How Tiny Giants Rule the World. in *Lesser Known Large dsDNA Viruses* (ed. Van Etten, J. L.) 1–42 (Springer, 2009). doi:10.1007/978-3-540-68618-7_1.
23. Normile, D. African swine fever marches across much of Asia. *Science* **364**, 617–618 (2019).
24. Salas, M. L. & Andrés, G. African swine fever virus morphogenesis. *Virus Research* **173**, 29–41 (2013).
25. Klose, T. *et al.* Structure of faustovirus, a large dsDNA virus. *Proceedings of the National Academy of Sciences of the United States of America* **113**, 6206–6211 (2016).
26. Bajrai, L. H. *et al.* Kaumoebavirus, a New Virus That Clusters with Faustoviruses and Asfarviridae. *Viruses* **8**, (2016).
27. Andreani, J. *et al.* Pacmanvirus, a New Giant Icosahedral Virus at the Crossroads between Asfarviridae and Faustoviruses. *Journal of Virology* **91**, (2017).
28. A novel Asfarvirus-like virus identified as a potential cause of mass mortality of abalone |

Scientific Reports. https://www.nature.com/articles/s41598-020-61492-3.

29. Properties of three iridovirus-like agents associated with systemic infections of fish. https://www.cabi.org/ISC/abstract/19922274750 (1992).
30. The organization of frog virus 3 as revealed by freeze-etching. *Virology* **138**, 287–299 (1984).
31. Zaghloul, H. A. H. M. Transcriptome Analyses of Ascovirus Genome Expression in Lepidopteran Larvae and Host Responses. (UC Riverside, 2018).
32. Cheng, X.-W., Wan, X.-F., Xue, J. & Moore, R. C. Ascovirus and its evolution. *Virol. Sin.* **22**, 137 (2008).
33. Virus evolution: how far does the double β-barrel viral lineage extend? | Nature Reviews Microbiology. https://www.nature.com/articles/nrmicro2033.
34. Raoult, D., Scola, B. L. & Birtles, R. The Discovery and Characterization of Mimivirus, the Largest Known Virus and Putative Pneumonia Agent. *Clin Infect Dis* **45**, 95–102 (2007).
35. Raoult, D. *et al.* The 1.2-Megabase Genome Sequence of Mimivirus. *Science* **306**, 1344–1350 (2004).
36. Ogata, H. & Claverie, J.-M. Unique genes in giant viruses: regular substitution pattern and anomalously short size. *Genome Res.* **17**, 1353–1361 (2007).
37. Ghedin, E. & Claverie, J.-M. Mimivirus relatives in the Sargasso sea. *Virol. J.* **2**, 62 (2005).
38. Viruses of the eukaryotic plankton are predicted to increase carbon export efficiency in the global sunlit ocean | bioRxiv. https://www.biorxiv.org/content/10.1101/710228v1.
39. Lunardini, V. J. Climatic warming and the degradation of warm permafrost. *Permafrost and Periglacial Processes* **7**, 311–320 (1996).
40. Abrupt changes of thermokarst lakes in Western Siberia: impacts of climatic warming on permafrost melting: International Journal of Environmental Studies: Vol 66, No 4. https://www.tandfonline.com/doi/abs/10.1080/00207230902758287.
41. Shi, T., Reeves, R. H., Gilichinsky, D. A. & Friedmann, E. I. Characterization of Viable Bacteria from Siberian Permafrost by 16S rDNA Sequencing. *Microb Ecol* **33**, 169–179 (1997).
42. Vishnivetskaya, T., Kathariou, S., McGrath, J., Gilichinsky, D. & Tiedje, J. M. Low-temperature recovery strategies for the isolation of bacteria from ancient permafrost sediments.

Extremophiles **4**, 165–173 (2000).

