NNT/NL : 0000AIXM0000/000ED000

THÈSE DE DOCTORAT

Soutenue à Aix-Marseille Université
le 02 décembre 2022 par

Sofia Rigou

Etude des *Pithoviridae* à travers leur pangénome et métagénomes

Discipline
Biologie santé

Spécialité
Génomique et Bioinformatique

École doctorale
ED62 - Sciences de la vie et de la santé

Laboratoire/Partenaires de recherche
Laboratoire d'Information Génomique et Structurale, UMR 7256

Composition du jury

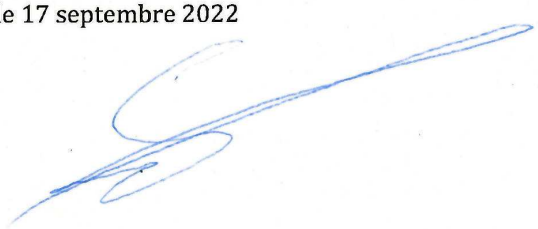
Elisabeth Herniou	Rapportrice et présidente du jury
CNRS, Institut de recherche sur la biologie de l'insecte	
Gwenaël PIGANEAU	Rapportrice
CNRS, Biologie Intégrative des Organismes Marins	
Jean-Michel CLAVERIE	Examinateur
CNRS, Information Génomique et Structurale	
François ENAULT	Examinateur
UCA, Laboratoire Microorganismes : Génome Environnement	
Matthieu LEGENDRE	Directeur de thèse
CNRS, Information Génomique et Structurale	

Affidavit

Je soussigné, Sofia Rigou, déclare par la présente que le travail présenté dans ce manuscrit est mon propre travail, réalisé sous la direction scientifique de Matthieu Legendre, dans le respect des principes d'honnêteté, d'intégrité et de responsabilité inhérents à la mission de recherche. Les travaux de recherche et la rédaction de ce manuscrit ont été réalisés dans le respect à la fois de la charte nationale de déontologie des métiers de la recherche et de la charte d'Aix-Marseille Université relative à la lutte contre le plagiat.

Ce travail n'a pas été précédemment soumis en France ou à l'étranger dans une version identique ou similaire à un organisme examinateur.

Fait à Marseille, le 17 septembre 2022



Liste de publications et participation aux conférences

1) Liste des publications[1] réalisées dans le cadre du projet de thèse :

    1. Sandra Jeudy, Sofia Rigou, Jean-Marie Alempic, Jean-Michel Claverie, Chantal Abergel & Matthieu Legendre, The DNA methylation landscape of giant viruses, 2020, *Nat Commun* 11, 265. https://doi.org/10.1038/s41467-020-16414-2
    2. Léo Blanca, Eugène Christo-Foroux, Sofia Rigou, Matthieu Legendre, Comparative Analysis of the Circular and Highly Asymmetrical *Marseilleviridae* Genomes, 2020, Viruses 12(11):1270, https://doi.org/10.3390/v12111270
    3. Sofia Rigou, Eugène Christo-Foroux, Sébastien Santini, Artemiy Goncharov, Jens Strauss, Guido Grosse, Alexander N Fedorov, Karine Labadie, Chantal Abergel, Jean-Michel Claverie, Metagenomic survey of the microbiome of ancient Siberian permafrost and modern Kamchatkan cryosols, 2020, *microLife*, Volume 3, https://doi.org/10.1093/femsml/uqac003
    4. Sofia Rigou, Sébastien Santini, Chantal Abergel, Jean-Michel Claverie, Matthieu Legendre, Past and present giant viruses diversity explored through permafrost metagenomics. 2022, *Nat Commun,* https://doi.org/10.1038/s41467-022-33633-x
    5. En préparation : Sofia Rigou, Alain Schmitt, Jean-Marie Alempic, Audrey Lartigue, Peter Vendlosky, Eugène Christo-Forroux, Chantal Abergel, Jean-Michel Claverie, Matthieu Legendre, Evolutionary history of the drastically transposon-invaded Pithoviridae genomes

2) Participation aux conférences[2] et écoles d'été au cours de la période de thèse :
    1. "Ancient giant viruses discovered through permafrost metagenomics", S. Rigou, C. Abergel, J-M Claverie, J. Strauss, M. Legendre. Symposium on Environmental and Agronomical Genomics, 29 octobre 2021
    2. "Ancient giant viruses discovered through permafrost metagenomics", S. Rigou, C. Abergel, J-M Claverie, J. Strauss, M. Legendre. New waves of thinking – Junior Scientists Microbiology Meeting of Marseille (JSM3), 5 mai 2022

[1] Cette liste comprend les articles publiés, les articles soumis à publication et les articles en préparation ainsi que les livres, chapitres de livre et/ou toutes formes de valorisation des résultats des travaux propres à la discipline du projet de thèse. La référence aux publications doit suivre les règles standards de bibliographie et doit être conforme à la charte des publications d'AMU.

[2] Le terme « conférence » est générique. Il désigne à la fois « conférence », « congrès », « workshop », « colloques », « rencontres nationales et/ou internationales » … etc.

Indiquer si vous avez fait une présentation orale ou sous forme de poster.

Remerciements

La recherche est un travail d'équipe. Une thèse n'échappe donc pas à cette règle. Je tiens donc évidement à remercier toutes les personnes avec qui j'ai pu échanger et travailler pour rendre cette recherche possible mais aussi toutes les personnes m'ayant apporté un appui pratique et moral pour l'écriture de cette thèse. Trois ans, ça passe vite, alors il faut les traverser bien accompagné pour arriver jusqu'à la ligne d'arrivée !

Un grand merci tout d'abord à Matthieu qui a été un directeur exemplaire (c'est le moment où c'est moi qui le note pour une fois c'est ça ?). Merci pour ton soutien scientifique et ta rigeur. Ça a été très rassurant de pouvoir discuter des résultats en allant dans le détail des choses. Merci pour nos débats bienveillants. Merci également à toute l'équipe de l'IGS, Chantal et Jean-Michel pour vos impulsions aux projets et présentations inspirantes. Sébastien, Lionel, Alain et Sandra ça a été un grand plaisir de travailler ensemble et d'apprendre de vos expériences. Merci à Estelle, Jean-Marie, Audrey, Eugène, Virginie et aussi à Aurélie Aufray de rendre tout ce travaille possible que ce soit par la découverte de nouveaux virus ou part un appui technique et administratif. N'oublions pas les rapportrices, Elisabeth Herniou et Gwenaël Piganeau et le comité de suivi de thèse avec Fabrice Armougom et François Enault qui prennent le temps de faire avancer la recherche des autres ! Ça c'est le sens du service public. Enfin, pour votre travail de relecture, merci à Thibault <3 <3, à Sandra et à toute l'équipe cassidaine ! Aline, Alain et Léonie vous êtes super.

Les remerciements viennent souvent avec une part de sentiments, c'est le moment où l'on remercie les proches d'avoir « été là pour nous »[1]. Merci à Thibault encore, de m'avoir supporté ces derniers temps à flux tendu. Hugo and Avi, what a funny team you make ! I hope our paths will meet again in Marseille in future. Aux belles rencontres marseillaises également, Magali, Malvina, Ysaline, Elo, Claire, les bibous, Lionel, Marie-Anne, Florie, Justine et j'en oublie mais votre présence me donne un souffle d'air frais. Ghis, Virginie, Fanny, Slim, Jéjé, Seb, vous m'avez apporté tellement de bonheur. Le club me manque les amis, promis je reviens vite ! Merci à Solène, Ainhoa, Kamel, Aïtana, Luc, et la famille (de sang cette fois mais vous aussi vous êtes le S) d'être venus me voir ! Merci aussi à Alain, Olivier, Anna, Alexandra pour nos discussions à refaire le monde à la pause café. A bientôt au Zoumaï !

[1] Euphémisme courrament employé pour signifier selon le contexte : écouter l'autre se plaindre et le consoler affectueusement alors qu'en fait on a envie de lui dire « arrête, tu commences à faire pitié là. Fragile… » ou simplement avoir gardé une relation d'amitié.

Résumé

Le pergélisol est un réservoir d'organismes anciens, où des amibes survivent jusqu'à plusieurs centaines de milliers d'années. En 2014, le virus géant d'amibe Pithovirus sibericum devient le premier virus ancien réactivé. Depuis, douze Pithoviridae ont été isolés dont six par le laboratoire IGS depuis le début de cette thèse. Les Pithoviridae intriguent de par leur morphologie unique, leurs génomes complexes et leur résistance au pergélisol.

En premier lieu, cette thèse décrit à travers la métagénomique, la place des virus géants au sein des microbiomes complexes du pergélisol russe. Des échantillons profonds de Yukechi, pourtant dominés par les bactéries, se sont avérés très riches en séquences proches des Pithoviridae. Parmi celles-ci se trouve Hydrivirus, dont le génome circularisé, donc complet, fait 1,6 Mégabases.

Dans une deuxième partie nous proposons d'étudier l'histoire évolutive des Pithoviridae. Cela a montré que les génomes circulaires des Pithoviridae, bien que complexes, sont conservatifs. Par ailleurs, trois génomes contiennent plus de 20% de séquences répétées. Ces régions génomiques se sont avérées être constituées de deux types de séquences à répétition Inversée qui se suivent séquentiellement.

Le troisième chapitre présente une analyse préliminaire d'un cycle infectieux chez l'amibe *A. castellanii*. L'objectif est de plonger plus précisément dans le fonctionnement de ces virus avec Cedratvirus kamchatka comme modèle. Nous constatons, dans un transcriptome largement dominé par les gènes nucléaires de l'hôte, que la quasi-totalité des gènes viraux prédits s'expriment.

Mots clés : génomique, virus géants, métagénomique, évolution

Abstract

Permafrost is a reservoir of ancient organisms, including amoebas surviving up to several hundred thousand years. In 2014, the giant amoeba-infecting Pithovirus sibericum became the first ancient virus to be reactivated. Since then, twelve *Pithoviridae* have been isolated including six by the IGS laboratory since the beginning of this thesis. The *Pithoviridae* are intriguing because of their unique morphology, complex genomes and resistance to permafrost.

Firstly, we will study, through metagenomics, the place of giant viruses in the complex microbiomes of Russian permafrost and surface layer. Although dominated by Bacteria, deep samples from Yukechi were found to be very rich in sequences close to the *Pithoviridae*. Among those lies Hydrivirus, whose circularized, thus complete, genome is of 1.6 Megabases.

Secondly, we will focus on the evolution of *Pithoviridae* through comparative genomics. This showed that the circular genomes of *Pithoviridae*, although complex, are conservative. Three genomes contain more than 20% repeated sequences. It appears that these regions are made up of two types of inverted repeats that follow each other sequentially.

We will finally present preliminary analyses of an infectious cycle in the host, *A. castellanii*. The aim is to study more precisely the functioning of these viruses using Cedratvirus Kamchatka as a model. In a transcriptome largely dominated by host nuclear genes, we find that almost all predicted viral genes are expressed.

Keywords: genomics, giant viruses, metagenomics, evolution

Glossaire

Bin (bioinformatique) : Anglissime, Groupe de séquences métagénomiques regroupées car il est estimé qu'elles appartiennent au même organisme

Binning (bioformatique) : Action de générer des « bins »

Contig : Séquence génomique issue de l'assemblage de lectures

Cryosol : Sols des régions froides composé du pergélisol en profondeur et de la couche active en surface

Exocytose : Fusion d'une vésicule avec la membrane plasmique d'une cellule permettant l'expulsion du contenu de la vésicule

Gène accessoire (bioinformatique) : Gène absent chez une partie des génomes du groupe étudié

Gène cœur (bioinformatique) : Gène présent chez tous les génomes du groupe étudié

Kyste : se dit aussi cyste, Microorganisme entouré d'une coque protectrice (enveloppe ou matrice) et en état de dormance lui permettant de survivre en cas de stress environnemental

Métagénomique : synonyme de génome environnementale, Analyse de séquences ADN d'un échantillon complexe donné après séquençage de l'ADN dit total

MOI : Anglissisme, Multiplicity of Infection, Multiplicité d'infection, Nombre de particules virales par cellule appliqué lors d'une expérience d'infection au laboratoire

ORF : Anglissisme, Open Reading Frame, cadre ouvert de lecture, Séquence nucléotidique susceptible d'être traduite en protéine donc commençant par un codon start et se terminant par un codon stop. Par abus de langage, le terme est aussi utilisé pour désigner des protéines prédites par bioinformatique.

Orthogroupe : Groupe de gènes ayant un ancêtre commun

Orthologie : Lien de parenté entre deux gènes

Pangénome : Ensemble des gènes d'un groupe d'organisme apparentés d'intérêt

Phagocytose : Processus par lequel des cellules ingère un objet extérieur dans un « phagosome » suite au contact de cet objet avec la cellule

Richesse (écologie) : Nombre d'espèces présentes dans un milieu donné

Scaffold (bioinformatique) : Anglissisme, Séquence génomique recréée à partir de l'assemblage de contigs

Séquençage : Définition de la suite d'unités composant un biopolymère linéaire. Pour l'ADN cela se traduit par la révélation de la suite d'adénines (A), thymines (T), guanines (G) et cytosines (C).

Séquence répétée (ADN) : Suite de nucléotides d'une taille définie présente plusieurs fois dans un même génome

Tégument : Tissu ou structure située à l'interface enter le milieu extérieur et le milieu intérieur d'un organisme

Transcriptomique : Etude de l'ensemble des ARN messagers présents dans un système donné via le séquençage ARN

Transfert de gène horizontal : Passage de matériel génétique contenant au moins un gène vers le génome d'un organisme n'étant pas le descendant du donneur. S'oppose au transfert vertical.

Virion : Forme extracellulaire du virus utilisé ici pour désigner également les particules virales intracellulaires matures donc supposées identiques aux formes extracellulaires.

Table des matières

A mes grands pères,
deux passionnés de science

Introduction

I. La notion changeante de virus

La nécessité de définir le terme virus vient de leur diversité. Beaucoup de virus n'ayant aucun gène orthologue (c'est-à-dire issu d'un ancêtre commun), leur mécanisme de réplication variant et même leur matériel génétique ne reposant pas sur le même type de molécule, nous pouvons en conclure que les virus n'ont probablement pas une origine unique. Le besoin de définir cette notion vient alors de l'absence de lien de parenté entre tous les virus qui aurait pu être une définition suffisante.

Le terme « virus » nous est parvenu jusqu'aux langues modernes par le latin et avant lui par le sanskrit[2] où il désigne un poison ou un venin. Il pouvait aussi signifier « suc, jus, humeur » ou encore « puanteur »[3]. Le terme virus sera plus tard progressivement associé à la contagion et à la maladie. On retrouve même une définition qui semble avant-gardiste dans le *Guidon en françois* dès 1478 : « substance qui recèle l'agent du contage[4] et est capable de transmettre la maladie »[5]. Bien que contredite par la religion, l'idée que la maladie puisse se transmettre par contact date de bien avant la théorie des germes puisqu'on la retrouve chez plusieurs auteurs et pratiques médicales au moyen-âge et avant (Ober et Aloush 1982). Sans en avoir fait l'observation, Marcus Terentius Varro au 1er siècle avant J.C. donne même un caractère vivant à l'agent pathogène en les qualifiant de « petits animaux » (Ober et Aloush 1982). Depuis la première observation de microorganismes vers la preuve que ce sont eux qui causent les maladies infectieuses, de Leeuwenhoek à Pasteur puis Koch, deux siècles d'expériences passent pour prouver la théorie des germes et lui donner son caractère scientifique, reproductible. Le caractère vivant des agents pathogènes peine à être accepté et le mot « virus » est souvent utilisé à la fin du 19ème siècle pour désigner de manière ambigüe un corps chimique nocif et inerte, différent d'un germe vivant. Le terme a été à la foi utilisé comme synonyme de germe et comme antagoniste, parfois de manière controversée (Kostyrka 2018).

C'est dans le cadre des études sur la théorie des germes et pour tenter de résoudre les contradictions entre les maladies pour lesquelles on arrive à cultiver ou non la bactérie responsable que l'on entreprend de distinguer les agents pathogènes vivants et non vivants. En effet, Paul Bert remarque que l'oxygène comprimé permet de tuer les êtres vivants et cela lui permet de faire la différence entre les fermentations du fait de microorganismes, et celles du fait d'agents chimiques. En 1877 il applique ces mêmes expériences à des maladies et observe que le venin de scorpion ainsi que « le virus du vaccin et le pus de la morve » résistent

[2] « viṣa » signifiait bien poison mais aussi « quelque chose d'actif », d'après le dictionnaire sanskrit anglais Monier-Williams consulté sur le site sanskrit.inria.fr le 7 juillet 2022

[3] Dictionnaire latin-français Gaffiot consulté en ligne sur gaffiot.fr le 7 juillet 2022

[4] Contage vient du mot *contagium* et signifie contagion

[5] Etymologie de virus d'après le Centre National de Ressources Textuelles et Lexicales, cnrtl.fr, consulté le 7 juillet 2022

à l'épreuve de l'oxygène comprimé, c'est-à-dire, restent nocifs. Ces maladies sont donc l'effet de « substances diastasiques[6] » et non pas d'êtres vivants (R. 1877). L'étude des toxines bactériennes complique cette distinction. En effet, du sang contaminé par *Bacillus anthracis* peut se révéler toujours porteur de maladie après application de l'oxygène comprimé, tuant les bactéries (R. 1877). On pense donc à un virus associé à la bactérie et pouvant causer des symptômes similaires avant que le mécanisme des toxines ne soit réellement compris. Un flou artistique sur la définition de virus règne à cette époque. Nous pouvons cependant remarquer que le terme a parfois été utilisé par élimination, pour désigner un agent invisible de la maladie pour lequel aucun microorganisme n'est visible au microscope, peut-être une « espèce de fermens chimique soluble » (R. 1877).

A partir des années 1880, les méthodes de filtration seront largement améliorées avec l'invention du filtre de Chamberland dont le rôle est de purifier l'eau. C'est en utilisant cet outil que Dimitri Ivanovski va transcender les travaux d'Adolf Eduard Maydolf Mayer sur la maladie « mosaïque du tabac ». Il montrera que l'extrait d'une feuille malade reste infectieux même après passage par le filtre de Chamberland. Ivanokski, n'arrivant pas à cultiver la bactérie responsable qui pourrait expliquer la présence de toxines, conclut à une erreur ou un problème de méthode de la culture de bactéries (Kostyrka 2018). Quelques années plus tard, Martinus Beijerinck n'hésite pas aller au-delà des connaissances établies et conclue à l'existence d'un « contagium vivum fluidum ». De plus cet agent traverse une goutte d'agar pour infecter une feuille ce qui permet donc selon lui de prouver la fluidité de l'agent infectieux (Kostyrka 2018). Les expériences de Friedrich Löffler et Paul Frosch en 1898 ont pu contredire cette affirmation car, travaillant sur des filtres plus fins, ils ont réussi à stopper la maladie de la fièvre aphteuse touchant le bétail (le foot-and-mouth disease virus) lui redonnant son caractère particulaire. Revenons aux expériences de Beijerinck qui ne sauraient se résumer au « contagium vivum fluidum ». La sève extraite est extrêmement infectieuse même après infection successive de différents plants. Un agent se reproduit donc dans la plante de tabac. Beijerinck constate également que, ce qui plus tard sera appelé virus, ne peut pas croitre hors de la plante et en conclut que ce virus aurait besoin de la division de cellules pour sa propre multiplication. Est-ce le signe d'une dépendance à la cellule ou le virus est-il créé par la cellule ? Cela n'est pas clair pour tous les chercheurs de l'époque surtout quand on sait que le virus de l'herpès peut ressurgir comme créé *de novo* après un stimuli (Boycott 1928). Par ailleurs, un consensus scientifique ne s'est pas établi tout de suite sur la dépendance à la cellule car plusieurs chercheurs ont prétexté avoir réussi ce que l'on cherchait à faire depuis des décennies : cultiver le virus en l'absence de cellules. Peut-être était-ce le fait d'une contamination ou une fausse identification comme virus (Eagles 1933) ? Quoi qu'il en soit, les premières études sur le virus de la mosaïque du tabac à partir de filtrats ainsi que l'absence d'entité exogène visible au microscope à partir d'échantillons de sang malade vont associer la qualité de petit au virus et on parlera alors de « virus filtrables ». Ceci est pour les différencier des autres virus ou « agents visibles des maladies infectieuses » jusqu'à ce que progressivement, le mot virus ne signifie plus que « virus filtrable » (Lwoff 1953). Les premiers virus dont on essaye de déterminer la taille seront le virus de la mosaïque du tabac et le bactériophage d'Herelle, tous deux de petite taille (Rivers 1927). D'après Rivers, qui d'ailleurs

[6] Agent qui catalyse une réaction chimique, ferment

faisait partie de ceux qui affirmaient qu'ils étaient des parasites obligatoires, la taille semble être un critère pour accorder ou non le statut de vivant aux virus :

> *« Il peut être dit que beaucoup de virus sont probablement de taille suffisante pour exister dans un état vivant et que d'autres sont probablement assez petits pour satisfaire les demandes de ceux qui insistent sur le fait qu'ils ne possèdent pas de vie. »* [7]

La première cristallisation des protéines du virus de la mosaïque du tabac en 1935, donnant des cristaux plus infectieux que le jus de feuilles malades, contribuera à une vision inerte, autocatalytique des virus (Stanley 1935). Peu après, c'est la présence d'acides nucléiques qui est prouvée par diffraction de rayons X. Alors qu'il commence à être bien accepté que les virus sont bien des organismes à part et non le produit d'une cellule, les premières constatations de la lysogénie viennent encore compliquer le scénario et sèment le doute parmi les chercheurs. Ils font face à une maladie à la fois contagieuse et héréditaire mais aussi corpusculaire ce qui est à première vue antinomique. On doute de l'existence même du virus d'Herelle. C'est en trois décennies de recherche que la nature des bactériophages lysogéniques fut élucidée. Le coup de grâce est souvent attribué à André Lwoff et l'induction d'un cycle lytique, notamment par rayonnements UV, qui permit de dire exactement quelles bactéries étaient porteuses du prophage et surtout qu'en l'absence de primo-infection, il n'y a pas de prophage. Des études génétiques montrèrent ensuite l'intégration du prophage dans le génome bactérien (Galperin 1987).

Dans « The concept of viruses », André Lwoff tente une définition des virus en opposition aux organismes cellulaires. Ceux-ci n'ont qu'un type d'acide nucléique et non deux, il se multiplient comme les acides nucléiques et non par la croissance et la division cellulaire et ils n'ont pas de système de lipman c'est-à-dire le métabolisme nécessaire pour fournir de l'ATP ou d'autres molécules pour stocker de l'énergie. Il évoque alors le gradient de taille qui existe entre les virus et les bactéries (Lwoff 1957). Pourtant, la taille à une importance capitale dans leur définition qui est alors donnée par Bawden : « parasites obligatoires pathogènes avec au moins une dimension inférieure à 200 µm » [8].

Dans les décennies qui suivirent on chercha moins à définir les virus qu'à les classer. La diversité virale connue permet à David Baltimore en 1971 de créer une classification qui sera largement utilisée jusqu'à aujourd'hui (Figure 1). Celle-ci se base sur le matériel génétique du virion et le mode de réplication dans la cellule.

[7] Traduction de « [In general, however,] it can be said that many viruses are probably of sufficient size to exist in a living state, and that others are probably small enough to satisfy the demands of those who insist that they are not possessed of life. »

[8] Traduction de « obligatory parasitic pathogens with at least one dimension of less than 200 mµ. »

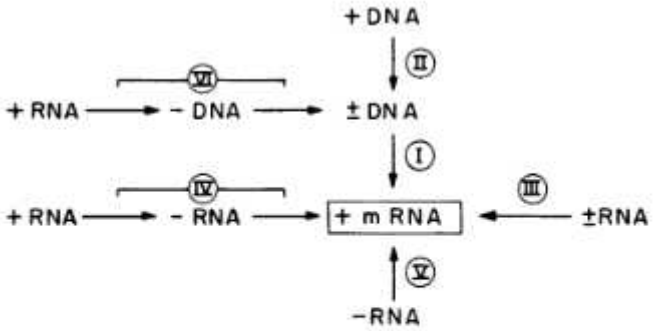


FIG. 1. *Examples: I = T4 phage, vaccinia virus; II = ΦX174; III = reovirus; IV = RNA phages, poliovirus; V = vesicular stomatitis virus, Newcastle disease virus; VI = RNA tumor viruses.*

Figure 1 - Classification de Baltimore selon sa première définition en 1971 (Baltimore 1971)

Cette classification de Baltimore illustre bien l'énoncé moderne, que l'on peut qualifier de réducteur, «la virologie est devenue une science de la biologie moléculaire»[9] (Norrby 2008). D'autres classifications se basent sur des critères morpho-génétiques (Norrby 1983):

- Le type d'acide nucléique
- La symétrie cubique ou hélicoïdales du virion
- La présence ou non d'enveloppe
- Les paramètres dépendant du type : diamètre pour les capsides hélicoïdales comme celle du virus de la mosaïque du tabac, nombres de capsomères dans le cas d'une capside icosaédrique.

Enfin, la classification du Comité International de la Taxonomie des Virus (ICTV) qui fait autorité dans le domaine, va classer les virus de la même manière qu'elle classe les organismes ; sur des critères d'histoire évolutive commune ; de monophylie. L'ICTV ira jusqu'à donner des noms espèces aux virus. Avant 2013, les espèces virales n'étaient pas supposées monophylétiques mais à présent, l'espèce associée à un virus a pour but de déterminer un groupe à l'histoire évolutive commune qui les distingue des autres. Le but est surtout de donner une cohérence taxonomique entre les codes utilisés pour étudier le monde cellulaire et viral et ainsi classer les virus dans des phylum, classes etc… Bien que les virus aient le droit à leur nom d'espèce, l'ICTV donne récemment une définition de virus en tant qu'élément génétique mobile qui nous rappelle la phrase de Norrby. Ce matériel génétique doit coder une protéine majeure de capside[10]. Cette définition provient en fait surtout d'un besoin pratique

[9] Traduction de « virology has become a molecular biological science »

[10] La définition donnée dans le code de l'ICTV de mars 2021 va comme suit: « Viruses *sensu stricto* are defined operationally by the ICTV as a type of MGEs that encode at least one protein that is a major component of the virion encasing the nucleic acid of the respective MGE and therefore the gene encoding the major virion protein itself; or MGEs that are clearly demonstrable to be members of a line of evolutionary descent of such major virion protein-encoding entities. Any monophyletic group of MGEs that originates from a virion protein-encoding ancestor should be classified as a group of viruses. ».
En français : « Les virus au sens stricte sont définis opérationnellement par l'ICTV comme un type d'Eléments Génétiques Mobiles qui codent pour au moins une protéine étant un composant majeur du virion enveloppant l'acide nucléique du l'EGM respectif et pour cela, le gène lui-même codant pour la protéine majeure du virion; ou des EMG ayant clairement démontré qu'ils étaient membres de la lignée évolutive descendante d'une telle

de classification puisque c'est la phylogénie de la protéine majeure de capside qui est utilisée dans la nouvelle classification taxonomique virale officielle.

Un débat ouvert depuis des décennies est celui de la définition du virus dans son cycle complet. Le virion devrait-être considéré comme la graine, le véhicule de l'information génétique du virus pour reproduction et non comme le virus en lui-même (Claverie 2006). Lwoff répondait déjà à ce débat en 1957 « le prophage[11] et la phase végétative sont des parties du 'bactériophage' tout comme la particule infectieuse qui est donc privée de sa suprématie. »[12] (Lwoff 1957). Cette question a son importance puisque d'elle découle en partie la réponse à la question de « un virus est-il vivant ? ». En effet, si on ne considère pas la phase cellulaire alors le virus parait être une chose inerte, dépourvue de métabolisme.

Une vision plus moderne des virus est de les considérer comme des organismes dépourvus de ribosomes : « nous proposons de diviser les entités biologiques en deux groupes d'organismes : les organismes encodant des ribosomes […], et les organismes encodant des capsides »[13] (Raoult et Forterre 2008). Les virus géants nous ont appris à ne pas se reposer sur la capacité à coder ou non une protéine car leur diversité génomique pourrait un jour leur permettre de se rapprocher dangereusement de l'encodage de ribosomes propres (Mizuno et al. 2019). De la même manière, des virus sans capside conventionnelle existent (Philippe et al. 2013). Une chose ne change pas dans la définition de virus : la nécessité pour le virus de se répliquer au sein d'une cellule hôte. Dans une lettre à Lwoff, Salvadore E. Luria tente de capter l'essence des virus dans une définition dont il dira qu'il n'est pas encore entièrement satisfait ; ce serait « un élément de matériel génétique capable de prendre une forme transmissible en s'incorporant dans un appareil de transmission synthétisé sous le contrôle du virus lui-même. »[14] (Luria 1957). L'idée est de faire ici la différence entre le prophage dont la réplication est passive et le virus ou prophage réactivé dont la réplication n'est pas sous le contrôle exclusif de l'hôte. Il me semble qu'une définition plus complète peut-être approchée en y ajoutant le caractère de parasite intracellulaire obligatoire. Cela donnerait alors [*ensemble contenant du* « matériel génétique capable de prendre une forme transmissible en s'incorporant dans un appareil de transmission synthétisé » *et assemblé de novo dans une cellule hôte* « sous le contrôle du virus lui-même »].

Avec tous ces éléments on comprend que la notion de virus découle davantage d'exemples et de découvertes progressives et non d'un changement de paradigme radical. Ces découvertes bien-sûr sont le fruit d'avancées technologiques majeures au cours du 19ème et 20ème siècle non ou brièvement évoquées ici. Dans les parties qui suivent nous verrons comment les grands virus et les virus géants troublent d'avantage les limites entre le monde cellulaire et viral.

entité codant pour une protéine majeure de virion. Tout groupe d'EMG monophylétiques qui provient d'un ancêtre encodant une protéine de virion devrait être classé comme un groupe de virus ».

[11] Un prophage est un bactériophage dont le génome a été intégré dans celui de l'hôte.

[12] Traduction de « the prophage and the vegetative phage are parts of the 'bacteriophage' as well as the infectious particle which is thus deprived of its supremacy »

[13] Traduction à partir de « we propose to divide biological entities into two groups of organisms: ribosome-encoding organisms, which include eukaryotic, archaeal and bacterial organisms, and capsid-encoding organisms , which include viruses. »

[14] Traduction de « an element of genetic material capable of assuming a transmissible form by incorporation into a transmission apparatus synthesized under the virus' own control ».

Avant cela, nous évoquerons ce qui a mené la virologie à être également une science de la biologie environnementale puisque l'intégration des virus dans notre compréhension des interactions écologiques change radicalement notre vision des réseaux trophiques, de l'épidémiologie et de l'évolution.

II. Du virus pathogène humain vers la virologie environnementale, deux disciplines intriquées

Les premiers modèles d'étude en virologie, nous l'avons vu, sont les virus humains et les virus d'intérêt économique qui affectent notre environnement proche : le bétail et les plantations. L'étude des virus d'animaux a permis non seulement d'appliquer les premiers vaccins dont on ne saurait dater la première apparition en Chine, mais aussi de comprendre le passage d'un virus d'une espèce à une autre et même l'origine virale de certains cancers dès le début du 20ème siècle.

La virologie environnementale nait à l'origine comme un moyen de comprendre et de prévenir les épidémies et non comme un moyen de comprendre les interactions écologiques. Depuis le 19ème siècle, il est connu que les maladies provenant entre autres de spores bactériennes peuvent être présents dans l'environnement et donc emprunter des routes de contamination diverses. L'étude de la poliomyélite a été fondatrice pour la virologie environnementale. A l'origine perçue comme une maladie miasmatique, théorie selon laquelle les maladies sont transmises à travers le mauvais air et la putréfaction, son épidémiologie a finalement montré que la maladie était contagieuse. Les ravages de ce virus en Scandinavie sont peu connus, et pourtant, la maladie a eu en Suède plus d'incidence qu'aux Etats-Unis. On estime que cette épidémie a commencé dès 1881 et culmine en 1911-1913 faisant 10 000 victimes. Des chercheurs suédois du laboratoire médical d'état montrent que des singes peuvent être infectés par le poliovirus à travers la nourriture et dessinent la route « oro-fécale ». Ces résultats présentés à une conférence aux Etats-Unis en 1912 seront par la suite moqués car l'équipe de Simon Flexner au Rockefeller Institute avait conclu précipitamment à une « route olfactive » du virus. Ils n'arrivaient pas à répliquer les résultats suédois du fait de l'espèce de singes utilisée, insensible à la route « oro-fécale » (Axelsson 2009). Cependant, le modèle du Rockefeller Institute resta dominant jusque dans les années 40, quand le virus fut retrouvé dans des échantillons d'eaux usées de grandes villes étasuniennes puis en Suède et que celui-ci fut injecté dans différents animaux (Metcalf, Melnick, et Estes 1995). On comprit alors que les virus étaient présents en grand nombre dans les eaux usées et que ceux-ci restaient infectieux pendant des semaines.

L'amélioration des cultures cellulaires a par la suite été d'une grande utilité en virologie. Celles-ci ont permis régulièrement d'isoler des entérovirus à partir d'échantillons environnementaux puis d'étudier leur résistance à différents stress et conditions de purification d'eau. D'autres méthodes de détection utilisées étaient la microscopie électronique immunologique, le dosage radio-immunologique ou encore la méthode immuno-enzymatique ELISA. La contamination par l'eau devint un sujet très important en virologie surtout après l'épidémie d'hépatite E à New Delhi qui fit 30 000 cas en deux mois entre 1955

et 1956 et qui n'avait pas pu être prévenue par les mesures d'assainissement (Cashdollar et Wymer 2013). Malgré les avancées, les méthodes d'isolement ou de détection de virus n'étaient pas assez sensibles considérant la haute infectiosité de certains entérovirus. La capacité des mollusques à concentrer les bactéries et virus a été comprise au fur et à mesure de contaminations par l'alimentation. Les huitres, les moules et autres bivalves sont étudiés par des essais d'infection sur culture cellulaire à la fois pour évaluer les risques d'infection par l'eau et de la consommation de fruits de mer (Richards 1985). Plus généralement, alors qu'une méthode pour concentrer le poliovirus existait déjà, la communauté internationale créa de nouvelles méthodes de concentration de virus y compris certaines basées sur la floculation (agglomération de particules en suspension), l'adsorption par affinité de charge sur filtre électronégatif ou électropositif puis élution et enfin, l'ultrafiltration de centaines de litres d'eau (Cashdollar et Wymer 2013).

Etant très sensible, la Réaction en Chaîne par Polymérase (PCR pour « polymerase chain reaction »), méthode rendue publique en 1985, vint considérablement simplifier et améliorer la détection de virus dans les échantillons. Elle a surtout l'avantage de détecter les virus non cultivables (Metcalf, Melnick, et Estes 1995). La PCR eut en réalité un effet double sur la virologie environnementale ; en plus de la détection plus précise de virus elle permit d'améliorer les méthodes de séquençage ADN (acide désoxyribonucléique) et ARN (acide ribonucléique) et donc, les débuts de la virologie environnementale motivée par l'écologie. De plus, cette avancée technologique permit aussi de séquencer des génomes viraux et cellulaires pour concevoir des amorces PCR plus spécifiques des virus recherchés.

Etonnamment, les premiers microbiomes étudiés sont issus de milieux marins et non humains. Tout commence par le métabarcoding avec le séquençage d'une poignée de séquences clonées de sources chaudes du parc Yellowstone (Pace et al. 1986) puis, de moins d'une quarantaine de clones de la petite sous-unité de l'ARN ribosomal (16S) provenant de l'océan pacifique à Hawaii. Des études métagénomiques du microbiote humain suivront, identifiant également des virus (Breitbart et al. 2003). Parmi les premiers métagénomes à proprement parler, la mer des Sargasses a été explorée par le Weatherbird II et le fameux voilier Sorcerer II. Déjà, des bactériophages avaient été identifiés (Venter et al. 2004). Forts de cette expérience, Craig Venter lance le Sorcerer II dans l'Expédition mondiale d'échantillonnage des océans[15] qui a fait le tour du monde. Pour comprendre cette ambition il faut revenir au premier séquençage du génome humain. En 1991, la communauté scientifique internationale démarre ce projet avec l'Institut National de Santé des Etats-Unis (NIH)[16] pour un budget de 2,7 milliards de dollars (dont 500 millions pour le séquençage). Craig Venter, ancien employé du NIH, décide de monter l'entreprise de séquençage Celera Genomics qui séquencera elle aussi, pour 10 fois moins cher, le génome humain (Gauthier et al. 2019). Ces différences de prix s'expliquent par une méthode employée différente mais il ne faut pas oublier que Celera genomics s'est largement basée sur des données du consortium international pour finaliser ce premier génome. L'approche utilisée pour le premier génome humain est celle de séquençage shotgun[17] hiérarchique ; un jeu minimal d'insert couvrant le génome est

[15] Traduction de « Global Ocean Sampling Expedition »
[16] Traduction de « U.S. National Institutes of Health »
[17] Signifie « fusil de chasse » en référence à la dispersion quasi-aléatoire du plomb des fusils, la volée de plomb

sélectionné avant d'être séquencé et assemblé. A l'inverse, suite aux avancées des algorithmes d'assemblage et à l'acquisition d'une puissance de calcul suffisante, Celera genomics applique la méthode de séquençage du génome entier (whole-genome shotgun WGS) (Gauthier et al. 2019). C'est donc pour continuer de prouver la puissance de cette méthode que l'homme d'affaires se lance dans l'aventure avec le Sorcerer II.

Le virome marin gagne progressivement en intérêt (Angly et al. 2006; Breitbart et al. 2002). Sans doute, le concept fondateur du shunt[18] viral explique cet engouement croissant. En étudiant les premières énumérations de virus à travers le monde allant de 10^8 à 10^{11} particules virales par litre d'eau de mer, les estimations du taux de lyse virale et des données de chimie de l'océan, les auteurs comprennent que les virus impactent fortement les réseaux trophiques et les grands cycles biogéochimiques (Wilhelm et Suttle 1999). Les virus permettent de libérer du carbone et des nutriments, y compris des nutriments limitants, là où la lyse a lieu donc majoritairement dans les couches productives de l'océan qui sont les plus compétitives en nutriments. La grande étude du virome marin de Sorcerer II trouve 3% de virus dans les échantillons filtrés à la taille microbienne mais conclut qu'ils sont largement sous-estimés. Les virus séquencés sont majoritairement des bactériophages. Parmi les virus d'eucaryotes toutes tailles confondues, les *Phycodnaviridae* sont largement majoritaires suivis des *Mimiviridae* (S. J. Williamson et al. 2008). Dans ce domaine on ne peut pas non plus ne pas citer le fameux voilier Tara qui a, à travers ses expéditions et son échantillonnage spécifiquement adapté à l'étude du virome marin, considérablement augmenté la diversité virale séquencée. Par l'étude de 145 échantillons, presque 200 000 populations virales, largement dominées par des séquences inconnues puis par les phages, ont été définies (Gregory et al. 2019). Ces populations ont été définies en appliquant un seuil de 95% d'identité à l'échelle nucléotidique et couvrant au moins 80% des contigs supérieurs à 10 kb. On comprend cependant qu'à travers cette méthode, le morcèlement des génomes viraux a pu augmenter la diversité virale estimée. Cette même étude propose un deuxième résultat important : les populations des couches profondes de l'océan (bathypélagique) correspondent souvent à des populations que l'on peut retrouver dans des zones moins profondes de l'océan. A l'inverse, les couches moins profondes ont des populations virales plus spécifiques. Cela s'expliquerait tout simplement par la sédimentation passive des virus vers le fond de l'océan (Gregory et al. 2019).

Malgré de nombreuses études, la diversité virale est telle qu'elle est difficile à appréhender. Les virus ne sont pas toujours inclus dans les modélisations écologiques et leurs hôtes restent souvent inconnus. Il en est de même des microbiotes viraux humains et particulièrement des viromes terrestres, riches et difficiles à étudier en partie du fait de l'hétérogénéité du sol (K. E. Williamson et al. 2017). Un petit nombre d'études dénombre entre 10^3 et 10^9 virus par gramme de sol sec (Figure 2).

[18] Shunt signifie dérivation, transfert

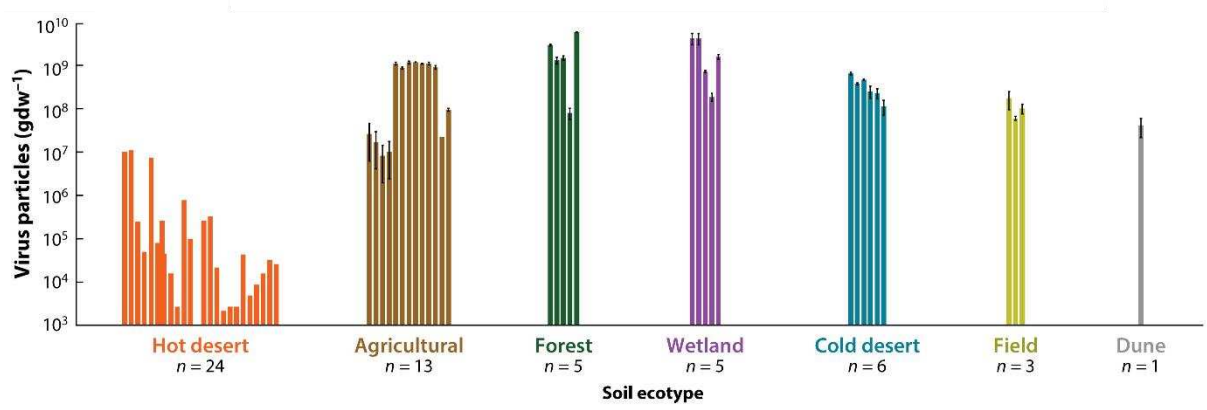


Figure 2 – Enumération de particules virales par gramme de sol d'après une méta-analyse (K. E. Williamson et al. 2017)
Les résultats du dénombrement des virus de neuf études ont été rassemblés par type d'environnement. Les différentes estimations proviennent de comptages directs par microscopie électronique à transmission, microscopie à fluorescence ou cytométrie en flux.

Bien que les objectifs soient différents, l'étude des virus dans le cadre de la santé humaine, du bétail et des plantations et la virologie environnementale ont bénéficié mutuellement des avancées technologies et conceptuelles de leurs disciplines à travers l'histoire. Les deux disciplines se retrouvent par exemple en épidémiologie dans l'étude de la permissivité des hôtes viraux et des sauts d'hôtes[19] (Woolhouse et Gowtage-Sequeria 2005) pour laquelle une exploration des milieux naturels est nécessaire. Ainsi nous avons vu que les deux histoires s'entremêlent et que les avancées de chaque discipline est largement due aux avancées technologiques de la fin du 20ème siècle jusqu'à aujourd'hui. La découverte des virus géants vient poser un problème à la virologie environnementale puisqu'on découvre qu'il ne suffit pas d'étudier les échantillons filtrés pour étudier un certain virome. S'il faut donner une définition elle peut être que les virus géants ont un génome supérieur à 300 000 bases et/ou une particule supérieure à 200/300 µm (Claverie et al. 2005) et donc se retrouvent le plus souvent dans la fraction cellulaire s'il y a filtration.

III. Découverte des virus géants

Avant la découverte des virus géants, des familles virales étonnaient déjà par leur complexité. C'est le cas des très étudiés *Poxviridae*, à l'origine de la variole qui a fait des ravages jusqu'à son éradication officiellement proclamée en 1979 grâce à la vaccination malgré l'existence de réservoirs de la variole du singe en Afrique (Reynolds et Damon 2012). Les *Poxviridae* connus ont un génome allant de 120 kilobases à 457 kilobases (Figure 4). Certaines espèces virales sont connues pour leur capacité à infecter plusieurs espèces de mammifères différents. Les *Phycodnaviridae* ont à l'inverse été moins étudiés avant les années 1990. C'est à cette époque

[19] Aussi appelé « saut d'espèce », le virus devient capable d'infecter une nouvelle espèce

que l'on séquence Paramecium bursaria chlorella virus (PBCV-1) et que l'on découvre son génome complexe de plus de 300 kb contenant de nombreux gènes aux fonctions inconnues (Etten et Meints 1999). L'étude de ces virus gagne également en intérêt scientifique quand il a été montré qu'un virus d'algue pouvait arrêter une efflorescence d'*Emiliana huxleyi* dans des conditions semi-contrôlées (mésocosme)[20] (Bratbak, Egge, et Heldal 1993). Ce résultat est majeur pour comprendre les cycles biologiques, le processus d'eutrophisation etc. L'étude a été répliquée en utilisant des techniques de détection différentes et avec succès (Jacquet et al. 2002; Pagarete et al. 2011). Le phénomène n'est cependant pas étudié à la hauteur de son importance puisque ce sont un petit nombre d'articles qui sont toujours cités. Les *Phycodnaviridae* ont et vont probablement gagner en popularité depuis que les études de métagénomique marine ont constaté leur ubiquité (p. 47). C'est très probablement ce qui s'est passé en 2020, année pendant laquelle 96 espèces et sous-espèces de *Phycodnaviridae* sont enregistrées dans le NCBI (Figure 3). Le premier isolement d'un *Phycodnaviridae* date de 1979 à partir du phytoplancton *Micromonas pusilla*. Aujourd'hui il y a 369 génomes classés dans cette famille sur le NCBI[21]. Au départ, certains *Mimiviridae* infectant les algues avaient à tort été placés dans ce groupe. Cette erreur consistant à classer les virus selon leur hôte et non selon leur histoire évolutive est récurrente dans l'histoire de la virologie mais sera rectifiée après la découverte des *Mimiviridae* infectant les protistes.