43. Yashina, S. *et al.* Reply to Oxelman et al.: On the taxonomic status of the plants regenerated from 30,000-y-old fruit tissue buried in Siberian permafrost. *PNAS* **109**, E2736–E2736 (2012).
44. Biagini, P. *et al.* Variola virus in a 300-year-old Siberian mummy. *New England Journal of Medicine* **367**, 2057–9 (2012).
45. Hueffer, K., Drown, D., Romanovsky, V. & Hennessy, T. Factors Contributing to Anthrax Outbreaks in the Circumpolar North. *EcoHealth* **17**, 174–180 (2020).
46. Kashuba, E. *et al.* Ancient permafrost staphylococci carry antibiotic resistance genes. *Microbial Ecology in Health and Disease* **28**, 1345574 (2017).
47. Abergel, C., Legendre, M. & Claverie, J.-M. The rapidly expanding universe of giant viruses: Mimivirus, Pandoravirus, Pithovirus and Mollivirus. *FEMS Microbiol. Rev.* **39**, 779–796 (2015).
48. Aherfi, S., Colson, P., Scola, B. L. & Raoult, D. Giant Viruses of Amoebas: An Update. *Frontiers in Microbiology* **7**, (2016).
49. Arslan, D., Legendre, M., Seltzer, V., Abergel, C. & Claverie, J.-M. Distant Mimivirus relative with a larger genome highlights the fundamental features of Megaviridae. *Proc. Natl. Acad. Sci. U.S.A.* **108**, 17486–17491 (2011).
50. Deeg, C. M., Chow, C.-E. T. & Suttle, C. A. The kinetoplastid-infecting Bodo saltans virus (BsV), a window into the most abundant giant viruses in the sea. *eLife* **7**, e33014 (2018).
51. Abrahão, J. *et al.* Tailed giant Tupanvirus possesses the most complete translational apparatus of the known virosphere. *Nature Communications* **9**, 749 (2018).
52. Mutsafi, Y., Shimoni, E., Shimon, A. & Minsky, A. Membrane Assembly during the Infection Cycle of the Giant Mimivirus. *PLOS Pathogens* **9**, e1003367 (2013).
53. Piacente, F. *et al.* The rare sugar N-acetylated viosamine is a major component of Mimivirus fibers. *J. Biol. Chem.* **292**, 7385–7394 (2017).
54. Piacente, F. *et al.* Characterization of a UDP-N-acetylglucosamine biosynthetic pathway encoded by the giant DNA virus Mimivirus. *Glycobiology* **24**, 51–61 (2014).
55. Abrahão, J. S. *et al.* Acanthamoeba polyphaga mimivirus and other giant viruses: an open field to outstanding discoveries. *Virol J* **11**, 120 (2014).

56. Zauberman, N. *et al.* Distinct DNA Exit and Packaging Portals in the Virus Acanthamoeba polyphaga mimivirus. *PLOS Biology* **6**, e114 (2008).
57. Fridmann-Sirkis, Y. *et al.* Efficiency in Complexity: Composition and Dynamic Nature of Mimivirus Replication Factories. *Journal of Virology* **90**, 10039–10047 (2016).
58. Mutsafi, Y., Zauberman, N., Sabanay, I. & Minsky, A. Vaccinia-like cytoplasmic replication of the giant Mimivirus. *PNAS* **107**, 5978–5982 (2010).
59. Kuznetsov, Y. G., Klose, T., Rossmann, M. & McPherson, A. Morphogenesis of Mimivirus and Its Viral Factories: an Atomic Force Microscopy Study of Infected Cells. *Journal of Virology* **87**, 11200–11213 (2013).
60. La Scola, B. *et al.* The virophage as a unique parasite of the giant mimivirus. *Nature* **455**, 100–104 (2008).
61. Claverie, J.-M. & Abergel, C. Mimivirus and its Virophage. *Annu. Rev. Genet.* **43**, 49–66 (2009).
62. Jeudy, S. *et al.* Exploration of the propagation of transpovirons within Mimiviridae reveals a unique example of commensalism in the viral world. *ISME J* **14**, 727–739 (2020).
63. Arantes, T. S. *et al.* The Large Marseillevirus Explores Different Entry Pathways by Forming Giant Infectious Vesicles. *Journal of Virology* **90**, 5246–5255 (2016).
64. Fabre, E. *et al.* Noumeavirus replication relies on a transient remote control of the host nucleus. *Nat Commun* **8**, 15087 (2017).
65. Thomas, V. *et al.* Lausannevirus, a giant amoebal virus encoding histone doublets. *Environmental Microbiology* **13**, 1454–1466 (2011).
66. Erives, A. J. Phylogenetic analysis of the core histone doublet and DNA topo II genes of Marseilleviridae: evidence of proto-eukaryotic provenance. *Epigenetics & Chromatin* **10**, 55 (2017).
67. Scheid, P., Zöller, L., Pressmar, S., Richard, G. & Michel, R. An extraordinary endocytobiont in Acanthamoeba sp. isolated from a patient with keratitis. *Parasitol Res* **102**, 945–950 (2008).
68. Scheid, P., Balczun, C. & Schaub, G. A. Some secrets are revealed: parasitic keratitis amoebae as vectors of the scarcely described pandoraviruses to humans. *Parasitol Res* **113**, 3759–3764 (2014).