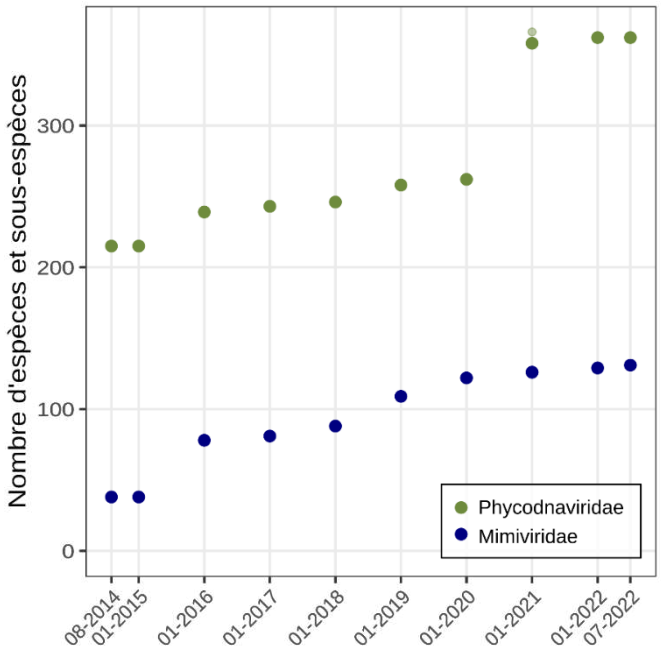


Figure 3 – Nombres d'espèces et sous-espèces enregistrées par le NCBI dans le temps Les données brutes montraient plus de *Phycodnaviridae* en 2021 qu'en 2022 (point vert clair) mais cela est dû à des doublons qui ont été retirés (point foncé). Le nombre d'espèces et de sous-espèces des deux familles virales proviennent des archives de la taxonomie du NCBI, disponible jusqu'à août 2014 sur leur site ftp.

[20] Au milieu de la baie est placé un aquarium flottant de plusieurs m3. Cet aquarium est ouvert à l'air et de l'eau de mer est pompée (filtré ou non selon l'expérience) et réinjectée dans ce « mésocosme ».
[21] Site consulté le 15 juillet 2022

A sa première observation, Mimivirus fut confondu avec une bactérie intracellulaire d'amibe et fut baptisée Bradford coccus malgré les tentatives infructueuses de répliquer son gène 16S. En 1998, Timothy Robotham cherche à protéger ses échantillons de la fermeture imminente de son laboratoire et envoie la bactérie à ses collaborateurs marseillais (Zimmer 2017). Elle fut ressortie quelques années plus tard et observée au microscope électronique. On aperçut alors sa forme icosaédrique. Le séquençage de ce dernier a laissé peu de doute : c'était un virus. Il avait un gène de capside et n'avait pas d'ARN ribosomaux (La Scola et al. 2003). Son génome fait presque 1,2 Mégabases. On y trouve tout types de fonctions possibles dont des gènes de la traduction : ARN de transfert, des facteurs de traduction et des aminoacyl-ARNt synthétases (Raoult et al. 2004). Le plus grand génome viral connu à l'époque était celui de PBCV-1, 4 fois plus petit. On comprend alors que la dépendance du virus à la machinerie cellulaire de l'hôte peut probablement être bien moindre que ce qui était possible jusqu'à présent.

Mimivirus fut la porte d'entrée vers un pan peu exploré de la virologie : l'étude des virus de protozoaires. Durant la dernière décennie, la famille des *Mimiviridae* a été considérablement enrichie. L'amibe *Acanthamoeba castellanii* s'est trouvée être l'hôte idéal pour isoler de nouveaux virus géants à partir d'échantillons environnementaux. Cela permet également la découverte d'autres familles virales étonnantes, le tout donnant un continuum de taille de génome entre les *Iridoviridae* et les géants *Pandoraviridae* (Figure 6). Deux de ces derniers ont été isolés à partir d'échantillons de sédiments océaniques et lacustres. Le plus grand à un génome de 2,5 Megabases. Parmi ces ORFs (cadres ouverts de lecture, Open Reading Frame), seul 16% ont une ressemblance avec les protéines déjà présentes dans les bases de données (Philippe et al. 2013). Ces virus ont également une morphologie encore jamais observée en virologie. Ils ont une forme d'amphore d'environ 1.2 µm de long avec une ostiole, par où son matériel est relargué dans la cellule pour l'infection. Ils ont un tégument assez particulier : des sucres particuliers ont été mis à l'évidence chez *Pandoravirus massiliensis* sur la couche externe du virion. Avant cette couche se trouve une membrane, une double couche d'épais tubules (Brahim Belhaouari et al. 2019) et enfin, une membrane lipidique au plus près du centre (Philippe et al. 2013).

Consécutivement, deux virus anciens sont pour la première fois réactivés depuis le pergélisol : Pithovirus sibericum (Legendre et al. 2014) puis Mollivirus sibericum (Legendre et al. 2015) l'année suivante. Les deux proviennent d'une couche datant d'il y a 30 000 ans. Pour faciliter l'échantillonnage, du pergélisol a été récupéré en forant horizontalement sur les bancs de la rivière Anuï, à 27 mètres sous la surface du sol. M. sibericum se présente comme des sphères d'environ 600 µm entourée de petits fibrilles à peine visibles. Il est le plus proche parent des Pandoravirus, et pourtant, sa morphologie est très différente et son génome est deux fois plus petit. P. sibericum présente en revanche le plus grand virion connu avec pourtant un génome de « seulement » 600 kilobases. L'histoire de la découverte des *Pithoviridae* rappelle beaucoup celle de Mimivirus. La première observation d'un Pithovirus date d'avant 1997. Une équipe allemande contrôlait la qualité de l'eau d'une usine de traitement des eaux via la présence d'amibes et la présence de microorganismes pathogènes au sein des amibes. Elle fit alors l'étrange constat ; celui d'un organisme inconnu dont on ne sait s'il ressemble plus à une archée (Hoffmann et al. 1997) ou à une bactérie (Michel, Müller, et al. 2003), et qui vit dans la cellule amibienne (Hoffmann et al. 1997; Michel, Müller, et al. 2003). La morphologie et la

description du cycle infectieux de KC5/2 correspondent parfaitement aux observations contemporaines d'un *Pithoviridae*. L'assemblage des virions au sein de l'usine virale, ici appelée le cytosol étendu, vu au microscope électronique aurait pu les mener vers la piste virale cependant il est vrai que la formation des virions est complexe et ne ressemble à rien de connu. De plus, la forme des *Pithoviridae* ne rappelle rien d'un virus. Malheureusement, toutes leurs tentatives d'extraction d'ADN et d'ARN ont été infructueuses (Michel, Müller, et al. 2003). En effet, l'enveloppe virale résiste aux méthodes de lyse classique.

En 2018, la découverte d'Orpheovirus vient ajouter une famille de virus ayant une taille de génome de l'ordre des Mégabases.

IV. Le gigantisme au sein de génomes viraux

Nous avons vu qu'au cours de l'histoire les virus ont été définis, de par leur découverte, comme étant des agents filtrables, c'est-à-dire petits. Ce constat est généralement vrai cependant, dès le début de la virologie, on comprend que ce groupe est extrêmement diversifié et qu'il sera difficile de les définir dans leur globalité. Comme pour venir défier les biologistes, la nature montre que les différences entre groupes d'organismes présentent des gradients, plutôt que des barrières faciles à appréhender.

Ainsi, il existe des bactéries très petites comme l'endosymbionte *Mycoplasma genitalium* et son rayon de 0.15 µm (Luisi et Stano 2011). A l'inverse, probablement le plus grand virion connu, celui de Pithovirus sibericum, fait 1.5 µm de long pour 0.5 µm de large. De la même manière, il y a un chevauchement entre la taille des génomes viraux et cellulaires (Figure 4). Des phages (*Myoviridae* et *Siphoviridae*), des virus d'animaux (*Poxviridae*, *Iridoviridae*, *Herpesviridae*, *Nimaviridae*, *Polydnaviriformidae* anciennement appelés *Polydnaviridae*), des virus d'algues (*Phycodnaviridae*) et surtout les virus d'amibes (*Marseilleviridae*, *Pithoviridae*, *Mimiviridae*, *Pandoraviridae*) surpassent certaines bactéries en taille de génome.

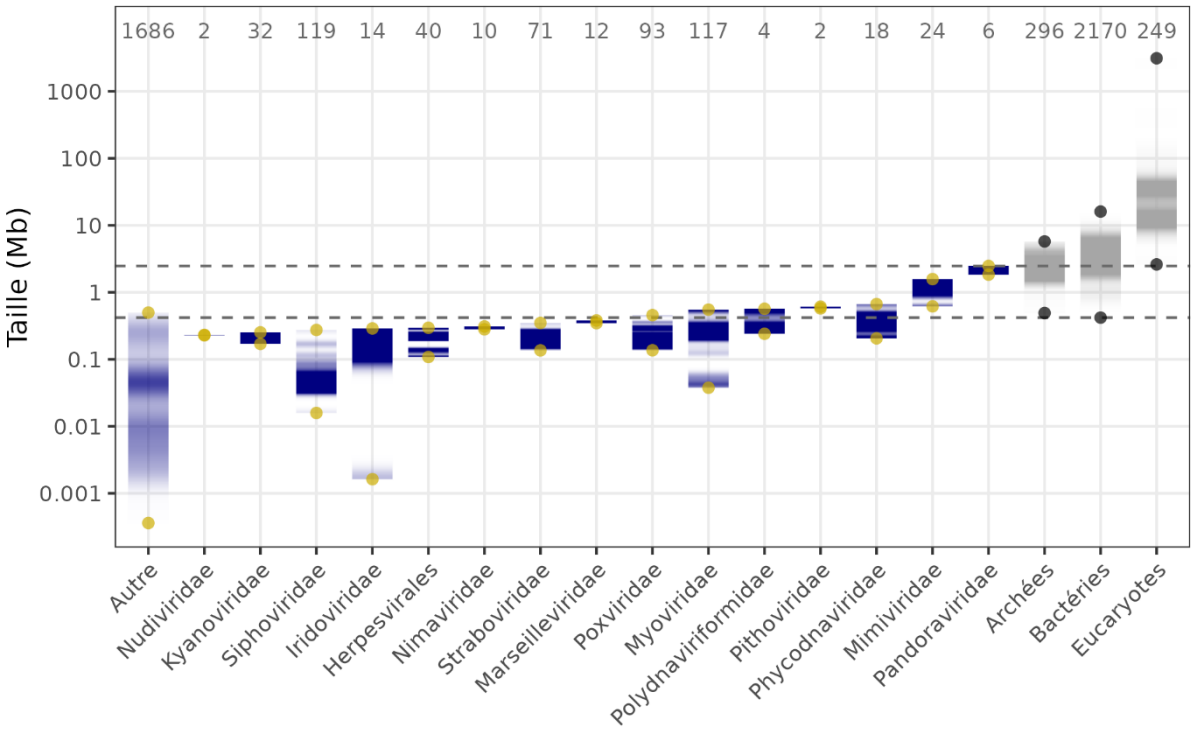


Figure 4 - Taille des génomes séparés par groupe viral et domaine cellulaire
Les groupes sont ordonnés selon leur plus grand représentant. S'ils n'ont aucun représentant supérieur à 200 Megabases, alors, ils ont été placés avec les virus « Autre ». Les deux lignes pointillées représentent la zone de chevauchement entre les tailles de génomes viraux et cellulaires. Les points jaunes (virus) ou noirs (cellulaires) montrent le plus grand et plus petit génome de la famille. Les tailles de génome utilisées sont issues des génomes complets parmi la base de données de génomes du NCBI. Les gradients bleus (virus) et gris (cellulaires) représentent la densité de données aux différentes tailles. Pour les bactéries et les virus, 2000 génomes ont été tirés au hasard. Ont également été ajoutés les 200 génomes bactériens les plus petits et les génomes viraux dont un membre de la famille dépasse les 200 kb. Les nombres en haut du graphique indiquent le nombre de génomes échantillonnés. L'axe des ordonnées est logarithmique.

Toutes les archées et bactéries qui ont une taille chevauchante avec celle des virus ne sont pas intracellulaires. Le gigantisme de certains génomes viraux pose un problème, non pas à cause de leur présumée « filtrabilité », mais à cause de leur parasitisme. En effet, le parasitisme obligatoire, comme la symbiose, implique différentes pressions de sélection qui peuvent mener à une réduction de génome chez les bactéries (Toft et Andersson 2010) et chez les eucaryotes (Sundberg et Pulkkinen 2015). Cela est dû à une dépendance de l'organisme à son hôte et à l'échange possible de molécules entre les deux qui donnent de la redondance et donc pourrait diminuer la pression de sélection sur certains gènes de l'organisme intracellulaire. Le parasite peut se permettre de réduire son génome pour gagner en productivité. Pour mettre à jour cette tendance avec les génomes viraux connus actuellement il nous faut croire que :

1) La règle est fausse

2) La règle est vraie sauf pour les virus géants qui sont contraints de choisir une autre stratégie
3) La règle est toujours vraie, les virus géants sont issus de virus encore plus grands

Choisir l'hypothèse n° 1 est tentant mais cela reviendrait à ignorer la grande quantité d'exemples allant dans le sens d'une évolution réductive (Moran et Plague 2004; Wernegreen 2005; Sundberg et Pulkkinen 2015; Toft et Andersson 2010). Dans les articles cités cela concerne les organismes parasitaires et symbiotiques. Ces derniers sont particulièrement touchés par la réduction génomique comme en témoigne la taille des génomes des plastes et mitochondries qui fait en général quelques centaines de kb. L'hypothèse n° 1 est de fait éliminée. Cependant, cela n'implique pas une relation de cause à effet directe entre le mode de vie parasitaire et la taille du génome. Notons qu'il y a aussi parmi les cellules, chevauchement entre les fourchettes de taille du mode de vie libre et intracellulaire obligatoire (Sundberg et Pulkkinen 2015; Toft et Andersson 2010).

Choisir l'hypothèse n° 2 c'est adhérer à la théorie expansionniste. Cela revient à dire que les avantages liés à un grand génome (robustesse grâce à la redondance génomique, plus grand contrôle du cycle, protection face aux défenses le l'hôte, hypothétiquement une diversité d'hôtes possibles…) dépassent le coût d'avoir un grand génome (réplication plus longue, plus grande probabilité d'erreurs, génome plus difficile à empaqueter…). Cela peut être comparé à des forces qui s'opposent ; la force qui pousse le virus à acquérir de nouveaux gènes serait supérieure à celle qui le pousse à en perdre. Ainsi le virus géant serait plus compétitif dans son environnement donné. Ce scénario est soutenu par l'analyse de duplications des virus géants montrant que plus un génome est grand, plus il possède de gènes dupliqués (Filée et Chandler 2008). Il est également soutenu par le principe de parcimonie. D'après les groupes viraux de la famille des *Nucleocytoviricota* (voir plus loin, p. 29) présentant du gigantisme et leur phylogénie, le gigantisme serait apparu plusieurs fois (Bäckström et al. 2019). Encore selon ce principe, l'observation de la présence ou de l'absence de gènes orthologues chez cette famille pointe vers une origine multiple du gigantisme (Koonin et Yutin 2018). Le défaut de l'approche parcimonieuse est que celle-ci pose le présupposé d'un coût du gigantisme qui serait égal aux bénéfices apportés. Il suppose que gagner un gène coûte la même chose que d'en perdre un et donc une probabilité égale de transmettre son génome à la génération suivante. Les *Pandoraviridae* possédant le plus grand génome viral connu allant jusqu'à 2.5 Megabases sont de bons modèles pour étudier le gigantisme. Il a été proposé qu'ils étaient capable de créer des gènes *de novo* à partir de régions intergéniques (Legendre et al. 2018). Cela montre que, chez cette famille virale, quelque chose pousse à créer de la diversité génique. Des résultats expérimentaux et bioinformatiques récents tendent à montrer que les *Pandoraviridae* ont accru leurs génomes successivement dans l'histoire grâce à l'acquisition de gènes préférentiellement à partir d'une extrémité de leurs génomes linéaires (Bisio et al. 2022).

Enfin, selon l'hypothèse n° 3, les virus géants subissent les mêmes pressions de sélection que tout parasite obligatoire et la force poussant à réduire leurs génomes est supérieure à celle les poussant à agrandir leurs génomes. Cette conclusion à l'avantage de réduire le nombre d'hypothèses. La base de données largement utilisée NCVOG qui a été créée par l'outil COG

montre qu'il y avait en 2014 plus de 3000 groupes de gènes orthologues au sein de la famille virale comprenant les virus géants. Avec les nouveaux virus découverts aujourd'hui, ce chiffre serait certainement beaucoup plus grand. Cela signifierait que pour soutenir une évolution majoritairement réductive, le premier *Nucleocytoviricota* devait avoir un génome au moins aussi grand que celui d'un procaryote. Mêlé à l'abondance de gènes ressemblant à des gènes cellulaires chez les virus géants, cela a mené des auteurs à conclure à un quatrième domaine de la vie. Cette théorie n'a pas été soutenue par différentes études phylogénétique. Cependant, on peut citer en faveur d'une évolution réductive, des expériences qui montrent des virus géants subissant d'importantes délétions. Mimivirus, avec son génome d'origine de 1.2 Megabases, a été utilisé pour infecter son hôte naturel, l'amibe *Acanthamoeba polyphaga*. Après 150 passages successifs, Mimivirus avait perdu presque 300 kilobases de son génome réparties aux deux extrémités. Les modifications ont impliqué d'importants changements morphologiques (Boyer et al. 2011). Lausannevirus, un *Marseilleviridae*, a perdu 13 kilobases de son génome après 144 passages d'infection de *Acanthamoeba castellanii* (Mueller et al. 2017), un hôte de laboratoire et hôte naturel suspecté.

Le constat de duplications et de transferts horizontal de gènes vers les virus géants contredit une vision purement réductive de leur évolution. De même, les pertes de gènes observées expérimentalement ou estimées par analyse d'orthologie ne correspond pas à une vision purement expansionniste. La théorie qui progresse ces dernières années est celle de l'évolution en « accordéon ». Les virus géants auraient traversé des périodes d'expansion génomiques et des périodes de réduction génomique pour devenir ce qu'ils sont aujourd'hui. Une étude plus précise de ces histoires évolutives fluctuances pouraient peut-être permettre un jour de trancher entre les hypothèses n°2 et n°3. On imagine que pour s'adapter à un milieu changeant, une forte diversité génomique a pu être avantageuse. Ainsi, quand Lausannevirus est soumis a 144 cycles d'infection de l'amibe mais cette fois en co-culture avec *Estrella lausannensis*, une bactérie intracellulaire d'amibe appartenant aux Chlamydiae, il n'y a plus eu qu'une culture sur huit qui présentait une délétion (Mueller et al. 2017). Cela suggère que le gigantisme sert dans la compétition. Enfin, une évolution en accordéon a été observée chez le virus vaccinia. Des cellules humaines ont été infectées par le virus, mal adapté aux défenses de l'hôte. Le gène K3L du virus n'inhibe pas correctement la protéine kinase R (PKR), puissant antiviral. Après seulement 4 ou 6 passages, le gène K3L était dupliqué chez le virus. Au 10ème passage, le gène était en moyenne présent 3 à 4 fois dans les génomes viraux, parfois avec une ou deux mutations. Certaines cultures l'avaient jusqu'à 15 fois. Cependant, quand ces virus sont mis à présent en culture avec des cellules de hamster, pour lesquelles elles peuvent plus facilement outrepasser les défenses, le nombre de copies du gène K3L diminue à nouveau (Elde et al. 2012). Il semblerait que l'expansion génomique soit une réaction à un environnement difficile alors que la contraction génomique soit une réponse à un environnement connu. C'est ce que nous dit aussi l'expérience menée sur Lausannevirus.

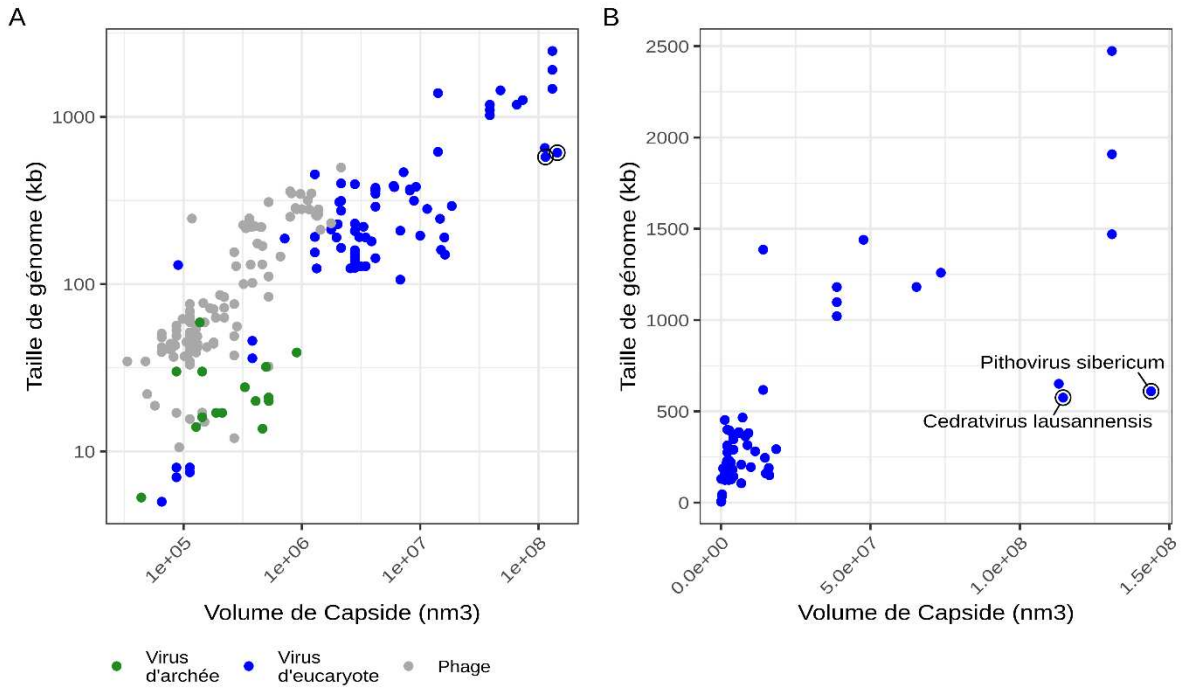


Figure 5 – Taille génomique associée au volume des particules virales (données issues de (Chaudhari, Inamdar, et Kondabagil 2021))
(A) Taille de génome et volume de capside sur échelles logarithmiques comparées pour plusieurs groupes. Les points entourés correspondent au *Pithoviridae*. (B) Données n'incluant que les virus d'eucaryotes et sans échelle logarithmique. Le point juste au-dessus de Cedratvirus lausannensis est Mollivirus sibericum.

Il semblerait y avoir une relation relativement linéaire (il faut passer aux échelles logarithmiques) entre le volume des particules virales et la taille de génome (Figure 5A). Chez les virus d'eucaryotes, les *Pithoviridae* cassent cette relation linéaire puisqu'ils ont un génome petit par rapport à leur volume de virion (Rodrigues et al. 2018) (Figure 5B). Le gigantisme chez les phages a été moins étudié que chez les virus de protozoaire. Néanmoins, la tendance actuelle est de dire que ces « jumbophages » proviennent de virus aux génomes plus petits et que cela a été permis par une capside plus grande (Nazir et al. 2021; Hua et al. 2017).

V. La famille des grands virus nucléocytoplasmiques

Les grands virus nucléocytoplasmiques forment un phylum viral appelé *Nucleocytoviricota*. Ce sont des virus d'eucaryotes à ADN double brin. Cela leur a valu leur ancien nom NCLDV pour « Nucleocytoplasmic Large DNA Viruses ». Ils ont la caractéristique de tous avoir un passage de leur cycle dans le cytoplasme de la cellule hôte, certains ont également une phase nucléaire. Il y a officiellement 1366 espèces et sous-espèces virales déclarées comme

appartenant à cette famille sur le site du Centre National pour l'Information Biotechnologique nord-américain (NCBI)[22] dont certains sont issus de la métagénomique.

Cette famille regroupe les *Pokkesviricetes* que sont les *Poxviridae* et les *Asfarviridae*. Elle regroupe également les *Phycodnaviridae* avec les *Pandoraviridae* et les *Molliviridae*, très divergents, mais qui se retrouve de manière récurrente associés aux virus d'algue par phylogénie (Yutin et Koonin 2013). Les plus proches parents de ces *Phycodnaviridae* sont les *Mimiviridae*. Enfin, la famille contient aussi les *Pimascovirales* que sont les *Marseilleviridae*, les alpha et Beta-*Iridoviridae*, les *Ascoviridae*, les *Pithoviridae* et Orpheovirus (Figure 6). Les *Pithoviridae* et Orpheovirus n'ont pas encore été formellement acceptés au sein des *Pimascovirales* mais ils partagent bien un lien de parentée (Figure 6). De plus, la lettre « P » de « *Pimascovirales* » vient à l'origine de *Pithoviridae* (Guglielmini et al. 2018). Les quatre grands clades de *Nucleocytoviricota* ; *Pokkesviricetes*, *Mimiviridae*, *Pimascovirales* et *Phycodnaviridae* (si l'on considère que les *Pandoravirus* en font partie), contiennent des familles de virus géants.

Les morphologies des virions sont diverses, la plupart étant icosaédriques, avec (*Mimiviridae*) ou sans (*Pandoraviridae*) fibrilles[23] et avec (Tupanvirus) ou sans (Mimivirus) queue (Figure 6). Les *Ascoviridae* et les *Poxviridae* sont ovoïdes et pourtant les deux ont la Protéine Majeurs de Capside (MCP) toujours présente dans leur capside. En revanche, les *Pandoraviridae* et les *Pithoviridae* sont plus grands et ont une forme d'amphore. La MCP n'est plus un composant de leurs virions, d'autant plus que les *Pandoraviridae* ont perdu le gène qui code pour cette protéine (Philippe et al. 2013).

La famille recouvre un grand panel d'hôtes eucaryotes qui sont répartis dans au moins 6 super-groupes (Needham et al. 2019). Les plus grands virus de la famille entrent dans la cellule par phagocytose. Les virus plus petits comme les *Poxviridae*, les *Iridoviridae* et les *Asfarviridae* entrent dans la cellule par endocytose ou macropinocytose (un type d'endocytose par invagination de la membrane cellulaire) (Rodrigues et al. 2016). Les *Marseilleviridae* ont également un virion plus petit que les *Mimiviridae*. Ils entrent dans la cellule par endocytose ou dans des vésicules contenant de nombreux virus qui sont alors phagocytées (Arantes et al. 2016). Cela a confirmé l'hypothèse selon laquelle le gigantisme du virion est une adaptation au mode d'entrée par phagocytose. L'amibe peut confondre un virus avec une bactérie de laquelle elle se nourrit et va ingérer le cheval de Troie. Les fibrilles de Mimivirus servent d'attache à la cellule amibienne (Rodrigues et al. 2016). Les *Phycodnaviridae* doivent à l'inverse traverser la paroi des algues ce qu'ils font grâce à des protéines embarquées qui vont dégrader cette paroi (Agarkova et al. 2021)… on voudrait presque le comparer à un métabolisme. Ensuite, le plus souvent leur membrane lipidique interne fusionne avec la membrane cellulaire (Rodrigues et al. 2016). Les *Poxviridae* peuvent eux aussi entrer dans la cellule par fusion.

[22] Selon la taxonomie au 1er juillet 2022

[23] Les fibrilles sont les structures protéiques qui couvrent la capside que l'on compare de manière imagée à des cheveux.

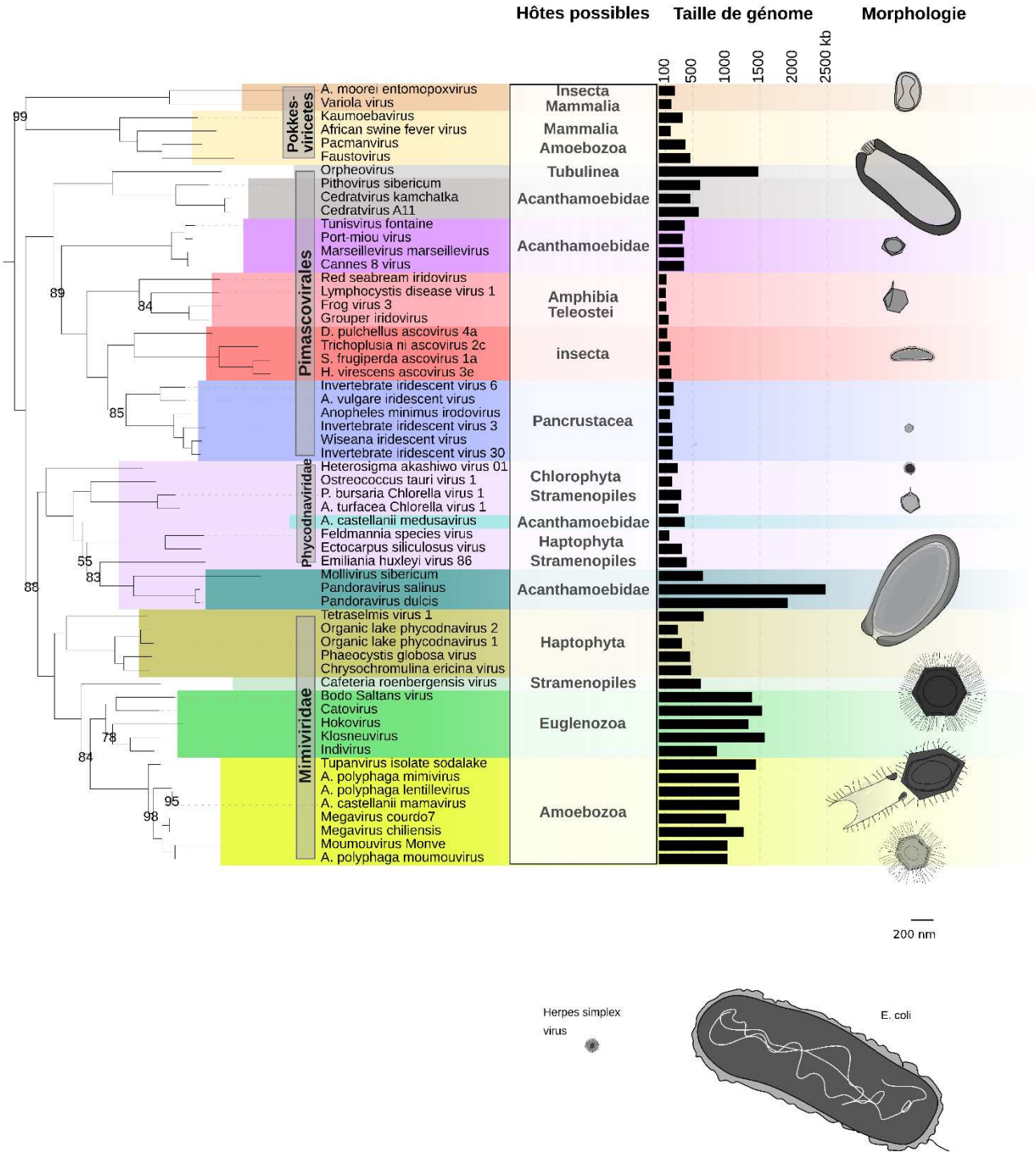


Figure 6 – Phylogénie des *Nucleocytoviricota*

Phylogénie calculée à partir de l'alignement de huit gènes marqueurs. Les valeurs de support des branches ont été calculés à partir de 1000 arbres. Uniquement les valeurs en dessous de 100% sont marquées. Les couleurs dépendent des différentes sous-familles. La morphologie des virions est à l'échelle. Tous les virions ont été dessinés à partir d'observations réelles de microscopie. Herpes simples virus et *E. coli* sont présentés pour comparaison.

Le cycle cellulaire des *Poxviridae*, des *Mimiviridae* et des *Pithoviridae* semble entièrement cytoplasmique. A l'inverse, les *Pandoraviridae*, les *Iridoviridae* et les *Asfarviridae* font la première étape de leur cycle dans le noyau de l'hôte. Les *Marseilleviridae* se placent en intermédiaire. Encodant une machinerie transcriptionnelle dans leurs génomes mais ces protéines n'étant pas présentes dans le virion, les *Marseilleviridae* ont besoin des protéines de l'hôte pour commencer la transcription. Ils y accèdent en rendant la membrane nucléaire perméable aux protéines et récupèrent ainsi la machinerie transcriptionnelle dans le cytoplasme. Le virus provoque également une forte invagination du noyau (Fabre et al. 2017). Tous les virus de la famille forment ce qu'on appelle des « usines virales » souvent proche du noyau. Dans cette usine virale aura lieu plusieurs étapes de la synthèse de nouveaux virions. Chez certaines familles comme les *Mimiviridae*, les virions se forment à la périphérie de l'usine virale alors que chez les *Pithoviridae*, ceux-ci se forment à l'intérieur. La libération de virions matures se fait par exocytose et par lyse cellulaire.

VI. Les Pithoviridae

Vue d'ensemble

En 2019, lorsque cette thèse a débuté, il y avait six *Pithoviridae* connus (Tableau 1) et au moins trois dont nous n'avons toujours pas le génome à ce jour. Tous ont été isolés à partir d'infections de *A. castellanii*. Il semblerait que l'eau douce soit un bon milieu pour trouver des *Pithoviridae* (Tableau 1). KC5/2 et Cedratvirus getuliensis ont également été isolés à partir d'eau douce (eau potable et eaux usées). Lurbovirus (c.-à-d. Cedratvirus lurbo), dont la séquence n'a pas non plus été rendue publique, provient du sol d'une rive de rivière près de Uppsala (Kördel et al. 2021).

Groupe	**Nom**	**Génome (kilobases)**	**% en GC**	**Milieu**	**Publication**
Pithovirus	*P. sibericum*	610	36	Permafrost sibérien de 30 000 ans	(Legendre et al. 2014)
	P. massiliensis	686	35	Eaux usées de la Ciotat	(Levasseur et al. 2016)
Cedratvirus	*C. A11*	589	43	Inconnu, Algérie	(Andreani et al. 2016)
	C. lausannensis	575	43	Usine de traitement de l'eau potable à Morsang-Sur-Seine	(Bertelli et al. 2017)
	C. zaza	561	43	Homogénat du champignon alpova sp. à Toulon	(Rodrigues et al. 2018)
	Brazilian cedratvirus	460	43	Eau de Belo horizonte	(Rodrigues et al. 2018)

Tableau 1 – *Pithoviridae* séquencés avant le début de ces travaux de thèse

Les *Pithoviridae* sont séparés en deux groupes bien distincts ; les *Pithovirus* et les *Cedratvirus*. Ces groupes se distinguent par leur histoire évolutive et par leur morphologie (voir p. 35). Notons également que le génome des Pithovirus est de quelques dizaines de kilobases plus grand.

La méthodologie pour isoler de nouveaux *Pithoviridae* varie peu. Lorsqu'il s'agit de sol, l'échantillon est mis en solution. Cette solution est mise en présence d'un fongicide puis laissé décantée. Le surnageant est récupéré. Celui-ci est centrifugé et les deux fractions servent à inoculer une culture de *A. castellanii*. Le milieu de culture utilisé contient des antibiotiques pour éviter les infections bactériennes. Enfin, la lyse cellulaire est évaluée en microscopie optique. Si un échantillon est positif, les virus sont séparés des cellules par centrifugation avant purification. Dans le cas de P. sibericum, la méthode a été réalisée avec succès en duplicata à partir de sous-échantillons différents (Legendre et al. 2014).

L'hôte

Le seul hôte ayant permis l'isolement de nouveaux *Pithoviridae* est *Acanthamoeba castellanii*. C'est une amibe du clade Amoebozoa. Ce pathogène opportuniste de l'homme peut causer des infections de la cornée et kératites si mis en contact avec les yeux comme à travers des lentilles de contact. Il peut également être la cause d'encéphalites chez les personnes immunodéprimées (Thomas et al. 2010). Les amibes du genre Acanthamoeba peuvent survivre dans de nombreux environnements. Elles sont très abondantes dans les sols et dans les environnements d'eau douce où elles vivent attachées au sédiment ou aux particules en suspension. Elles peuvent également survivre dans un environnement marin et dans l'atmosphère. Le transport de particules à travers l'atmosphère serait d'ailleurs une stratégie pour coloniser de nouveaux environnements (Rodríguez-Zaragoza 1994). Elles sont communes dans les eaux domestiques malgré les traitements (Thomas et al. 2010). Ces amibes se nourrissent de champignons, d'algues mais avant tout de bactéries et ont par conséquent un impact non négligeable sur l'écosystème à travers la prédation de ces dernières (Rodríguez-Zaragoza 1994) notamment sur les biofilms (Thomas et al. 2010). On trouve souvent des bactéries intracellulaires chez l'amibe, qu'elles soient symbiotiques ou parasitaires. Certaines de ces bactéries peuvent être des pathogènes humains et il a été proposé que les mécanismes de résistance à la phagocytose par l'amibe soient similaires aux mécanismes de résistance aux macrophages chez l'homme (Greub et Raoult 2004).

Quand elles sont soumises à un stress tel que le manque de nourriture, les changements de pH où l'exposition à un poison, les amibes peuvent passer de l'état trophozoïtes (Figure 7A) à l'état de cyste (Figure 7B). L'enkystement passe par la déshydratation et la formation de cette coque protectrice faite de parois cellulosiques (paroi de l'endocyste) et d'une membrane externe de protéines et de lipides (paroi de l'ectocyste) (Thomas et al. 2010). *A. castellanii* peut également survivre en anaérobiose. En réponse au manque d'oxygène (<1.1%), 77% des cellules se retrouvent enkystées. Remises dans un milieu oxygéné, celles-ci redeviennent trophozoïte (Turner, Biagini, et Lloyd 1997). De nombreuses amibes, notamment du genre *Acanthamoeba*, ont été isolées à partir d'échantillons de pergélisol Nord-sibériens datant du pléistocène tardif (Malavin et al. 2020).

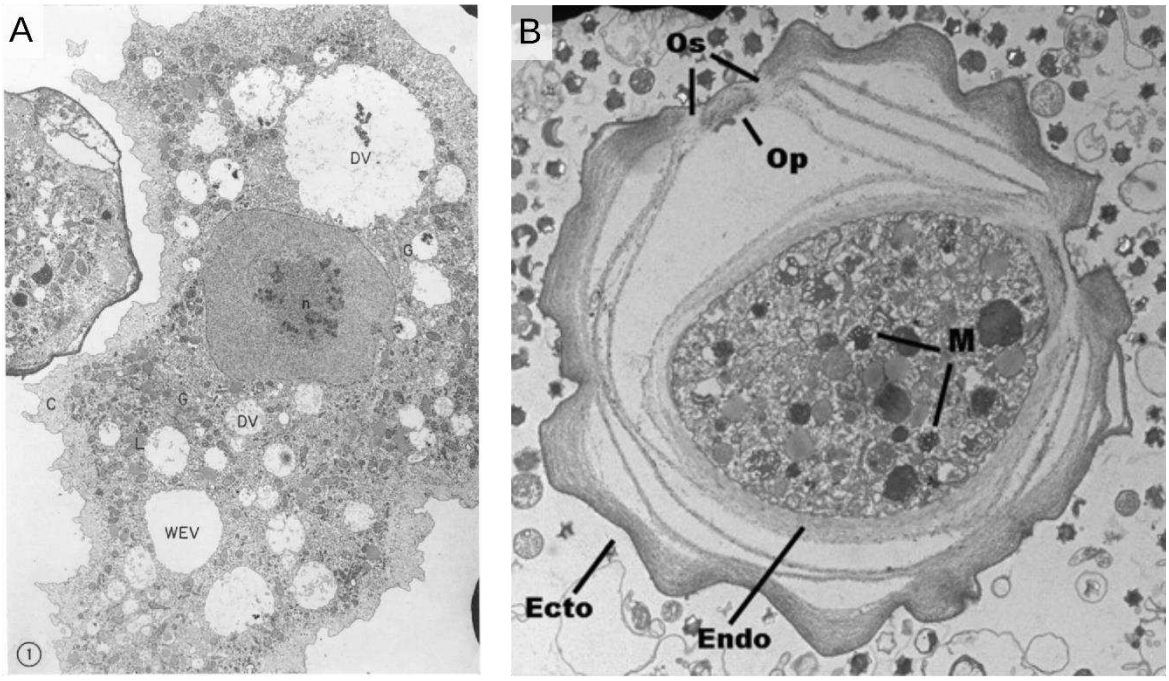


Figure 7 – *Acanthamoeba castellanii* vue au microscope électronique
(A) Trophozoïte de *A. castellanii* (grossissement 5900) provenant de (Bowers et Korn 1968) présentant un cortex « C », des vacuoles digestives « DV », une vésicule d'expulsion d'eau « WEV » et des gouttelettes lipidiques « L ». (B) Cyste d'*A. castellanii* (grossissement 4500) obtenu par (Thomas et al. 2010) avec 5 à 7 jours d'incubation de trophozoïtes à 33°C en solution saline. Il présente un endocyste « Endo », un ectocyste « Ecto », des ostioles « Os », un opercule « Op » et des mitochondries « M ».

A. castellanii possèderait 35 chromosomes et sa ploïdie exacte n'est pas connue. Elle a été estimée étant autour de 25 (Byers 1986). Une couverture inégale des différents chromosomes laisse penser à une aneuploïdie, c'est-à-dire un nombre inégal de copies par chromosome (Matthey-Doret et al. 2021).

La découverte du *Pithoviridae* KC5/2 dans une amibe *A. castellanii* environnementale nous donne une information précieuse : l'hôte naturel des *Pithoviridae* est le même que son hôte de laboratoire. En effet, nous pourrions croire que le mode d'isolement de nouveaux virus géants sur une culture amibienne force le parasitisme de cette espèce en particulier. L'étude de KC5/2 nous apprend également que ce virus est assez spécialiste des *Acanthamoeba* du groupe II (Hoffmann et al. 1997). Ce groupe définit par la morphologie de leurs cystes contient les expèces *A. mauritaniensis, A. castellanii, A. polyphaga, A. quina, A. divionensis, A. triangularis, A. lugdunensis, A. griffini, A. rhysodes, A. paradivionensis and A. hatchetti* (Khan 2006). KC5/2 infecte également *Comandonia operculata* (Michel, Schmid, et al. 2003, 5). Les découvreurs de Cedratvirus lausannensis soulignent la faible permissivité au changement d'hôtes puisque ce virus n'a presque pas réussi à infecter *Acanthamoeba comandoni*. Par ailleurs*, D. discoideum*, des cellules d'arthropode et de mammifère n'ont pas été permissives au virus (Bertelli et al. 2017). Ceci est en contradiction avec KC5/2 qui aurait infecté des cellules de mammifère (Michel, Schmid, et al. 2003). Aucune précision à ce sujet n'est donnée.

Morphologie du virion

Les virions de *Pithoviridae* font entre 0.9 et 1.5 µm de long et 0.5 µm de large. Les tailles varient selon l'espèce mais une très grande plasticité a souvent été remarquée au sein d'une même espèce. Les particules de *P. sibericum* peuvent aller de 1 µm à 2.2 µm de long (Figure 8A).

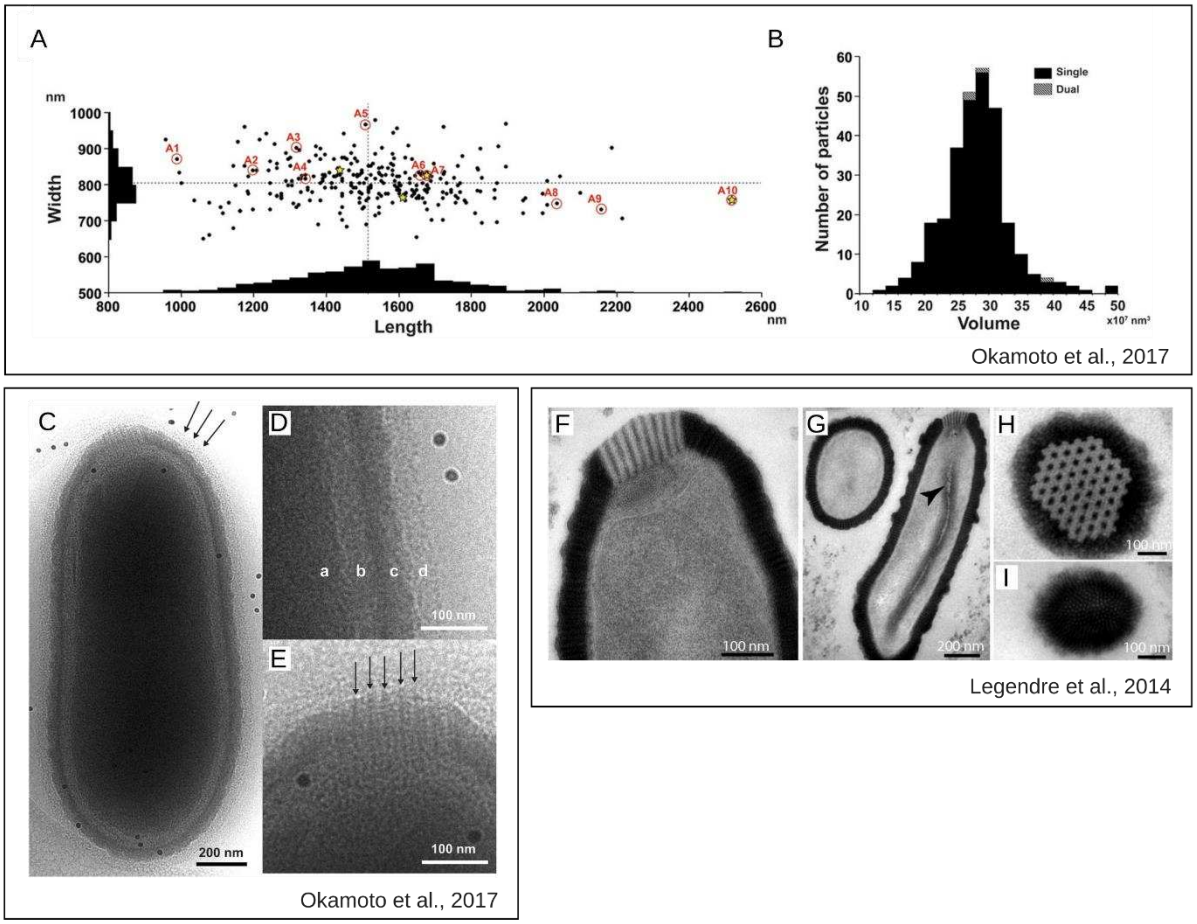


Figure 8 – Morphologie de *Pithovirus sibericum* vue de trois façons
(A) La première image (Okamoto et al. 2017) montre les dimensions des particules virales mesurée à partir d'images obtenues par cryo-microscopie électronique à haut voltage et (B) également de nombre de pores apicaux présents. (C-E) Les images du deuxième panneau ont été obtenues par cryo-microscopie électronique par filtration d'énergie. (C) On y voit la particule entière, les flèches indiquant le bord externe d'une couche à basse densité. (D) Un agrandissement de la couche externe montre le nucléoïde (a), l'espace intérieur (b), le tégument (c) et la couche basse densité (d). (E) Une troisième image montre le pore apical. (F-I) Le troisième panneau correspond à des images de microscopie électronique (Legendre et al. 2014). (F) On y voit une vue latérale du pore apical, (G) la particule entière selon deux vues perpendiculaires, la flèche indique une structure vue épisodiquement mais qui s'est plus tard avéré être un artéfact), (H) le bouchon au niveau du pore apical vu du dessus et enfin (I) la particule vue du dessous.

De l'intérieur vers l'extérieur le virus possède un nucléoïde homogène séparé par une membrane, un faussé, un tégument épais et très dense et une couche d'environ la même épaisseur que le tégument mais à très basse densité donc à peine visible (Figure 8D). Le tégument est fait de très claires stries perpendiculaires à la surface de la particule (Figure 9A et B). A l'apex de la particule se trouve un pore refermé par un bouchon. Ce bouchon est fait de stries moins denses et plus épaisses que le tégument, chacune séparées de 15 nm (Figure 8F et H). Vu du dessus cela donne une structure en nid d'abeille. Entre le bouchon et la membrane interne se trouve la racine du bouchon qui a aussi été qualifiée de membrane enroulée (Figure 8F). La grande différence morphologique entre les Pithovirus et les Cedratvirus est que ces derniers ont deux bouchons, chacun à une extrémité du virion, au lieu d'un (Figure 8B, Figure 9A).

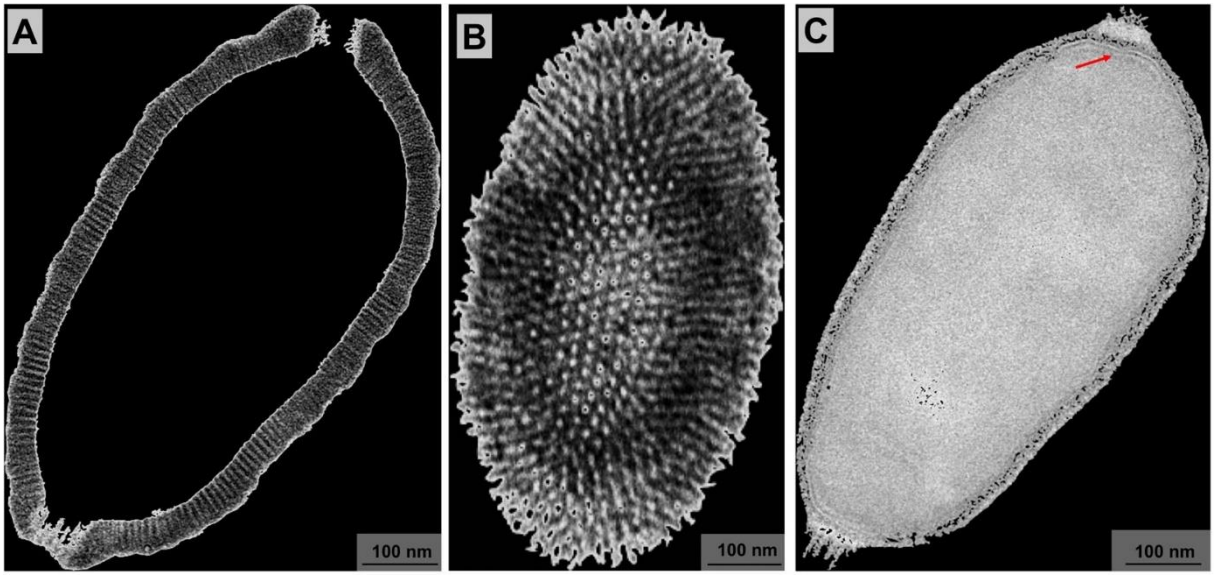


Figure 9 – Particule virale de *Cedratvirus getuliensis* observée au microscope électronique à transmission (dos Santos Silva et al. 2018)
Pour différencier ces structures, les images ont été retravaillées par les auteurs avec un logiciel spécialisé. (A) On y voit le tégument et les deux bouchons apicaux, (B) une coupe externe de la particule et (C) l'intérieur de la particule. La flèche rouge indique la membrane interne qui sépare le nucléoïde du reste.

Cycle infectieux

Dans sa vue d'ensemble, il y a peu de variations dans le cycle infectieux des *Pithoviridae*. L'entrée se fait par phagocytose. Cela a été prouvé par des observations mais aussi par l'annulation de l'infection en présence de cytochalasin D ou de chloroquine qui inhibent l'endocytose (dos Santos Silva et al. 2018). Une fois dans le phagosome, le virus perd le bouchon apical et relargue le nucléoïde toujours avec sa membrane qui va fusionner avec celle du phagosome (Figure 10).

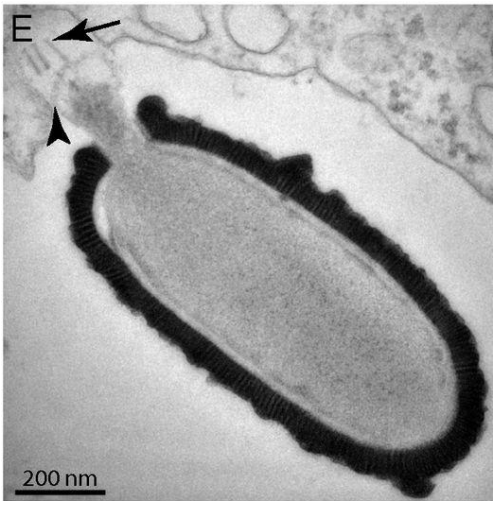


Figure 10 – Image de l'ouverture de la particule de Pithovirus sibericum vue par microscopie électronique (Legendre et al. 2014)
Les flèches indiquent des parties du bouchon qui a été expulsé et la membrane interne du virion sortie de la particule.

Après fusion des membranes, le contenu du nucléoïde est relargué dans le cytoplasme. Comme souvent chez les virus géants, il y a une phase d'éclipse où aucun changement morphologique n'est visible dans la cellule. Environ 4 heures après infection on aperçoit une grande zone riche en ADN par coloration au DAPI ou, en microscopie électronique, une zone homogène, claire et peu dense. C'est l'usine virale (Figure 11). Contrairement à celle de Mimivirus, les contours sont peu définis. On reconnait déjà une cellule infectée grâce à cette zone mais aussi par les vacuoles de la cellule : moins nombreuses mais en moyenne plus grandes.

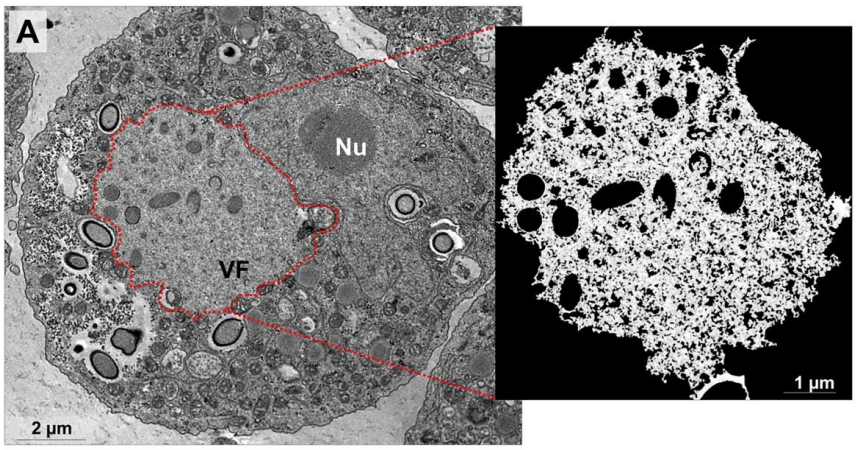


Figure 11 – Usine virale au sein d'une cellule infectée par C. getuliensis (dos Santos Silva et al. 2018)
L'usine virale « VF » transparente aux électrons a été entourée en rouge. Le noyau est indiqué « Nu ».

Dans l'usine virale qui prend une place considérable dans la cellule, commencent à se former les nouveaux virions. D'abord, on aperçoit des premiers fragments de membrane, souvent aux angles droits. Les bouchons, plus denses, commencent à se former au même moment. La particule se présente progressivement comme un cylindre vide. Celui-ci sera par la suite rempli. Parfois il semblerait qu'à l'inverse, le contenu interne et la membrane soient « tricottés » en même temps (Legendre et al. 2014). Il aurait été observé que le deuxième bouchon des *Cedratvirus* se rajoute après remplissage (dos Santos Silva et al. 2018). La dernière phase de maturation est l'épaississement du tégument. Les virions matures sont retrouvés à l'extérieur de l'usine virale, dans une vacuole ou non (Figure 12). Le virion mature a alors deux modes de sortie : l'exocytose ou la lyse cellulaire qui a lieu en général 10 à 12 heures après infection.

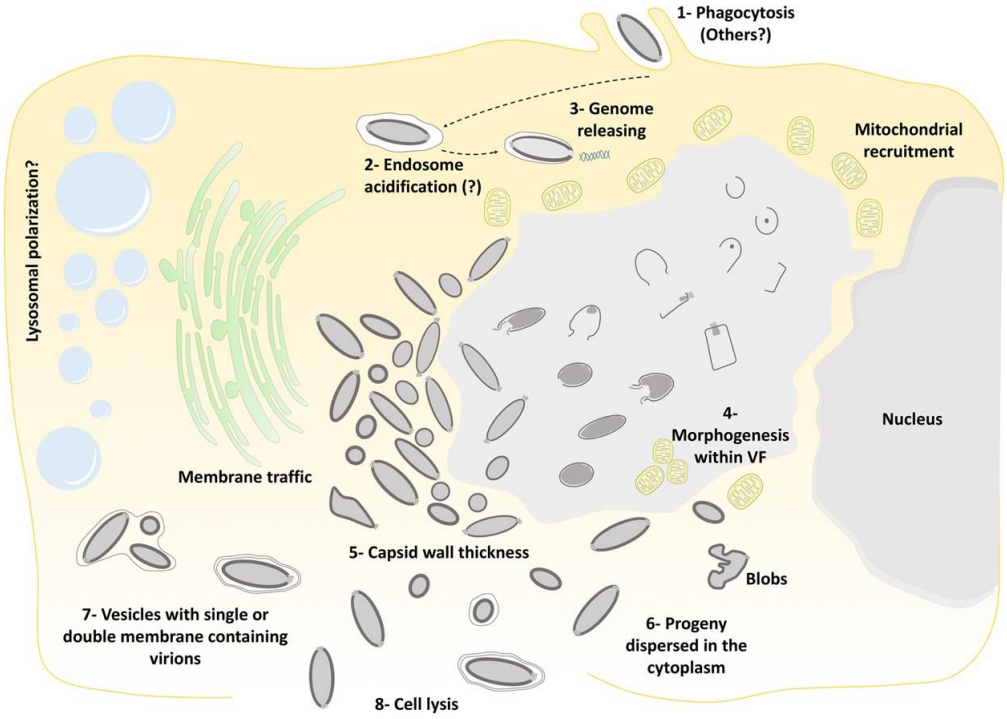


Figure 12 – Cycle infectieux de *Cedratvirus getuliensis* schématisé (dos Santos Silva et al. 2018)
1 – Phagocytose
2 – Acidification de l'endosome
3 – Relargage du génome
4 – Morphogénèse au sein de l'usine virale
5 – Epaississement du tégument
6 – Progéniture dispersée dans le cytoplasme
7 – Vésicule à simple ou double membrane contenant les virions
8 – Lyse cellulaire

Tardivement, on peut voir apparaitre des formes totalement aberrantes lors d'infections par un Pithovirus ou un Cedratvirus. Ce sont ces longs téguments qui prennent des formes étonnantes (Legendre et al. 2014). Du matériel ressemblant à la texture de l'intérieur du virion mature peut rester confiné dans ces « blobs » ou prendre une place diffuse dans l'usine à virion. Ces blobs peuvent aussi apparaitre dans le cytoplasme comme dans la Figure 12. Ils peuvent contenir des morceaux de bouchon.

Il a été estimé chez Cedratvirus lurbo que 9 particules sur 10 sont infectieuses. Surtout, les auteurs ont calculé un rendement allant de 220 à 660 particules par cellule infectée. Une analyse thermogravimétrique montre également que le virus est beaucoup plus riche en carbone résiduel que l'amibe (62 contre 28%). En prenant en compte toutes ces données, les auteurs estiment que 5 % à 12 % de la biomasse amibienne est convertie en virus après une infection. Pour corroborer ces résultats les auteurs ont également calculé ce paramètre en se basant sur des données de microcopie à rayons X. Ceci donne 6 à 17% de la biomasse convertie (Kördel et al. 2021).

Protéomique et transcriptomique

Seul les protéomes et transcriptomes de P. sibericum ont été étudiés.

La particule virale est complexe avec 159 protéines retrouvées dont deux tiers de fonction inconnue. La machinerie transcriptionnelle, elle est au complet (Legendre et al. 2014). Cela vient corroborer les observations montrant un cycle cellulaire entièrement cytoplasmique ; avec cette machinerie embarquée, le virus peut commencer la transcription rapidement après infection. Sans cela il serait dépendant de la machinerie du noyau de l'hôte, ne serait-ce que pour transcrire sa propre machinerie transcriptionnelle… La particule contient également des protéines glycosylées. Des protéines de l'hôte sont en revanche très peu abondantes (Legendre et al. 2014).

Le transcriptome de P. sibericum analysé tout au long du cycle montre que 82 % des régions codantes prédites sont transcrites. Il a également permis d'étudier les régions non traduites des transcrits. Ces régions sont très courtes, de 8 et 16 nucléotides au niveau des extrémités 5' et 3' respectivement. Dans cette petite région en 3' il y a souvent une séquence palindromique que les auteurs associent hypothétiquement à un signal de terminaison et de polyadénylation (Legendre et al. 2014) comme chez Mimivirus (Byrne et al. 2009).

Génomes

Les *Pithoviridae* ont un génome circulaire allant de 460 kilobases pour celui de *Brazilian cedratvirus* à 686 kilobases pour celui de *Pithovirus massiliensis*. Ces génomes codent pour 467 à 574 protéines chez *P. sibericum* et *C. A11* respectivement. Le pourcentage en guanine et en cytosine (GC%) est plus stable ; aux alentours de 36 chez les *Pithovirus* et de 43 chez les *Cedratvirus* (Tableau 1, p. 32). Le génome de Brazilian cedratvirus est très étonnant puisqu'il est beaucoup plus petit que ceux des autres *Cedratvirus* (alors C. zaza, C. A11 et C. lausannensis) mais aussi complètement réarrangé (Rodrigues et al. 2018).

Dès la découverte de P. sibericum, on remarque que 21.2% de son génome est composé de répétitions palindromiques non codantes et non transcrites de 150 paires de bases. Celles-ci sont régulièrement réparties et organisées dans des régions de 5000 paires de bases. Ces

régions sont très pauvres en guanines et en cytosines (Legendre et al. 2014). Les Cedratvirus, eux, sont dépourvus d'une telle séquence (Andreani et al. 2016).

Une étude de génomique comparative entre *P. sibericum* et *P. massiliensis* a été menée profitant du fait que l'un d'entre eux était un « fossile vivant ». Tous les gènes homologues ont été comparés deux à deux pour estimer la pression de sélection qui est appliquée sur eux. Cette pression de sélection est estimée par le ratio $\frac{dN}{dS}$ tel que :

$$dN = \frac{nombre\ de\ mutations\ non\ synonymes}{\text{nombres de positions où une mutation non synonyme est possible}}$$

$$dS = \frac{\text{nombres de mutations synonymes}}{\text{nombres de positions où une mutation synonyme est possible}}$$

Les taux de substitutions non synonymes sur substitutions synonymes semblent du même ordre que ceux des bactéries Neisseria et plus fable que chez les Rickettsia ou Helicobacter. En prenant en compte les 30 000 ans de *P. sibericum*, ils calculent un taux de substitutions de 2.6 x 10^{-5} par site et par an, c'est-à-dire dans la tranche haute par rapport aux autres organismes à ADN (acide désoxyribonucléique) double brin (y compris les animaux). En fonction du ratio $\frac{dN}{dS}$ cela donne un taux de mutations au niveau des acides aminés de 3 x 10^{-6} (Levasseur et al. 2016). Cette estimation est à prendre avec précaution car *P. sibericum et P. massiliensis* sont assez divergents (la ressemblance moyenne en nucléotides est de 83.6 %) et par conséquent, il est probable que bien plus que 30 000 ans les séparent. De plus, nous savons que *P. sibericum* a pu résister tout ce temps dans le pergélisol et cela nous dit que l'activité de ces virus peut être hétérogène incluant de longues périodes sans rencontrer une amibe.

Les clades voisins – Les *Pimascovirales*

Orpheovirus

Le plus proche virus de la famille des *Pithoviridae* n'a pas encore de groupe puisqu'il est pour le moment le seul représentant de la famille. Il s'agit d'Orpheovirus. C'est un virus géant au génome de 1.5 Mégabases et présentant encore une nouvelle morphologie rappelant une amphore. Il mesure de 0.9 à 1.1 µm de long. Il possède un génome riche en adénosine et thymine (AT) (75%) faiblement codant (66.4%). 47% de ses gènes n'ont aucun homologue connu.

A la place du dense tégument des Pithovirus on retrouve une couche tout aussi épaisse mais moins dense en composée de fibrilles (Andreani et al. 2018). Sa morphologie rappelle un peu plus celle des *Pandoravirus* que celle des *Pithoviridae*. Il aurait en revanche une couche interne référée comme capside. On voit aussi apparaître des virions défectueux en fin d'infection par Orpheovirus mais ceux-ci sont très différents des « blobs » des *Pithoviridae* (Souza et al. 2019).

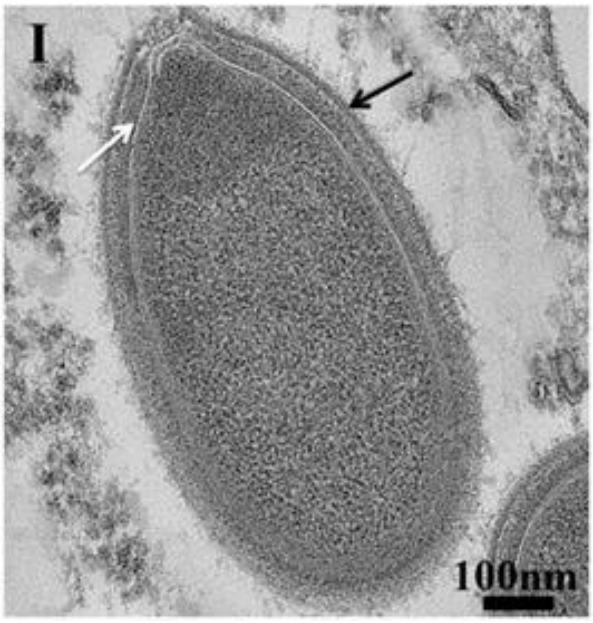


Figure 13 – Particule mature d'Orpheovirus vue après une infection de *Vermamoeba vermiformis* au microscope électronique (Andreani et al. 2018)
La flèche noire montre la membrane externe et la flèche blanche, la membrane interne.

Ce virus a été isolé sur une culture de *Vermamoeba vermiformis*, une amibe de la classe des Tubulinea. Après attachement, des pseudopodes vus par microscopie électronique à transmission indiquent un mode d'entrée dans la cellule par phagocytose. Une fois dans la cellule, le contenu du virion est relargué à partir de l'ostiole. Une usine virale est formée proche du noyau dont la membrane reste intacte. Comme chez les *Pithoviridae*, des virions à différents stades de maturation sont visibles dans l'usine virale. Leur formation commence par une fine structure appelée croissant[24]. Ces structures sont étendues et remplies avec leur contenu interne. Les feuillets viennent ensuite s'ajouter en périphérie du nouveau virion. Les virions matures apparaissent dans des vacuoles dans le cytoplasme. La sortie se fait par exocytose et lyse cellulaire (Souza et al. 2019).

Orpheovirus a donc une morphologie et un génome très différent des *Pithoviridae* puisque sur ses 1200 gènes, 946 n'ont pas d'homologue chez les *Cedratvirus* ou *Pithovirus* connus à l'époque (Andreani et al. 2018). On peut cependant souligner la grande ressemblance avec le cycle infectieux des *Pithoviridae*.

Les *Pimascovirales*

Les *Pimascovirales* comprend également les *Marseilleviridae*, *Iridoviridae* et *Ascoviridae* (Figure 14). Récemment, des auteurs réalisèrent que la famille des *Iridoviridae* était paraphylétique (Toenshoff et al. 2018). Cela n'a pas encore été ratifié par l'ICTV. Pris dans leur ensemble, la famille n'a d'ailleurs que 26 gènes cœur. Pour que le groupe soit monophylétique il suffirait de classer les *Ascoviridae* au sein des *Iridoviridae* puisque ceux-ci ont un ancêtre commun. En attendant, les *Iridoviridae* sont séparés en deux groupes. Le premier, les alpha-,

[24] Traduction de « crescent »

infectent les vertébrés (les poissons osseux, les amphibiens et reptiles) et le second, les beta-, infectent les invertébrés (crustacés et insectes).

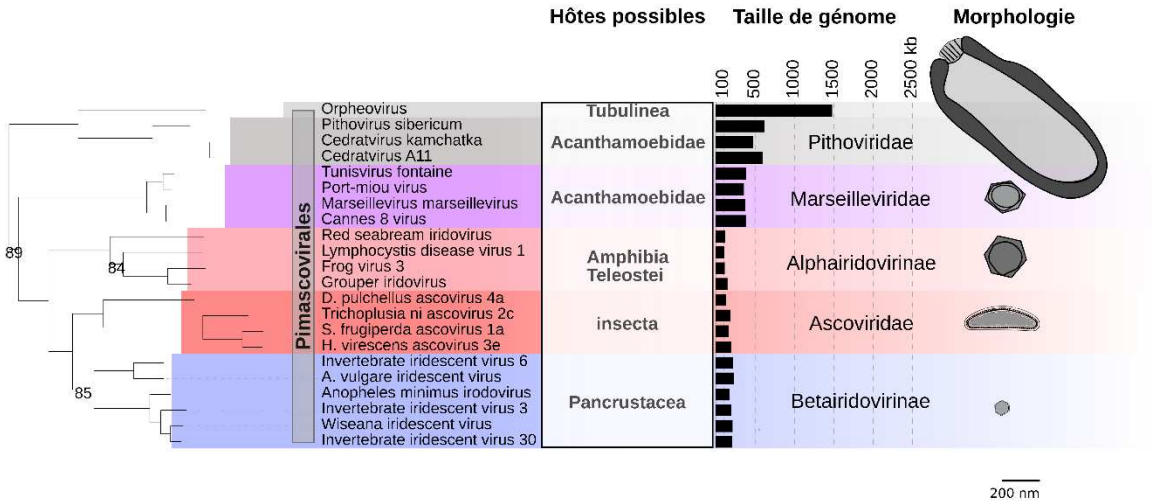


Figure 14 – Phylogénie et diversité des *Pimascovirales*, zoom de la Figure 6

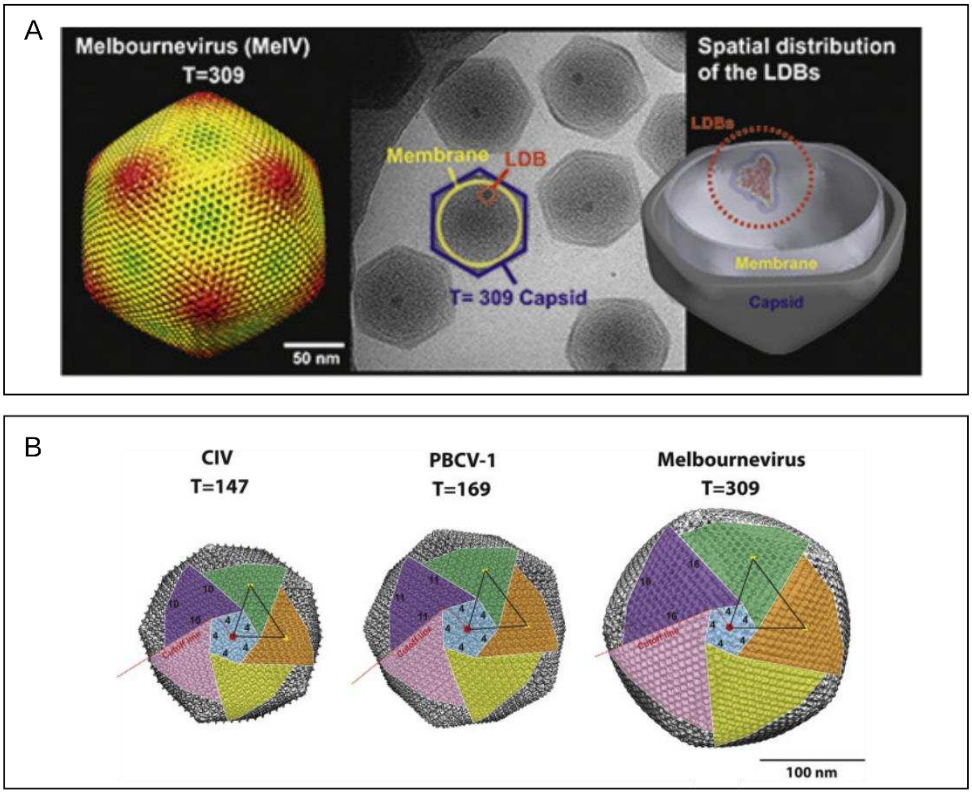


Figure 15 – Structure du virion de Melbournevirus et comparaison avec CIV et PBCV-1 (Okamoto et al. 2018)
Les images ont été obtenues par cryo-microscopie électronique. (A) Les images brutes annotées ainsi que la reconstitution 3D montrent un grand corps dense (Large and Dense Body, LDB) à l'intérieur de la particule. (B) Melbournevirus a un nombre de capsomère (trois protéines de capsides formant un triangle) supérieur à Chilo iridescent *virus* (CIV) et PBCV-1.

Les *Marseilleviridae* sont également des virus d'amibes isolés pour la première fois grâce à *Acanthamoeba polyphaga*. La plupart proviennent de milieux d'eau douce. Ils ont une particule icosaédrique d'environ 250 nm de diamètre (Figure 15A).

Les *Alpha-* et *Betairidovirinae* ont un diamètre de 120 à 350 nm avec une capside icosaédrique (exemple de CIV, Figure 15B) (Tsai et al. 2005). Ils peuvent être ou non enveloppés et possèdent une membrane interne dont les lipides sont différents de ceux de la membrane plasmique (Chinchar, Yu, et Jancovich 2011; İnce et al. 2018). Les *Ascoviridae* ont également la protéine majeure de capside comme protéine majoritaire dans leur virion (Asgari et al. 2021) mais leur morphologie est différente. Les virions mesurent environ 130 nm de large par 350 nm de long. Ils ont une forme ovoïde (Figure 16B) aplatie sur un côté (Figure 16D). L'ADN et les protéines à l'intérieur de la particule sont séparés du reste par une bicouche lipidique. On y trouve ensuite une couche protéique et enfin un couche lipidique externe (Figure 16D).

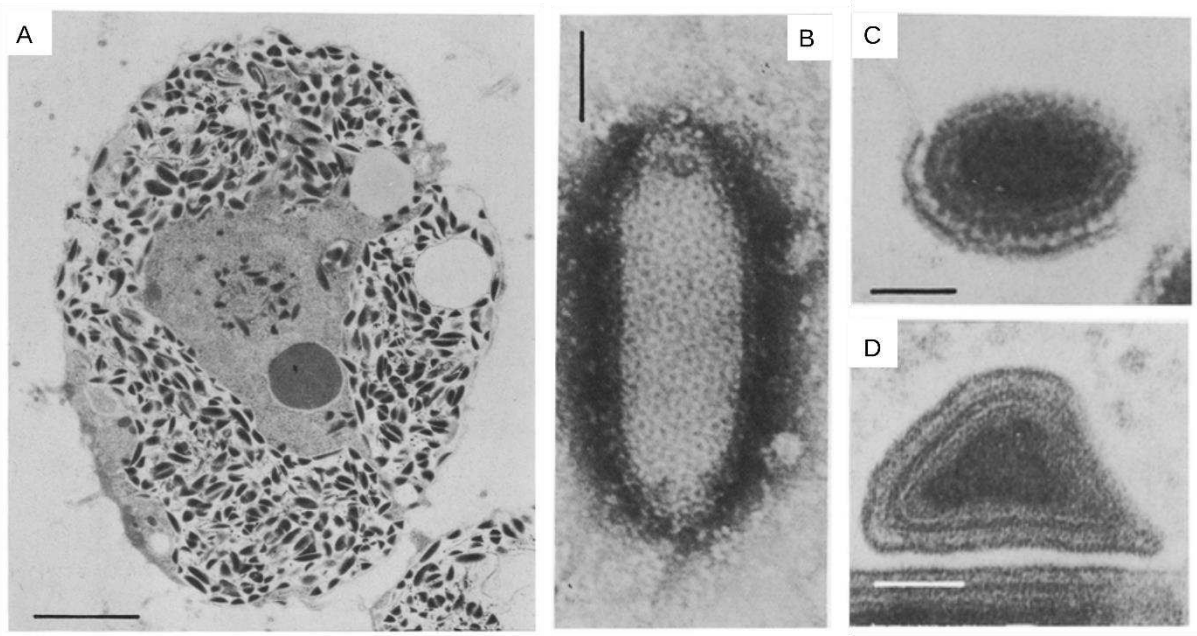


Figure 16 – Observations de Ascovirus et de sa cellule hôte au microscope électronique (Federici, Vlak, et Hamm 1990)
(A) Coupe ultrafines à travers de vésicules virales dans l'hémolymphe de *Trichoplusia ni* avec barre d'échelle de 2 µm. (B) Particule virale dans l'hémolymphe de *S. frugipeda* avec barre d'échelle correspondant à 100 nm. (C) Coupe transversale ultrafines à travers une particule virale pendant enveloppement de l'isolat provenant de *S. frugiperda*. (D) Coupe transversale ultrafine d'un virion dans *T. ni*. Les barres d'échelle correspondent à 50 nm.

Chez Frog Virus 3 l'entrée dans la cellule se fait par endocytose. La première phase du cycle est nucléaire, pourtant, le cycle se base largement sur la machinerie du virus pour la transcription et la réplication. La deuxième phase se passe dans le cytoplasme où l'ADN est méthylé et concaténé. Cet ADN concaténé, répliqué sera par la suite clivé pour obtenir des génomes uniques. Au sein de ce qui ressemble à une usine virale nommé « site d'assemblage », les capsides sont formées. La synthèse des protéines nécessaires se fait à l'extérieur de ce site dépourvu de ribosomes. L'ADN viral est ensuite empaqueté dans les

capsides. Les virions matures sont souvent rassemblés dans des réseaux paracristallins ressemblant à un nid d'abeille. Pour sortir, les virions matures bourgeonnent à partir de la membrane plasmique (Chinchar, Yu, et Jancovich 2011). Chez les *Betairidovirinae*, les virus peuvent aussi former des réseaux paracristallins dans la cellule et sortent après la lyse. Ils sont alors nus (İnce et al. 2018). Le cycle infectieux des *Ascoviridae* est différent et très étonnant puisque les virus étudiés utilisent en général comme vecteur une guêpe endoparasite qui les transmet dans ses œufs. Le virus peut être délétère ou non pour la larve parasitoïde mais il est bien délétère pour la larve parasitée (Asgari et al. 2021). L'infection provoque une invagination de l'enveloppe nucléaire et une croissance cellulaire extraordinaire qui multiplie par 5 ou par 10 son diamètre. Le noyau s'élargit jusqu'à rupture de sa membrane (Asgari et al. 2021). On comprend ainsi comment le cycle des *Marseilleviridae,* cytoplasmique mais avec échanges à travers la membrane nucléaire (p. 32), se place en intermédiaire entre les cycles en partie nucléaires des virus décrits ici et des *Pithoviridae*, cytoplasmiques.

Le génome circulaire des *Marseilleviridae* est long de 340 à 390 kilobases et code pour 450 à 550 protéines (Blanca et al. 2020). Ces virus ont également été étudiés pour leur surprenante structure génomique ; leur génome est compacté dans des nucléosomes. Les *Marseilleviridae* codent pour des histones fusionnées. Ainsi, les sous-unités H2A et H2B mais aussi H3 et H4 sont codées dans un même gène, formant une seule protéine. De plus, les nucléosomes sont stables dans la capside mais pas dans le cytoplasme de la cellule hôte. Cela permet de décondenser l'ADN après infection et de rapidement commencer la transcription (Liu et al. 2021).

Etonnamment, le génome des *Iridoviridae* et *Ascoviridae* sont linéaires et possèdent des répétitions terminales inversées qui permettent la jonction d'extrémités non homologues. Cela sert la réplication. On dit qu'ils ont un génome circulairement permuté. Ce phénomène a été démontré chez Frog Virus 3 (Chinchar, Yu, et Jancovich 2011) et est suspecté chez un *Ascoviridae* (Bideshi et al. 2006). Les génomes de ces trois groupes de virus contiennent entre 100 et 200 kilobases environ. Ce sont donc des virus complexes et leurs gènes sont transcrits de manière régulée avec des gènes très précoces, précoces différés et tardifs (Majji et al. 2009).

VII. Génomique des *Nucleocytoviricota*

Les virus isolés

Le génome des virus géants est aussi variable que leur morphologie. Ils peuvent être circulaires ou linéaires avec des répétitions terminales inversées de part et d'autre de leurs génomes. Ces régions font généralement plusieurs kilobases. Les génomes linéaires peuvent être circulairement permutés comme chez les *Iridoviridae* ou non. Chez les *Phycodnaviridae* et chez les *Asfarviridae* il a été montré que la fin des génomes linéaires se termine par une liaison covalente donnant des structures en tête d'épingle (Figure 17). Même au sein d'un même ordre, les *Pimascovirales*, nous avons vu que plusieurs conformations génomiques pouvaient être représentées.

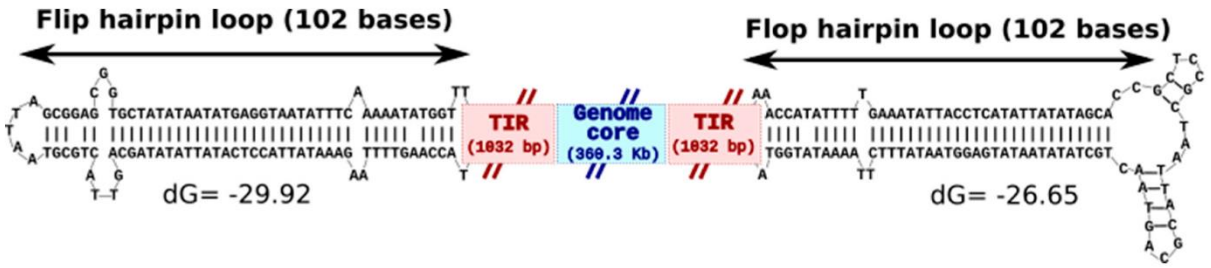


Figure 17 – Structure du génome de Kamoebavirus LCC10 schématisé (Geballa-Koukoulas et al. 2021)
Les TIR sont les répétitions terminales (« Terminal Inverted Repeats »)

Très peu de fonctions sont partagées entre tous les *Nucleocytoviricota*. Les fonctions les plus fréquemment trouvées sont le facteur de transcription VLTF3, l'ATPase d'empaquetage, l'hélicase, l'ADN polymérase, des protéines kinases et la disulfide oxidoreductase Erv1/Alr (Mönttinen et al. 2021). La protéine majeure de capside est aussi considérée comme un gène cœur de cette famille virale bien qu'elle soit absente chez les *Pandoraviridae* et très divergente chez les *Pithoviridae*. Les *Pimascovirales* partagent aussi une désoxynucléotide monophosphate kinase, les deux sous-unités alpha et beta de l'ARN (acide ribonucléique) polymérase et une endonucléase XPG (Mönttinen et al. 2021).

La grande surprise des virus géants a été la quantité inattendue de fonctions typiquement cellulaires encodées dans leurs génomes. Les gènes de traduction sont les plus parlants car ceux-ci étaient considérés un temps comme pouvant discriminer le monde cellulaire et viral. Les virus géants peuvent avoir leurs propres ARN de transfert et aminoacyl-ARNt synthétases. Ce n'est pas le cas des *Pithoviridae*. Dans cette catégorie se trouvent également de nombreux facteurs de traduction. Notamment chez les Mimivirus, un facteur de terminaison de la traduction est exprimé (Jeudy et al. 2012). *Nucleocytoviricota* possèdent également des gènes impliqués dans le métabolisme des carbohydrates et lipides. Des voies de synthèse de glycanes quasi complètes ont été retracées chez des *Mimiviridae* et *Phycodnaviridae* (Speciale et al. 2022). De plus, des gènes impliqués dans le métabolisme énergétique ont été retrouvés : la fermentation et le cycle de l'acide citrique. La rhodopsine est également encodée chez des *Mimiviridae* et *Phycodnaviridae* (Needham et al. 2019). Tout cela s'explique en général par les nombreux transferts de gènes horizontaux qu'il y a eu au cours de leur histoire évolutive. Les *Nucleocytoviricota* sont de loin le groupe avec le plus d'échanges avec les eucaryotes et ceci dans les deux sens comme expliqué par une étude qui a analysé 201 génomes eucaryotes et 109 milles génomes viraux. L'analyse des transferts de gènes s'est faite par phylogénie. La plupart des échanges ont eu lieu avec le super-groupe SAR (Stramenopile, Alveolata et Rhizaria), les Opisthocontes et les protistes Haptista (Figure 18).

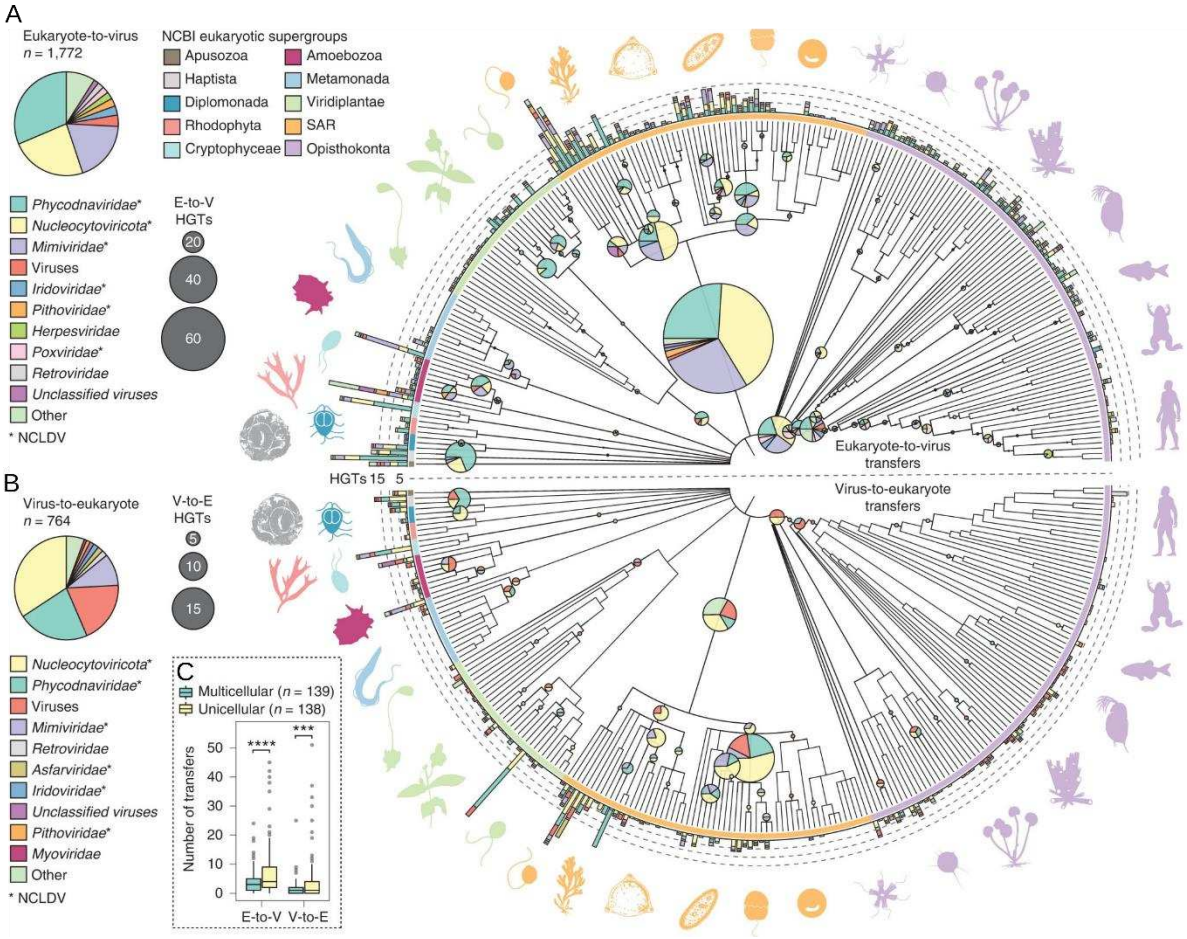


Figure 18 – Transferts de gène horizontaux entre virus et eucaryotes (Irwin et al. 2022)
La phylogénie des eucaryotes est représentée avec par-dessus des camemberts représentants le nombre d'échanges entre les virus et cette lignée d'eucaryote. Les barres à la périphérie de la phylogénie représentent le nombre de gènes issus d'un échange impliquant ce génome en particulier à un moment de son histoire. Les échanges vont dans le sens eucaryote à virus (A) ou virus à eucaryote (B). Le nombre d'événements de transfert de gène impliquant un organisme unicellulaire ou multicellulaire a été comparé par un test t de Welch (C).

Le mécanisme le plus flagrant de transfert de gène d'un virus vers un eucaryote est l'endogénéisation. Ce n'est que très récemment que l'on a appris que les *Nucleocytoviricota* pouvaient être intégrés au génome de leur hôte. Parmi 67 génomes d'algues chlorophytes étudiés précédemment, 24 contenaient des traces de cette famille virale (Moniruzzaman et al. 2020). Le plus grand fragment trouvé est celui d'un ancien *Mimiviridae* de 1.9 Mégabases intégré dans le génome de *Tetrabaena socialis* (Moniruzzaman et al. 2020) ! 133 kilobases d'un génome de Mollivirus a également été intégré en trois parties dans le génome de *Acanthamoeba castelanii* (Maumus et Blanc 2016). Ce phénomène mis en lumière par le séquençage de génomes eucaryotes ne nous dit pas s'il est le fruit d'une anomalie ou si les *Nucleocytoviricota* peuvent effectivement présenter un cycle lysogénique.

Les *Nucleocytoviricota* n'échangent pas qu'avec les eucaryotes. Bien que la nature de l'échange ne soit pas systématiquement confirmée par phylogénie, un très grand nombre de gènes de virus géants ont une similarité avec des gènes bactériens. Il y a peu d'échanges en

revanche avec les archées (Mönttinen et al. 2021). L'explication donnée est souvent que les virus d'amibe partagent leur milieu avec d'autres organismes dont des bactéries endoparasitaires. Ce milieu où tout se rencontre est donc propice à l'échange de gènes. Il convient tout de même de se méfier des analyses basées uniquement sur BLAST puisque le peuplement des bases de données étant dominé par les génomes bactériens, il y a de plus grandes chances de retrouver des homologues potentiels au sein de ce groupe qu'au sein d'autres domaines.

Les gènes aux fonctions connues et dont l'histoire a été retracée ne sont que la partie émergée de l'iceberg. La grande majorité des ORF des virus géants n'ont pas de fonction connue et/ou pas d'homologue connu hors de sa famille. Lorsque l'on découvre une nouvelle famille virale, Il est fréquent que la majorité de ses gènes n'aient pas de ressemblance avec ceux des bases de données. Le reste des résultats de BLAST est réparti entre virus, eucaryotes et bactéries (Figure 19).