69. Legendre, M. *et al.* Thirty-thousand-year-old distant relative of giant icosahedral DNA viruses with a pandoravirus morphology. *PNAS* **111**, 4274–4279 (2014).
70. Levasseur, A. *et al.* Comparison of a Modern and Fossil Pithovirus Reveals Its Genetic Conservation and Evolution. *Genome Biol Evol* **8**, 2333–2339 (2016).
71. Andreani, J. *et al.* Cedratvirus, a Double-Cork Structured Giant Virus, is a Distant Relative of Pithoviruses. *Viruses* **8**, 300 (2016).
72. Frontiers | Orpheovirus IHUMI-LCC2: A New Virus among the Giant Viruses | Microbiology. https://www.frontiersin.org/articles/10.3389/fmicb.2017.02643/full.
73. Michel, R., Müller, K.-D., Schmid, E., Zöller, L. & Hoffmann, R. Endocytobiont KC5/2 induces transformation into sol-like cytoplasm of its host Acanthamoeba sp. as substrate for its own development. *Parasitol Res* **90**, 52–56 (2003).
74. Scheid, P. A strange endocytobiont revealed as largest virus. *Current Opinion in Microbiology* **31**, 58–62 (2016).
75. Legendre, M. *et al.* 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 (2015).
76. Maumus, F. & Blanc, G. Study of Gene Trafficking between Acanthamoeba and Giant Viruses Suggests an Undiscovered Family of Amoeba-Infecting Viruses. *Genome Biology and Evolution* **8**, 3351 (2016).
77. Malavin, S. & Shmakova, L. Isolates from ancient permafrost help to elucidate species boundaries in Acanthamoeba castellanii complex (Amoebozoa: Discosea). *bioRxiv* 755348 (2019) doi:10.1101/755348.
78. Siddiqui, R. & Khan, N. A. Biology and pathogenesis of Acanthamoeba. *Parasites & Vectors* **5**, 6 (2012).
79. G, T. & Ea, J. Isolation of cellulose from the cyst wall of a soil amoeba. *Biochimica et biophysica acta* vol. 63 https://pubmed.ncbi.nlm.nih.gov/13985444/ (1962).
80. Dornas, F. P. *et al.* Isolation of new Brazilian giant viruses from environmental samples using a panel of protozoa. *Frontiers in Microbiology* **6**, (2015).
81. Seligman, A. M., Wasserkrug, H. L. & Hanker, J. S. A NEW STAINING METHOD (OTO) FOR ENHANCING CONTRAST OF LIPID-CONTAINING MEMBRANES AND

DROPLETS IN OSMIUM TETROXIDE-FIXED TISSUE WITH OSMIOPHILIC THIOCARBOHYDRAZIDE (TCH). *The Journal of Cell Biology* **30**, 424 (1966).

82. Bankevich, A. *et al.* SPAdes: A New Genome Assembly Algorithm and Its Applications to Single-Cell Sequencing. *Journal of Computational Biology* **19**, 455–477 (2012).
83. Wick, R. R., Judd, L. M., Gorrie, C. L. & Holt, K. E. Unicycler: Resolving bacterial genome assemblies from short and long sequencing reads. *PLOS Computational Biology* **13**, e1005595 (2017).
84. Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. https://genome.cshlp.org/content/27/5/722.short.
85. Li, H. Minimap and miniasm: fast mapping and de novo assembly for noisy long sequences. *Bioinformatics* **32**, 2103–2110 (2016).
86. Hoff, K. J. & Stanke, M. Predicting Genes in Single Genomes with AUGUSTUS. *Current Protocols in Bioinformatics* **65**, e57 (2019).
87. Besemer, J. & Borodovsky, M. GeneMark: web software for gene finding in prokaryotes, eukaryotes and viruses. *Nucleic Acids Res* **33**, W451–W454 (2005).
88. BRAKER1: Unsupervised RNA-Seq-Based Genome Annotation with GeneMark-ET and AUGUSTUS | Bioinformatics | Oxford Academic. https://academic-oup-com.insb.bib.cnrs.fr/bioinformatics/article/32/5/767/1744611.
89. Haas, B. J. *et al.* Automated eukaryotic gene structure annotation using EVidenceModeler and the Program to Assemble Spliced Alignments. *Genome Biol* **9**, R7 (2008).
90. Henschel, R. *et al.* Trinity RNA-Seq assembler performance optimization. in *Proceedings of the 1st Conference of the Extreme Science and Engineering Discovery Environment: Bridging from the eXtreme to the campus and beyond* 1–8 (Association for Computing Machinery, 2012). doi:10.1145/2335755.2335842.
91. Gs, S. & E, B. Automated generation of heuristics for biological sequence comparison. *BMC Bioinformatics* **6**, 31–31 (2005).
92. Bringans, S. *et al.* Deep proteogenomics; high throughput gene validation by multidimensional liquid chromatography and mass spectrometry of proteins from the fungal wheat pathogen Stagonospora nodorum. *BMC Bioinformatics* **10**, 301 (2009).