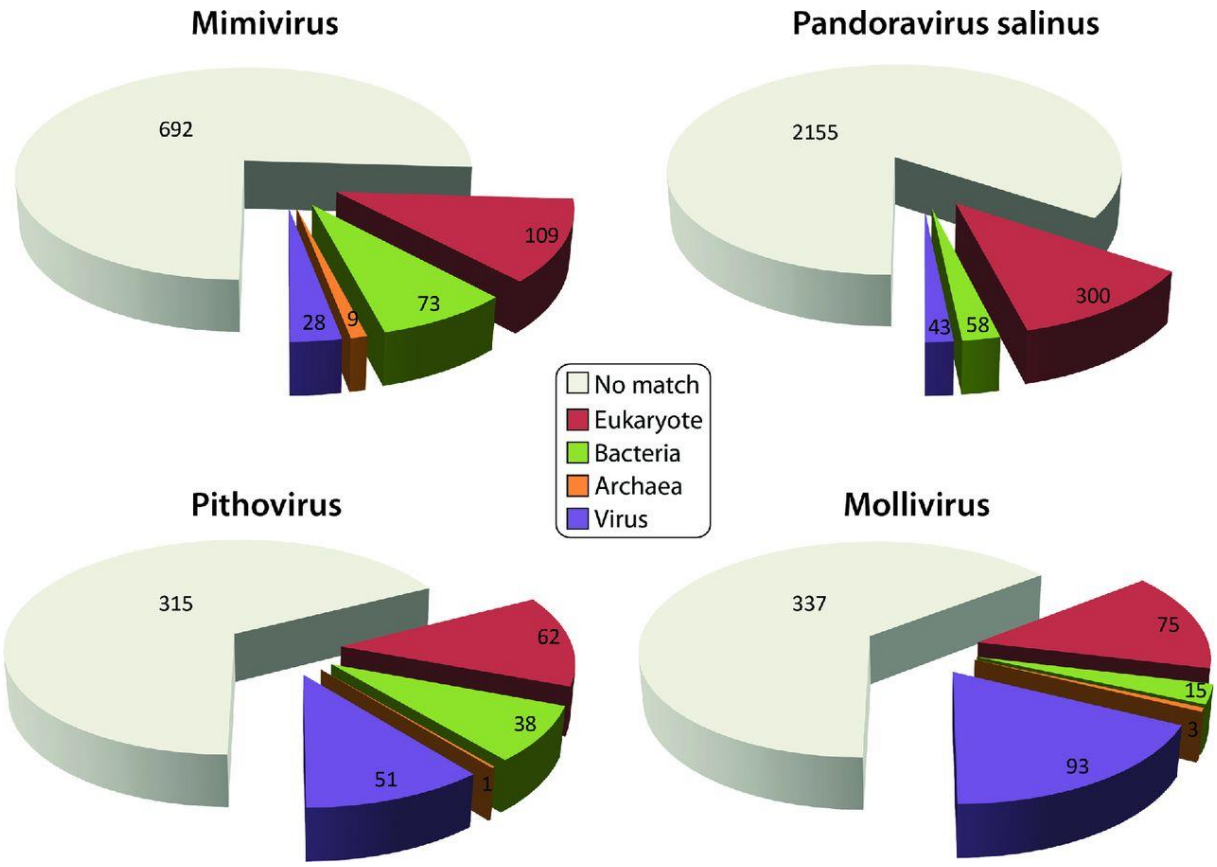


Figure 19 – Distribution des meilleurs résultats d'alignement des protéines prédites de génomes de quatre virus géants à l'heure de leur publication (Abergel, Legendre, et Claverie 2015)
Ces génomes, premiers représentants de leurs familles respectives, ont été alignés par BLASTP contre la base de données non redondante avec un seuil de e-valeur de 1e-5. Les publications originales sont celles de Mimivirus (Raoult et al. 2004), Pandoravirus (Philippe et al. 2013), Pithovirus (Legendre et al. 2014) et Mollivirus (Legendre et al. 2015).

Apports des données méta-omiques

La métagénomique a tout d'abord apporté ce constat : les *Mimiviridae* et les *Phycodnaviridae* sont ubiquitaires et très abondants dans l'océan. Peu après la découverte de Mimivirus, il est révélé qu'une partie de ses plus proches homologues résident dans la mer des Sargasses (Ghedin et Claverie 2005). En 2008, 0.7% du jeu de données virales de l'expédition de Sorcerer II (« Global Ocean Sampling ») pointait vers Mimivirus et 5% vers d'autres virus d'eucaryote (S. J. Williamson et al. 2008). Un peu plus tard, les données récoltées par Tara Océans estiment à 10^2 à 10^3 génomes par ml la quantité de *Nucleocytoviricota* dans la zone minimum d'oxygène l'océan et à 10^4 à 10^5 dans la zone photique. Cette étude comprend 17 échantillons provenant de 5 mers et océans différents (Hingamp et al. 2013).

Au-delà d'être ubiquitaires, les *Nucleocytoviricota* sont aussi très divers. Parmi, d'autres données du projet Tara océan (jeu de données « pôle à pôle »), les échantillons de l'océan Arctique se sont montrés les plus divers en *Nucleocytoviricota*. La diversité de ces virus est d'ailleurs plus efficacement séquencée à partir la fraction filtrée de l'ordre du picomètre : 0.22 à 3 µm (Endo et al. 2020). Malheureusement, les virus géants ne peuvent pas être étudiés à travers les immenses jeux de données du virome marin total du projet Tara Océan car ceux-ci proviennent d'échantillons de la fraction de taille inférieure à 0.22 µm (Brum et al. 2015). En 2020, deux études basées sur deux jeux de données métagénomiques différents, provenant de fractions microbiennes de milieux variés, publient des milliers de séquences de *Nucleocytoviricota*. L'une des deux multipliant par 11 la diversité connue (Schulz et al. 2020). Dans les deux cas la conclusion est la même : les échantillons aquatiques recèlent d'une grande richesse[25] en *Mimiviridae* et *Phycodnaviridae*. Les *Pimascovirales* sont également présents mais moins divers (Schulz et al. 2020; Moniruzzaman et al. 2019). En augmentant le nombre de séquences de *Nucleocytoviricota* identifiées, les auteurs trouvent aussi de nouvelles fonctions virales comme celle de la glycolyse (Moniruzzaman et al. 2019). Cependant, une annotation des gènes très permissive ou des « bins »[26] contaminés par des génomes cellulaires peuvent fausser ces résultats. Dès les premières études sur les *Mimiviridae* issus de la métagénomique, des auteurs remarquent que la diversité de ces virus est plus grande que celle des bactéries en se basant sur le nombre d'Unités Taxonomiques Opérationnelles (OTU) recréés à partir des séquences des deux sous-unités de l'ARN polymérase. Ceci est d'autant plus impressionnant que les contraintes de sélection négative sur ces deux gènes semblent plus fortes chez les *Mimiviridae* que chez les bactéries (Mihara et al. 2018). Enfin, une étude d'une baie japonaise surveillée régulièrement car sujette à des efflorescences d'algues toxiques, nous apprend sans surprise que la diversité de *Mimiviridae* dépend de la saison, tout comme la diversité d'organismes cellulaires. Il y a une cyclicité dans les communautés mais celle-ci évolue et ne revient pas strictement à l'identique par rapport à ce qu'elle était à la même époque, l'année d'avant. Cela a été constaté en évaluant des dissimilarités de communautés par l'indice de Sørensen-Dice d'échantillons deux à deux en

[25] Signifie diversité en écologie, nombre d'espèces

[26] Un bin est un groupe de séquences métagénomiques regroupées selon leur contenu en k-mers et/ou leur couverture pour former une portion ou un génome complet provenant potentiellement d'un seul organisme.

comparaison avec le nombre de jours qui séparent ces deux échantillons. Un modèle est proposé où la diversité virale change le plus d'une année à l'autre suivie par la communauté d'eucaryotes et enfin, de procaryotes, communauté plus conservatrice (Prodinger et al. 2022).

Pour assurer cette abondance et diversité, ces virus sont nécessairement actifs. C'est ce que confirment des études de métatranscriptomique. Encore une fois, les *Phycodnaviridae* et les *Mimiviridae* se sont avérés les plus actifs parmi les *Nucleocytoviricota* du courant californien d'après un suivi de 60 heures en échantillonnant toutes les 4 heures à 23 m de profondeur. Six virus ont présenté une activité cyclique et plus forte la nuit (Ha, Moniruzzaman, et Aylward 2021). Sans conclure quantitativement, une étude trouve par métatranscriptomique et single-cell que *Aureococcus anophagefferens Virus* est actif au sein de son hôte durant et pendant le déclin d'une efflorescence (Moniruzzaman et al. 2017). Une étude du lac Taihu trouve au contraire, que les *Nucleocytoviricota* et des virus à ARN infectant des protistes sont plus actifs durant une efflorescence de cyanobactéries du genre *Mycrocystis*. A l'inverse du modèle « tuer le vainqueur » (« Kill the winner »), des infections virales pourraient permettre au compétiteur de l'hôte de se développer. Encore une fois, les *Mimiviridae* et les *Phycodnaviridae* sont les plus actifs mais il y a de la transcription de *Pimascovirales* et d'*Asfarviridae* aussi (Pound et al. 2020).

Nous avons vu que le rôle des *Nucleocytoviricota* dans les milieux aquatiques commence à être connu bien que de nombreuses zones d'ombre demeurent. A l'inverse, leur rôle en milieu terrestre à très peu été exploré. Nous pouvons citer une étude qui a séquencé du sol de forêt et assemblé plusieurs virus géants très divergents ayant jusqu'à 80% de gènes sans homologue dans les bases de données mondiales. Les auteurs ont dévoilé deux *Pimascovirales* proches des *Pithoviridae*, un *Phycodnaviridae* et plusieurs *Mimiviridae*. Parmi ceux-ci, Hyperionvirus atteint la taille encore inégalée pour ce groupe de 2.4 Megabases, malheureusement en plusieurs morceaux (Schulz et al. 2018). Des *Pithoviridae* ont surtout été séquencés à partir de sédiments marins à plus de 3000 mètres de profondeur. Des grands génomes proches des *Pithoviridae* et Orpheovirus mais aussi des *Marseilleviridae* sont identifiés (Figure 20) avec notamment des ARN de transfert encore jamais retrouvés chez ces familles virales (Bäckström et al. 2019).

De nombreuses études de métagénomique tentent de dessiner le spectre d'hôte eucaryotes des virus géants en se basant sur leur co-occurrence. Cependant les données sont souvent complexes à analyser et les résultats difficiles à appréhender avec souvent de nombreux liens significatifs entre un virus et des hôtes potentiels complètement différents. La complexité des relations biotiques peut même mener à des fausses identifications de vecteurs d'une espèce virale (Peterson et al. 2020). Le pouvoir prédictif des analyses de co-occurrence n'étant pas connu, celles-ci ne seront pas décrites ici. A l'inverse, les relations bactéries-bactériophage peuvent être estimées de manière plus fiable à travers l'analyse des séquences CRISPR[27] (« Clustered Regularly Interspaced Short Palindromic Repeats »). Une méthode applicable chez les virus d'eucaryotes est l'analyse des transferts de gènes horizontaux. Cette analyse a

[27] Les « Clustered Regularly Interspaced Short Palindromic Repeats » traduit en « Courtes répétitions palindromiques groupées et régulièrement espacées » est un système permettant aux bactéries de garder dans leur ADN des séquences d'infections passées pour pouvoir s'en défendres

d'ailleurs confirmé les *Acanthamoebidae* comme hôte des *Pithoviridae* isolés. Etonnamment, sur la base de génomes isolés et métagénomiques, les auteurs prédisent de nombreux hôtes possibles pour les *Mimiviridae* dont les Anthoathecata (un ordre d'hydrozoaire), les champignons, les arthropodes, les animaux et différents protistes (Schulz et al. 2020). Ce type d'analyse n'est cependant pas sans bruit.

Figure 20 – Phylogénie de l'ADN polymérase de *Nucleocytoviricota* métagénomiques (Bäckström et al. 2019)
Les séquences en oranges appartenant à Loki's Castle métagénomes sont celles référées plus haut provenant de sédiments marins profonds. Les séquences marquées par une étoile appartiennent à un métagénome au sein d'un bin.

Toutes ces questions ont de l'importance non seulement pour dévoiler la diversité de la « matière noire » virale mais aussi pour comprendre le fonctionnement de l'écosystème : les interactions des organismes en fonction des conditions environnementales et leurs liens avec les cycles biogéochimiques. Les études métagénomiques des virus géants nous apprennent également que nous pouvons pendant longtemps passer à côté d'un groupe biologique abondant si nous ne savons pas qu'il existe. Ce n'est ici pas la technologie qui a limité notre compréhension du monde viral mais l'absence de représentants de ces familles virales dans les bases de données permettant de classer les séquences métagénomiques.

VIII. Le microbiome du pergélisol et ses enjeux

Le pergélisol dans un monde anthropisé

Le pergélisol, comme son nom l'indique, désigne un sol gelé en permanence. La définition pratique des géologues est de parler de pergélisol à partir du moment où un sol est gelé pendant au moins deux ans. En Arctique ce sol peut être gelé depuis 1 à 3 millions d'années. Il existerait même du pergélisol plus ancien en Antarctique (Jansson et Taş 2014). Il n'y a pas d'étude qui retrace de manière systématique l'âge du pergélisol antarctique mais nous savons qu'il peut atteindre une profondeur de 1000 m (Decker et Bucher 1977) et il est estimé que le pergélisol ait pu se maintenir depuis le milieu du miocène (Gilichinsky et al. 2007); période allant de 23 à 5 millions d'années avant notre ère.

Le pergélisol couvre 15% de la partie émergée de l'hémisphère nord (Obu 2021) mais il en existe également également sous les mers et océans. Le pergélisol est présent principalement dans les régions polaires et dans les régions de très haute montagne. La couche de surface n'est pas du pergélisol à proprement parler car elle dégèle l'été. Cette couche dite « active » peut éventuellement concentrer une forte activité microbienne et permettre la croissance de végétaux. La matière organique est ensuite capturée par le pergélisol. Si le pergélisol progresse avec la sédimentation on dit qu'il est syngénétique. A l'inverse, si le pergélisol se forme après le sol qu'il occupe c'est-à-dire qu'il est plus récent que le sol et que le pergélisol est plus instable, il est dit épigénétique. C'est le cas typiquement du pergélisol du type alass. Ces structures démontrant une instabilité du pergélisol se forment à l'endroit d'anciens lacs. Ces lacs dits thermokarstiques, se forment suite à la fonte de glace. Cela produit un affaissement de terrain puis le drainage du lac permet le regel des sédiments (Figure 21). On trouve beaucoup d'alass en Sibérie car le pergélisol de cette région est souvent riche en glace. C'est typique du pergélisol de type Yedoma, celui-ci, riche également en matière organique, s'est formé durant le Pléistocène (c'est la période du quaternaire allant de 2.58 millions d'années à 12 milles ans avant le présent).

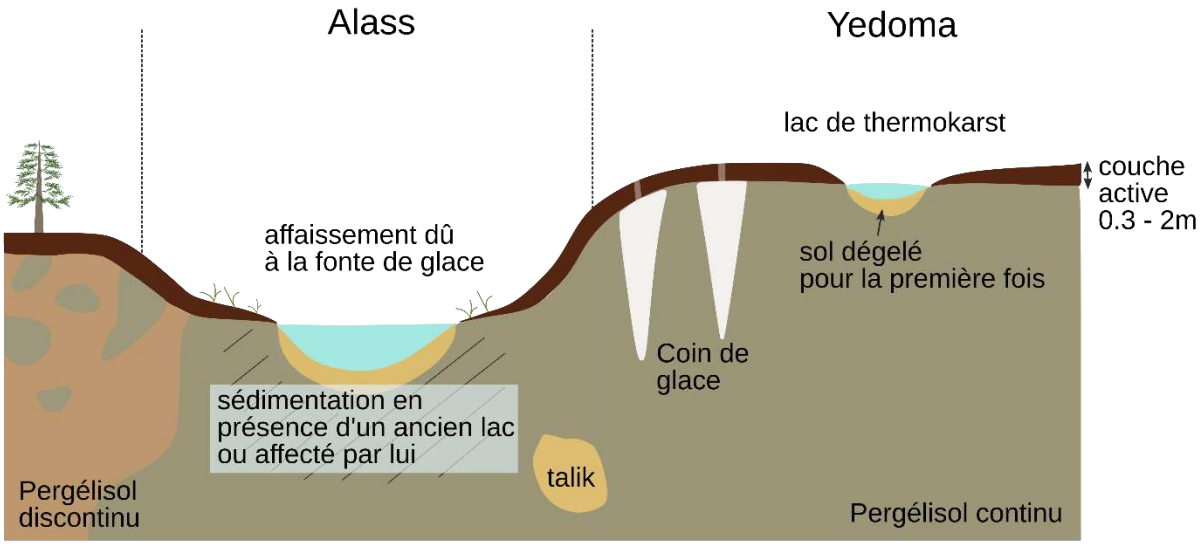


Figure 21 – Différents structures et types de pergélisol rencontrés en Sibérie

Ces dernières décennies, le pergélisol s'est en moyenne réchauffé rapidement. C'est vrai en particulier pour le pergélisol continu d'Arctique et d'Antarctique qui ont subi une augmentation de 0.39 et 0.37°C respectivement de 2007 à 2016. Les plus grandes variations observées sont en Sibérie d'après un suivi de 123 points GPS répartis globalement dans du pergélisol (Figure 22, (Biskaborn et al. 2019)). La dégradation du pergélisol de surface se traduit plutôt par l'approfondissement de la couche active. Entre 1956 et 1990 il est estimé que la couche active a progressé de 20 cm en Russie (Frauenfeld et al. 2004). La dégradation du pergélisol profond se manifeste notamment par la formation de talik (sédiment dégelé au sein de pergélisol continu) et de lacs de thermokarst ((Grosse, Jones, et Arp 2013),exemple: Figure 21). Cette dégradation fragilise le pergélisol qui est lié par la glace. Cela a pour conséquence d'accroitre de manière importante les risques d'érosion des côtes, d'éboulement de terrain, de chutes de pierres en montagne… Elle a aussi pour conséquence d'activer une boucle de rétroaction positive sur le réchauffement climatique. La matière organique stockée dans le pergélisol devient disponible quand ce dernier se dégrade. Sa décomposition par l'activité microbienne provoque la libération de gaz à effet de serre ; le CO_2 et le méthane. Celle-ci a surtout lieu dans les talik, dans la couche active qui avance chaque année et au fond des lacs de thermokarst. Ces lacs peuvent provoquer un changement d'une respiration aérobie vers une respiration anaérobie qui a lieu chez la communauté microbienne vivant sous le lac ce qui favorise d'autant plus la libération de méthane (Grosse, Jones, et Arp 2013). En Yakoutie centrale, deux types de sédiments sous un lac thermokarstique ont été incubés pendant un an : un provenant d'un alass et un de type Yedoma. Ce dernier type a produit trois fois plus de méthane et 1.5 fois plus de CO2 que les échantillons d'alass. Cela s'explique probablement par la plus grande quantité de matière organique disponible dans le sol dégelé pour la première fois depuis la formation de pergélisol (Jongejans et al. 2021).

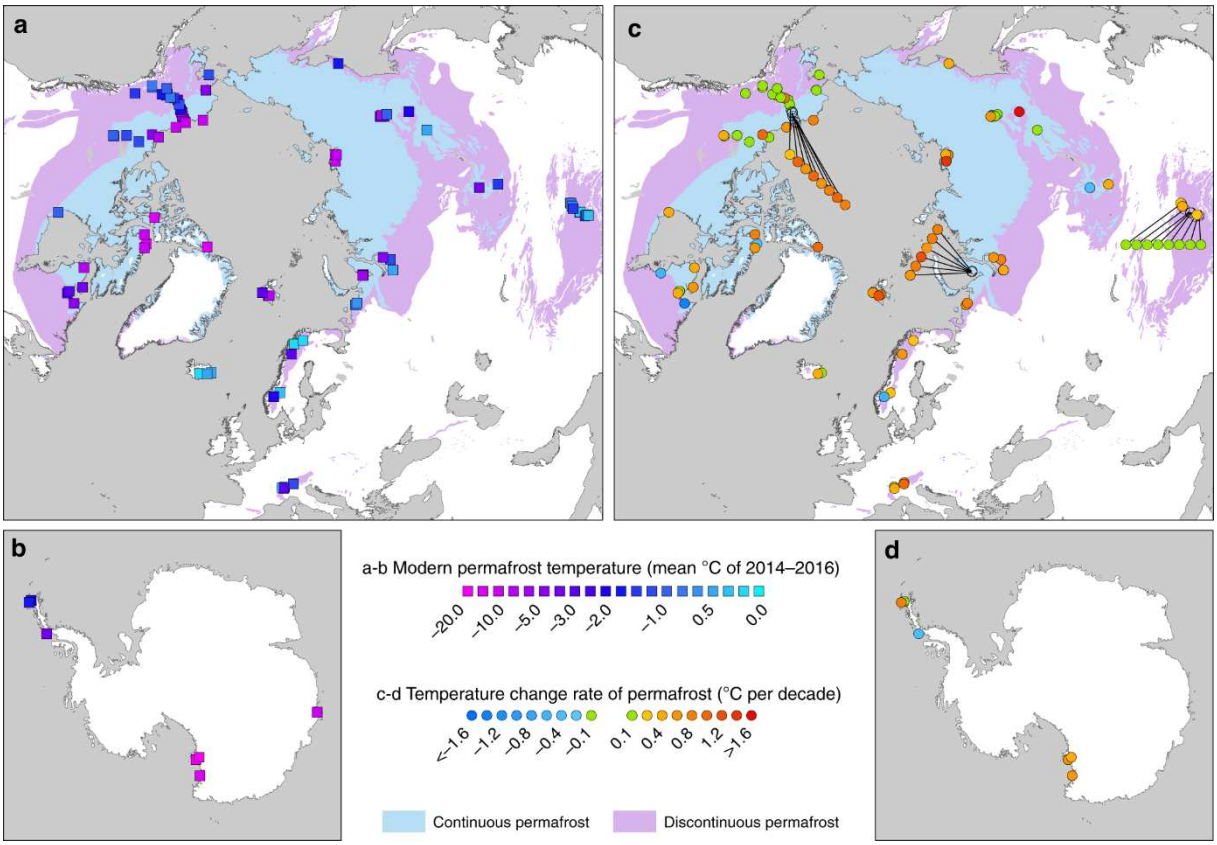


Figure 22 – Changements de température du pergélisol mesurés entre 2007 et 2016 à la profondeur où l'amplitude annuelle de température est proche de zéro (Biskaborn et al. 2019)
Les valeurs données correspondent à la différence de température à 2m de profondeur. Les températures moyennes entre 2014 et 2016 mesurées dans l'hémisphère Nord (a) et l'Antarctique (b). Les variations de température entre 2007 et 2016 ont été calculées pour les deux hémisphères également (c et d).

Le microbiome du pergélisol

Le microbiome de pergélisol est différent de celui de la couche active à commencer par un nombre de cellules largement plus important dans la couche active (20 fois plus selon une étude) (Jansson et Taş 2014). Bien que les archées aient un fort impact sur la production de méthane, ces environnements restent dominés par les bactéries. Une étude du pergélisol du pléistocène récent en Alaska compte de l'ordre de 10^7 cellules dont 14 à 26% sont en vie selon l'âge de l'échantillon (Burkert et al. 2019). Le postulat qu'une membrane plasmique préservée indique une cellule vivante. Les basses températures peuvent-être délétères car elles peuvent diminuer la flexibilité des membranes plasmiques qui se retrouvent gélifiées et ainsi plus facilement ouvertes. Cela peut également dénaturer les protéines et les rendre moins flexibles et stabiliser les acides nucléiques ce qui inhibe la réplication, la transcription et la traduction (Jansson et Taş 2014). Une des stratégies pour se protéger du froid est donc l'enkystement ou la sporogénèse. Les Firmicutes, les Actinobactéries et les Chlamydiae ont tendance à être présents sous forme d'endospore dans le pergélisol d'Alaska (Burkert et al. 2019). Plus le pergélisol est ancien, moins on y trouvera de cellules en vie. Un consensus scientifique s'est

accordé sur la possibilité de maintenir une activité métabolique basale dans le pergélisol même à très basses températures. Pour ce qui est de savoir si la croissance cellulaire est possible, deux camps s'affrontent : l'un dépeignant le microbiome du pergélisol comme une communauté de survivants (Abramov, Vishnivetskaya, et Rivkina 2021) et l'autre semant un doute en se basant sur les expériences en laboratoire de croissance d'organisme psychrophiles anaérobies (Jansson et Taş 2014). Ce microbiote est-il une image très dégradée mais fidèle du microbiote ancien de la couche active avant que le sol ne devienne du pergélisol ? Est-ce qu'au contraire certains organismes peuvent croître dans les conditions abiotiques de ce milieu et donc introduire de la nouveauté génétique au microbiote d'origine ? Si la croissance de bactéries jusqu'à -15°C est possible (Mykytczuk et al. 2013), celle-ci n'est probablement pas commune dans le pergélisol compte-tenu de la disponibilité des ressources et du déclin de diversité et de cellules vivantes avec l'âge du pergélisol (Burkert et al. 2019; Wu et al. 2021). De plus, le microbiome n'est pas dominé par des psychrophiles qui ne comprend qu'une petite fraction selon les analyses métagénomiques et les cultures cellulaires (Abramov, Vishnivetskaya, et Rivkina 2021). Des adaptations au froid sont néanmoins nécessaires pour survivre. Il a été proposé que l'état de dormance (dans une endospore ou un cyste) ne soit plus la stratégie la plus efficace après plusieurs centaines de milliers d'années piégés dans ce sol gelé à cause des dommages accumulés à l'ADN. A l'inverse, les survivants à de si longues périodes seraient les organismes capables de maintenir une réparation de l'ADN efficace (Johnson et al. 2007). Une stratégie utilisée par les bactéries du pergélisol peu profond pour leur métabolisme énergétique est potentiellement la fermentation. Des gènes de fermentation du pyruvate, du lactate ou du propanoate ont été retrouvés dans 14 métagénomes sur 91, la plupart appartenant aux Bacteroidetes (Wu et al. 2021). Une autre étude du pergélisol peu profond de Svalbard trouve, au contraire, que la majorité de l'abondance bactérienne provient d'organismes aérobies (Xue et al. 2020). Nous pouvons donc considérer que, en grande partie, le pergélisol permet d'étudier des organismes anciens servant d'objet d'étude pour comprendre l'écosystème du pergélisol, tenter d'élucider les mécanismes de survie à cet environnement et étudier une biodiversité passée.

Un grand nombre d'organismes a été isolé à partir de la couche active et du pergélisol. Historiquement, ces organismes ont été étudiés en tant que modèles d'adaptation à ce que pourrait être la vie extra-terrestre et notamment sur Mars (Vishnivetskaya et al. 2003; Aszalós et al. 2020). Plus récemment y ont été découvert des bactéries productrices d'antibiotiques ce qui pourrait avoir un grand intérêt compte tenu de la crise de résistance aux antibiotiques que nous vivons (Efimenko et al. 2018). Des amibes ont également été régulièrement isolées entre autres par une équipe russe du laboratoire de cryologie du sol à Pushchino. Ces amibes Discosea, Evosea ou Heterolobosea peuvent avoir jusqu'à 600 mille à un million d'années (Malavin et al. 2020). Les seuls protistes qui survivent dans le pergélisol sont ceux capables de former des spores ou des cystes. Il est également possible de cultiver des organismes photosynthétiques à partir du pergélisol profond comme en témoigne ces dizaines de souches d'algues vertes ou de cyanobactéries trouvées à jusqu'à 56 mètres de profondeur. Ces organismes pourraient justement avoir un intérêt pour l'exobiologie (Vishnivetskaya et al. 2003). Le premier organisme pluricellulaire ancien à avoir été cultivé est une plante, *Silene Stenophylla*, regénérée à partir de tissus de fruit vieux de 30 000 ans (Yashina et al. 2012).

C'est notamment cette surprise scientifique qui a inspiré le laboratoire Information Génomique et Structurale à prendre contact avec cette équipe de Pushchino puis avec l'Alfred Wegner Institute en Allemagne pour tenter de découvrir des virus ayant également survécu au pergélisol. C'est ainsi que furent isolés les premiers virus anciens que nous avons évoqué plus haut : P. sibericum et M. sibericum. On découvre ainsi que des virus peuvent survivre de très longues périodes et que la perturbation de pergélisol peut représenter un risque sanitaire en plus de tous les risques déjà encourus. Ces deux virus ne sont pas pathogènes pour l'homme et, pour des raisons évidentes, personne n'a tenté de réactiver des virus anciens sur des cultures cellulaires humaines ou de mammifères. C'est donc par analogie que la découverte de P. sibericum a fait couler beaucoup d'encre sur la possibilité qu'une future épidémie proviennent d'un virus ancien inconnu ou de la réactivation de la variole. Très peu d'études se sont penchées sur le virome du pergélisol. Une étude de la couche active de pergélisol au nord de la Suède avec près de 200 échantillons (dont des triplicats) n'a retrouvé parmi les séquences virales classifiées que des phages. Ces phages se répartissent équitablement entre les *Myoviridae*, *Siphoviridae* et *Podoviridae*. Les auteurs identifient également des *Caudovirales*. Etonnamment, dans les échantillons de marais, milieu étudié le plus souvent dégelé, près de 15% de séquences proches de Tectivirus ont été identifiées (Emerson et al. 2018). Les virus géants n'ont pas pu être pris en compte puisque les échantillons ont été mis en suspension et filtrés à 0.4 µm ou moins pour pouvoir analyser le virome non contaminé par des séquences cellulaires.

Le risque sanitaire lié à la fonte du pergélisol n'est pas hypothétique car il y a en effet eu une épidémie d'anthrax en 2016 en Sibérie qui peut être reliée à ce déséquilibre. Plus de 2000 rênes et un enfant sont morts de cette maladie dont la dernière incidence connue datait de 1941, 70 ans plus tôt (Carlson et al. 2018). La cause principale identifiée est la température de l'été 2016, 20 à 100% plus importante que celle des 30 étés passés, ayant fait ressurgir des spores de *Bacillus anthracis* dans un contexte où le gouvernement russe avait arrêté de faire vacciner les rênes depuis 2007 (Liskova et al. 2021). Malgré le fait que ces régions soient relativement peu en contact avec les humains, il est possible d'y trouver des pathogènes qui peuvent passer à l'homme ou aux espèces d'élevage à travers les espèces sauvages. Il en est de même pour l'agriculture. La fonte liée aux fortes températures n'est pas la seule cause de réexposition à d'anciens pathogènes. Les constructions, industries dont surtout l'exploitation d'hydrocarbures, et la construction de pipelines souvent enterrés sont aussi source d'altération du pergélisol (Vasiliev, Dzhaljabov, et Leonovich 2021) et donc potentiellement de réactivation d'anciens organismes. Au même titre que la déforestation a été une cause de la réémergence du flavivirus de l'encéphalite japonaise (Mackenzie et Williams 2009) ou de la maladie de Lyme (LoGiudice et al. 2003), empiéter sur les paysages naturels peut mettre en contact les humains avec des pathogènes potentiels par des mécanismes divers. Contrairement à ce qui peut être dit, la métagénomique n'est pas un outil approprié pour prévenir les risques en « détectant de nouveaux pathogènes »[28] puisque les mécanismes évolutifs d'émergence de nouveaux pathogènes ne sont pas actuellement prévisibles. De plus, pour un pathogène très virulent, une petite quantité non détectable par métagénomique (Martínez-Puchol et al. 2021) suffit à provoquer une épidémie. C'est un outil puissant en

[28] D'après l'outil IDseq https://czid.org/

revanche pour détecter la présence de pathogènes connus abondants et pour comprendre le fonctionnement d'un écosystème.

IX. Objectifs de la thèse

Cette thèse s'inscrit dans la continuité du projet France-Génomique « Permagenomics » et l'isolement de virus géants dans le pergélisol par le laboratoire Information Génomique et Structurale (IGS) en prenant l'angle d'étude des *Pithoviridae*. Le but de mes travaux est donc de caractériser cette famille virale encore très peu connue. Deux grands axes de recherche se dessinent.

Les *Pithoviridae* sont-ils diversifiés et répandus dans l'environnement ?

Nous allons commencer par répondre à cette question en recherchant la diversité des *Pithoviridae* en comparaison avec le microbiome total et les autres *Nucleocytoviricota* au sein d'un milieu où nous savons qu'il est possible d'isoler ces virus : le pergélisol sibérien. Ainsi la première partie traitera des « *Pithoviridae* au sein du microbiome du pergélisol ». Nous allons étudier des échantillons des deux types de pergélisol : Alass et Yedoma provenant de Yakoutie centrale datant du Pléistocène récent. Ces échantillons ont été obtenus par une collaboration avec le Alfred Wegner Institute, spécialistes de l'étude du pergélisol. Nous allons également étudier des échantillons de la couche active d'une région où le pergélisol est discontinu : le Kamchatka. Ceux-ci proviennent d'une collaboration avec Alexander Morawitz, explorateur amateur. L'outil idéal pour répondre à cette question est la métagénomique car elle permet d'accéder rapidement à la diversité totale d'un échantillon sans a priori et de pouvoir évaluer l'abondance de certains groupes comparativement aux autres ainsi que d'étudier les fonctions génomiques encodées par ces différents groupes.

Quelle est l'histoire évolutive de cette famille virale ? Comment a-t-elle marqué les séquences génomiques et les grands mécanismes fonctionnels des *Pithoviridae* ?

Le deuxième axe de recherche est celui de la génomique comparative avec une ouverture sur l'étude du cycle infectieux d'un nouveau *Pithoviridae*. Dans un deuxième chapitre intitulé « Génomique comparative des *Pithoviridae* » nous nous proposons d'étudier les génomes de douze *Pithoviridae* en comparaison également avec leurs plus proches familles. Certains de ces *Pithoviridae* ont été isolés, assemblés et annotés par un travail commun de membres de l'IGS et moi-même puis publiés dans le cadre de cette thèse. Nous allons tenter d'élucider les mécanismes qui ont forgé leur évolution, aujourd'hui visibles dans leurs génomes. Quelle part de leur génome est partagée ou spécifique ? Les séquences sont-elles bien conservées ? Il y a-t-il des motifs dans les séquences ADN permettant de dévoiler des mécanismes évolutifs ? Toutes ces questions devraient nous permettre de plonger dans l'histoire des *Pithoviridae*. Ainsi, la meilleure compréhension des génomes des *Pithoviridae* devrait nous aider à interpréter l'étude du dernier chapitre : « le cycle infectieux de Cedratvirus Kamchatka », un virus isolé à partir de la

couche active d'échantillons de Kamchatka. Ce cycle est étudié à travers la transcriptomique et l'imagerie.

Chapitre 1. Les *Pithoviridae* au sein du microbiome du pergélisol

I. Contexte

Plusieurs virus géants ayant été isolés à partir d'échantillons de cryosol, dont notamment les deux virus anciens : Pithovirus sibericum et Mollivirus sibericum, le laboratoire s'est intéressé à la métagénomique générale du pergélisol dans une optique exploratoire. Ce travail se focalise sur le microbiome du pergélisol russe dans deux régions : le Kamchatka et la Sibérie (Figure 23). Les échantillons sont issus d'un travail de collaboration avec Guido Grosse et Yens Strauss de l'Alfred Wegner Institute en Allemagne qui ont fait parvenir des échantillons de pergélisol profond de Yakoutie centrale ainsi que leur expertise dans le domaine jusqu'à Marseille. Ce projet a également été permis par Alexander Morawitz apportant des échantillons de surface du Kamchatka.

La première partie de l'analyse sur le microbiome total du pergélisol inclus également l'analyse d'un échantillon profond du nord-est sibérien, celui qui a donné la première impulsion au projet puisqu'il date de 2012, moment de la toute première collaboration avec l'institut de problèmes physicochimiques et biologiques des sciences du sol de l'Académie des Sciences russes, pionniers dans la réactivation d'organismes anciens. Cet échantillon a été écarté de l'analyse des virus géants car son séquençage étant antérieur à celui des autres échantillons, les données métagénomiques produites sont moins importantes et les lectures plus petites.

Le Kamtchatka est une région volcanique où le pergélisol est discontinu. Le climat est tempéré froid. La Yakoutie centrale et le nord-est sibérien connaissent un climat arctique allant jusqu'à 24h de nuit en hiver avec jusqu'à presque -40 °C et -35 °C. Les variations de température annuelles en Yakoutie sont extrêmes puisque l'été il peut faire 20 °C (Figure 23). Cette petite région contient de nombreux lacs thermokarstiques qui couvrent 13% de la zone. Pour ce qui est du pergélisol, 66.4% correspond à une structure géologique de type Yedoma et 20.6 % à du pergélisol de type Alass (Windirsch et al. 2020).

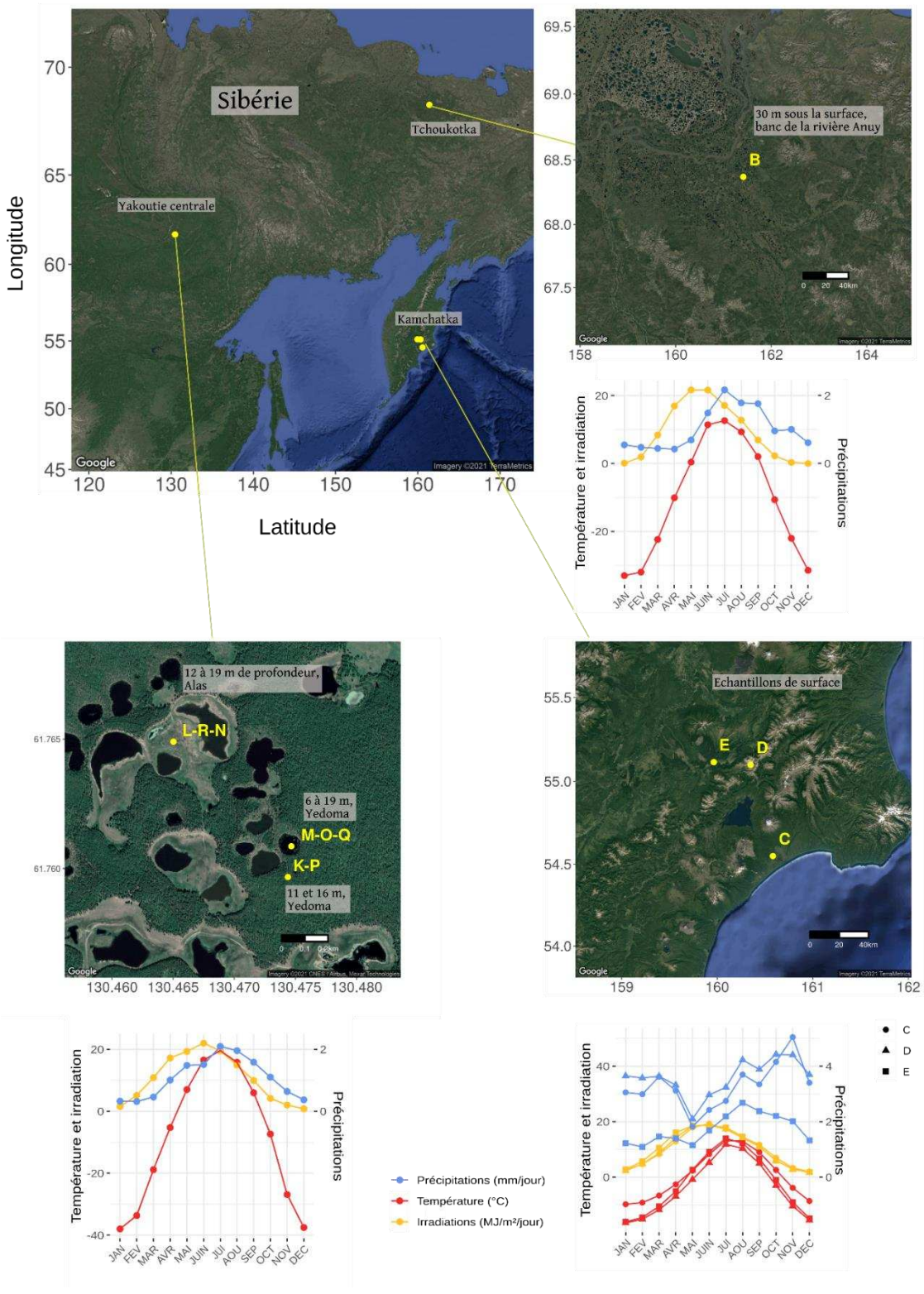


Figure 23 – Carte des échantillons étudiés en métagénomique par le projet Permagenomics et climat des régions associées

Les graphiques représentent le climat moyen entre 2011 et 2021 à la localisation des points GPS des échantillons étudiés. Les données de précipitation, de température à 2 m et d'irradiation (courtes longueurs d'ondes) ont été récupérées sur le site power.larc.nasa.gov le 4 août 2022. Les trois points GPS en Yakoutie centrale étant très proches, le climat présenté n'a pas donné de variation d'un site à l'autre.

Le plan d'échantillonnage, le sous-échantillonnage à partir des carottes, l'extraction d'ADN ainsi que le séquençage, en collaboration avec Karine Labadie du Genoscope, ont été réalisés dans le cadre de la thèse d'Eugène Christo-Foroux (Eugène Christo-Foroux 2020) démarrée en 2017 à l'IGS. Par la suite, le nettoyage des lectures, le premier assemblage des données métagénomiques, l'alignement des lectures suivi de la détection des cadres ouverts de lecture (ORFing) ont été réalisés par Sébastien Santini. A partir de tout cela j'ai pu prendre la suite de concert avec l'équipe dirigée par Jean-Michel Claverie puis Matthieu Legendre, pour réaliser l'analyse taxonomique et fonctionnelle des échantillons ainsi que les analyses complémentaires. Cela nous a permis dans un premier temps d'obtenir une vision globale de nos échantillons et de nous intéresser spécifiquement à la fraction bactérienne, la plus abondante, dans la partie «Le microbiome du pergélisol, article I» (p. 60).

Dans un second temps, nous avons cherché à découvrir de nouvelles séquences virales. Pour ce faire nous avons d'abord amélioré le processus d'assemblage pour rallonger les séquences étudiées et élaboré une méthode pour l'extraction de séquences de virus géants dans un microbiome complexe. L'enjeu de cette étape est que les virus géants possèdent de nombreuses signatures cellulaires dans leurs séquences et que, par conséquent, les méthodes de taxonomie classique échouent souvent à les classer. Dans la deuxième partie «Les virus géants du pergélisol, article II» (p.80) nous présentons donc spécifiquement les *Nucleocytoviricota* retrouvés dans la partie séquencée des échantillons avec une analyse également fonctionnelle et comparative avec d'autres milieux.

II. Le microbiome du pergélisol, article I

microLife, 2022, **3**, 1–15
DOI: 10.1093/femsml/uqac003
Advance access publication date: 7 April 2022
Research Article

Metagenomic survey of the microbiome of ancient Siberian permafrost and modern Kamchatkan cryosols

Sofia Rigou[1], Eugène Christo-Foroux[1], Sébastien Santini[1], Artemiy Goncharov[2], Jens Strauss[3], Guido Grosse[3,4], Alexander N. Fedorov[5,6], Karine Labadie[7], Chantal Abergel[1], Jean-Michel Claverie[1,*]

[1]IGS, Information Génomique & Structurale (UMR7256), Institut de Microbiologie de la Méditerranée (FR 3489), CNRS, Aix Marseille University, Marseille, 13288, France
[2]Department of Molecular Microbiology, Institute of Experimental Medicine, Saint Petersburg, Russia, Department of Epidemiology, Parasitology and Disinfectology, Northwestern State Medical Mechnikov University, Saint Petersburg, 195067, Russia
[3]Permafrost Research Section, Alfred Wegener Institute, Helmholtz Centre for Polar and Marine Research, 14473 Potsdam, Germany
[4]Institute of Geosciences, University of Potsdam, 14478 Potsdam, Germany
[5]Melnikov Permafrost Institute, 677010 Yakutsk, Russia
[6]BEST International Centre, North-Eastern Federal University, 677027 Yakutsk, Russia
[7]Genoscope, Institut François Jacob, CEA, Université Paris-Saclay, Évry, 91000, France
*__Corresponding author:__ Laboratoire Information Génomique et Structurale (IGS) UMR7256, Aix Marseille Université, CNRS, Parc Scientifique de Luminy – 163 Avenue de Luminy, 13288, Marseille cedex 09, France. Tél: 04 13 94 67 77; E-mail: Jean-Michel.Claverie@univ-amu.fr
One sentence summary: Permafrost bacteria appear to constitute an enormous reservoir of antibiotic resistance genes.
Editor: Julian Parkhill

Abstract

In the context of global warming, the melting of Arctic permafrost raises the threat of a reemergence of microorganisms some of which were shown to remain viable in ancient frozen soils for up to half a million years. In order to evaluate this risk, it is of interest to acquire a better knowledge of the composition of the microbial communities found in this understudied environment. Here, we present a metagenomic analysis of 12 soil samples from Russian Arctic and subarctic pristine areas: Chukotka, Yakutia and Kamchatka, including nine permafrost samples collected at various depths. These large datasets (9.2×10^{11} total bp) were assembled (525 313 contigs > 5 kb), their encoded protein contents predicted, and then used to perform taxonomical assignments of bacterial, archaeal and eukaryotic organisms, as well as DNA viruses. The various samples exhibited variable DNA contents and highly diverse taxonomic profiles showing no obvious relationship with their locations, depths or deposit ages. Bacteria represented the largely dominant DNA fraction (95%) in all samples, followed by archaea (3.2%), surprisingly little eukaryotes (0.5%), and viruses (0.4%). Although no common taxonomic pattern was identified, the samples shared unexpected high frequencies of β-lactamase genes, almost 0.9 copy/bacterial genome. In addition to known environmental threats, the particularly intense warming of the Arctic might thus enhance the spread of bacterial antibiotic resistances, today's major challenge in public health. β-Lactamases were also observed at high frequency in other types of soils, suggesting their general role in the regulation of bacterial populations.

Keywords: metagenomics, Kamchatka, Yakutia, Siberia, permafrost thaw, antibiotic resistance genes

Introduction

Two of the front-page societal concerns are global warming and the increasing frequency of emerging infectious diseases, eventually turning into pandemics (Morens *et al.* 2020, Zhou *et al.* 2020). It turns out that both predicaments are partly linked. Through its influence on the ecology of living organisms vectoring infectious pathogens, the warming climate contributes to spreading endemic diseases from tropical regions into temperate areas of the globe, such as Western Europe (Hertig 2019, Fuehrer *et al.* 2020, Liu *et al.* 2020). This phenomenon is further amplified by the encroachment into pristine areas (e.g. tropical forests) where expanding human activities generate new risks through imponderable contacts with (mostly uncharted) microbial environments and their associated wildlife hosts (Bradley and Altizer 2007, Keita *et al.* 2014, Plowright *et al.* 2017, Valentine *et al.* 2019).

While such dangers are mostly pointed out as coming from the South, more recent concerns have been raised that new plagues could also come from the Arctic, through the release of infectious agents until now trapped in perennial frozen soils (i.e. permafrost) up to 1.5 km deep and 2–3 million years old (Revich and Podolnaya 2011, Revich *et al.* 2012, Parkinson *et al.* 2014, Huber *et al.* 2020).

Climate warming is particularly noticeable in the Arctic where average temperatures are increasing twice as fast as in temperate regions (Cohen *et al.* 2014). One of the most visible consequences is the widespread thawing of permafrost at increasing depths (Biskaborn *et al.* 2019, Turetsky *et al.* 2019) and the rapid erosion of permafrost bluffs (Fuchs *et al.* 2020), a phenomenon most visible in Siberia where deep continuous permafrost underlays most of the North Eastern territories.

The 'ice age bug' concern has been periodically brought back to the public attention. For instance, when an exceptionally hot summer triggered local outbreaks of anthrax on Yamal Peninsula, Northwest Siberia, in 2016, a deeper than usual summer season thaw of soils above the permafrost layer (i.e. the 'active layer') exhumed infectious *B. anthracis* endospores buried in the frozen ground for 75 years (Timofeev *et al.* 2019). Historically frequent

Received: November 30, 2021. **Revised:** March 11, 2022. **Accepted:** March 25, 2022

Table 1A. Deep permafrost samples analyzed in this study.

Sample	IGS code	Latitude (°)	Longitude (°)	Locality	Depth (m)	Dating years BP	Type
P 1084T	B	68.37016	161.41555	Stanchikovskiy Yar	NA	35 000	Wall sampling, Yedoma bluff
Yuk15 Yed1/Yed.6	K	61.75967	130.47438	Yukechi Alas landscape	16	>49 000	Core, Yedoma upland
Yuk15 Yed1/Yed.7	P				11	>36 000	
Yuk15 Alas1/Ala.11	L	61.76490	130.46503		12	>28 000	Core, from a drained thermokarst lake basin (Alas)
Yuk15 Alas1/Ala.10	N				16	>45 000	
Yuk15 Alas1/Ala.9	R				19	>42 000	
Yuk15 YuL15/Y1	O	61.76086	130.47466		19	>40 000	Core, Yedoma upland, under lake (4 m water depth)
Yuk15 YuL15/Y2	M				16	>48 500	
Yuk15 YuL15/Y4	Q				6	53[a]	

[a]: This soil sample was not frozen (talik). O, M, Q sample data from: Jongejans LL, Grosse G, Ulrich M *et al.* (2019). Radiocarbon ages of talik sediments of an alas lake and a Yedoma lake in the Yukechi Alas, Siberia. PANGAEA, https://doi.org/10.1594/PANGAEA.904738.

outbreaks of anthrax killed 1.5 million reindeer in Russian North between 1897 and 1925 and human cases of the disease have occurred in thousands of settlements across the Russian North (Hueffer *et al.* 2020).

The capacity of deeper (hence much older) permafrost layers to preserve the integrity of 'live' (i.e. infectious) viral particles was demonstrated by the isolation of two previously unknown giant DNA viruses: Pithovirus (Legendre *et al.* 2014) and Mollivirus (Legendre *et al.* 2015). Those Acanthamoeba-infecting viruses were isolated (and cultivated) from a wall-sampled layer of a well-studied permafrost bluff (the Stanchikovsky Yar) (sample B, Table 1A) radiocarbon dated to 35 000 years before present (BP) (Legendre *et al.* 2014). This first demonstration that eukaryotic DNA viruses could still be infectious after staying dormant since the late Pleistocene came after many other studies (Vishnivetskaya *et al.* 2006, Hinsa-Leasure *et al.* 2010, Graham *et al.* 2012) going back to 1911 (reviewed by Gilichinsky and Wagener 1995). It was shown that ancient bacteria could be revived from even much older permafrost layers, up to half a million years (Johnson *et al.* 2007). Contamination by modern bacteria has been suspected in the case of older isolates (Willerslev *et al.* 2004). A flower plant was also revived from 30 000-year-old frozen tissues (Yashina *et al.* 2012), as well as various fungi (from Antarctica) (Kochkina *et al.* 2012) and amoebal protozoans (Shmakova *et al.* 2016, Malavin and Shmakova 2020). There is now little doubt that permafrost is home to a large diversity of ancient microorganisms that can potentially be revived upon thawing. Many of them, in particular among viruses, are unknown. If detected, the follow-up questions are: Are those microorganisms a threat for today's society? Could some of them represent an infectious hazard for humans, animals or plants? Can we identify microbes the metabolisms of which would accelerate the emission of greenhouse gases (e.g. methane and carbon dioxide) upon thawing?

Beyond the direct infectious risk represented by viable microorganisms released from thawing ancient permafrost layers, the persistence of a multitude of genomic DNA fragments from all the organisms that were present in the soil at various times is also to be considered. Although it originated from dead (or even extinct) organisms, the permafrost DNA content constitutes a historical library of genetic resources, from which useful elements (genes) can be reintroduced into modern organisms through bacterial transformation and/or lateral gene transfer in protozoa or viruses. Of particular interest are DNA fragments coding for bacterial virulence factors or antibiotic resistance genes (ARGs) the origin of which predates by far human history (Finley *et al.* 2013, Perron *et al.* 2015, Potron *et al.* 2015, Wright 2019).

To address these questions, a first step is to evaluate the DNA in frozen soils. The technique of choice (which avoids the risk of reactivating unknown dangerous pathogens) for this task is metagenomics. In contrast with targeted approaches, such as metabarcoding (i.e. PCR-amplified subset of environmental DNA), it offers the possibility of unexpected discoveries in all domains of life and provides an estimate of the relative abundances of the various organisms, although at the cost of a lower sensitivity. Finally, the metagenomic analysis of DNA obtained from up to 60 g of frozen soil samples (Table S1, Supporting Information) is much less sensitive to eventual contaminations by minuscule amounts of ambient DNA or contemporary microorganisms and more suitable for site-to-site comparisons than culture-based approaches.

Here, we report the analysis of nine subsurface Siberian permafrost samples, and of three modern surface soils from pristine area in Kamchatka. All the permafrost samples were taken from carbon-rich, ice-rich frozen soils called Yedoma deposits (Schirrmeister *et al.* 2013, Strauss *et al.* 2017) found in vast regions of northeast Siberia (Lena and Kolyma river basins in Yakutia) (Fig. S1, Supporting Information) and known for well-preserved Late Pleistocene megafauna remains (mammoths, wholly rhinoceros). Yedoma permafrost is of special interest as it is prone to rapid thaw processes such as thermokarst and thermo-erosion (Nitzbon *et al.* 2020) releasing not only carbon as greenhouse gases (Schneider von Deimling *et al.* 2015, Turetsky *et al.* 2020) but also its formerly freeze-locked microbial content, in which we detected a high abundance of β-lactamases genes carried in the genomes (or plasmids) of a large phyletic diversity of bacteria.

Methods

Sample collection and preparation

We sampled Yedoma and thermokarst sediments from Central Yakutia from the Yukechi Alas landscape ~50 km southeast of Yakutsk (Table 1A). The Yukechi landscape is characterized by Yedoma uplands and drained as well as extant lake basins, so-called Alases, indicating active permafrost degradation processes (Fedorov and Konstantinov 2003). Field work took place in March

2015 during a joint Russian-German drilling expedition. We used a Russian URB2-47 drilling rig with a steel core barrel without a core catching system. Core barrels had different diameters, starting with 15.7 cm for the upper core segments and narrowing down to 8 cm for lower core segments, a setup preventing to get stuck in the deep borehole. To avoid contaminations, coring was conducted as dry drilling and no drilling fluid was used. The retrieved core segments were pushed out from the core barrel with air pressure only. The core segments were visually described and immediately stored in plastic bags or core foil and sealed tightly with duct tape. The wrapped and sealed cores were put into opaque and thermally insulated hard case expanded polypropylene boxes (thermoboxes) and stored outside, where they were kept frozen (or refroze slowly if the sediment material was originally unfrozen) at outside temperature around −10°C to −20°C. During the entire transportation, the core material was kept in the dark and frozen in the thermoboxes. Temperatures below freezing during the entire storage and transport chain were recorded with a temperature data logger in one of these boxes from the drilling site to the laboratories in Potsdam, Germany.

In total, three long (all ∼20m below surface/lake ice) permafrost sediment and unfrozen thaw bulb (talik) cores were retrieved: one from Yedoma deposits (61.75967°N, 130.47438°E), one below a lake on the Yedoma upland (61.76086°N, 130.47466°E) and one from the adjacent drained Yukechi alas basin (61.76490°N, 130.46503°E) (Table 1A; Fig. S1, Supporting Information). Biogeochemical characterization (Windirsch *et al.* 2020, Ulrich *et al.* 2021) as well as greenhouse gas production (Jongejans *et al.* 2021) is published. In Potsdam, the frozen cores were handled in a freezer laboratory at −10°C. They were split lengthwise using a band saw and were subsequently subsampled. After cleaning the splitting surfaces by scratching with cleaned knifes, the core stratigraphies were described visually and photographed. The sampling core halves were cut into subsamples and divided for further analysis, while the archive core halves were wrapped and sealed again for potential future additional analysis. For the samples used in this study, we took the inner part of the cores only to avoid any other sources of contamination on the outer core margins. After extraction, the subsamples were sent frozen with dry ice to Marseille, France, for further analyses.

The Stanchikovsky Yar sample (sample B) from which two live viruses were previously isolated (Legendre *et al.* 2014, 2015), and three modern surface soils from pristine cold regions (Kamchatka) were also analyzed for comparison (Table 1B). Figure S1 (Supporting Information) indicates the locations of these sampling sites.

DNA extraction

Samples were processed in BSL2 conditions using either the Miniprep (0.25 g of starting material) or Maxiprep (20 g of starting material) version of the DNeasy PowerSoil Kit from Qiagen (Venlo, Netherlands). Extractions were repeated until ∼1 μg (or more) of raw DNA was obtained for each sample (Table S1, Supporting Information). Samples B, C, D, E were processed using the miniprep kit, following the manufacturer's protocol and by adding dithiothréitol (DTT) to the C1 buffer at a 10 mM final concentration. All other samples (K, L, M, N, O, P, Q, R) were processed using the Maxiprep kit, including the addition of DTT to the C1 buffer at a 10 mM final. Samples were ground for 20 s in an MP FastPrep homogenizer (MP Biomedicals, Santa Ana, California, USA) at a speed of 4 m/s then incubated for 30 min at 65°C and finally ground again for 20 s at 4 m/s. After elution in 5 ml of elution buffer the extracted raw DNA was concentrated on a silica column from the Monarch Genomic DNA purification Kit from New England Biolabs (Ipswich, MA, USA). The DNA contents were then quantified using Qubit (Thermo Fisher Scientific, Waltham, MA, USA) and NanoDrop (Thermo Fisher Scientific) assays (Table S1, Supporting Information) prior sequencing.

DNA sequencing, assembly and annotation

Each metagenomic sample was sequenced with DNA-seq paired-end protocol on Illumina (San Diego, CA, USA) HiSeq 4000 platform at the French National Sequencing Center 'Genoscope' (https://jacob.cea.fr) producing datasets of 2 × 150 bp read length, except for sample B (2 × 100 bp read length) (Table S1, Supporting Information). Sequencing libraries were prepared as previously described in Alberti *et al.* (2017) using an 'on beads' protocol. Briefly, when available, 250 ng of genomic DNA were sonicated using the E210 Covaris instrument (Woburn, Massachusetts, USA) and the NEBNext DNA Modules Products (New England Biolabs) were used for end-repair, 3′-adenylation and ligation of NextFlex DNA barcodes (Bioo Scientific Corporation, Austin, TX, USA). After two consecutive 1× AMPure XP cleanups (Beckman Coulter, Brea, CA, USA), the ligated product were amplified by 12 cycles of PCR using Kapa Hifi Hotstart NGS library Amplification kit (Kapa Biosystems, Wilmington, MA, USA), followed by 0.6× AMPure XP purification. Finally, libraries were quantified by quantitative PCR (qPCR) (MxPro; Agilent Technologies, Santa Clara, CA, USA) and library profiles were evaluated using an Agilent 2100 Bioanalyzer (Agilent Technologies).

Reads were quality checked with FASTQC (v0.11.7) (Andrews 2014). Identified contaminants were removed and remaining reads were trimmed on the 3′ end using 30 as quality threshold with BBTools (v38.07) (BBMap 2022). Assemblies of filtered datasets were performed using MEGAHIT (v1.1.3) (Li *et al.* 2015) with the following options: k-list 33,55,77,99,127-min-contig-len 1000. All filtered reads were then mapped to the generated scaffolds using bowtie2 (Langmead and Salzberg 2012) with the –very-sensitive option. The mapping of the contigs was used to calculate the mean coverage of each contig. To ensure a reliable taxonomic assignation, only scaffolds longer than 5 kb were considered and submitted to metaGeneMark (v3.38) (Zhu *et al.* 2010) with the default metagenome-style model. The predicted genes were then compared with the RefSeq database (Li *et al.* 2021) using the Diamond version of blastP (v0.9.31) (Buchfink *et al.* 2015) with an *E*-value < 10^{-5}, identity percentage threshold of 35 and retaining only the best hits. The taxonomy of each scaffold was then inferred with a custom-made script applying a last common ancestor (LCA) method. At each rank, recursively, the taxonomy was conserved only if over half of the annotated genes presented the same keyword. Unclassified and ambiguous contigs were screened against RefSeq with Diamond blastX (–sensitive option and *E*-value < 10^{-5}), in order to detect exons. Non-overlapping best hits were used. The above taxonomical protocol was applied and only eukaryotic scaffolds were conserved. An alternative taxonomical annotation at the single read level was computed by Kraken (v2) (Wood *et al.* 2019) using its standard database. This annotation was found to be much less comprehensive and was not used further.

The diversity and sequencing effort of each sample was estimated by Nonpareil (v3.304) (Rodriguez-R *et al.* 2018) on forward reads over 35 bp using the *k*-mer algorithm.

Contig taxonomy and coverage data were used for beta-diversity estimation together with the *k*-mer frequency of reads. To do so, 15 bp *k*-mer frequency was retrieved for forward and re-

Table 1B. Cold surface soils samples analyzed in this study.

Sample	IGS code	Latitude (°)	Longitude (°)	Locality	Type
P2	C	54.54972	160.58194	Kamchatka, Kronotsky river bank	Vegetation free soil
P4	D	55.09861	160.34944	Kamchatka, Kizimen volcano	Tundra soil
P5	E	55.11500	159.96333	Kamchatka, Shapina river bank	Vegetation free soil

verse reads by Gerbil (v1.12) (Erbert *et al.* 2017). An in-house script was used to merge the produced files and compute the cosine distance between samples.

The Pfam domains (Nov. 2021 database version; Mistry *et al.* 2021) of all open reading frames (ORFs) were searched with InterProScan (v.5.39-77.0; Jones *et al.* 2014). The non-overlapping matches with E-value $< 10^{-5}$ were retrieved with an in-house script.

β-Lactamase gene detection and analyses

ORFs with a best Diamond blastP match to a β-lactamase and β-lactamase-related proteins in the NCBI RefSeq database were retrieved and analyzed (Table S2, Supporting Information). Functional domains were inferred by InterProScan (v5) against Pfam, ProSitePatterns and by Batch CD-search (Marchler-Bauer and Bryant 2004) against the COG database (Galperin *et al.* 2021) (E-value $< 10^{-5}$) (Table S3, Supporting Information). ORFs at least 100 amino acids in length exhibiting a β-lactamase domain according to Pfam and COG (except for class D solely defined by COG domain) were kept for further analyses. We attempted to identify the β-lactamase-encoding contigs as potential plasmids using PlasClass (Pellow *et al.* 2020) and PlasFlow (Krawczyk *et al.* 2018). However, the results were found to be unreliable and not used further.

To normalize the β-lactamase count by a biologically relevant metric, we estimated the number of bacteria in our metagenomic dataset using 120 bacterial single copy genes from the Genome Taxonomy database using GTDB-Tk (Chaumeil *et al.* 2020) as reference (Table S4, Supporting Information).

For comparison, the abundance of β-lactamases was also computed in all of the terrestrial metagenomics datasets available from the IMG/M database from June 2021 (Chen *et al.* 2021). One thousand eight hundred two datasets were initially downloaded and then checked for contig length, coverage and quality. Four hundred thirty-four datasets from known biomes and of sufficient sizes (at least one contig over 10 kb and at least three bacteria equivalent sequenced) were retained and analyzed with the exact same protocol used to detect ORFs, β-lactamases and bacterial single-copy genes in our own dataset (Table S5, Supporting Information).

All statistical analyses were done using v4.1.1 of R (R Core Team 2021). Bray–Curtis dissimilarities were calculated with package vegan (v2.5-7) (Oksanen *et al.* 2020) and principal coordinate analysis performed with package stats4 (v4.1.1).

Results

DNA extraction yield and quality

Large variations were observed in the DNA contents of the different soil samples despite their similar macroscopic appearances (black/brownish fine silt and sandy compact soils), including between surface samples (taken from vegetation-free spots). Using the same extraction protocol, up to 75 times more DNA could be recovered from similarly dated ancient soils from different sites (Table S1, Supporting Information, for instance sample B vs P). A lesser range of variation (10 times) was observed between surface samples (samples C and D) already suggesting that the sample age was not the main cause of these differences. Interestingly, very similar total numbers of usable base pairs and good quality reads were determined from the 250 ng of purified DNA used for sequencing) (Table S1, Supporting Information). Except for sample B (sequenced as shorter reads on a different platform), we observed no correlation between the initial DNA content of the sample and the number of usable good quality reads (Table S1, Supporting Information), indicating that, once purified, all DNAs exhibited a similar quality.

Sample diversity

In contrast, the assembly of a similar number of sequenced reads from the different datasets resulted in a highly variable number of contigs (≥5 kb). As these values are linked to the complexity of the sample (species richness, number of different species; and their relative abundance, species evenness), this suggested that the samples exhibit globally different microbial population compositions and structures. As our samples are expected to be variable in richness and evenness while the number of reads is quite similar (except for sample B), the sequencing effort is expected to be different for each sample. This was confirmed by the Nonpareil tool that calculates the sequencing effort and an alternative to the Shannon diversity index computed from the *k*-mer diversity of reads (Rodriguez-R *et al.* 2018). Prior to any taxonomical annotation we could thus estimate the sample diversity as well as the estimated sample coverage (i.e. the proportion of the sample diversity actually sequenced) that ranges from 0.39 (sample E) to 0.9 (sample L) (Fig. S2, Supporting Information). Given the same number of reads for two samples, a larger coverage corresponds to a lesser diversity. According to Fig. 1, the richest is sample E and the poorest is sample L. There are large differences between samples, including among the Kamchatka surface samples that appear the most diverse. The Nonpareil diversity values do not significantly correlate with the amounts of DNA recovered from one gram of soil (Table S1, Supporting Information). In contrast and as expected, the diversity values negatively correlate with the median coverage of the contigs in each sample (Pearson correlation of -0.75, $P < 0.005$).

When focusing on a given borehole, we found that the diversity appeared to increase with sample depth in a given borehole (Fig. 1). However, diversity values were quite variable for samples taken at similar depths across different boreholes, in line with the lack of a consistent relationship between the diversity and the age (from radiocarbon dating) of the sample (Table 1A).

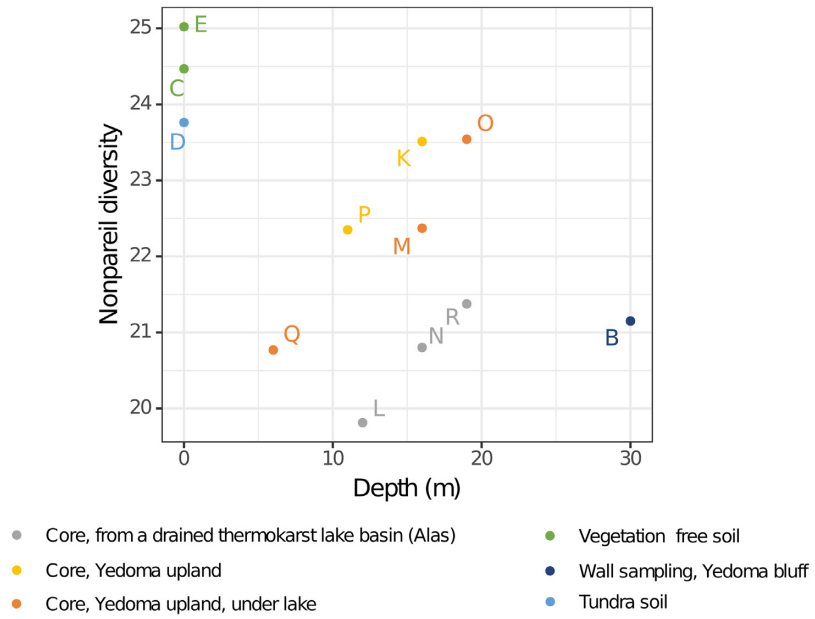


Figure 1. Nonpareil diversity compared with sample depth. Nonpareil diversity was computed on reads over 35 bp in size. This metric is correlated to the Shannon index (Rodriguez-R *et al.* 2018).

Global species content

As depicted in Fig. S3A (Supporting Information), bacteria account for over 90% of the estimated abundance in most samples and 87% in average but they represent 95% of contigs (>5 kb). This is true for all modern cryosols and most ancient permafrost layers despite their different origins and DNA contents. Even higher proportions of bacterial sequences were computed by Kraken2 (Wood *et al.* 2019) (Fig. S3B, Supporting Information). However, we found this method to be less sensitive and was not used further in this study. Four samples, L, N, R and Q, appear to be poorer in bacterial sequences (≤77%) due to their higher proportions of unclassified and Archaea sequences (Fig. S3A, Supporting Information).

In all samples combined, Archaea represent 3.2% of the abundance but only 1.8% of the total contig number. In most samples, they are very rare except in sample Q (unfrozen talik) and three samples from different depths in the same borehole (L, R, N: respectively, 12, 16 and 19 m) where they represent 6.5–11.3% of the total abundance (Fig. S3A, Supporting Information). In the samples where they are in sizable amounts, the archaeal population is dominated by Bathyarchaeota (a recently described clade of methanogens belonging to the TACK phylum) (Evans *et al.* 2015), Nitrososphaeria (a class of chemolithoautotrophic ammonia oxidizing Archaea belonging to the Thaumarchaea and globally distributed in soils) (Hatzenpichler 2012) and Methanomicrobia (a class of methanogens belonging to the Euryarchaeota and previously noticed to be enriched in the sediment microbiomes associated with the rhizosphere of macrophytes) (Behera *et al.* 2020) (Fig. S4, Supporting Information).

Eukaryotic DNA represents <1% of contigs and 0.5% of the total abundance, except for two subsurface samples from the same borehole (R, N) with 1% and 2% of eukaryotes respectively (Fig. S3, Supporting Information). The Viridiplantae (land plants and green algae) and Unikont (mostly Fungi, Metazoan and Amoebozoa) constitute most of the represented clades except for samples E and O (Fig. S5, Supporting Information). The latter only exhibits 68 total eukaryotic contigs, 21 of which were attributed to the Bacillariophyceae (the class of diatoms). With 30 other contigs most likely originating from green algae, this suggests that this 40 000-year-old layer originated from a lake bottom deposit (Table 1A) (Fig. S5, Supporting Information).

Finally, according to the LCA assignment method, viruses were predicted to only represent 0.4% of the global coverage of all samples (Fig. S3, Supporting Information), with one sample (R) standing out by reaching 2.3%. Such low representation prompted us to reanalyze our datasets using VirSorter, a leading software tool specifically dedicated to the identification of viruses in metagenomics data (Roux *et al.* 2015). This new protocol increased the average abundance of virus to 2.68%, with only a handful of samples exhibiting values higher than 2%, including the R sample at 8.6% (Fig. S6, Supporting Information). The newly annotated viral (mostly phage) contigs originated from contigs attributed to Bacteria by the LCA method (Fig. S6B, Supporting Information). The family distributions between theses samples is quite variable but, in general, the most abundant viruses are Mimiviridae, Pithoviridae (the prototype of which was previously isolated from ancient permafrost; Legendre *et al.* 2014), Caudovirales, Phycodnaviridae and Ascoviridae, all families known to gather giant or large DNA viruses. A detailed analysis of these virus populations will be published elsewhere.

A high diversity that points out syngenetically trapped communities

The global community analysis and an analysis of the *k*-mer frequencies of reads in each dataset do not consistently put samples from the same borehole close to each other (L-N-R, P-K, Q-M-O) (Fig. S7, Supporting Information). These analyses do not allow a meaningful clustering of samples originating from the different boreholes, or even distinguish surface samples from permafrost samples (Fig. S7, Supporting Information).

Focus on the bacterial populations

The relative abundances of phyla representing >0.5% of the total sum of coverages are depicted in Fig. 2. Actinobacteria and Proteobacteria are the two most abundant phyla. Although

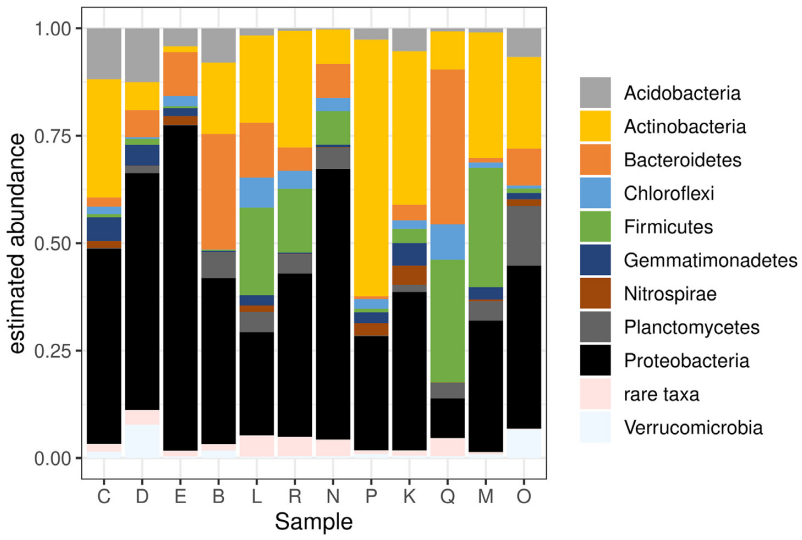


Figure 2. Bacterial phyla abundances across samples. The abundance is estimated as the fraction of the total sum of coverages.

Proteobacteria appear globally dominant (40% of the estimated abundance and over 75% in sample E), it is not always the most abundant in every sample. The most extreme cases include sample P in which Actinobacteria are largely dominant, and sample Q in which Bacteroidetes is the leading phylum. The relative abundance of Firmicutes, Planctomycetes, Acidobacteria and Verrumicrobia varies strongly across samples from up to 28%, 14%, 13% and 7.8%, respectively, almost down to zero. None of these characteristics appears to correlate with the geographic location, depth or age of the samples.

Within a given phylum, such as the Proteobacteria (Fig. S8, Supporting Information), very distinct distributions across classes could be seen between different modern cryosols (C, D, E) or ancient permafrost layers either from the same borehole (L-R-N, Q-M-O) or from similar depth (16 m: M and N; 19 m: R and O). Species from the Alpha, Beta and Delta/Epsilon divisions constitute the bulk of the Proteobacteria populations, although with strong variations. For instance, Betaproteobacteria are quasi absent from the R and N samples, as are Alphaproteobacteria from sample E.

At lower taxonomical ranks the distributions become scattered, as well as less informative given the smaller proportion of non-ambiguously annotated contigs (66.3% at the phylum level, 43.6% at the order level, but only 16.4% at the genus level). The most abundant order overall is Rhizobiales that is present in all samples (Fig. 3). Unclassified Actinobacteria are the second most abundant group suggesting that the cryosol Actinobacteria are very different from their cultured relatives. In samples taken individually, Rhizobiales is the most abundant order in samples N and R where they represent 53% and 23% of the bacterial abundance, respectively. Burkholderiales are the most abundant order in samples E, D and O and Clostridiales are the most represented in samples Q and O. In the other samples, unclassified Actinobacteria represent the major group. In some samples (C, D, E, B, L, O), the high abundance of Proteobacteria is divided between multiples orders in the phylum rather than due to a predominant one. In samples R, N, P, K, M on the other hand, Proteobacteria are mostly represented by Rhizobiales (Fig. 3). Some of the orders found also in deep samples are known to be aerobic. This is the case for Vicinamibacteraceae (Huber and Overmann 2018) and Gaiellales (Albuquerque and da Costa 2014) but these together constitute only 2.4% of the bacterial abundance. Most groups present in these cryosols are thus facultative anaerobes and chemotrophs.

The analysis suggests a lot of variability across samples, making it illusory to expect that a tractable number of bacterial orders could be used to classify the various types of cryosol and permafrost into recurrent taxonomical types. While surface samples and most deep samples form separate clusters (Fig. 3), we should be careful in drawing conclusions as the surface samples come from separate locations. When samples from the same borehole do cluster together (L-N-R and P-K), their dominant bacterial communities still exhibit some variations (Fig. 3). Interestingly, the large distance separating sample Q from the same borehole samples M and O is consistent with its unfrozen state as part of an under lake talik (Table 1A). Statistical analysis of similarities (ANOSIM) confirms these observations (Fig. S7, Supporting Information) showing that the distances between samples from the same borehole are not significantly lower than the distances between samples from different boreholes ($R = 0.388$, $P < 0.069$).

As a complimentary attempt to interpret the microbiome differences between samples, we screened the Pfam database, identified a total of 9481 motifs associated with specific protein families and/or molecular functions, and computed their distribution across samples. We expected that this type of analysis could reveal ecological/environmental constraints shared by various samples in spite of their different microbial communities. A list of the 50 most abundant Pfam motifs globally found in our dataset is given in Table S6 (Supporting Information). This list includes 7 of the 20 most abundant motifs found in the whole Pfam database, supporting the validity of our annotation protocol. Figure 4A presents a Bray–Curtis dissimilarity map computed from the Pfam counts in the samples. This map confirmed the functional similarity of the L-N-R samples, of the P-K samples (with the insertion of M from a different borehole, but from a similar depth and age as K) and to a lesser extent of two surface soil samples (D, E). The rest of the samples exhibited unrelated functional patterns, with the unfrozen talik-derived sample Q as the most dissimilar.

Table S6 (Supporting Information) also indicated the presence of three Pfam motifs, found in proteins involved in bacterial antibiotic resistance: TetR-like transcription regulators (PF00440, ranked 29th/9481), metallo-β-lactamases (PF00753, ranked 44th)

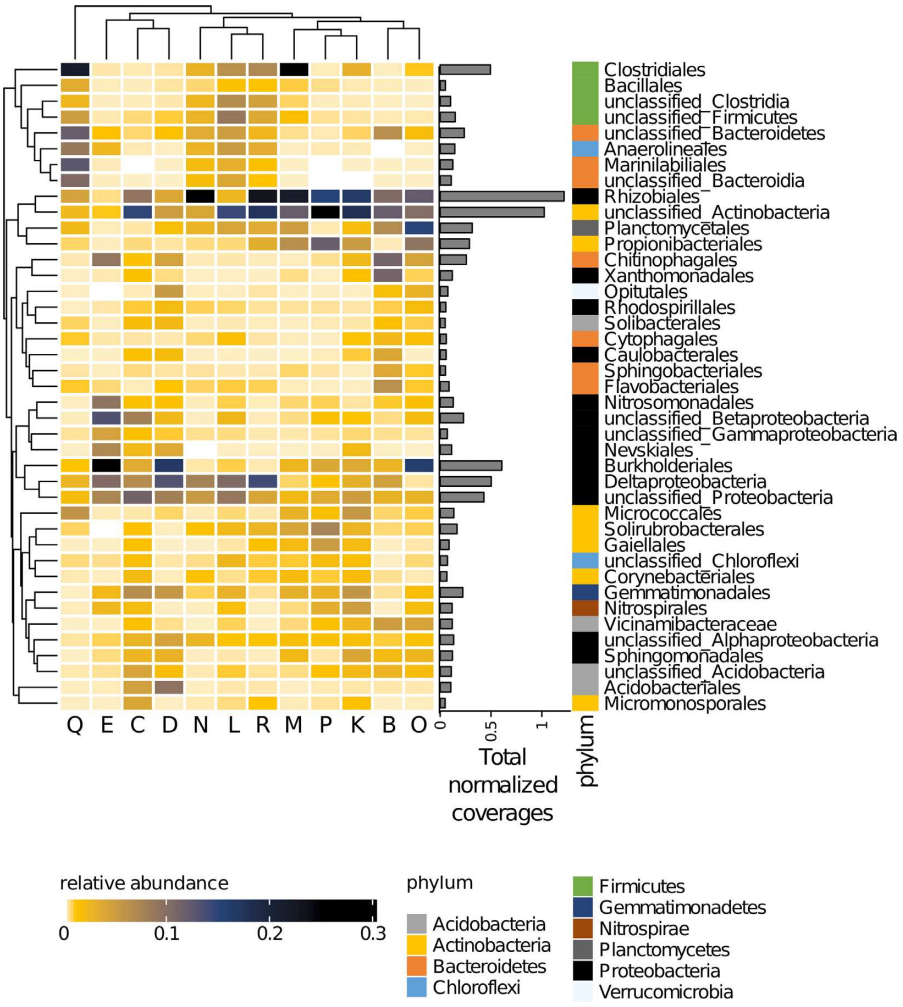


Figure 3. Bacterial abundance across samples. Samples and most abundant bacterial orders (over 0.5% of the total normalized coverage) were clustered using the Bray–Curtis distance. The abundance is estimated by the sum of coverages (normalized to 1 in each sample).

and glyoxalases (PF00903, ranked 46th). This prompted us to focus our next analysis on Pfam motifs putatively associated with virulence factors, the results of which are presented as a heat map in Fig. 4B. In this graph, three different protein families associated with bacterial antibiotic resistance appear to be largely dominant. The top one corresponds to a large family of enzymes catalyzing different reactions, including glyoxalases, dioxygenases and bleomycin/fosfomycin resistance proteins. Unfortunately, this Pfam motif is not specific enough to distinguish the minority of proteins truly involved in antibiotic inactivation from the rest of the family. Fortunately, the next most abundant functional families (β-lactamases and metallo-β-lactamases) are much better defined and for the most part correspond to enzymes active against various types of β-lactam antibiotics. Their distribution is further analyzed in the next section. Finally, the less abundant TetR (tetracycline resistance) family in the heat map is defined by an ambiguous motif found in a large diversity of DNA-binding proteins most of which are not involved in antibiotic resistance.

Bacteria carrying β-lactamase genes

Given the central role of extended-spectrum β-lactamases (ESBL) in the spread of multiple bacterial antibiotic resistances (Cantón *et al.* 2012, Potron *et al.* 2015, Wright 2019) we specifically investigated the presence and diversity of these proteins/enzymes in our dataset. Although the ancient microorganisms trapped in the pristine permafrost layers studied here have little probability to come in direct contact with human beings, the risk remains that they might contribute to new β-lactam resistances once in contact with modern bacterial species.

A total of 1071 β-lactamase genes were detected by diamond blastP against RefSeq and confirmed by a domain search (Fig. S9, Supporting Information). Among them, 1066 were found in bacterial contigs, 4 in an archaeal contigs and 1 in an ambiguous contig. Figure 5 presents the numbers, origins and classes of bacterial β-lactamases. Although these numbers are small compared with the total number of contigs (Table S1, Supporting Information), they nevertheless suggest that, in average, most (87%) of the sampled Bacteria harbor one copy of a β-lactamase gene, once normalized by the frequency of 120 known single copy genes detected in each sample (Fig. 5; Table S4, Supporting Information).

The normalized β-lactamase copy-number ranges from 0.39 [confidence interval: 0.22–0.55] per bacteria (sample N) to 1.75 [confidence interval: 1.54–1.97] (sample K) (Fig. 5). We noticed that the lowest prevalence of β-lactamase-encoding bacteria corresponds to samples from the same boreholes (L-R-N and Q-M), suggesting an actual ecological constraint. In most cases, we identified a single β-lactamase copy per contig, except for 17 contigs carrying several occurrences. In 8 of these, two β-lactamase genes

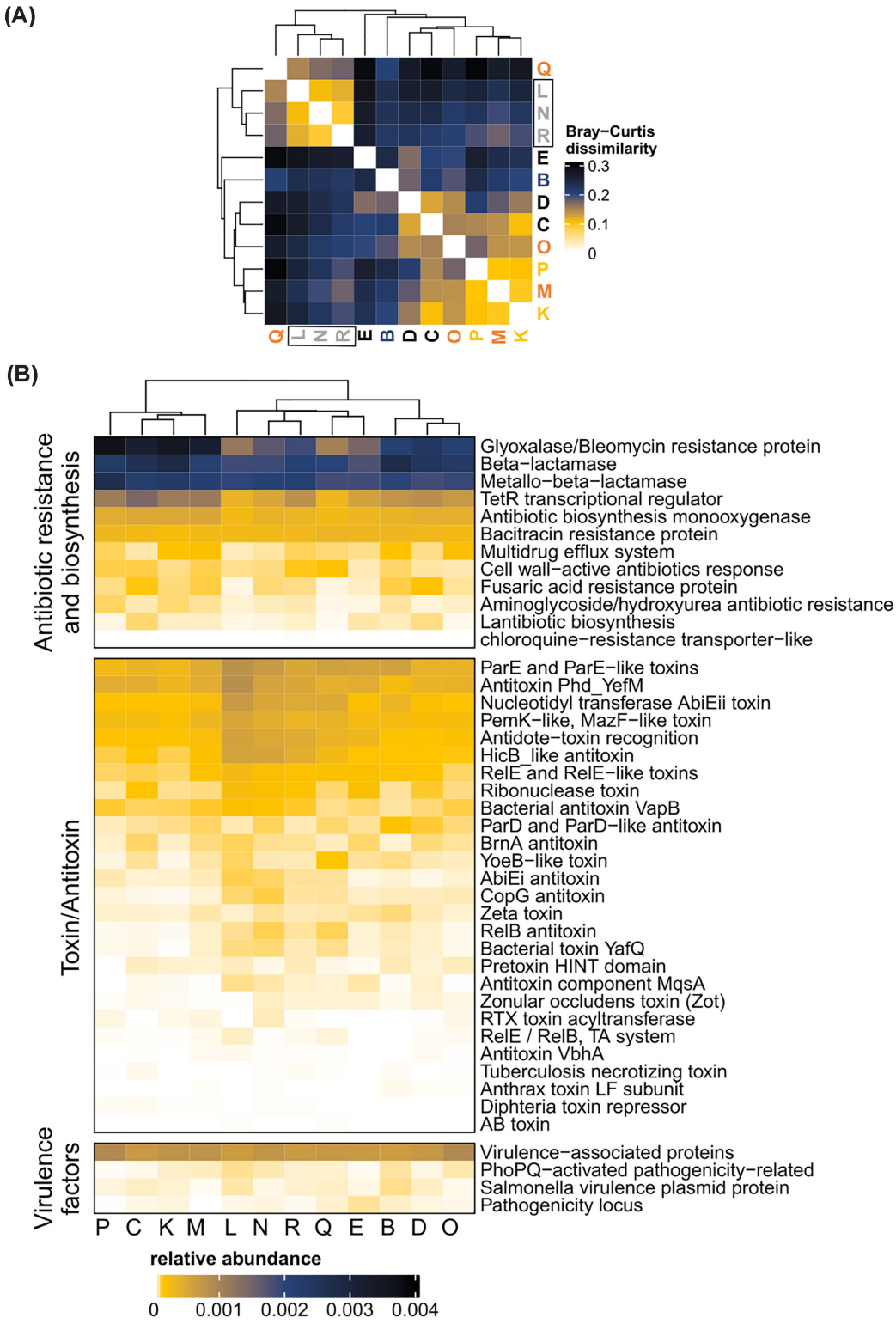


Figure 4. Analysis of the functional similarity between samples. **(A)** Bray–Curtis dissimilarity matrix computed from all Pfam counts in bacterial contigs. **(B)** Heat map of the abundance of Pfam motifs potentially associated with bacterial virulence factors.

were located next to each other and in 14 contigs, the two or three β-lactamase genes were found <10 ORFs away from each other.

Despite the large fluctuation in the average β-lactamase copy number, the relative proportions of the four enzyme classes remain fairly stable across samples. Class C (52–83%) was always the most represented, followed by class B (14% in average), then Class A and Class D (10%) (Fig. 5). All classes were present in every sample except for sample E (no class A β-lactamase).

Finally, we investigated the distribution of β-lactamase genes among the various bacterial phyla, as one given group of bacte-

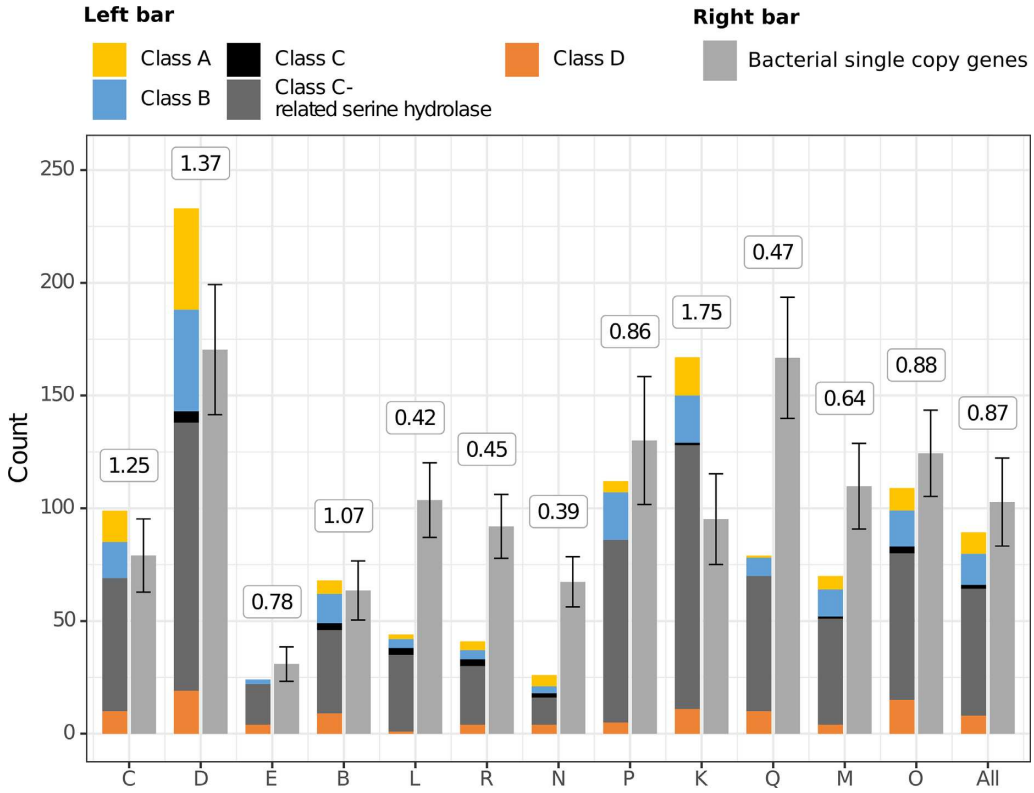


Figure 5. Bacterial β-lactamase classes and copy numbers. The distribution of the four main Ambler's classes of β-lactamases in the various samples is shown by the left bars. 'Class C-related serine hydrolase' stands for proteins with a β-lactamase AmpC domain and whose best diamond blastP matches were annotated 'β-lactamase-related serine hydrolase'. The number of bacterial single copy genes is shown by the adjacent gray right bar (mean count out of 120 reference single copy genes, with error bars corresponding to standard deviation). The boxes on top of each barplot indicate the average β-lactamase copy number per bacterial cell.

ria could concentrate most of β-lactamase genes. We thus simply compared the relative abundance of each bacterial phylum within the set of all cognate contigs with that of those exhibiting a β-lactamase gene. Bacteroidetes appeared to be the most enriched group in β-lactamase genes, while Planctomycetes exhibited the lowest proportion. Although the two distributions (Fig. 6) are different, they remain significantly correlated (Pearson coefficient > 0.8) pointing out that the occurrences of β-lactamase genes in these cryosols were shared among the diverse phyla constituting these microbiomes.