93. Marchler-Bauer, A. & Bryant, S. H. CD-Search: protein domain annotations on the fly. *Nucleic Acids Res* **32**, W327–W331 (2004).
94. HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment | Nature Methods. https://www.nature.com/articles/nmeth.1818.
95. Emms, D. M. & Kelly, S. OrthoFinder: phylogenetic orthology inference for comparative genomics. *Genome Biology* **20**, 238 (2019).
96. Li, Y. *et al.* Distinct Effects on Diversifying Selection by Two Mechanisms of Immunity against Streptococcus pneumoniae. *PLOS Pathogens* **8**, e1002989 (2012).
97. Katoh, K., Asimenos, G. & Toh, H. Multiple Alignment of DNA Sequences with MAFFT. in *Bioinformatics for DNA Sequence Analysis* (ed. Posada, D.) 39–64 (Humana Press, 2009). doi:10.1007/978-1-59745-251-9_3.
98. Suyama, M., Torrents, D. & Bork, P. PAL2NAL: robust conversion of protein sequence alignments into the corresponding codon alignments. *Nucleic Acids Res* **34**, W609–W612 (2006).
99. Yang, Z. *PAML: a program package for phylogenetic analysis by maximum likelihood*. (1997).
100. Xia, X. An Improved Implementation of Codon Adaptation Index. *Evol Bioinform Online* **3**, 117693430700300030 (2007).
101. Goncharov, A. E. *et al.* WHOLE-GENOME SEQUENCING AS A TOOL FOR COMPREHENSIVE ASSESSMENT OF THE PATHOGENIC POTENTIAL OF ANCIENT ARCTIC MICROBIOMES. *Инфекция и иммунитет* vol. 8 512–513 https://www.iimmun.ru/iimm/article/view/865 (2018).
102. Genetic structure and biological properties of the first ancient multiresistance plasmid pKLH80 isolated from a permafrost bacterium | Microbiology Society. https://www.microbiologyresearch.org/content/journal/micro/10.1099/mic.0.079335-0.
103. Kraken: ultrafast metagenomic sequence classification using exact alignments | Genome Biology | Full Text. https://genomebiology.biomedcentral.com/articles/10.1186/gb-2014-15-3-r46.
104. Association of mitochondria DNA with viral DNA in purified preparations of poxviruses.

Virus Research **11**, 165–174 (1988).

105. Hughes, A. L., Irausquin, S. & Friedman, R. The evolutionary biology of poxviruses. *Infection, Genetics and Evolution* **10**, 50–59 (2010).
106. Claverie, J.-M. & Abergel, C. Giant viruses: The difficult breaking of multiple epistemological barriers. *Stud Hist Philos Biol Biomed Sci* **59**, 89–99 (2016).
107. Jeudy, S. *et al.* The DNA methylation landscape of giant viruses. *Nat Commun* **11**, 1–12 (2020).
108. Dupuy, C., Periquet, G., Bézier, A. & Drezen, J.-M. Les polydnavirus : des virus qui pratiquent le transfert de gènes depuis 100 millions d'années. *Med Sci (Paris)* **26**, 125–127 (2010).
109. Microbial Minimalism: Cell. https://www-cell-com.insb.bib.cnrs.fr/cell/fulltext/S0092-8674(02)00665-7?_returnURL=https%3A%2F%2Flinkinghub.elsevier.com%2Fretrieve%2Fpii%2FS0092867402006657%3Fshowall%3Dtrue.
110. Designing minimal genomes using whole-cell models | Nature Communications. https://www-nature-com.insb.bib.cnrs.fr/articles/s41467-020-14545-0.
111. Tamames, J. *et al.* The frontier between cell and organelle: genome analysis of Candidatus Carsonella ruddii. *BMC Evol Biol* **7**, 181 (2007).
112. Nikolic, J. *et al.* Negri bodies are viral factories with properties of liquid organelles. *Nature Communications* **8**, 58 (2017).
113. [Giant viruses that create their own genes]. - Abstract - Europe PMC. https://europepmc.org/article/med/30623766.
114. Koonin, E. V. *et al.* Global Organization and Proposed Megataxonomy of the Virus World. *Microbiol Mol Biol Rev* **84**, e00061-19, /mmbr/84/2/MMBR.00061-19.atom (2020).