Comparing β-lactamase gene occurrences in the permafrost versus other soils types

To compare the β-lactamase gene proportion found in the above Russian cryosols to other types of soils, we extended our analysis to over 1500 terrestrial metagenomic assemblies publicly available from the JGI. As soil communities are usually very complex (e.g. Rodriguez-R *et al.* 2018), their assembly into large contigs is usually challenging (Alteio *et al.* 2020). Accordingly, only 434 datasets were found suitable to be further analyzed by our protocol. If the average number of β-lactamase genes per bacteria was again found to be close to one, it was quite variable (mean: 1.09 ± 0.77) (Fig. 7). Interestingly, if our samples did not stand out, other permafrost samples from Alaska and north Sweden ranked high in the list (Fig. 7). The samples with most occurrences of β-lactamase genes per bacteria were agricultural soils followed by riparian soils. There is significantly more β-lactamase copy number in agricultural samples than in other terrestrial soils except for the JGI permafrost samples and riparian soils (Fig. 7). Bog and forest soils also exhibited β-lactamase gene copy numbers close to one. The environments exhibiting the smallest β-lactamase gene copy numbers are shale carbon reservoirs, uranium contaminated soils and mire (a kind of wetland) (Fig. 7).

Discussion

A variable DNA content

We found the quantities of DNA recovered from the various cryosols to be highly variable, from >10 μg/g to 0.02 μg/g, thus a factor of 500 (Table S1, Supporting Information). Most of the low yielding samples are from ancient (subsurface) permafrost layers, albeit without a strict correlation with age. The Stanchikovsky Yar sample (B) exhibited a particularly high DNA content. However, once purified, the recovered DNA exhibited similar sequencing yields (i.e. number of reads/ng) and assembled into contigs of sizes similar to that of other samples (Table S1, Supporting Information). This suggests that the variation in DNA content is not due to its degradation, but reflects real differences in the global amount of microorganisms per gram of sample. Such a result is significant because it supports the hypothesis that genes encoded by the ancient DNA could be recycled within contemporary microorganisms through transformation.

An unexpected increase of diversity with sample depth

The nonpareil diversity (a good correlate of α-diversity and of the Shannon index) (Rodriguez-R *et al.* 2018) was found to increase with depth for each of the three boreholes. This is in

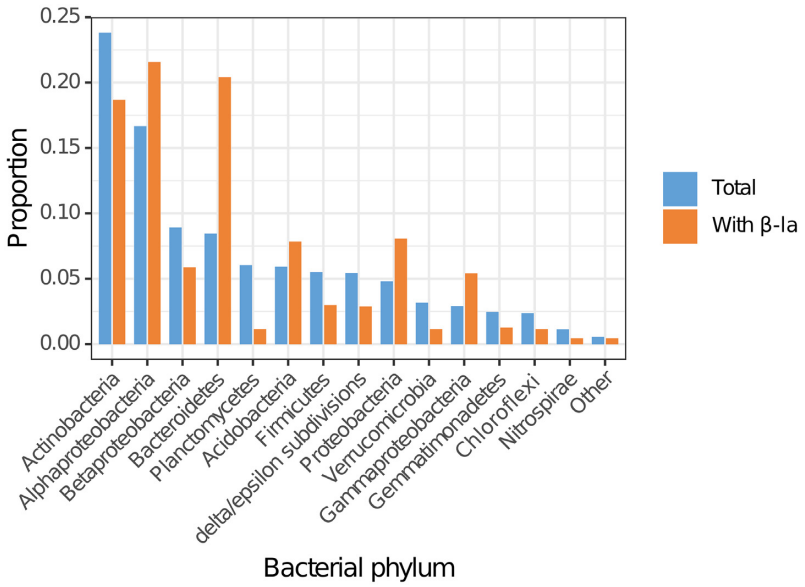


Figure 6. Side-by-side comparison of the proportion of bacterial phyla in all samples with the distribution of β-lactamase carrying bacteria. These proportions are globally correlated (Pearson correlation coefficient > 0.8), indicating that β-lactamase genes do not originate from a single dominant phylum. Bacteroidetes appeared to be the most enriched phylum in β-lactamases.

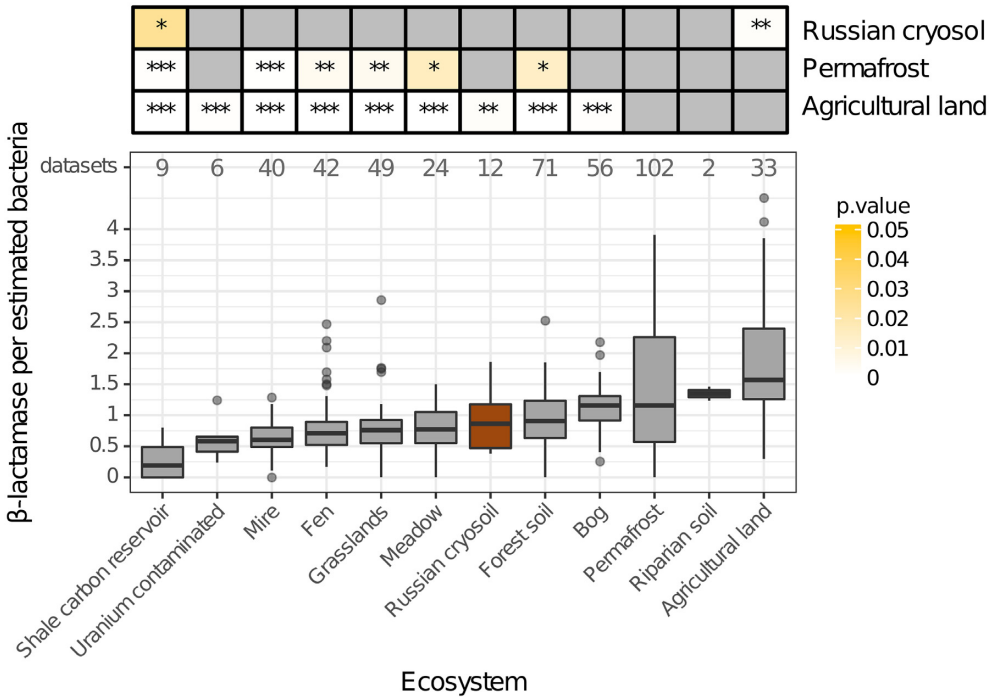


Figure 7. β-Lactamase gene copy numbers in available terrestrial metagenomes. The values are computed as the ratios between the β-lactamase gene counts and the average number of single copy genes identified in each sample by GTDB-Tk. Boxes are ranked according to the median of the β-lactamase gene copy number. Only datasets with three or more estimated entire bacteria sequenced were included. The brown box corresponds to the 12 new metagenomes analyzed in this work. The *P*-values of the pairwise comparison (Wilcoxon test) between our samples, agricultural lands and permafrost to other datasets are shown in the above heat map. Empty gray boxes correspond to nonsignificant differences.

contrast to the trend previously reported for culturable permafrost microorganisms (Gilichinsky *et al.* 1992). As argued by Abramov *et al.* (2021), microorganisms trapped in permafrost do not multiply and, because of gradual death, their diversity is expected to decrease over time. They claimed that, although growth as subzero temperature remains possible for some microorganisms (for instance, Panikov and Sizova 2007, Mykytczuk *et al.* 2013), the nutrient level and diffusion is too low to sustain growth as seen in the laboratory. The trend exhibited in our data between age and diversity is thus probably due to actual variations in diversity between syngenetically trapped microbiomes.

Scarcity of eukaryotes

A striking finding of our taxonomic analyses is the very low representation of eukaryotic organisms, only present in trace amounts in all our samples except for two deep samples (R, N) (Figs S3 and S5, Supporting Information). This result is unexpected given the fact that the warm spring–summer period sees the growth of a dense vegetation cover, accompanied by a variety of mosses and fungi, and a proliferation of insects, which should leave lasting traces in the metagenome of permafrost, once buried. Furthermore, the most abundant bacteria we found belong to the order Rhizobiales, which are often in symbiosis with plants. The Siberian permafrost harbored also members of Burkholderiales (6% in average, up to 27%), Chitinophagales (3% and up to 9%) and Cytophagales (0.7%), all of which are enriched in fungal–bacterial networks (Bonito *et al.* 2019). A previous culture-based study of Siberian, Canadian and Antarctic permafrosts estimated the count of fungal cells around 10^2 per gram of dry soil, distributed among 49 species (Ivanushkina *et al.* 2005). As fungi are in form of spores in permafrost (Vorobyova *et al.* 2001), part of their diversity might remain unseen because of difficulties to extract their DNA. The proportion of eukaryotic cells might also be much lower than that of bacteria. In such case, their sequencing coverage might be insufficient to assemble contigs of a least 5 kb in size. The larger size of eukaryotic genomes, the presence of introns and their lower coding density due to large intergenic regions also strongly decrease the likelihood of their identification by similarity searches in protein databases. However, *k*-mer-based read classification (Fig. S3B, Supporting Information) is consistent with an overall low proportion of eukaryotes. The remaining possibility is thus that eukaryotic DNA is rapidly recycled into microbial DNA during the seasonal decomposition process that takes place in the superficial active layer before being perennially frozen in deeper permafrost.

Variable bacterial community composition

As expected, bacteria constituted most of the DNA sources (Fig. S3, Supporting Information). Compared with barcoding methods, metagenomics allows an exploration of the sample diversity that is, in principle, only limited by a threshold of minimal abundance for a given species. This threshold results from three main factors: (i) the proportion of cells (or viral particles) of a given species (mostly unicellular) in each sample, (ii) the efficiency with which their DNA can be extracted (e.g. sporulating vs nonsporulating species) and (iii) our capacity to identify them by sequence similarity searches in reference databases. The last factor involves variations in protein-coding density (high in viruses and prokaryotes, much lower in most eukaryotes except for some fungi and protozoans), the fraction of the database corresponding to known members of the various phyla (high for Bacteria, much lower for viruses and Archaea), the level of sequence divergence among homologues within different domains (very high in viruses, lower in eukaryotes). Thus, even for a similar coverage (i.e. number of redundant reads per genome position) in a metagenomics mixture, bacteria are more likely to be identified than archaeal and eukaryotic organisms, viruses being the least likely to be identified (i.e. taxonomically unclassified). The comparisons of organism frequencies across domains are thus much less reliable than within a given domain. Yet, the predominance of the bacterial domain is unlikely to be due to a bioinformatic/annotation bias. To estimate the abundance of a given group we used the sum of coverages of the contigs attributed to the group. In theory, a better estimation would have been the number of attributed reads relative to the genome size (Nayfach and Pollard 2016). In the absence of information on the genome sizes, the sum of coverages thus remains a good approximation.

Globally, the phyla found to dominate the microbiomes in our cryosol and permafrost samples do not differ from those already reported as most abundant in previous analyses of temperate soils (Janssen 2006) or permafrosts (Tripathi *et al.* 2019, Xue *et al.* 2019). Those are Proteobacteria (40%), Actinobacteria (22%), Bacteroidetes (10%), Firmicutes (9%), Acidobacteria (5%), Planctomycetes (4%) and Chloroflexi (3%) (Fig. 2). However, even the most common phyla do exhibit very different proportions between samples: 76–9% for Proteobacteria and 59.7–1.4% for Actinobacteria (Fig. 2). We noticed that the same phyla dominate both in the surface cryosols and the permafrost sample except for members of the Firmicutes that are mostly found below the surface (<1.3% at the surface, up to 28.5% in permafrost). In contrast to a previous study (Tripathi *et al.* 2019), we did not detect a significant amount of Crenarchaeota. On the opposite, we detected Bathyarchaeota, absent from this same study (Fig. S4, Supporting Information).

The high taxonomical variability depicted here confirms that no significant mixing occurred between the distinct microbiomes of the permafrost layers, in particular for the Q, M, O samples extracted from the same borehole (Figs 2 and 4). The weak clustering of samples L, N, R and samples P, K using the Bray–Curtis dissimilarity might solely reflect the influence of the soil type or environmental conditions shaping bacterial communities.

Although dead cells would result in altered DNA and thus a lower efficiency of sequencing and assembling large enough contigs, it remains possible that most of the bacteria identified in our study might be dead. A previous metagenomic study of ancient Alaskan permafrost estimated that only 15–18% of cells were alive. However, the DNA depletion to remove dead cells did not significantly changed the relative proportion of bacterial groups (from phylum to family) or the phylogenetic diversity (Burkert *et al.* 2019).

There is a growing concern that taxonomy alone is insufficient to interpret the complexity of soil processes (Baldrian 2019) given the already high diversity of metabolisms and lifestyles encountered in a given bacterial group. A good example of such a difficulty is the presence of Anaerolineales species in all our samples (Fig. 3). Although their anaerobic metabolism is consistent with life in deep permafrost, all known isolates exhibit optimal growth temperature in the 37–55°C range (Sekiguchi *et al.* 2003, Yamada *et al.* 2006) or come from hot environments (Konishi *et al.* 2012). On the other hand, these detected Anaerolineales species appear to cohabit with close relatives to known psychrophilic bacteria such as *Devosia psychrophila* and *Dyadobacter psychrophilus*, the presence of which is not unexpected in cryosols.

Thawing permafrost as a source of ARGs

Our study indicated the presence of the four main classes of β-lactamases (Ambler 1980), the evolution of which can rapidly lead to ESBL in the context of strong antibiotic selection in clinical settings. Class A EBSL includes cephalosporinases and six types of dreaded carbapenemases (Sawa *et al.* 2020). Class B β-lactamases are metalloenzymes using zinc at their active center compared with a serine residue for other classes. They hydrolyze carbapenems and degrade all β-lactam agents except monobactams (Sawa *et al.* 2020). Class C β-lactamases derives from the ampC gene carried on the genome of many Enterobacteriaceae. Variants of these enzymes are known to reduce sensitivity to carbapenems (Sawa

et al. 2020). Finally, class D β-lactamases, also known as oxacillinases (OXAs) evolved from degrading an extended spectrum of cephalosporins to hydrolyze carbapenems (Sawa *et al.* 2020). However, from their sequences alone (except for near 100% identical residues) it is impossible to predict the substrate range and the potential clinical risk associated with the diverse β-lactamases identified in our dataset.

The presence of antibiotic-resistant bacteria carrying various types of β-lactamases has been previously reported in pristine Arctic and Antarctic surface cryosols and ancient permafrost (Allen *et al.* 2009, D'Costa *et al.* 2011, Petrova *et al.* 2014, Perron *et al.* 2015, Kashuba *et al.* 2017, Van Goethem *et al.* 2018, Haan and Drown 2021). However, these previous studies did not evaluate the proportion of bacteria carrying a β-lactamase gene or exhibiting a β-lactam resistance in the global population of soil bacteria, most of which are unculturable (or eventually dead).

The number of identified β-lactamase genes was divided by the estimated number of distinct bacteria to obtain a normalized gene count. This original approach provides a biologically relevant quantification that can be generalized to target other enzymes and detect specifically enhanced metabolic features in prokaryotic communities. The usual approach normalizes the ORF count by the number of reads, and is less biologically relevant than the ORF count per bacteria.

We found few previous studies against which comparing our results about the presence of β-lactamases in a quantitative manner. Almost all investigations of environmental 'resistomes' were performed on populations of culturable bacteria sampled near the surface and selected for their antibiotic resistance (e.g. Haan and Drown 2021). A few others were restricted to a given bacterial species (e.g. *Escherichia coli*) (Pormohammad *et al.* 2019). We found one study analyzing the proportion of β-lactam-resistant bacteria within different agricultural soils, untreated or amended with manure (Udikovic-Kolic *et al.* 2014). This proportion was found to vary from 0.67% to 7.4% in untreated versus manure-amended soils. Most of these resistant bacteria presumably encoded a β-lactamase, thus in a much lesser proportion than found in our cryosol samples. Another metagenomic analysis of the distribution of ARGs in 17 pristine Antarctic surface soils estimated the relative frequency of ARG to all genes in the [1.3–4.4 10^{-5}] range, which approximately correspond to 17% of soil bacterial encoding at least one ARG (Van Goethem *et al.* 2018) (assuming an average content of 3850 genes per bacterium; diCenzo and Finan 2017). Moreover, this fraction includes all ARG (e.g. multidrug resistance efflux pump, aminoglycoside acetyltransferase/nucleotidylyltransferase, aminocoumarin resistant alanyl-tRNA synthetase, etc) of which β-lactamases only constitute a small proportion.

The high proportion of β-lactamase-encoding bacteria found in our 12 cryosol samples initially appeared significantly greater than previously reported in the few publications concerning other pristine environments, undisturbed by anthropogenic activities. However, following the analysis of 434 supplementary datasets, the proportion of β-lactamases in our 12 samples came out near the average of its value for a variety of other soils, and for instance twice less than in some American agricultural lands (Fig. 7). Other permafrost samples turned out to be richer in β-lactamase than ours, some even overlapping with the level of cultivated soils (Fig. 7). The comparison of β-lactamase copy numbers between biomes revealed statistically significant differences (Fig. 7). Thus, β-lactamases are not distributed evenly among diverse environments under the influence of unknown factors that remain to be identified in future studies.

The function of β-lactamases in soil microbial communities is not clear, but our results might help building new hypotheses. Several soil microorganisms produce antibiotic β-lactams, including members of the Actinobacteria (e.g. *Streptomyces* species) (Ogawara 2016). It turns out that these bacteria are among the most abundant in the cryosol microbiomes (Fig. 2). Among eukaryotes, fungi (also abundant in our samples, Fig. S5, Supporting Information) can also produce many complex β-lactam-related compounds. They are the original source of two foundational β-lactam antibiotics: penicillin and cephalosporin (Brakhage *et al.* 2009). The high β-lactamase abundance might thus result from the selective advantage they confer to their bearers in the middle of a biochemical war between soil microbes.

Alternatively, β-lactamase may also play a role in a less violent scenario, outside of the simplistic 'war' paradigm. Below the minimal inhibitory concentration, antibiotics may modulate bacterial gene expression, thus acting as signaling molecules rather than lethal weapons. In such context, β-lactamases might interfere with quorum sensing (Yim *et al.* 2007). Another possibility is that some of these genes encode a broader function (such as carboxylesterases or DD-carboxypeptidases active on β-lactams (Higgins *et al.* 2001, Nan *et al.* 2019, Pandey *et al.* 2020) providing a fitness gain in addition to protecting against soil antibiotics).

Unexpectedly, this study revealed a new potential danger, due to the abundance of β-lactamases found encoded in a large phyletic diversity of cryosol bacteria. The DNA from these permafrost bacteria, dead or alive, thus constitutes an immense reservoir of historical ARGs. Their transfer to contemporary pathogenic bacterial in a clinical setting might further contribute to the antibiotic resistance crisis, now considered one of the biggest public health challenges of our time.

In a context of global warming, ancient permafrost becomes an antibiotic resistance flavored ice cream ready to be consumed without moderation by all passing bacteria.

Acknowledgements

We are deeply indebted to our volunteer collaborator, Alexander Morawitz, for collecting the Kamchatka soil samples. We thank Dr Audrey Lartigue for technical advice about DNA extraction and acknowledge the computing support of the PACA BioInfo platform. We thank M. Ulrich (DFG project #UL426/1-1) and P. Konstantinov for helping with fieldwork at the Yukechi site, as well as the Alfred Wegener Institute and Melnikov Permafrost Institute logistics for field support and sample acquisition.

Supplementary data

Supplementary data are available at *FEMSML* online.

Contributions

J-MC and CA initiated the project and designed the study; SR and SS performed the bioinformatic analyses; J-MC wrote the initial version of the manuscript; EC-F processed the sample and performed DNA extraction; KL supervised DNA sequencing; and GG, JS and ANF collected the deep core samples from the Yukechi site. All authors contributed to the writing of the final manuscript.

Data availability

The data underlying this article are available in the EMBL-EBI database at http://www.ebi.ac.uk/, and can be accessed with

the study number PRJEB47746 (secondary accession: ERP132049), and samples accession numbers: ERS7649018, ERS7649019, ERS7649020, ERS7649021, ERS7649022, ERS7649023, ERS7649024, ERS7649025, ERS7649026, ERS7649027 and ERS7649028.

Funding

This work was supported by the Agence Nationale de la Recherche grant (ANR-10-INBS-09-08) to J-MC and the CNRS Projets de Recherche Conjoints (PRC) grant (PRC1484-2018) to CA. EC-F was supported by a PhD grant (DGA/DS/MRIS #2017 60 0004). GG and JS were funded by a European Research Council starting grant (PETA-CARB, #338335) and the Helmholtz Association of German Research Centres (HGF) Impulse and Networking Fund (ERC-0013).

Conflict of interest statement. None declared.

References

Abramov A, Vishnivetskaya T, Rivkina E. Are permafrost microorganisms as old as permafrost? *FEMS Microbiol Ecol* 2021;**97**:fiaa260.

Alberti A, Poulain J, Engelen S *et al.* Viral to metazoan marine plankton nucleotide sequences from the Tara Oceans expedition. *Sci Data* 2017;**4**:170093.

Albuquerque L, da Costa MS. The family Gaiellaceae. In: Rosenberg E *et al.* (eds). *The Prokaryotes: Actinobacteria*. Berlin, Heidelberg: Springer, 2014, 357–60.

Allen HK, Moe LA, Rodbumrer J *et al.* Functional metagenomics reveals diverse beta-lactamases in a remote Alaskan soil. *ISME J* 2009;**3**:243–51.

Alteio LV, Schulz F, Seshadri R *et al.* Complementary metagenomic approaches improve reconstruction of microbial diversity in a forest soil. *mSystems* 2020;**5**:e00768–19.

Ambler RP. The structure of beta-lactamases. *Philos Trans R Soc Lond B Biol Sci* 1980;**289**:321–31.

Andrews S. FastQC: a quality control tool for high throughput sequence data. 2019. http://www.bioinformatics.babraham.ac.uk/projects/fastqc/.

Baldrian P. The known and the unknown in soil microbial ecology. *FEMS Microbiol Ecol* 2019;**95**:fiz005.

BBMap. 2018. https://sourceforge.net/projects/bbmap/.

Behera P, Mohapatra M, Kim JY *et al.* Benthic archaeal community structure and carbon metabolic profiling of heterotrophic microbial communities in brackish sediments. *Sci Total Environ* 2020;**706**:135709.

Biskaborn BK, Smith SL, Noetzli J *et al.* Permafrost is warming at a global scale. *Nat Commun* 2019;**10**:264.

Bonito G, Benucci GMN, Hameed K *et al.* Fungal–bacterial networks in the populus rhizobiome are impacted by soil properties and host genotype. *Front Microbiol* 2019;**10**:481.

Bradley CA, Altizer S. Urbanization and the ecology of wildlife diseases. *Trends Ecol Evol* 2007;**22**:95–102.

Brakhage AA, Thön M, Spröte P *et al.* Aspects on evolution of fungal beta-lactam biosynthesis gene clusters and recruitment of trans-acting factors. *Phytochemistry* 2009;**70**:1801–11.

Buchfink B, Xie C, Huson DH. Fast and sensitive protein alignment using DIAMOND. *Nat Methods* 2015;**12**:59–60.

Burkert A, Douglas TA, Waldrop MP *et al.* Changes in the active, dead, and dormant microbial community structure across a pleistocene permafrost chronosequence. *Appl Environ Microbiol* 2019;**85**:e02646–18.

Cantón R, Akóva M, Carmeli Y *et al.* Rapid evolution and spread of carbapenemases among Enterobacteriaceae in Europe. *Clin Microbiol Infect* 2012;**18**:413–31.

Chaumeil P-A, Mussig AJ, Hugenholtz P *et al.* GTDB-Tk: a toolkit to classify genomes with the genome taxonomy database. *Bioinformatics* 2020;**36**:1925–7.

Chen I-MA, Chu K, Palaniappan K *et al.* The IMG/M data management and analysis system v.6.0: new tools and advanced capabilities. *Nucleic Acids Res* 2021;**49**:D751–63.

Cohen J, Screen JA, Furtado JC *et al.* Recent Arctic amplification and extreme mid-latitude weather. *Nat Geosci* 2014;**7**:627–37.

D'Costa VM, King CE, Kalan L *et al.* Antibiotic resistance is ancient. *Nature* 2011;**477**:457–61.

diCenzo GC, Finan TM. The divided bacterial genome: structure, function, and evolution. *Microbiol Mol Biol Rev* 2017;**81**:e00019–17.

Erbert M, Rechner S, Müller-Hannemann M. Gerbil: a fast and memory-efficient *k*-mer counter with GPU-support. *Algorithms Mol Biol* 2017;**12**:9.

Evans PN, Parks DH, Chadwick GL *et al.* Methane metabolism in the archaeal phylum bathyarchaeota revealed by genome-centric metagenomics. *Science* 2015;**350**:434–8.

Fedorov A, Konstantinov P. Observations of surface dynamics with thermokarst initiation, Yukechi site, Central Yakutia. In: Phillips M, Springman SM, Arenson LU (eds). *Permafrost, Swets & Zeitlinger*, Lisse, Netherlands, 2003, 240–3.

Finley RL, Collignon P, Larsson DGJ *et al.* The scourge of antibiotic resistance: the important role of the environment. *Clin Infect Dis* 2013;**57**:704–10.

Fuchs M, Nitze I, Strauss J *et al.* Rapid fluvio-thermal erosion of a yedoma permafrost cliff in the Lena River Delta. *Front Earth Sci* 2020;**8**:336.

Fuehrer H-P, Schoener E, Weiler S *et al.* Monitoring of alien mosquitoes in Western Austria (Tyrol, Austria, 2018). *PLoS Negl Trop Dis* 2020;**14**:e0008433.

Galperin MY, Wolf YI, Makarova KS *et al.* COG database update: focus on microbial diversity, model organisms, and widespread pathogens. *Nucleic Acids Res* 2021;**49**:D274–81.

Gilichinsky D, Wagener S. Microbial life in permafrost: a historical review. *Permafr Periglac Process* 1995;**6**:243–50.

Gilichinsky DA, Vorobyova EA, Erokhina LG *et al.* Long-term preservation of microbial ecosystems in permafrost. *Adv Space Res* 1992;**12**:255–63.

Graham DE, Wallenstein MD, Vishnivetskaya TA *et al.* Microbes in thawing permafrost: the unknown variable in the climate change equation. *ISME J* 2012;**6**:709–12.

Haan TJ, Drown DM. Unearthing antibiotic resistance associated with disturbance-induced permafrost thaw in interior Alaska. *Microorganisms* 2021;**9**:116.

Hatzenpichler R. Diversity, physiology, and niche differentiation of ammonia-oxidizing archaea. *Appl Environ Microbiol* 2012;**78**:7501–10.

Hertig E. Distribution of anopheles vectors and potential malaria transmission stability in Europe and the mediterranean area under future climate change. *Parasit Vectors* 2019;**12**:18.

Higgins CS, Avison MB, Jamieson L *et al.* Characterization, cloning and sequence analysis of the inducible ochrobactrum anthropi AmpC β-lactamase. *J Antimicrob Chemother* 2001;**47**:745–54.

Hinsa-Leasure SM, Bhavaraju L, Rodrigues JLM *et al.* Characterization of a bacterial community from a Northeast Siberian seacoast permafrost sample. *FEMS Microbiol Ecol* 2010;**74**:103–13.

Huber I, Potapova K, Ammosova E *et al.* Symposium report: emerging threats for human health: impact of socioeconomic and climate

change on zoonotic diseases in the Republic of Sakha (Yakutia), Russia. *Int J Circumpolar Health* 2020;**79**:1715698.

Huber KJ, Overmann J. Vicinamibacteraceae fam. nov., the first described family within the subdivision 6 acidobacteria. *Int J Syst Evol Microbiol* 2018;**68**:2331–4.

Hueffer K, Drown D, Romanovsky V *et al.* Factors contributing to anthrax outbreaks in the circumpolar north. *Ecohealth* 2020;**17**:174–80.

Ivanushkina NE, Kochkina GA, Ozerskaya SM. Fungi in ancient permafrost sediments of the Arctic and Antarctic regions. In: Castello J, Rogers S (eds). *Life in Ancient Ice*. Princeton, NJ: Princeton Press. 2005, 127–40.

Janssen PH. Identifying the dominant soil bacterial taxa in libraries of 16S rRNA and 16S rRNA genes. *Appl Environ Microbiol* 2006;**72**:1719–28.

Johnson SS, Hebsgaard MB, Christensen TR *et al.* Ancient bacteria show evidence of DNA repair. *Proc Natl Acad Sci USA* 2007;**104**:14401–5.

Jones P, Binns D, Chang H-Y *et al.* InterProScan 5: genome-scale protein function classification. *Bioinformatics* 2014;**30**:1236–40.

Jongejans LL, Liebner S, Knoblauch C *et al.* Greenhouse gas production and lipid biomarker distribution in Yedoma and Alas thermokarst lake sediments in Eastern Siberia. *Global Change Biol* 2021;**27**:2822–39.

Kashuba E, Dmitriev AA, Kamal SM *et al.* Ancient permafrost staphylococci carry antibiotic resistance genes. *Microb Ecol Health Dis* 2017;**28**:1345574.

Keita MB, Hamad I, Bittar F. Looking in apes as a source of human pathogens. *Microb Pathog* 2014;**77**:149–54.

Kochkina G, Ivanushkina N, Ozerskaya S *et al.* Ancient fungi in Antarctic permafrost environments. *FEMS Microbiol Ecol* 2012;**82**:501–9.

Konishi M, Nishi S, Takami H *et al.* Unique substrate specificity of a thermostable glycosyl hydrolase from an uncultured anaerolinea, derived from bacterial mat on a subsurface geothermal water stream. *Biotechnol Lett* 2012;**34**:1887–93.

Krawczyk PS, Lipinski L, Dziembowski A. PlasFlow: predicting plasmid sequences in metagenomic data using genome signatures. *Nucleic Acids Res* 2018;**46**:e35.

Langmead B, Salzberg SL. Fast gapped-read alignment with bowtie 2. *Nat Methods* 2012;**9**:357–9.

Legendre M, Bartoli J, Shmakova L *et al.* Thirty-thousand-year-old distant relative of giant icosahedral DNA viruses with a pandoravirus morphology. *Proc Natl Acad Sci USA* 2014;**111**:4274–9.

Legendre M, Lartigue A, Bertaux L *et al.* In-depth study of mollivirus sibericum, a new 30,000-y-old giant virus infecting acanthamoeba. *Proc Natl Acad Sci USA* 2015;**112**:E5327–35.

Li D, Liu C-M, Luo R *et al.* MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. *Bioinformatics* 2015;**31**:1674–6.

Li W, O'Neill KR, Haft DH *et al.* RefSeq: expanding the prokaryotic genome annotation pipeline reach with protein family model curation. *Nucleic Acids Res* 2021;**49**:D1020–8.

Liu Y, Lillepold K, Semenza JC *et al.* Reviewing estimates of the basic reproduction number for dengue, Zika and chikungunya across global climate zones. *Environ Res* 2020;**182**:109114.

Malavin S, Shmakova L. Isolates from ancient permafrost help to elucidate species boundaries in Acanthamoeba castellanii complex (Amoebozoa: Discosea). *Eur J Protistol* 2020;**73**:125671.

Marchler-Bauer A, Bryant SH. CD-Search: protein domain annotations on the fly. *Nucleic Acids Res* 2004;**32**:W327–31.

Mistry J, Chuguransky S, Williams L *et al.* Pfam: the protein families database in 2021. *Nucleic Acids Res* 2021;**49**:D412–9.

Morens DM, Daszak P, Markel H *et al.* Pandemic COVID-19 joins history's pandemic legion. *mBio* 2020;**11**:e00812–20.

Mykytczuk NCS, Foote SJ, Omelon CR *et al.* Bacterial growth at −15°C; molecular insights from the permafrost bacterium *Planococcus halocryophilus* Or1. *ISME J* 2013;**7**:1211–26.

Nan F, Jiang J, Wu S *et al.* A novel VIII carboxylesterase with high hydrolytic activity against ampicillin from a soil metagenomic library. *Mol Biotechnol* 2019;**61**:892–904.

Nayfach S, Pollard KS. Toward accurate and quantitative comparative metagenomics. *Cell* 2016;**166**:1103–16.

Nitzbon J, Westermann S, Langer M *et al.* Fast response of cold ice-rich permafrost in northeast siberia to a warming climate. *Nat Commun* 2020;**11**:2201.

Ogawara H. Self-resistance in *Streptomyces*, with special reference to β-lactam antibiotics. *Molecules* 2016;**21**:605.

Oksanen J, Blanchet FG, Friendly M *et al. V e gan: community ecology packag e*. 2022. https://cran.r-project.org.

Pandey SD, Jain D, Kumar N *et al.* MSMEG_2432 of *Mycobacterium smegmatis* mc^2155 is a dual function enzyme that exhibits DD-carboxypeptidase and β-lactamase activities. *Microbiology* 2020:**166**:546–53.

Panikov NS, Sizova MV. Growth kinetics of microorganisms isolated from alaskan soil and permafrost in solid media frozen down to −35°C: growth kinetics in frozen media. *FEMS Microbiol Ecol* 2007;**59**:500–12.

Parkinson AJ, Evengard B, Semenza JC *et al.* Climate change and infectious diseases in the Arctic: establishment of a circumpolar working group. *Int J Circumpolar Health* 2014;**73**:25163.

Pellow D, Mizrahi I, Shamir R. PlasClass improves plasmid sequence classification. *PLoS Comput Biol* 2020;**16**:e1007781.

Perron GG, Whyte L, Turnbaugh PJ *et al.* Functional characterization of bacteria isolated from ancient Arctic soil exposes diverse resistance mechanisms to modern antibiotics. *PLoS One* 2015;**10**:e0069533.

Petrova M, Kurakov A, Shcherbatova N *et al.* Genetic structure and biological properties of the first ancient multiresistance plasmid pKLH80 isolated from a permafrost bacterium. *Microbiology* 2014;**160**:2253–63.

Plowright RK, Parrish CR, McCallum H *et al.* Pathways to zoonotic spillover. *Nat Rev Microbiol* 2017;**15**:502–10.

Pormohammad A, Nasiri MJ, Azimi T. Prevalence of antibiotic resistance in *Escherichia coli* strains simultaneously isolated from humans, animals, food, and the environment: a systematic review and meta-analysis. *Infect Drug Resist* 2019;**12**:1181–97.

Potron A, Poirel L, Nordmann P. Emerging broad-spectrum resistance in *Pseudomonas aeruginosa* and *Acinetobacter baumannii*: mechanisms and epidemiology. *Int J Antimicrob Agents* 2015;**45**:568–85.

R Core Team. 22-04-2022. *R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing*, Vienna, Austria. https://www.R-project.org

Revich BA, Podolnaya MA. Thawing of permafrost may disturb historic cattle burial grounds in East Siberia. *Glob Health Action* 2011;**4**:8482.

Revich BA, Tokarevich N, Parkinson AJ. Climate change and zoonotic infections in the Russian Arctic. *Int J Circumpolar Health* 2012;**71**:18792.

Rodriguez-R LM, Gunturu S, Tiedje JM *et al.* Nonpareil 3: fast estimation of metagenomic coverage and sequence diversity. *mSystems* 2018;**3**:e00039–18.

Roux S, Enault F, Hurwitz BL *et al.* VirSorter: mining viral signal from microbial genomic data. *PeerJ* 2015;**3**:e985.

Sawa T, Kooguchi K, Moriyama K. Molecular diversity of extended-spectrum β-lactamases and carbapenemases, and antimicrobial resistance. *J Intensive Care* 2020;**8**:13.

Schirrmeister L, Froese D, Tumskoy V *et al.* Yedoma: late pleistocene ice-rich syngenetic permafrost of Beringia. In: *Encyclopedia of Quaternary Science*. 2nd edn. 2013, 542–52.

Schneider von Deimling T, Grosse G, Strauss J *et al.* Observation-based modelling of permafrost carbon fluxes with accounting for deep carbon deposits and thermokarst activity. *Biogeosciences* 2015;**12**:3469–88.

Sekiguchi Y, Yamada T, Hanada S *et al.* Anaerolinea thermophila gen. nov., sp. nov. and Caldilinea aerophila gen. nov., sp. nov., novel filamentous thermophiles that represent a previously uncultured lineage of the domain bacteria at the subphylum level. *Int J Syst Evol Microbiol* 2003;**53**:1843–51.

Shmakova L, Bondarenko N, Smirnov A. Viable species of flamella (Amoebozoa: Variosea) isolated from ancient Arctic permafrost sediments. *Protist* 2016;**167**:13–30.

Strauss J, Schirrmeister L, Grosse G *et al.* Deep yedoma permafrost: a synthesis of depositional characteristics and carbon vulnerability. *Earth Sci Rev* 2017;**172**:75–86.

Timofeev V, Bahtejeva I, Mironova R *et al.* Insights from *Bacillus anthracis* strains isolated from permafrost in the tundra zone of Russia. *PLoS One* 2019;**14**:e0209140.

Tripathi BM, Kim1 HM, Jung JY *et al.* Distinct taxonomic and functional profiles of the microbiome associated with different soil horizons of a moist tussock tundra in Alaska. *Front Microbiol* 2019;**10**:1442.

Turetsky MR, Abbott BW, Jones MC *et al.* Carbon release through abrupt permafrost thaw. *Nat Geosci* 2020;**13**:138–43.

Turetsky MR, Abbott BW, Jones MC *et al.* Permafrost collapse is accelerating carbon release. *Nature* 2019;**569**:32–4.

Udikovic-Kolic N, Wichmann F, Broderick NA *et al.* Bloom of resident antibiotic-resistant bacteria in soil following manure fertilization. *Proc Natl Acad Sci USA* 2014;**111**:15202–7.

Ulrich M, Jongejans LL, Grosse G *et al.* Geochemistry and weathering indices of Yedoma and alas deposits beneath thermokarst lakes in Central Yakutia. *Front Earth Sci* 2021;**9**:704141.

Valentine MJ, Murdock CC, Kelly PJ. Sylvatic cycles of arboviruses in non-human primates. *Parasit Vectors* 2019;**12**:463.

Van Goethem MW, Pierneef R, Bezuidt OKI *et al.* A reservoir of 'historical' antibiotic resistance genes in remote pristine Antarctic soils. *Microbiome* 2018;**6**:40.

Vishnivetskaya TA, Petrova MA, Urbance J *et al.* Bacterial community in ancient siberian permafrost as characterized by culture and culture-independent methods. *Astrobiology* 2006;**6**:400–14.

Vorobyova E, Minkovsky N, Mamukelashvili A *et al.* Micro-organisms and biomarkers in permafrost. In: Paepe R, Melnikov VP, Elfi Van Overloop E *et al.* (eds). *Permafrost Response on Economic Development, Environmental Security and Natural Resources*. Springer: Dordrecht, Netherlands, 2001, 527–41.

Willerslev E, Hansen AJ, Poinar HN. Isolation of nucleic acids and cultures from fossil ice and permafrost. *Trends Ecol Evol* 2004;**19**:141–7.

Windirsch T, Grosse G, Ulrich M *et al.* Organic carbon characteristics in ice-rich permafrost in alas and Yedoma deposits, Central Yakutia, Siberia. *Biogeosciences* 2020;**17**:3797–814.

Wood DE, Lu J, Langmead B. Improved metagenomic analysis with Kraken 2. *Genome Biol* 2019;**20**:257.

Wright GD. Environmental and clinical antibiotic resistomes, same only different. *Curr Opin Microbiol* 2019;**51**:57–63.

Xue Y, Jonassen I, Øvreås L *et al.* Bacterial and Archaeal metagenome-assembled genome sequences from Svalbard permafrost. *Microbiol Resour Announc* 2019;**8**:e00516–9.

Yamada T, Sekiguchi Y, Hanada S *et al. Anaerolinea thermolimosa* sp. nov., *Levilinea saccharolytica* gen. nov., sp. nov. and *Leptolinea ta r divitalis* gen. nov., sp. nov., novel filamentous anaerobes, and description of the new classes *An aer o lineae classis* nov. and *Caldilineae classis* nov. in the bacterial phylum *Chloroflex i*. *Int J Syst Evol Microbiol* 2006;**56**:1331–40.

Yashina S, Gubin S, Maksimovich S *et al.* Regeneration of whole fertile plants from 30,000-y-old fruit tissue buried in Siberian permafrost. *Proc Natl Acad Sci USA* 2012;**109**:4008–13.

Yim G, Wang HH, Davies J. Antibiotics as signalling molecules. *Philos Trans R Soc Lond B Biol Sci* 2007;**362**:1195–200.

Zhou P, Yang X-L, Wang X-G *et al.* A pneumonia outbreak associated with a new coronavirus of probable bat origin. *Nature* 2020;**579**:270–3.

Zhu W, Lomsadze A, Borodovsky M. *Ab initio* gene identification in metagenomic sequences. *Nucleic Acids Res* 2010;**38**:e132.

Supplementary material

Fig S1: Sample locations. *Overview of sampling locations (left map) with details on the sampling sites from Yukechi (upper right) and Kamchatka (lower right)*

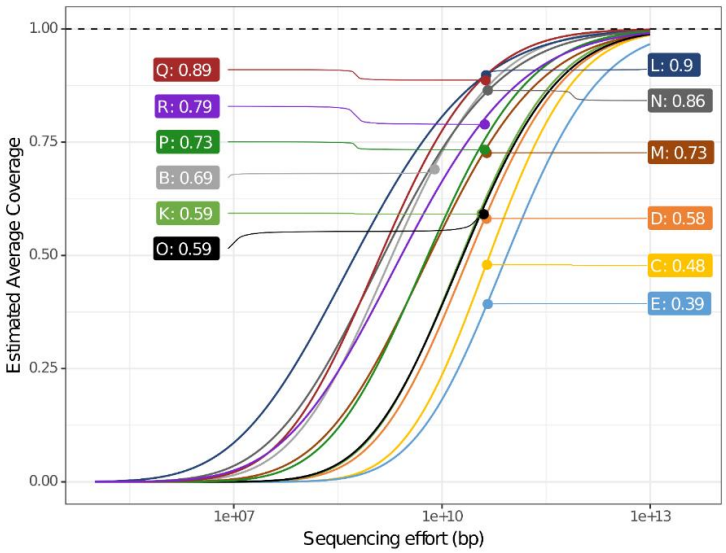


***Fig S2**: **Nonpareil curves and estimated dataset coverages.** Each point represents the estimated coverage (i.e proportion of the read diversity that was sequenced) for each dataset also given in the colored boxes. An estimated coverage of 1 would be reached in theory if all organisms from each sample had been sequenced.*

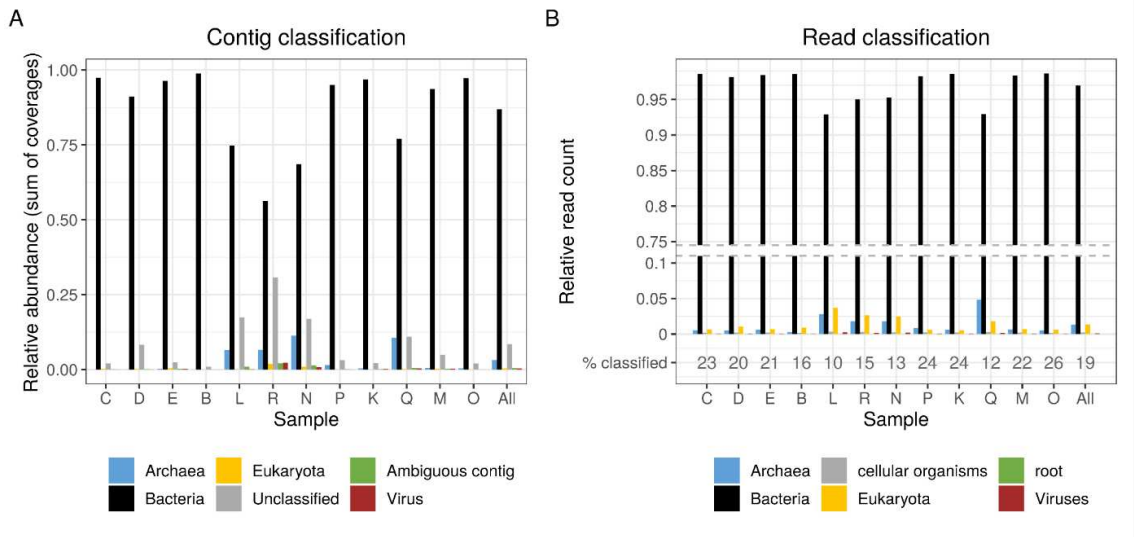


***Fig. S3**: **Relative domain abundance.** (A) Computed by the Lowest Common Ancestor (LCA) method. Contigs predicted as "Ambiguous" probably originate from various phages the detection of which is known to be challenging in metagenomics data. (B) Using a read-level algorithm (Kraken). Notice the small percentage (% classified) of used reads.*

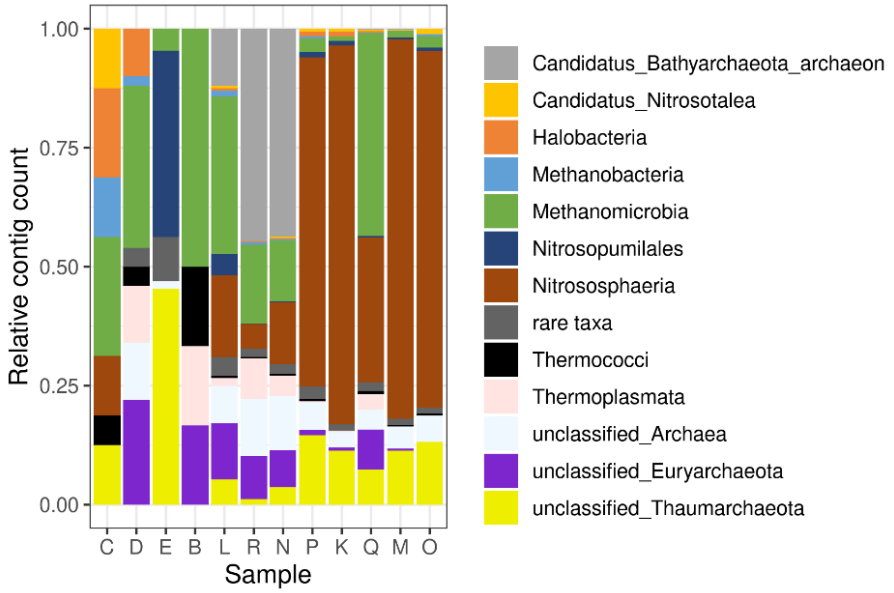


***Fig. S4**: **Detected Archaeal clades.** This bar graph should be interpreted with cautions as archaeal contigs constitute less than 1% of all contigs in most samples (Fig. S3). In the sample where they are in sizable amounts, the archaeal population is dominated by Bathyarcheota (a recently described clade of methanogens belonging to the TACK phylum), Methanomicrobia (a class of methanogens belonging to the Euryarchaeota and previously noticed to be enriched in macrophytes rhizosphere sediments), or Nitrososphaeria (a class of chemolithoautotrophic ammonia oxidizing archaea belonging to Thaumarchaea phylum and exhibiting global distribution in soils.*

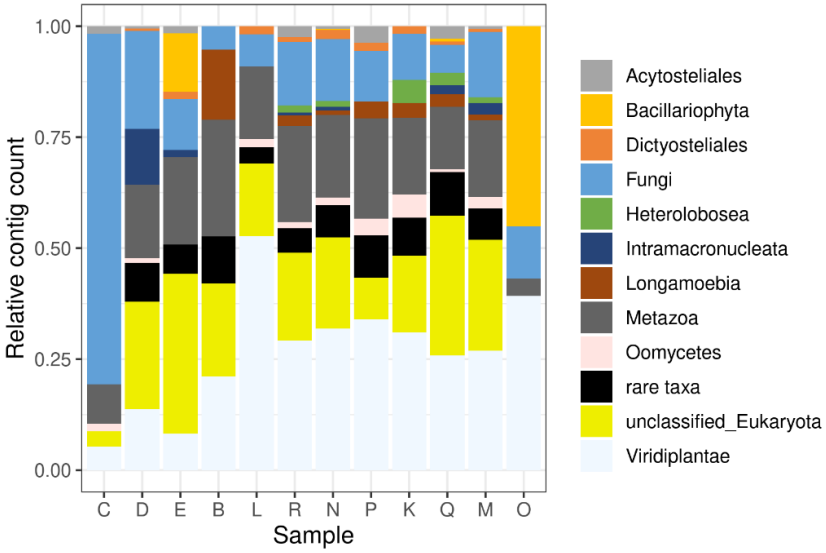


***Fig. S5**: **Detected eukaryotic clades.** This bar graph should be interpreted with cautions as eukaryotic contigs constitute less than 1% of all contigs in most samples (Fig. S3). In the sample where they are in sizable amounts (R and N), the eukaryote population is dominated by viridiplantae (land plants and green algae), metazoans and fungi.*

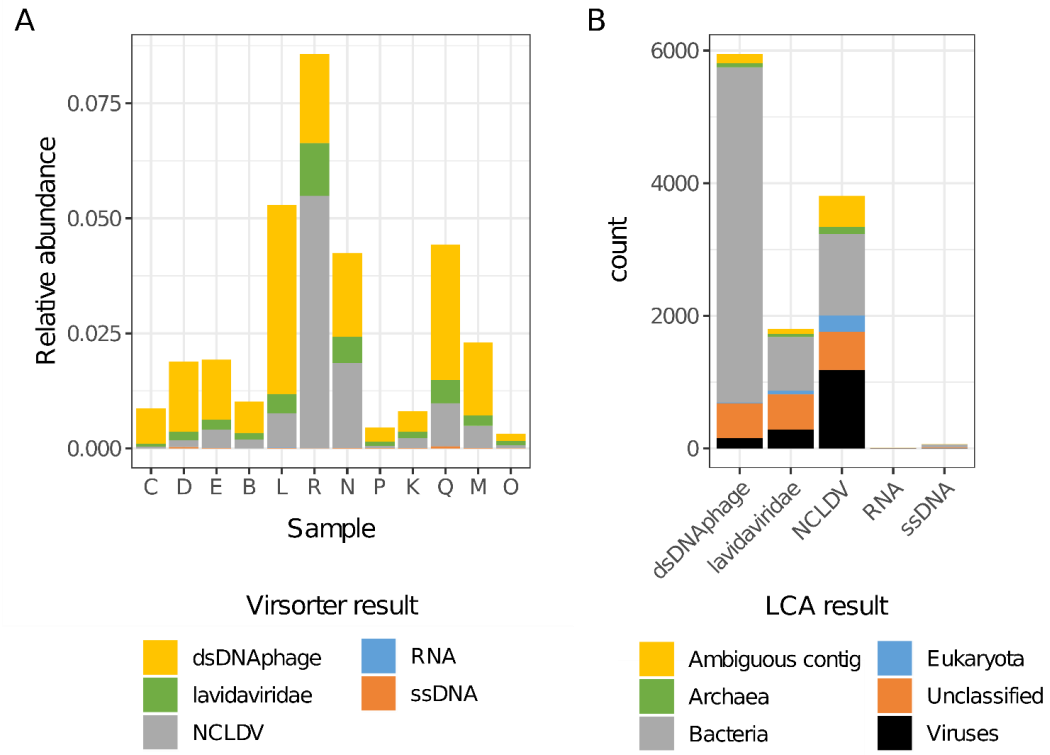


***Fig. S6**: **Viral contig classification by VirSorter**. (A) Viral contigs were retrieved applying a conservative threshold (0.9) as advised by the developers. On the left panel the predicted viral type is given for each dataset and its relative abundance is calculated as the sum of coverages divided by the total sample coverage. (B) Initial domain assignment by the LCA method of viral contigs predicted by VirSorter. Most differences involve contigs initially annotated as bacterial or unclassified.*

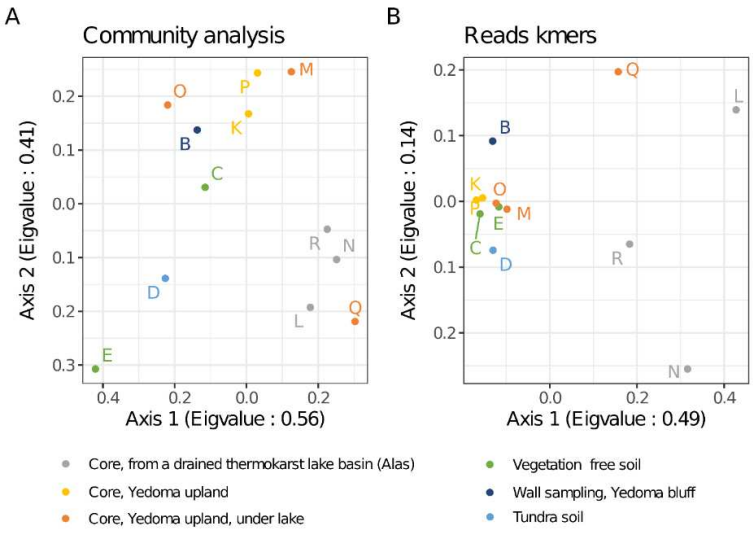


Fig S7: Principal coordinate analysis based on reference dependent and independent methods. *(A) The PCoA was calculated on the community data from all domains at the order level. The group abundance is estimated by the sum of coverages of each contig. (B) The calculation was done on the frequency of kmers of size 15 in reads of each dataset.*

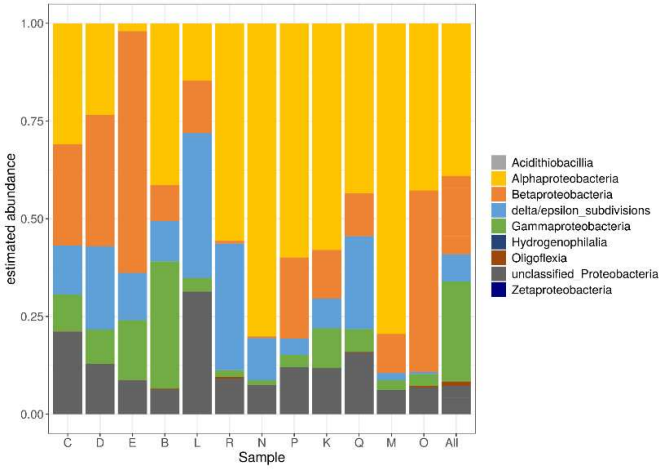


Fig. S8: Relative distributions of the main clades within the dominant Proteobacteria phylum. *Samples originating from various layers of the same borehole may exhibit very different profiles (e.g. L-R-N, Q-M-O), demonstrating the absence of significant mixing of their microbiome.*

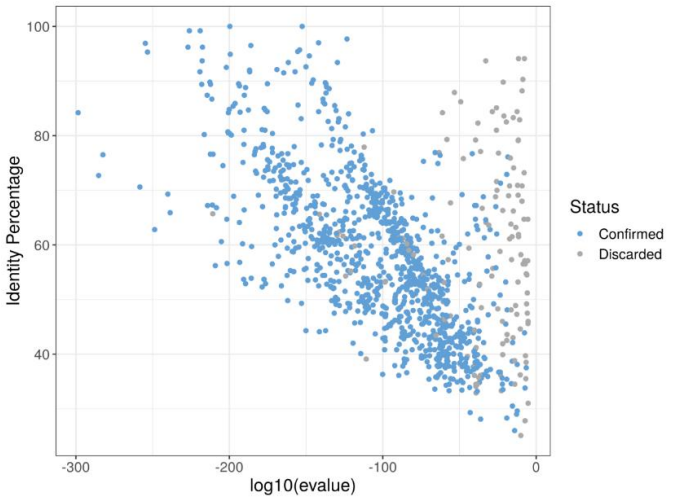


Fig. S9: Similarity (% identical residues) and matching P-value of predicted ORFs with a best match with a β-lactamase in the Refseq database using Diamond blastP. *ORFs were then discarded or confirmed based on size and the presence of a bona fide β-lactamase domain (See Material and methods). The bulk of predicted sample β-lactamases are in the unambiguous [30%-90%] identity range.*

Cet article présente la grande diversité microbienne des différents échantillons. Sur les trois forages réalisés pour accéder aux échantillons profonds, deux présentent des échantillons très divers mais dont la communauté bactérienne est tout de même plus ressemblante au sein d'un même forage qu'entre différents forages. Cependant, les ordres bactériens dominants ne sont pas les mêmes. Il en va de même pour les trois échantillons de surface provenant du Kamchatka. Un troisième forage (Q-M-O) présente des échantillons totalement différents. La métagénomique n'a pas permis de déceler de pathogène connu avec une grande confiance. En revanche, nous estimons qu'en moyenne presque toutes les bactéries séquencées présentent un gène codant pour une β-lactamase. Il est certain que la présence de paralogues fait augmenter cette moyenne.

Dans cette grande hétérogénéité, nous constatons que quatre échantillons (L-N-R et Q) ont moins de 77% de bactéries. Ce sont également les échantillons les plus riches en séquences inconnues puis en archées. Les contigs non classés de l'échantillon R représentent notamment 31% de l'abondance séquencée totale. Ceci pourrait indiquer la présence de virus. Nous avons testé cette hypothèse en appliquant un outil spécialisé, VirFinder, qui se base sur la fréquence de k-mers pour classer des séquences métagénomiques virales. La Figure S6 montre alors que les méthodes de taxonomie classiques sont mauvaises pour prédire la diversité virale. En effet : de nombreuses séquences classées comme *Nucleocytoviricota* (NCLDV) par VirFinder ont été classées par la méthode du dernier ancêtre commun comme bactérie. Il y a plusieurs raisons probables à cela :

- La grande proportion de gènes ou de fonctions typiquement cellulaires chez les virus géants
- La grande diversité des virus qui empêche parfois de reconnaitre les homologues par BLASTP car trop divergents
- La plus grande représentation des bactéries dans les bases de données

Cela nous montre l'importance d'étudier les virus environnementaux et surtout les virus géants avec un protocole adapté.

III. Les virus géants du pergélisol, article II

nature communications

Article

https://doi.org/10.1038/s41467-022-33633-x

Past and present giant viruses diversity explored through permafrost metagenomics

Received: 4 February 2022

Accepted: 27 September 2022

Published online: 07 October 2022

Check for updates

Sofia Rigou[1], Sébastien Santini[1], Chantal Abergel[1], Jean-Michel Claverie[1] & Matthieu Legendre[1]

Giant viruses are abundant in aquatic environments and ecologically important through the metabolic reprogramming of their hosts. Less is known about giant viruses from soil even though two of them, belonging to two different viral families, were reactivated from 30,000-y-old permafrost samples. This suggests an untapped diversity of *Nucleocytoviricota* in this environment. Through permafrost metagenomics we reveal a unique diversity pattern and a high heterogeneity in the abundance of giant viruses, representing up to 12% of the sum of sequence coverage in one sample. *Pithoviridae* and *Orpheoviridae*-like viruses were the most important contributors. A complete 1.6 Mb *Pithoviridae*-like circular genome was also assembled from a 42,000-y-old sample. The annotation of the permafrost viral sequences revealed a patchwork of predicted functions amidst a larger reservoir of genes of unknown functions. Finally, the phylogenetic reconstructions not only revealed gene transfers between cells and viruses, but also between viruses from different families.

Permafrost, soil remaining continuously frozen for at least 2 years, covers 15% of the Northern hemisphere[1] and gathers complex communities of living organisms and variable soil types. The microbial community of the surface cryosol is in some cases subject to freezing and thawing of the soil every year[2] whereas communities from deeper layers are trapped as the sediments are deposited (syngenetic permafrost) or as the sediment freezes (epigenetic permafrost). Pleistocene permafrost has been shown to harbor up to 5×10^7 cells per wet gram of soil, a fifth of which are alive[3]. The permafrost has thus the ability to preserve organisms for tens if not hundreds of thousands years and acts as a huge reservoir of ancient microorganisms. For instance, it has been shown that numerous bacteria isolated from permafrost samples remained viable[4,5], even potentially for up to 1.1 million years[6]. Even in low biomass-containing frozen environments such as glacier ice, metagenomics approaches have recently revealed hundreds of distinct bacterial genera[7]. Unicellular[8–10] and even multicellular[11,12] eukaryotes can also be preserved for thousands of years and be revived from such frozen environments.

Besides cellular organisms, metagenomics studies have revealed bacteriophages communities archived in surface[13] or deeper[7] glacier ice, the majority of which being taxonomically unassigned. Due to the high bacterial abundance[14], bacteriophages are expected to be the most abundant viruses in the permafrost. However, in the unfiltered size fraction, the eukaryotic viruses *Nucleocytoviricota* (formerly known as Nucleocytoplasmic large DNA viruses or NCLDVs) are also highly represented[14]. This phylum gathers large double-stranded DNA viruses such as *Pokkesviricetes* (*Poxviridae* and *Asfarviridae*) as well as all the known giant viruses (i.e., viruses visible by light microscopy): the *Megaviricetes* (*Phycodnaviridae*, *Mimiviridae,* and *Pimascovirales*). More importantly, among *Nucleocytoviricota*, the two giant viruses *Pithovirus sibericum* and *Mollivirus sibericum*, were reactivated from a 30,000-y-old permafrost sample on *Acanthamoeba castellanii*[15,16]. Taking into account the presence of numerous protists (in particular ameba) in permafrost[9], many more giant viruses probably exist in such environments.

Recently, several studies specifically targeting environmental viruses have started to grasp the diversity and gene content of the *Nucleocytoviricota*[17–19]. They seem to be widespread in aquatic environments. More specifically, *Mimiviridae* (in particular the proposed *Mesomimivirinae* sub-family[20]) and *Phycodnaviridae* are major contributors of the marine viromes all over the world, as revealed by thousands of metagenome-assembled viral genome (MAG)

[1]Aix–Marseille University, Centre National de la Recherche Scientifique, Information Génomique & Structurale (Unité Mixte de Recherche 7256), Institut de Microbiologie de la Méditerranée (FR3479), 13288 Marseille Cedex 9, France. e-mail: legendre@igs.cnrs-mrs.fr

sequences[17–19]. They have also been found active by metatranscriptomics at the surface layer of the ocean[21] and bloom-forming bays[22,23]. In addition to these two major groups, *Asfarviridae*, *Ascoviridae*, *Iridoviridae* and *Marseilleviridae* have been found active by marine metatranscriptomics[24]. Importantly, the genomes of giant viruses code for various auxiliary metabolic genes, making them capable of reprogramming their host's metabolism and thus, to potentially play an important role in global biogeochemical cycles[17,18,25].

The *Nucleocytoviricota* ecological functions and diversity in terrestrial samples are far less known, with the exception of *Klosneuvirinae* sequences recovered from forest soil samples[26] and of *Pithoviridae* sequences assembled from the Loki's castle deep sea sediments sequences[27]. The overwhelming proportion of *Nucleocytoviricota* metagenomic sequences of marine origin as compared to terrestrial is most likely due to the difficulty at revealing their hidden diversity in these environments[26]. Indeed, the high proportion of closely related strains in soil communities notoriously hampers sequence assembly, making soil metagenomic studies challenging[28,29].

Current giant viruses' metagenomic and metatranscriptomic studies rely on the detection of *Nucleocytoviricota* core genes[17,18,22,26,27]. However, among the handful of core genes, some of them are highly divergent or even completely absent from certain viral families. For instance, the Major Capsid Protein (MCP), often used as a marker gene to detect *Nucleocytoviricota* within metagenomic and metatranscriptomic assemblies[18,24], is absent from *Pandoraviridae*[30] and only present in a divergent form in *Pithoviridae*[15]. Thus, the probability to detect these types of non-icosahedral giant viruses is drastically lowered.

Although two distinct non-icosahedral giant viruses were initially isolated from permafrost samples[15,16], little is known about the diversity of *Nucleocytoviricota* in this type of environment. Here, we propose an analysis of these viruses from eleven permafrost samples ranging from the active layer up to 49,000-y-old sediment[14]. We show that the permafrost has a high viral diversity. Although the samples are very heterogeneous in *Nucleocytoviricota* content, they can reach up to an estimated relative abundance of 12% of the sequenced organisms (from sequence coverage). We found here that *Pithoviridae* and *Orpheoviridae*-like families followed by *Mimiviridae* are the main contributors of the giant virus diversity of the deep permafrost.

Results

Cryosol metagenomes assemblies

We gathered permafrost and surface cryosol raw metagenomic data from a previous study[14] on the three surface samples from Kamchatka (C-D-E, Supplementary Table 1) and eight deep samples from the Yukechi Alas area dated from 53 to over 49,000-y-old, four of which are syngenetic (Supplementary Table 1). Importantly, Cedratvirus kamchatka[31] and Mollivirus kamchatka[32] were isolated from the mentioned Kamchatka surface samples.

Previous analysis of this dataset[14] showed that prokaryotes are the most abundant (90% of the total coverage). Accordingly, the assembly of the reads (Supplementary Table 2) predominantly revealed bacterial contigs (mean = 94%, sd = 7%) according to the Lowest Common Ancestor (LCA) taxonomy based on BLASTP results against RefSeq. The samples with least bacterial contigs (N and R) still contain 80% of those, along with archaeal (10 and 7% respectively), unclassified (5%), viral (2%), and eucaryotic (1.3% and 1.6% respectively) contigs. Owing to the majority of bacterial contigs we reasoned that CheckM[33] could be applied to assess the overall validity of our assembly procedure of the complete dataset. This resulted in clean contigs with very few potential chimeras (0.004%) and no strain level chimera (Supplementary Fig. 1A). Next, as is custom in metagenomics studies, we performed a binning of the contigs to obtain less fragmented assemblies[34]. This revealed potential chimeras (Supplementary Fig. 1A). We thus chose not to consider bins as unique organisms but instead we used binning as a procedure to decrease complexity in our datasets. More precisely, the reads were first separated according to the bin they belonged to. Next, a second de novo assembly was made within each bin. This resulted in significantly longer scaffolds and a larger total assembly (Supplementary Table 2) while keeping contamination at a negligible level (on average 0.005% potential chimeras and again none at the strain level, Supplementary Fig. 1A). Thus, our method significantly gained in reliability by lowering the proportion of chimeras in comparison to conventional binning, while providing longer assembled sequences compared to standard assemblies.

To further validate this strategy, we applied the same assembly method on three complex mock communities generated by a previous study[35]. Aligning the reference genomes used in that study on the resulting assembled sequences revealed a similar pattern: a clean first assembly, a noisier binned assembly, and a clean final assembly (Supplementary Fig. 1B). The proportion of chimeras in the final scaffolds accounts for only 0.2%.

Discriminating *Nucleocytoviricota* in metagenomic samples

From the permafrost dataset we then sought to filter *Nucleocytoviricota* sequences. Our method is based on the detection of both *Nucleocytoviricota* genes (including the ones specific to the non-icosahedral *Pithoviridae* and *Pandoraviridae*) and cellular ones. We used a control metagenomic mimicking database containing reference *Nucleocytoviricota* genomes, cellular genomes randomly sampled from GenBank in addition to ameba and algae genomes (known hosts of *Nucleocytoviricota*) as well as ameba-hosted intracellular bacteria (*Babela massiliensis* and *Parachlamydia acanthamoebae*). Clearly the combination of the cellular and viral gene counts showed a very distinct pattern for *Nucleocytoviricota* compared to cellular genomic sequences (Fig. 1a). Using this control database, we computed the optimal parameters discriminating *Nucleocytoviricota* sequences (slope = 0.1, intercept = 1; Fig. 1), yielding high classification performance (sensitivity = 87.47% and specificity ≥99.53%; Supplementary Fig. 2). Taking proportions instead of viral and cellular ORFs counts did not yield better results (Supplementary Fig. 3). For comparison, we also tested the ViralRecall tool (35) that confirmed 1848 out of the 1973 (94%) scaffolds detected by our pipeline. Further controls for contamination in the *Nucleocytoviricota* dataset involved a search for ribosomal sequences, none of which were found. Manual functional annotation of all potential *Nucleocytoviricota* scaffolds allowed the identification of 7 scaffolds potentially belonging to intracellular bacteria, a phage and a nudivirus. All these sequences were removed. At the end, we identified 1966 *Nucleocytoviricota* scaffolds ranging from 10 kb up to 1.6 Mb in the permafrost dataset, corresponding to 1% of all scaffolds over 10 kb in size (Fig. 1b). Applying CheckM specifically fueled with viral HMM profiles made from low-copy NCVOGs (44) on this final sequence dataset resulted in virtually no contamination (mean = 0.0047%, s.d. = 0.027%) and strain heterogeneity (mean = 0.066%, s.d. = 1.8%).

As previously mentioned, *Nucleocytoviricota* metagenomic studies often rely on the MCP as a bait, making it hard, if not impossible, to catch some of the non-icosahedral viruses. By adding *Pithoviridae* and *Pandoraviridae* HMMs to the original profiles[18] and VOG's HMMs, we gained 5% (n = 110) more scaffolds that were mainly unclassified or from *Pithoviridae* and divergent *Pithoviridae* families (see further for phylogenies).

Heterogeneous *Nucleocytoviricota* abundance in cryosols

The permafrost samples were very heterogeneous in *Nucleocytoviricota* relative abundance (Fig. 2) and number of scaffolds, ranging from 2 to 721 scaffolds, found in samples O (core permafrost under a lake in Yedoma, frozen for 40,000 years) and R (core permafrost under a drained thermokarst lake, frozen for over 42,000 years), respectively. *Nucleocytoviricota* scaffolds corresponded to 12% of the R sample

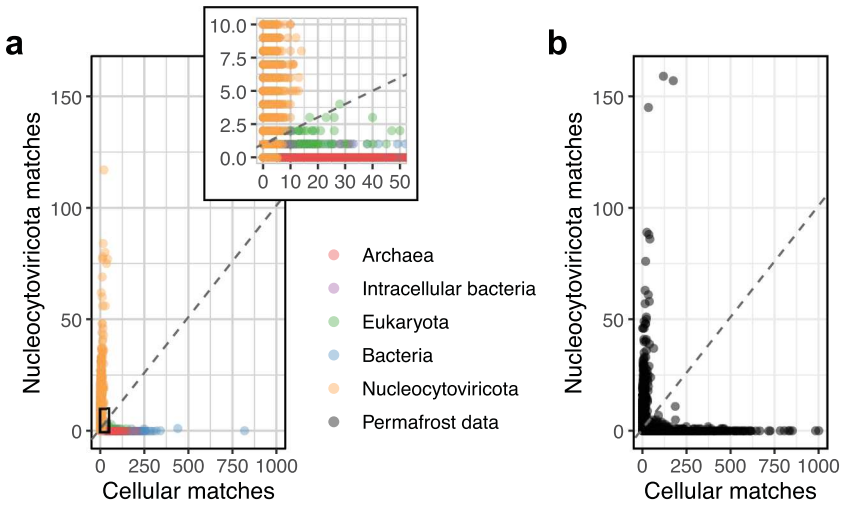


Fig. 1 | Viral scaffolds filtering. Each point corresponds to one scaffold. Viral matches (*y*-axis) were counted as the number of ORFs matching a *Nucleocytoviricota*-specific HMM at an *E* value of 10^{-10}. These HMMs come from a previous study[18] to which were added specific HMMs from the VOG database and HMMs constructed on *Pandoraviridae* and *Pithoviridae* genomes. Cellular matches (x-axis) are the number of DIAMOND BLASTP matches against the cellular RefSeq database with a threshold of 35% of sequence identity and an *E* value ≤ 10^{-5}. The dashed lines represent the chosen threshold excluding all point under or on the line. **a** Control dataset. The inset is a zoom of the bottom-left corner of the plot. For clarity, 1 bacterial point with over 1000 cellular matches and 1 viral match are not shown. **b** Permafrost data. For clarity, 5 points with over 1000 cellular matches are not shown. Source data are provided as a Source Data file.

sequence coverage (Fig. 2) and 17% of total reads mapped on scaffolds over 10 kb (4% of all raw reads). This sample was also the richest in eukaryotes with mostly Streptophyta (35%), Dikarya (14%), Platyhelminthes (9%), Eumycetozoa (8%) and Longamoebia (7%). Interestingly, amebas (Longamoebia) are on average 46.7 times more abundant in this sample than in the other ones (Supplementary Fig. 4A).

The relative proportion of giant viruses (Fig. 2) showed a strong correlation to the ones of Eukaryota. Precisely, Spearman correlation coefficients of $\rho = 0.72$ for the sum of coverages (two-sided correlation test p value = 0.017, Fig. 2) and $\rho = 0.83$ for the number of scaffolds (two-sided correlation test p value = 0.003) were observed. Such correlation could be explained by host-parasites dynamics. We therefore looked for potential co-occurrences of viral and eukaryotic families. Despite working with only 11 samples, we found significant associations (Supplementary Fig. 4B), including two *Pithoviridae*-like viruses with Entamoebidae. More surprisingly, we also found two other *Pithoviridae*-like associated with Hydrozoa. HGT between *Mimiviridae* and this eukaryotic class has already been observed[36]. Finally, two other *Pithoviridae*-like were also found associated with Cryptomonadaceae. Although these eukaryotes are not known to be infected with giant viruses, metagenomics co-occurrence analyses showed association between cryptophytes and *Mimiviridae*[22] as well as virophages[37].

Nucleocytoviricota scaffolds could also correspond to endogenized viruses in eukaryotes (GEVE), as previously shown in green algae[38]. This hypothesis is plausible as 57% (193 out of 338) of the GEVE pseudocontigs (see Methods) were captured by our *Nucleocytoviricota* detection method. To explore this possibility, we thus checked for endogenization signs in the viral scaffolds using ViralRecall[39] (example in Supplementary Fig. 5) but none was found. In addition, *Nucleocytoviricota* largely outnumber eukaryotes with a 4:1 *Nucleocytoviricota*/Eukaryota ratio in the sum of coverages (mean = 4.06, s.d. = 4.22) and number of scaffolds (mean = 4.40, s.d. = 3.34). Altogether, this suggests that most of the discovered permafrost *Nucleocytoviricota* scaffolds correspond to bona fide unintegrated viruses.

Exploration of the sequence diversity

To further investigate which viral families were present in the samples, we conducted a phylogenetic analysis based on 7 marker genes (Supplementary Data 1) and a curated database produced by a former study[40]. We excluded the transcription elongation factor TFIIS as its phylogeny breaks well-established clades (*Alphairidovirinae*, *Ascoviridae*, *Asfarviridae*, *Pimascovirales*, Supplementary Fig. 6). It should also be noted that the primase D5 revealed an unexpected grouping of the Cedratviruses with *Phycodnaviridae* instead of *Pithoviridae*, suggesting that this gene was acquired from an unknown source in Cedratviruses (Supplementary Fig. 6). We first classified permafrost scaffolds containing at least three of the seven marker genes to avoid split genomes in the tree. This resulted in 37 classified scaffolds (corresponding to 16.5% of the 72 Mb of total *Nucleocytoviricota* identified sequences) with 21 scaffolds within the *Pithoviridae* and *Orpheoviridae*-like clades, 8 in the *Megamimiviridae* clade and the rest associated to *Klosneuviridae*, *Phycodnaviridae* and one *Asfarviridae* (Fig. 3a).

However, filtering scaffolds with less than three marker genes only reveals the ones representing a substantial portion of the viral genome and thus probably under-estimate the true diversity of viral families. Indeed, counts derived from single markers (Fig. 3b) show that *Pithoviridae* and *Orpheoviridae*-like sequences might be particularly under-estimated as they lack the packaging ATPase and contain a highly divergent MCP. In addition, they contain a substantially lower fraction of duplicated marker genes than *Megamimivirinae* and *Klosneuvirinae* (Fig. 3b). We thus also performed a classification of all scaffolds containing at least one marker gene. This increased the taxonomically classified dataset to 369 *Nucleocytoviricota* scaffolds (40.1% of the *Nucleocytoviricota* sequences). Again, *Pithoviridae* and *Orpheoviridae*-like viral families were the most diverse, followed by *Mimiviridae* (Supplementary Fig. 7). In contrast, *Marseilleviridae*, *Alphairidovirinae*, *Betairidovirinae*, and *Ascoviridae* were completely absent in our samples. Interestingly, unclassified sequences do not encode for more ORFans (ORFs with no similar sequence in the public databases) than classified sequences (Supplementary Fig. 8A). This suggests that these sequences are not more divergent to known relatives than any other *Nucleocytoviricota* sequence but remained unclassified due to the lack of the marker genes.

We further confirmed the observed taxonomy pattern from individual marker genes phylogenies (Supplementary Fig. 9) and the best BLASTP matches of the unclassified sequences against the nr database (Supplementary Fig. 8B). Finally, an alternative phylogeny of the bins (instead of scaffolds) probably noisier but representing 85.4%

https://doi.org/10.1038/s41467-022-33633-x

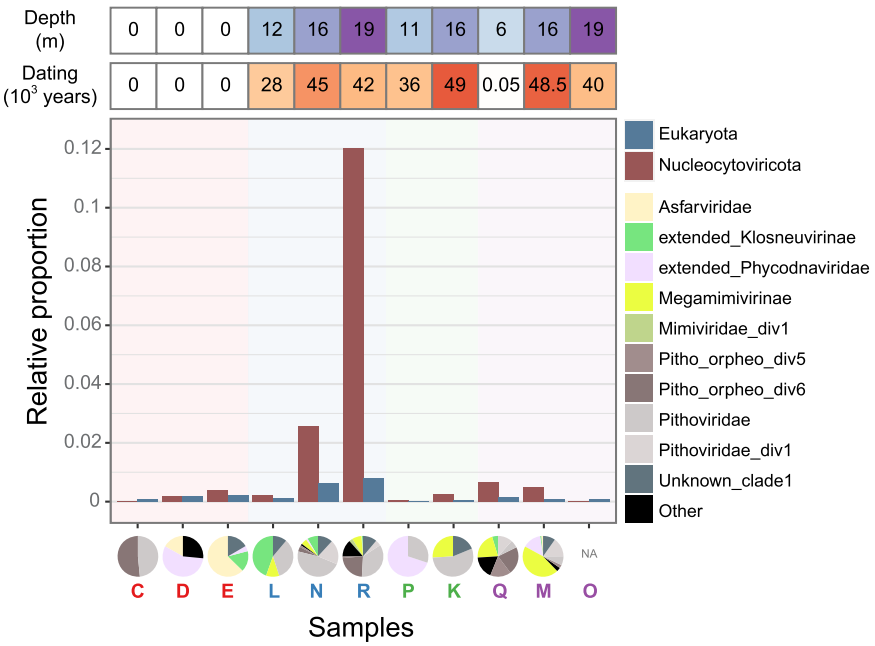


Fig. 2 | Relative abundance of *Nucleocytoviricota* and Eukaryota across samples. The relative abundance is calculated as the sum of scaffold coverages belonging to the given group divided by the total sample coverage among scaffolds ≥10 kb. Sample names in red are surface samples from Kamchatka while samples in blue, green and purple indicate that they come from three different forages in the Yukechi Alas area. The pie charts indicate the taxonomy of the *Nucleocytoviricota* in different samples (see further for phylogeny) whose abundance has also been estimated from the scaffold coverages. Only classified scaffolds were considered. Source data are provided as a Source Data file.

of the total *Nucleocytoviricota* sequences confirms the pattern (Supplementary Fig. 10). Altogether, these results clearly support *Pithoviridae* and *Orpheoviridae*-like as the most diverse families in our samples.

Most viruses are specific to the sample they were recovered from, in particular the ones from surface samples (Supplementary Fig. 11). Surprisingly, we also found viruses that were common to samples from close locations in Central Yakutia but from different ages (samples K, L, M, N, P, Q, and R; Supplementary Table 1). As the samples are unlikely contaminated[14], this indicates that part of the viral community was maintained over time.

Enrichment of *Pithoviridae* and *Orpheoviridae*-like genomes in the Permafrost

Not only *Pithoviridae* were unexpectedly diverse (Fig. 3, Supplementary Figs. 7 and 10), they were also the most abundant *Nucleocytoviricota* according to their normalized coverage (Fig. 2). *Pithoviridae*/*Orpheoviridae*-like families appear in all samples and are particularly abundant in samples R and N (Fig. 2). The single most covered (i.e., abundant) sequences in five samples (C, N, R, K, and Q) come from these, and from Extended_phycodnaviridae, *Megamimivirinae* and *Klosneuvirinae* in the other samples.

The *Pithoviridae* diversity and abundance observed in the Siberian permafrost (in particular in samples R and N) could either highlight the enrichment of this viral family in this environment or represent the improvement in our method for the detection of non-icosahedral viruses. To compare the diversity found in the presented samples to other soil environments, we applied the same detection method to 1835 terrestrial datasets collected from the JGI IMG/M database[41]. The vast majority of these terrestrial samples exhibited no *Nucleocytoviricota* sequences and few contigs over 10 kb in general, probably due to the difficulty at assembling sequence data from these complex environments. Comparatively, the diversity of *Pithoviridae* and *Orpheoviridae* observed in the cryosol samples is unique as they were significantly enriched in these viruses, followed by forest soil (Supplementary Fig. 12). Noteworthy, Pandoravirus-like sequences were found in sand and a 900 kb contig grouping next to *Pandoraviridae* and *Molliviridae* in peat permafrost samples.

Large viral genome fragments from the deep permafrost

Although our strategy to exclude conventional binning was primarily designed to capture high confidence MAGs at the price of completeness, we were still able to recover large *Nucleocytoviricota* genomes in single scaffolds with no apparent chimera (see "Methods"). Eight of them, assembled from 16 m to 19 m deep permafrost samples (R, N and M, Supplementary Table 1) dating from 42,000 to 49,000 years, reached over 500 kb (Fig. 4). The largest one of 1.6 Mb, referred to as "Hydrivirus", is likely complete as it was successfully circularized. Although these large scaffolds are deeply sequenced (with an average coverage in between 14 and 72), they do not belong to the most abundant viruses in their samples (the highest coverages are of 53, 181, and 1572 in samples M, N, R respectively).

These MAGs vary in divergence from known genomes, having from 22% up to 72% of ORFans for Unknown Permafrost:M_b2437_k1 (Fig. 4). As is common for newly discovered giant viruses, their genomes also match cellular genes from all domains of life (with very few Archaea). The four largest scaffolds were classified within *Pithoviridae*/*Orpheoviridae*-like families. Two are putative *Megamimivirinae* (Mimivirus Permafrost:R_b548_k1 and Mimivirus Permafrost:R_b2349_k1) and one is a putative *Klosneuvirinae*. Finally, the Unknown Permafrost:M_b2437_k1 scaffold is placed near the root of the tree (Fig. 3) and its evenly distributed viral best BLASTP matches have no specific family standing out (Fig. 4). Taking its scaffold phylogeny (Fig. 3 and Supplementary Fig. 7) together with its high ORFan content suggests that it belongs to a *Nucleocytoviricota* viral family with no isolate so far.

The complete 1.6 Mb Hydrivirus genome reaches a size similar to the isolated Orpheovirus[42]. The other 715 to 855 kb scaffolds are slightly larger than isolated *Pithoviridae* (ca. 600 kb)[15,43,44]. However, they were not circularized as expected for a *Pithoviridae* genome structure[15] and are thus potentially even larger. Still, in the four of them, most of the core genes are present (Supplementary Data 1).

https://doi.org/10.1038/s41467-022-33633-x

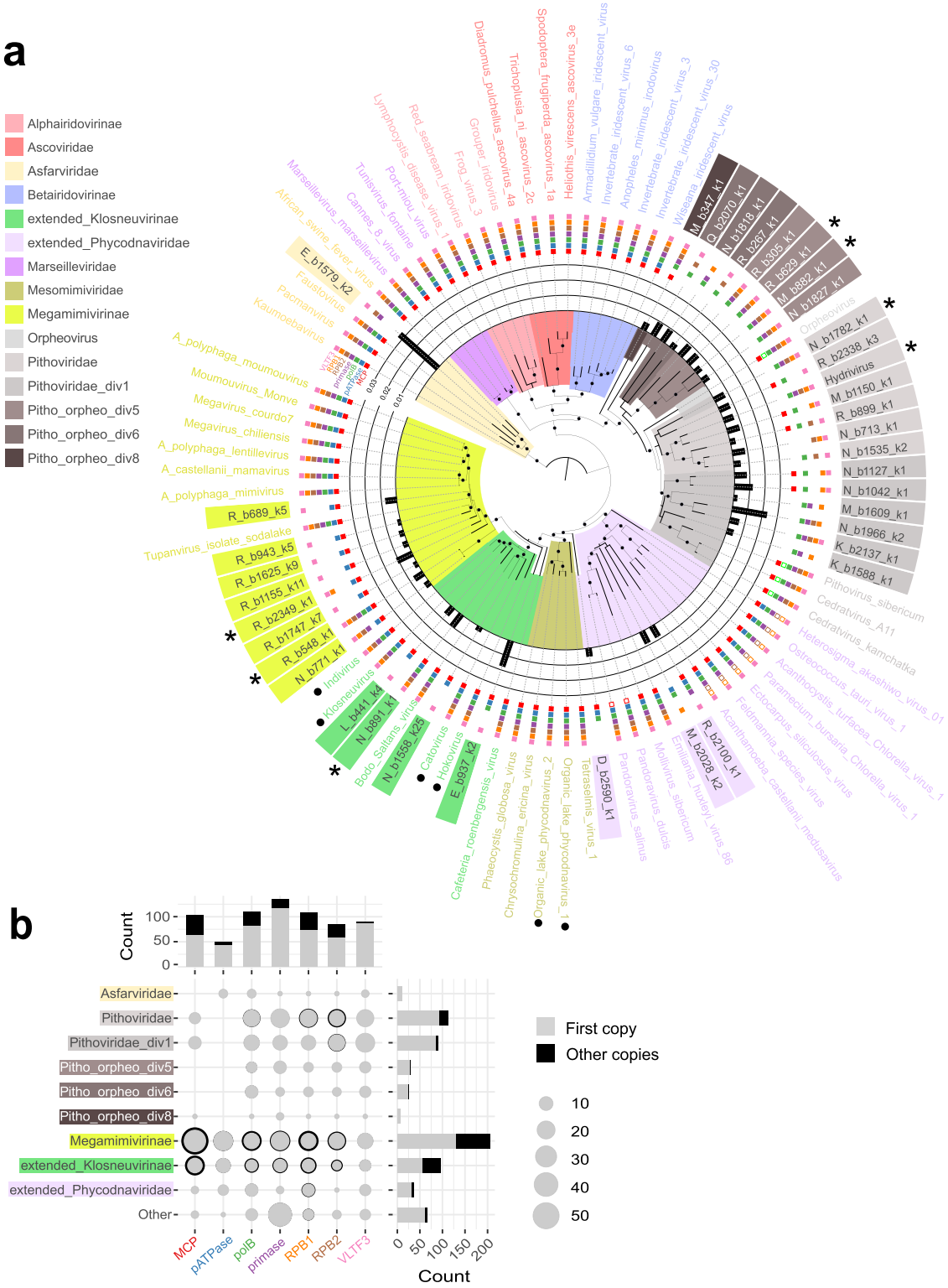


Fig. 3 | *Nucleocytoviricota* diversity in all samples. a Consensus of 1000 bootstrapped trees calculated by IQ-TREE through a partitioned analysis on 7 marker genes. The tree was performed on permafrost sequences (colored background labels) with more than three marker genes and on reference sequences (white background). The stars indicate the large (>500 kb) scaffolds identified in this study. The models used were the following: LG + F + I + G4 for the packaging ATPase, the RNA polymerase subunits, VT + F + I + G4 for the DNA polymerase, VT + F + G4 for VLTF3 and LG + F + G4 for the MCP. Black dots represent branch bootstrap support ≥ 90%. One should note that reference genomes coming from bins of previous metagenomic studies (marked with a black dot) are less reliable than the genomes of isolated viruses. The colored clades were manually created to be monophyletic. The marker genes used for this phylogeny are indicated as colored squares. Empty squares correspond to marker genes absent from the reference genomes. Black bars show the relative mean coverage of the scaffolds (%). The Extended_phycodnaviridae group includes *Pandoraviridae* and Mollivirus. The Extended_klosneuvirinae group includes the Cafeteria roenbergensis virus. **b** Total marker gene count associated to the taxonomy of scaffolds with at least one marker gene. Total counts of each viral clade and each marker gene are shown as barplots on the right and top panels, respectively. Source data are provided as a Source Data file.

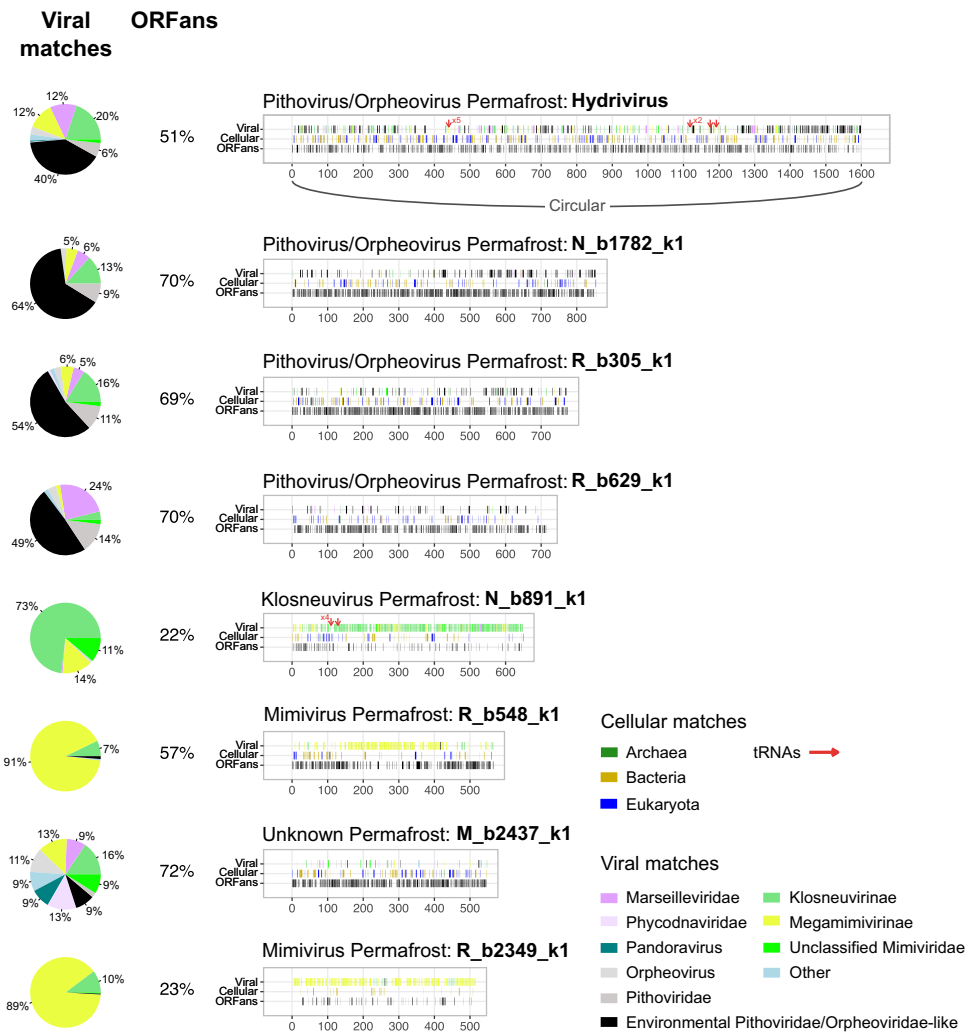


Fig. 4 | Gene content of the large genomes recovered from ancient permafrost samples. For each genome, the position of ORFans (ORFs with no match in the nr database), cellular and viral matches are recorded along the genome. The positions of tRNAs are also shown as red arrows. The pie charts present the proportion and taxonomy of viral matches with slices ≥5% labeled. Groups that match less than 5% of the Unknown Permafrost:M_b2437_k1 scaffold were gathered in "other" except for *Pithoviridae*. The environmental *Pithoviridae*/*Orpheoviridae*-like category contains metagenomic sequences from Bäckström et al.[26,27]. The Hydrivirus genome was circularized. Source data are provided as a Source Data file.

Furthermore, except for Pithovirus/Orpheovirus Permafrost:R_b629_k1, all the *Pithoviridae*-like large genomes and Klosneuvirus Permafrost:N_b891_k1 have a near complete base excision repair system.

Functions encoded in the permafrost *Nucleocytoviricota* sequences

To get insight into the functions encoded by the permafrost *Nucleocytoviricota* we manually annotated a total of 64,648 viral ORFs over 50 amino acids that were assigned to functional categories. The distribution follows the one of *Nucleocytoviricota* references (Supplementary Fig. 13), with most of the predicted proteins (81%) being of unknown function (as compared to 64% in reference genomes, Supplementary Fig. 13). We searched for significantly enriched Pfam and Gene Ontology annotations in the permafrost viral datasets compared to references but found none after false discovery p-value correction apart from a couple of core function (Supplementary Data 2). We also did not find specific functional enrichment when comparing samples to each other within the same viral families (Supplementary Data 3). Likewise, when mixing all viral families together, ecological parameters do not discriminate samples based on Pfam annotations (Supplementary Fig. 14). Altogether, this indicates that viral genome content and ecological parameters are not directly correlated or, more likely, that the high proportion of genes with unknown functions and the limited number of samples prevent this from being revealed at this time.

As expected from their reference counterpart, permafrost *Nucleocytoviricota* encode auxiliary metabolic genes that are scattered within the different viral families (Supplementary Figs. 15 and 16). In addition, they encode functions not previously observed, such as ATP synthases subunit F (in the N_b713_k2 Pithoviridae_div1 sequence), as well as truncated hemoglobins in 3 permafrost *Pithoviridae*-like (R_b2567_k1, M_b1150_k2 and N_b1127_k2) and in an Extended_phycodnaviridae sequence (M_b2028).

Looking at highly shared functions (i.e., present in most families) among the reference genomes and permafrost MAGs, we identified the known core genes (Fig. 5) with the exception of the mRNA capping enzyme, absent from the *Iridoviridae*/*Ascoviridae* clade. The patatin phospholipase, suspected to be conserved among *Nucleocytoviricota*[45], is confirmed as a core gene, only absent from *Alphairidoviridae* (Fig. 5). Conversely the A32-like packaging ATPase presumably encoded by a "core" gene in large DNA viruses is no longer a universal *Nucleocytoviricota* marker gene, as it is not only lacking from the reference *Pithoviridae* genomes[46] but also absent from all

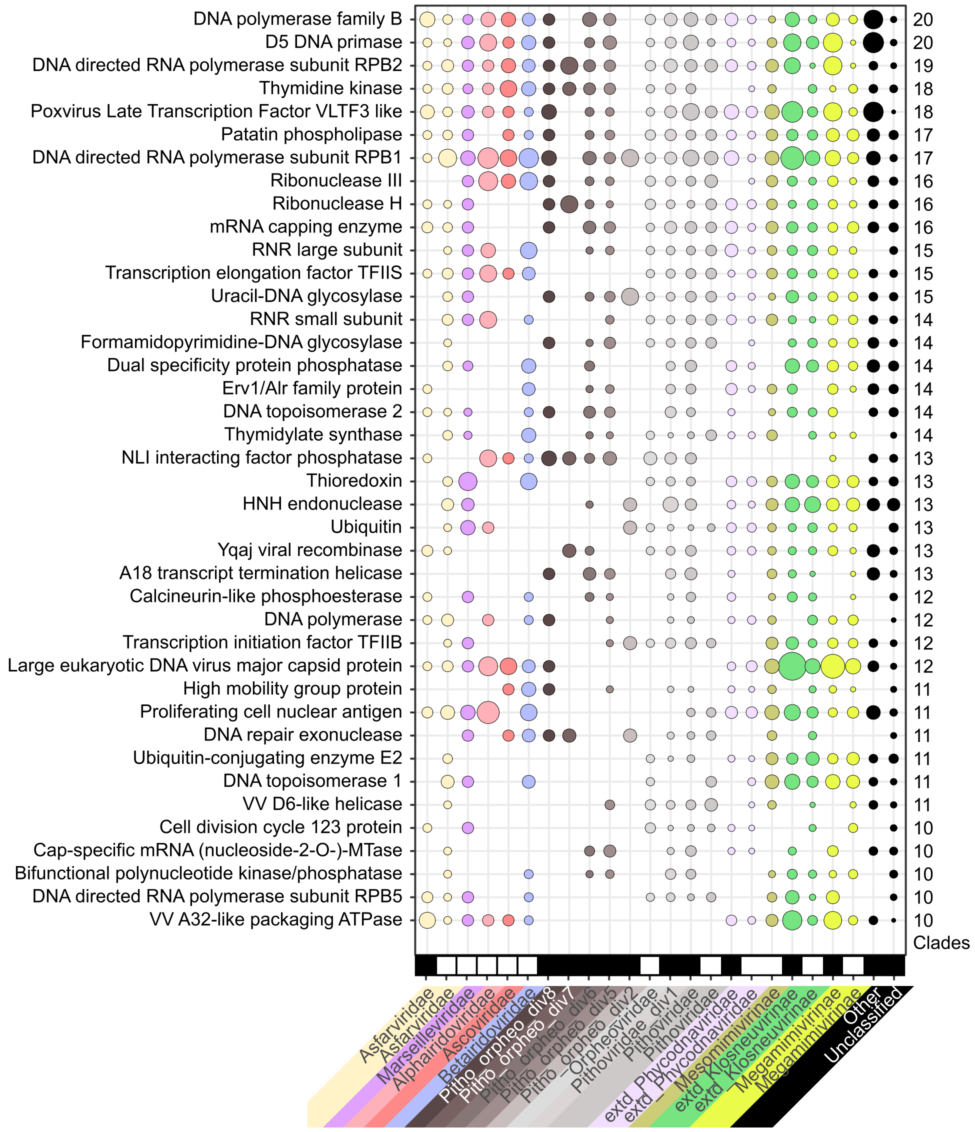


Fig. 5 | Comparison of most shared functions among metagenomic and reference *Nucleocytoviricota* families. Functions were selected among the annotations found in at least 10 clades. Metagenomic sequences are marked as black rectangles at the bottom of the plot while blank spaces correspond to reference genomes. Groups with less than 250 ORFs were marked as "Other". The size of the bubbles represents the normalized ORFs counts (i.e., ORF counts/total number of ORFs in the group). The right-most column indicates the number of distinct clades having the function. The lines are sorted according to this value. Source data are provided as a Source Data file.

clades ranging from Pitho-orpheo_div8 to *Pithoviridae* (Fig. 5). Overall, our analysis highlights a patchwork of functions encoded by these viruses (Supplementary Figs. 15 and 16).

Virally-encoded translation-related genes

Virally-encoded translation-related genes are a landmark of giant viruses. They were found in a large spectrum of viral families and might give clues on their evolution and interaction with cellular organisms. We thus specifically analyzed the translation-related genes in the permafrost data and found 20 different types of virally-encoded aminoacyl-tRNA synthetases (aaRSs). As previously observed in other *Klosneuvirinae*[47], the Klosneuvirus Permafrost:N_b891_k large genome fragment (Fig. 4) encodes an expanded translation-related gene repertoire (10 translation initiation factors, 4 translation elongation factors, a translation termination factor, 11 different aaRSs and 5 tRNAs clustered together). More surprisingly, ten different types of aaRSs were also found in the Pithoviridae_div1 clade, including 7 different ones in the Hydrivirus genome (Fig. 4). This virus also encodes 9 tRNA, 3 translation initiation and elongation factors, and a translation termination factor.

We then investigated the phylogeny of the different types of aaRSs found in our datasets that revealed entangled evolutionary pathways between viruses and cellular organisms (Supplementary Figs. 17–19). In most cases, the viral aaRSs were likely acquired by HGT from Eukaryotes (tryptophan, leucine, glutamine, threonine, methionine, isoleucine, arginine, aspartate, serine and phenylalanine) (Supplementary Figs. 17 and 19). In rare cases, we detected a possible HGT from an Archaea to a virus as for the glycine- and tyrosine-tRNA synthetases (Supplementary Fig. 18). Genes have also passed from Bacteria to *Nucleocytoviricota*, as for the glycine-tRNA synthetase of Hydrivirus and the valine-tRNA synthetase of a permafrost *Megamimivirinae*. For the latter, the bacterial sources were Rickettsiales, which are endosymbionts of ameba[48], and thus probably share the same host. The source of the tryptophan-tRNA synthetase in Hydrivirus is less clear, but a duplication event occurred probably at the same locus right after the gene was acquired (Supplementary Fig. 17).

While the vast majority of *Nucleocytoviricota* genes have no identifiable homologs, the ones with cellular homologs usually deeply branch in the phylogenetic trees[17,49], in accordance with their suspected ancient origin[40,50]. We found here several viral aaRSs that belong to divergent families tightly clustered together within the cellular homologs (Supplementary Fig. 19). So not only viral aaRSs are of cellular origin, spanning all domains of life, they were also probably exchanged between viruses of different families.

Discussion

Recent large-scale metagenomic data analyses strikingly revealed that *Nucleocytoviricota* are widespread in various environments[17,18,26,27]. Our analysis of cryosol samples confirmed this ubiquity. Nevertheless, we highlighted an important heterogeneity in *Nucleocytoviricota* proportions across the samples, in agreement with the already observed heterogeneity at various scales (domain, phylum, class, order and functional annotation) for all domains of life[14]. This heterogeneity is probably the testimony of the absence of mixing between layers but also of a spatially heterogeneous microbiome. It has been pointed out that the bacterial community of agricultural soil changes at a centimeter-level[51]. Thus, the heterogeneity we observe might translate a sampling bias although probably attenuated by the large amount of soil from each sample (20 g) used for DNA extraction. In any case, such heterogeneity includes eukaryotes which likely strongly influences the abundance of *Nucleocytoviricota*. The co-occurrence analysis of *Nucleocytoviricota* and eukaryotes performed in this study linked *Pithoviridae*-like clades to ameba. More surprisingly others were associated with Cryptomonadaceae and Hydrozoa. This widens the possible host range of *Pithoviridae* in the same way *Mimiviridae* infect various distant clades[52]. However, co-occurrence might translate indirect correlation and not direct virus-host interactions.

Importantly, our samples are among the most enriched in *Nucleocytoviricota*, reaching up to 12% of the estimated abundance of sequenced organisms. Moreover, the relative DNA sequence coverage (Fig. 2) suggests that they outnumber their hosts, in the same way bacteriophages often outnumber bacteria in the ocean[52,53]. This high abundance is the result of a high diversity in the samples, as it does not come from a single virus that would be at the origin of all sequenced reads since the individual maximum relative coverage never exceeds 0.3%. Furthermore, by taking advantage of the permafrost's ability to preserve ancient organisms, we showed that some *Nucleocytoviricota* strains have been present in the surface community for a long time (Supplementary Fig. 11). Considering only syngenetic permafrost samples, we found *Nucleocytoviricota* shared in samples of up to 13,000 years difference. This indicates that they probably are important players of this particular area of central Yakutia.

The *Nucleocytoviricota* diversity explored in this study revealed large genomic sequences of unknown families, such as the Permafrost:M_b2437_k1 scaffold (Fig. 4). In addition, we identified many divergent *Pithoviridae*-like sequences which constitute new clades within the *Pimascovirales*. In contrast, *Megamimivirinae*, *Klosneuvirinae* and *Mesomimivirinae* were the groups with the least ORFans within the permafrost sequences. These groups are thus better sampled than all other viral families found in this study (Supplementary Fig. 8A). Overall, the high ORFan content in our dataset probably explains the paucity of functions significantly enriched in permafrost samples compared to reference sequences. In addition, the patchwork-like pattern of *Nucleocytoviricota* functions might also blur the statistical signal.

Importantly, the 1.6 Mb Hydrivirus genome recovered by our method is complete. So, together with Pandoraviruses[30], Orpheoviruses[42], Klosneuviruses[47], and Mimiviruses[54], Hydrivirus is another example of a viral genome largely over 1 Mb. The nature of the evolutionary forces pushing some viruses to retain or acquire so many genes remains a matter of debate[55–58]. Horizontal gene transfers from cellular hosts is hypothesized by some authors to account for their large gene content[47,59]. We indeed found examples of cellular genes gained by HGT in this study (Supplementary Figs. 17–19) but this only accounts for a small proportion of their gene content, with the vast majority having no identifiable cellular homolog. Gene duplication, on the other hand, a well-known source of functional innovation since the pioneering work of Susumu Ohno[60], may contribute to the genome inflation of giant viruses[49,61]. Another possible source of genetic innovation is the de novo gene creation from intergenic regions[49,62]. The present work expanded the *Nucleocytoviricota* families' pangenomes, in particular the *Pithoviridae*-like with an overwhelming proportion of ORFans. Part of these genes might originate from de novo gene creation, a hypothesis that remains to be further tested.

The functional annotation performed in this work highlights the paucity of functions strictly shared between *Nucleocytoviricota*. This includes proteins thought to be central for viral replication/transmission, like the A32 Packaging ATPase[46] which is absent from the entire *Pithoviridae*-like clade (Fig. 5). Likewise, the MCP is not encoded in the *Pandoraviridae* genomes[30]. Our work also highlights a patchwork of functions and independent cases of HGT from Eukaryotes to viruses but also between viruses belonging to different families (Supplementary Figs. 17–19). This is probably the testimony of coinfections, as members of the *Marseilleviridae, Mimiviridae, Pithoviridae, Pandoraviridae* and *Molliviridae* families can infect the same host. Although endogenization may blur the counts, it was estimated by single-cell sequencing that as much as 37% of cells carry 2 or more viruses[63], thus promoting gene exchanges between viruses. In line with this hypothesis, it was recently showed that DNA methylation, widespread in giant viruses, is mediated by methyltransferases and Restriction-Modification systems that are frequently horizontally exchanged between viruses from different families[31].

The functional patchwork, the gene exchanges between viruses of different families and the very few shared genes may challenge the monophyly of the recently established *Nucleocytoviricota* phylum by the International Committee on Taxonomy of Viruses (ICTV)[64]. Except for the DNA primase of Cedratviruses, our trees of seven marker genes would indeed indicate a shared ancestry of the different *Nucleocytoviricota* families analyzed in this work (Supplementary Fig. 6). However, when cellular genes are integrated to the phylogenetic trees, only three of the five most shared genes strictly support the monophyly of the *Nucleocytoviricota*[65]. These are the viral late transcription factor 3, the Holliday junction resolvase and the A32 packaging ATPase genes. The latter has also been shown to be exchanged between *Mimiviridae* and Yaravirus, an Acanthamoeba infecting virus that does not belong to the phylum[65,66]. The other core genes such as the DNA polymerase is separated by several cellular clades between *Pokkesviricetes* and *Megaviricetes*[67]. Likewise the two largest subunits of the RNA polymerase of *Asfarviridae* and *Mimiviridae* have a different history than the other *Nucleocytoviricota*[40]. These examples question the consistency of the phylum.

The objective of this study was to assess the diversity of large DNA viruses in permafrost. Our analyses revealed an unexpected number of unknown viral sub-groups and clades, some among of the previously established families of the *Nucleocytoviricota* phylum. The phylogenetic diversity recovered from the ancient permafrost translated into an intricate functional patchwork amidst a majority of anonymous genes of unknown functions.

Methods

No approval by board/committee and institution was required for this study.

Data preparation

Illumina sequencing reads from all samples (Supplementary Table 1) were assembled into contigs using MEGAHIT (v1.1.3)[68] and then binned

using Metabat2[69] (v2.15) with a minimal contig length of 1500 and bin length of 10,000. Reads corresponding to each contig were retrieved and gathered from their respective bins using an in-house script. The read subsets were then reassembled using SPAdes[70] (v3.14) in default mode or with the "–meta" option. Reads were mapped on the resulting scaffolds ≥10 kb using Bowtie 2[71] (v2.3.4.1) with the "–very-sensitive" option and filtered with SAMtools (-q 3 option). Reads ≤30 nucleotides were discarded. Scaffold relative abundance was estimated as the mean scaffold coverage divided by the total sample coverage. Bins, contigs and scaffolds were verified with CheckM[33] (v1.1.2) using the lineage workflow. CheckM was also applied on a custom set of HMMs made from the NCVOGs database[45] using the "analyze" and "qa" tools. NCVOGs with 1.1 copies or less in average were used to construct the HMM profiles for a low-copy NCVOG database.

The validation of the method was performed on three datasets from a previous study[35]. We used high complexity mock communities with strain diversity within each species (ani100_cHIGH_stTrue_r0, ani100_cHIGH_stTrue_r1 and ani100_cHIGH_stTrue_r2) on which we applied the same assembly procedure (see previous paragraph). We then aligned the resulting contigs and scaffolds to the corresponding reference genomes with BLASTN (from BLAST + v2.8.1, options -evalue 1e-10 -perc_identity 99)[72]. Then matches ≥99.99% of identity were cut if overlapping with previous better matches and kept if they were ≥500 nucleotides using an in-house script. Bins with only one contig were not considered to assess the level of chimerism of bins or of the second assembly. With these data we assessed the proportion of chimeras (a contig, bin or scaffold matching different genomes) at each assembly step.

Control database preparation

Reference *Nucleocytoviricota* were chosen following a former phylogenetic study[40]. The corresponding genomes were gathered from the NCBI repository. Lausannevirus, Melbournevirus, Ambystoma tigrinum virus, Infectious spleen and kidney necrosis virus, Invertebrate iridovirus 22, Invertebrate iridovirus 25 and Singapore grouper iridovirus were removed to avoid an overrepresentation of their families. We added the genomes of *A. castellanii* medusavirus (AP018495.1), Bodo saltans virus (MF782455.1), Cedratvirus kamchatka (MN873693.1) and Tetraselmis virus 1 (KY322437.1). Genomes from Archaea, Eukaryota, and Bacteria (Supplementary Data 4) were retrieved from GenBank. For each genome, non-overlapping sequences were cut with an in-house script following a distribution similar to our dataset to simulate metagenomic contigs. Genes were then predicted by GeneMark (v3.36)[73] using the metagenomic model. For the *Nucleocytoviricota* phylogeny, core genes previously identified[40] were used in addition to the ones found by PSI-BLAST[74]. We also added Amsacta moorei entomopoxvirus (AF250284.1), Variola virus (NC_001611.1) and Cyprinid herpesvirus 2 (MN201961.1) as outgroup.

Nucleocytoviricota specific profiles databases

The database constructed by Schulz et al.[18] was completed with specific signatures of *Pithoviridae* using the genomes of Cedratvirus A11[44], Cedratvirus kamchatka[31], Cedratvirus lausannensis[75], Cedratvirus zaza[76], Brazilian cedratvirus[76], Pithovirus massiliensis[43], Pithovirus sibericum[15], Orpheovirus[42], all the metagenomic *Pithoviridae* (with the exception of Pithovirus LCPAC101) released from one study of Loki's Castle hydrothermal vents[27], the divergent *Orpheoviridae*/*Pithoviridae* SRX247688.42[17], the GVMAG-S-1056828-40[18] and other Cedratvirus/Pithovirus sequences (Supplementary Data 5). For *Pandoraviridae* we gathered sequences from Pandoravirus braziliensis[77], Pandoravirus celtis[62], Pandoravirus dulcis[30], Pandoravirus inopinatum[78], Pandoravirus macleodensis[49], Pandoravirus neocaledonia[49], Pandoravirus pampulha[77], Pandoravirus quercus[62], Pandoravirus salinus[30], Mollivirus kamchatka[32] and Mollivirus sibericum[16]. The ORFs were then predicted using GeneMark (v4.32) with the "–virus" option and ORFs ≥50 amino acids were kept. Orthogroups were calculated with OrthoFinder[79] and HMM profiles were built using the Hmmer suite[80] (v3.2.1) for each one. HMMs were further aligned to the RefSeq protein database (from March 2020) using the same suite. Only HMMs specific to *Pithoviridae*, *Orpheoviridae*, *Pandoraviridae* or Molliviruses with *E* value ≤ 10^{-10} were kept to complete the database. To these were added *Nucleocytoviricota*-specific VOG orthogroups (https://vogdb.org/).

Retrieving viral sequences

The *Nucleocytoviricota*-specific profile database was searched against the control and permafrost ORFs using hmmsearch (with *E* value <10^{-10}). To check for cellular signatures, all the ORFs were aligned to the RefSeq protein database using DIAMOND BLASTP (v0.9.31.132) with the "–taxonlist 2,2759,2157" option and hits ≥ 35% sequence identity and *E* value <10^{-5}. On the control metagenomic simulated dataset, the number of false positives and false negatives were assessed according to the cellular and viral matches for each group (*Nucleocytoviricota*, Archaea, Bacteria, Eukaryota). We set the threshold at less than 1% of false eukaryotic positives. The same threshold was applied to the permafrost data to retrieve viral scaffolds. For comparison, we also tested ViralRecall (v2.0) with the "–contiglevel" option, contigs with a score >0 were considered as viral.

Functional annotation

All the ORFs ≥ 50 amino acids were queried against the nr database (from June 2020) using BLASTP, the VOG database using hmmsearch, the Pfam database using InterProScan (v.5.39-77) and against EggNOG[81] using the online version of Emapper-1.03. For all, the *E* value threshold was set to 10^{-5}. Functional annotations of each predicted protein were defined manually, first based on the matching domains annotations, then by considering the full sequence alignments (BLAST, EggNOG and VOG). EggNOG categories were also set manually for each gene. When existing, the functional annotations of reference viral genomes (see control database preparation) were retrieved from GenBank. Grouper iridovirus, Heliothis virescens ascovirus 3e and Invertebrate iridescent virus 6 were manually reannotated using the same protocol as for the permafrost ORFs.

Functional enrichment analysis

The Pfam and GO term annotations were retrieved from the InterProScan output for statistical analysis. Each Pfam annotation was compared either to the references or between samples within taxonomical groups using fisher exact tests to search for enriched functions. The p-values were corrected for multiple testing using the Benjamini & Hochberg FDR correction. Biological Processes GO terms were analyzed using the topGO package with the "weight" algorithm. Samples were also compared to each other based on viral Pfam annotations with all viral families together applying Bray-Curtis dissimilarity for clustering. We used the nonpareil diversity of the complete sequence data of each sample computed in Rigou et al.[14]. Lastly, in order to search for complete or near complete functional pathways in the large genome fragments recovered we screened them for KEGG annotations using BlastKOALA[82].

Contamination control

The functional annotation step helped to remove non-*Nucleocytoviricota* scaffolds based on the presence of typical viral/phage genes or with ORFs consistently matching cellular organisms. The scaffolds were checked for the presence of ribosomes using Barrnap (v0.9)[83]. Finally, we checked for possible GEVEs (Giant Endogenous Viral Elements) in our curated scaffolds. We made pseudo-contigs from the GEVEs identified by Moniruzzaman et al.[38] and applied our method on them. As 57% (193 out of 338) of the GEVEs peudo-contigs were caught, we proceeded to check for endogenization signs in our permafrost scaffolds. This was done by plotting the domain of the best BLASTP

hits as well as the VOG matches for each scaffold with the results of the ViralRecall (v2.0) rolling score using default parameters[39]. Scaffolds with at least one region with a negative ViralRecall score were visually inspected.

Large genomes assembly verification and circularization

The eight largest MAGs (≥500 kb) were scrutinized for possible chimeric assemblies. First, we checked visually that there was a single trend in the coverage along the scaffolds in log scale. Then we used the Integrative Genome Viewer[84] to scrutinize the positions where the coverage dropped under 3× (mainly due to ambiguous bases added during scaffolding). In each case, read pairs overlapped the low coverage intervals. For circularization, we created a model contig concatenating both ends of the MAG, mapped the reads using Bowtie 2 and visually checked the uniformity of the coverage at the junctions using the Integrative Genome Viewer.

Abundance estimation and mapping

The relative mean coverage of the scaffolds calculated from the mapping data described above were used as estimators of the scaffold abundance in the sample. The taxonomy of all non-viral scaffolds was retrieved using the same Lowest Common Ancestor methodology than previously published for the same dataset[14]. For co-occurrence analysis, the abundance of pairs of eukaryotes and viruses present in at least two samples were compared by spearman correlations. The resulting *p* values were corrected using Benjamini & Hochberg FDR correction.

For in-between sample comparisons of viruses, read longer than 30 nucleotides were mapped to viral sequences from all samples with Bowtie 2 and a minimum quality filter of 30 was applied with SAMtools. Then, only scaffold with more than 10 kb covered was considered.

Phylogenetic analysis

For the selected marker genes, individual gene trees were built from reference genomes only. Multiple alignments were performed using MAFFT (v7.407)[85], removal of divergent regions with ClipKIT[86] and models estimations[87] and tree inference using IQ-TREE (v1.6.12)[88] (options "-bb 1000"[89], "-bi 100" and "-m MFP"). The best model was VT + F + R4 for the TFIIS tree, LG + F + G4 for the MCP and LG + F + R5 for all the other marker genes. A global tree was calculated by a partitioned analysis[90] to include genomes with missing data. In addition, individual gene trees were computed with the same options and models.

To identify the marker genes in the permafrost data, PSI-BLAST was used to align reference marker genes to the viral ORFs (initial *E* value $\leq 10^{-5}$). Next, in order to reduce the number of paralogs of the marker genes, we defined a second stringent *E* value threshold the following way: *E* values of all second matches for scaffolds with multiple copies were sorted in ascending order, then the stringent threshold was defined based on the first quartile (Supplementary Table 3). Finally, only the best match per scaffold was kept for phylogenetic reconstruction if it was better than the stringent threshold for this gene.

The 7 marker genes were aligned using PASTA[91], clipped with ClipKIT and concatenated by Catsequences[92]. The global tree was then inferred by IQ-TREE with ultrafast bootstraps options "-bb 1000" "-bi 200" and "-spp -m MFP" that calculates the best model per marker gene. Tree visualization was handled using FigTree (http://tree.bio.ed.ac.uk/software/FigTree/) and the Itol web server[93].

Terrestrial *Nucleocytoviricota* distribution

We downloaded 1835 terrestrial assemblies from the JGI IMG/M[41] database (Supplementary Data 6) from March 2021, of which we kept only contigs ≥10 kb reducing the analysis to 1502 datasets. The ORFs were predicted using GeneMark using the metagenomic model as previously. *Nucleocytoviricota* sequences were extracted as described above (see Retrieving viral sequences). The same method than previously described (see Phylogenetic analysis) was applied to search for marker genes for phylogeny. Reference and metagenomic marker genes were aligned using MAFFT with the "—auto" option. Amsacta moorei entomopoxvirus, Variola virus and Cyprinid herpesvirus 2 were included in the analysis. The alignments were clipped with ClipKIT and concatenated for a partitioned analysis. Empirical models for each partition were inferred by Modelestimator[94]. Finally, the trees were computed using IQ-TREE (with options -bb 1000 -bi 200).

Phylogenetic analyses of translation-related genes

A dataset of proteins was built using a combination of *Nucleocytoviricota* ORFs, corresponding BLAST matched proteins from the nr database and reference proteins from specific databases. The latter includes UniProt reviewed proteins of domains IPR001412 (class I aminoacyl-tRNA synthetases) and IPR006195 (class II aminoacyl-tRNA synthetases). The multiple alignments were performed using PASTA[91] or MAFFT[85] and trimmed with ClipKit[86]. The tree was then computed by IQ-TREE[88] with options -bb 5000 -bi 200 -m TEST.

Reporting summary

Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

Large genome fragments and annotations were deposited to the EBI under the study PRJEB47746 with the following accessions: ERS10539964, ERS10539963, ERS10539962, ERS10539961, ERS10539960, ERS10539959, ERS10539958, ERS10539957. Accession codes of complete datasets can be found in Supplementary Table 1. In addition, previously published public data used for analysis includes: Genbank NR (from June 2020), GVMAGs (https://figshare.com/s/14788165283d65466732 and https://genome.jgi.doe.gov/portal/GVMAGs/GVMAGs.home.html), VOG orthogroups (https://vogdb.org/), Refseq protein database (from March 2020), EggNOG (v5), GEVEs (https://zenodo.org/record/3975964#.XzFj0hl7mfZ), JGI IMG/M (database 2021, https://img.jgi.doe.gov/). Source data are provided with this paper.

Code availability

Custom scripts codes[95] used in this study can be accessed here: https://doi.org/10.6084/m9.figshare.20101850.

References

1. Obu, J. How much of the earth's surface is underlain by permafrost? *J. Geophys. Res. Earth Surf.* **126**, e2021JF006123 (2021).
2. Mackelprang, R. et al. Metagenomic analysis of a permafrost microbial community reveals a rapid response to thaw. *Nature* **480**, 368–371 (2011).
3. Burkert, A., Douglas, T. A., Waldrop, M. P. & Mackelprang, R. Changes in the active, dead, and dormant microbial community structure across a Pleistocene permafrost chronosequence. *Appl. Environ. Microbiol.* **85**, e02646–18 (2019).
4. 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).
5. Hinsa-Leasure, S. M. et al. Characterization of a bacterial community from a Northeast Siberian seacoast permafrost sample. *FEMS Microbiol. Ecol.* **74**, 103–113 (2010).
6. Liang, R. et al. Predominance of anaerobic, spore-forming bacteria in metabolically active microbial communities from ancient Siberian permafrost. *Appl. Environ. Microbiol.* **85**, e00560–19 (2019).
7. Zhong, Z.-P. et al. Glacier ice archives nearly 15,000-year-old microbes and phages. *Microbiome* **9**, 160 (2021).
8. Turchetti, B. et al. Psychrophilic yeasts in glacial environments of Alpine glaciers. *FEMS Microbiol. Ecol.* **63**, 73–83 (2008).

9. Malavin, S., Shmakova, L., Claverie, J.-M. & Rivkina, E. Frozen Zoo: a collection of permafrost samples containing viable protists and their viruses. *Biodivers. Data J.* **8**, e51586 (2020).
10. Vishnivetskaya, T. A. et al. The resistance of viable permafrost algae to simulated environmental stresses: implications for astrobiology. *Int. J. Astrobiol.* **2**, 171–177 (2003).
11. Yashina, S. et al. Regeneration of whole fertile plants from 30,000-y-old fruit tissue buried in Siberian permafrost. *PNAS* **109**, 4008–4013 (2012).
12. Shmakova, L. et al. A living bdelloid rotifer from 24,000-year-old Arctic permafrost. *Curr. Biol.* **31**, R712–R713 (2021).
13. Bellas, C., Anesio, A. & Barker, G. Analysis of virus genomes from glacial environments reveals novel virus groups with unusual host interactions. *Front. Microbiol.* **6**, 656 (2015).
14. Rigou, S. et al. Metagenomic survey of the microbiome of ancient Siberian permafrost and modern Kamchatkan cryosols. *microLife* uqac003. https://doi.org/10.1093/femsml/uqac003 (2022).
15. 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).
16. Legendre, M. et al. In-depth study of *Mollivirus sibericum*, a new 30,000-y-old giant virus infecting Acanthamoeba. *PNAS* **112**, E5327–E5335 (2015).
17. Moniruzzaman, M., Martinez-Gutierrez, C. A., Weinheimer, A. R. & Aylward, F. O. Dynamic genome evolution and complex virocell metabolism of globally-distributed giant viruses. *Nat. Commun.* **11**, 1–11 (2020).
18. Schulz, F. et al. Giant virus diversity and host interactions through global metagenomics. *Nature* **578**, 432–436 (2020).
19. Endo, H. et al. Biogeography of marine giant viruses reveals their interplay with eukaryotes and ecological functions. *Nat. Ecol. Evol.* 1–11. https://doi.org/10.1038/s41559-020-01288-w. (2020).
20. Gallot-Lavallée, L., Blanc, G. & Claverie, J.-M. 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 (2017).
21. Ha, A. D., Moniruzzaman, M. & Aylward, F. O. High transcriptional activity and diverse functional repertoires of hundreds of giant viruses in a coastal marine system. *mSystems* **6**, e0029321 (2021).
22. Moniruzzaman, M. et al. Virus-host relationships of marine single-celled eukaryotes resolved from metatranscriptomics. *Nat. Commun.* **8**, 16054 (2017).
23. Gann, E. R., Kang, Y., Dyhrman, S. T., Gobler, C. J. & Wilhelm, S. W. Metatranscriptome library preparation influences analyses of viral community activity during a brown tide bloom. *Front. Microbiol.* **12**, 664189 (2021).
24. Pound, H. L. et al. The "Neglected Viruses" of Taihu: abundant transcripts for viruses infecting eukaryotes and their potential role in phytoplankton succession. *Front. Microbiol.* **11**, 338 (2020).
25. Needham, D. M. et al. A distinct lineage of giant viruses brings a rhodopsin photosystem to unicellular marine predators. *Proc. Natl Acad. Sci. USA* **116**, 20574–20583 (2019).
26. Schulz, F. et al. Hidden diversity of soil giant viruses. *Nat. Commun.* **9**, 1–9 (2018).
27. Bäckström, D. et al. Virus genomes from deep sea sediments expand the ocean megavirome and support independent origins of viral gigantism. *mBio* **10**, e02497–18 (2019).
28. Alteio, L. V. et al. Complementary metagenomic approaches improve reconstruction of microbial diversity in a forest soil. *mSystems* **5**, e00768–19 (2020).
29. Sczyrba, A. et al. Critical assessment of metagenome interpretation-a benchmark of metagenomics software. *Nat. Methods* **14**, 1063–1071 (2017).
30. Philippe, N. et al. Pandoraviruses: amoeba viruses with genomes up to 2.5 Mb reaching that of parasitic eukaryotes. *Science* **341**, 281–286 (2013).
31. Jeudy, S. et al. The DNA methylation landscape of giant viruses. *Nat. Commun.* **11**, 2657 (2020).
32. Christo-Foroux, E. et al. Characterization of *Mollivirus kamchatka*, the first modern representative of the proposed *Molliviridae* family of giant viruses. *J. Virol.* **94**, e01997–19 (2020).
33. Parks, D. H., Imelfort, M., Skennerton, C. T., Hugenholtz, P. & Tyson, G. W. CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome Res.* **25**, 1043–1055 (2015).
34. Chen, L.-X., Anantharaman, K., Shaiber, A., Eren, A. M. & Banfield, J. F. Accurate and complete genomes from metagenomes. *Genome Res.* **30**, 315–333 (2020).
35. Parks, D. H. et al. Evaluation of the microba community profiler for taxonomic profiling of metagenomic datasets from the human gut microbiome. *Front. Microbiol.* **12**, 643682 (2021).
36. Leclère, L. et al. The genome of the jellyfish *Clytia hemisphaerica* and the evolution of the cnidarian life-cycle. *Nat. Ecol. Evol.* **3**, 801–810 (2019).
37. Roux, S. et al. Ecogenomics of virophages and their giant virus hosts assessed through time series metagenomics. *Nat. Commun.* **8**, 858 (2017).
38. Moniruzzaman, M., Weinheimer, A. R., Martinez-Gutierrez, C. A. & Aylward, F. O. Widespread endogenization of giant viruses shapes genomes of green algae. *Nature* **588**, 141–145 (2020).
39. Aylward, F. O. & Moniruzzaman, M. ViralRecall—a flexible command-line tool for the detection of giant virus signatures in 'Omic Data. *Viruses* **13**, 150 (2021).
40. Guglielmini, J., Woo, A. C., Krupovic, M., Forterre, P. & Gaia, M. Diversification of giant and large eukaryotic dsDNA viruses predated the origin of modern eukaryotes. *Proc. Natl Acad. Sci. USA* **116**, 19585–19592 (2019).
41. Chen, I.-M. A. et al. The IMG/M data management and analysis system v.6.0: new tools and advanced capabilities. *Nucleic Acids Res.* **49**, D751–D763 (2021).
42. Andreani, J. et al. Orpheovirus IHUMI-LCC2: a new virus among the giant viruses. *Front. Microbiol.* **8**, 2643 (2017).
43. Levasseur, A. et al. Comparison of a modern and fossil pithovirus reveals its genetic conservation and evolution. *Genome Biol. Evol.* **8**, 2333–2339 (2016).
44. Andreani, J. et al. Cedratvirus, a double-cork structured giant virus, is a distant relative of pithoviruses. *Viruses* **8**, 300 (2016).
45. Yutin, N., Wolf, Y. I., Raoult, D. & Koonin, E. V. Eukaryotic large nucleo-cytoplasmic DNA viruses: clusters of orthologous genes and reconstruction of viral genome evolution. *Virol. J.* **6**, 223 (2009).
46. Koonin, E. V. & Yutin, N. Evolution of the large nucleocytoplasmic DNA viruses of eukaryotes and convergent origins of viral gigantism. *Adv. Virus Res.* **103**, 167–202 (2019).
47. Schulz, F. et al. Giant viruses with an expanded complement of translation system components. *Science* **356**, 82–85 (2017).
48. Schulz, F. et al. A Rickettsiales symbiont of amoebae with ancient features. *Environ. Microbiol.* **18**, 2326–2342 (2016).
49. Legendre, M. et al. Diversity and evolution of the emerging Pandoraviridae family. *Nat. Commun.* **9**, 2285 (2018).
50. Nasir, A. & Caetano-Anollés, G. A phylogenomic data-driven exploration of viral origins and evolution. *Sci. Adv.* https://doi.org/10.1126/sciadv.1500527 (2015).
51. O'Brien, S. L. et al. Spatial scale drives patterns in soil bacterial diversity. *Environ. Microbiol* **18**, 2039–2051 (2016).
52. Bergh, O., Børsheim, K. Y., Bratbak, G. & Heldal, M. High abundance of viruses found in aquatic environments. *Nature* **340**, 467–468 (1989).

53. Cochlan, W. P., Wikner, J., Steward, G. F., Smith, D. C. & Azam, F. Spatial distribution of viruses, bacteria and chlorophyll a in neritic, oceanic and estuarine environments. *Mar. Ecol. Prog. Ser.* **92**, 77–87 (1993).
54. Raoult, D. et al. The 1.2-megabase genome sequence of Mimivirus. *Science* **306**, 1344–1350 (2004).
55. Claverie, J.-M. & Abergel, C. Giant viruses: the difficult breaking of multiple epistemological barriers. *Stud. Hist. Philos. Biol. Biomed. Sci.* **59**, 89–99 (2016).
56. Filée, J. Genomic comparison of closely related Giant Viruses supports an accordion-like model of evolution. *Front. Microbiol.* **6**, 593 (2015).
57. Krupovic, M. & Koonin, E. V. Polintons: a hotbed of eukaryotic virus, transposon and plasmid evolution. *Nat. Rev. Microbiol* **13**, 105–115 (2015).
58. 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).
59. Moreira, D. & Brochier-Armanet, C. Giant viruses, giant chimeras: the multiple evolutionary histories of Mimivirus genes. *BMC Evol. Biol.* **8**, 12 (2008).
60. Ohno, S. The creation of a new gene from a redundant duplicate of an old gene. In *Evolution by Gene Duplication* (ed. Ohno, S.) 71–82 (Springer, 1970).
61. Suhre, K. Gene and genome duplication in *Acanthamoeba polyphaga* Mimivirus. *J. Virol.* **79**, 14095–14101 (2005).
62. Legendre, M. et al. Pandoravirus Celtis illustrates the microevolution processes at work in the giant Pandoraviridae genomes. *Front. Microbiol.* **10**, 430 (2019).
63. Munson-McGee, J. H. et al. A virus or more in (nearly) every cell: ubiquitous networks of virus–host interactions in extreme environments. *ISME J.* **12**, 1706–1714 (2018).
64. Koonin, E. V. et al. Global organization and proposed megataxonomy of the virus world. *Microbiol. Mol. Biol. Rev.* **84**, e00061–19 (2020).
65. Mönttinen, H. A. M., Bicep, C., Williams, T. A. & Hirt, R. P. The genomes of nucleocytoplasmic large DNA viruses: viral evolution writ large. *Microb Genom.* **7**, 000649 (2021).
66. Boratto, P. V. M. et al. Yaravirus: a novel 80-nm virus infecting *Acanthamoeba castellanii*. *Proc. Natl Acad. Sci. USA* **117**, 16579–16586 (2020).
67. Kazlauskas, D., Krupovic, M., Guglielmini, J., Forterre, P. & Venclovas, Č. Diversity and evolution of B-family DNA polymerases. *Nucleic Acids Res.* **48**, 10142–10156 (2020).
68. Li, D., Liu, C.-M., Luo, R., Sadakane, K. & Lam, T.-W. MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. *Bioinformatics* **31**, 1674–1676 (2015).
69. Kang, D. D. et al. MetaBAT 2: an adaptive binning algorithm for robust and efficient genome reconstruction from metagenome assemblies. *PeerJ* **7**, e7359 (2019).
70. Nurk, S., Meleshko, D., Korobeynikov, A. & Pevzner, P. A. metaSPAdes: a new versatile metagenomic assembler. *Genome Res.* **27**, 824–834 (2017).
71. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**, 357–359 (2012).
72. Camacho, C. et al. BLAST+: architecture and applications. *BMC Bioinform.* **10**, 1–9 (2009).
73. Besemer, J., Lomsadze, A. & Borodovsky, M. GeneMarkS: a self-training method for prediction of gene starts in microbial genomes. Implications for finding sequence motifs in regulatory regions. *Nucleic Acids Res.* **29**, 2607–2618 (2001).
74. Altschul, S. F. et al. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* **25**, 3389–3402 (1997).
75. Bertelli, C. et al. Cedratvirus lausannensis - digging into Pithoviridae diversity. *Environ. Microbiol.* **19**, 4022–4034 (2017).
76. Rodrigues, R. A. L. et al. Morphologic and genomic analyses of new isolates reveal a second lineage of Cedratviruses. *J. Virol.* **92**, e00372–18 (2018).
77. Aherfi, S. et al. A large open pangenome and a small core genome for giant Pandoraviruses. *Front Microbiol* **9**, 1486 (2018).
78. Scheid, P. A strange endocytobiont revealed as largest virus. *Curr. Opin. Microbiol* **31**, 58–62 (2016).
79. Emms, D. M. & Kelly, S. OrthoFinder: phylogenetic orthology inference for comparative genomics. *Genome Biol.* **20**, 238 (2019).
80. Eddy, S. R. A new generation of homology search tools based on probabilistic inference. In *Genome Informatics 2009*, 205–211 (Imperial College Press, 2009). https://doi.org/10.1142/9781848165632_0019.
81. Huerta-Cepas, J. et al. eggNOG 5.0: a hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses. *Nucleic Acids Res.* **47**, D309–D314 (2019).
82. Kanehisa, M., Sato, Y. & Morishima, K. BlastKOALA and GhostKOALA: KEGG tools for functional characterization of genome and metagenome sequences. *J. Mol. Biol.* **428**, 726–731 (2016).
83. GitHub - tseemann/barrnap: Bacterial ribosomal RNA predictor. *GitHub* https://github.com/tseemann/barrnap.
84. Robinson, J. T. et al. Integrative genomics viewer. *Nat. Biotechnol.* **29**, 24–26 (2011).
85. Katoh, K. & Standley, D. M. MAFFT Multiple Sequence Alignment Software Version 7: improvements in performance and usability. *Mol. Biol. Evol.* **30**, 772–780 (2013).
86. Steenwyk, J. L., Iii, T. J. B., Li, Y., Shen, X.-X. & Rokas, A. ClipKIT: a multiple sequence alignment trimming software for accurate phylogenomic inference. *PLoS Biol.* **18**, e3001007 (2020).
87. Kalyaanamoorthy, S., Minh, B. Q., Wong, T. K. F., von Haeseler, A. & Jermiin, L. S. ModelFinder: fast model selection for accurate phylogenetic estimates. *Nat. Methods* **14**, 587–589 (2017).
88. Nguyen, L.-T., Schmidt, H. A., von Haeseler, A. & Minh, B. Q. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol. Biol. Evol.* **32**, 268–274 (2015).
89. Hoang, D. T., Chernomor, O., von Haeseler, A., Minh, B. Q. & Vinh, L. S. UFBoot2: improving the Ultrafast Bootstrap Approximation. *Mol. Biol. Evol.* **35**, 518–522 (2018).
90. Chernomor, O., von Haeseler, A. & Minh, B. Q. Terrace aware data structure for phylogenomic inference from supermatrices. *Syst. Biol.* **65**, 997–1008 (2016).
91. Mirarab, S., Nguyen, N. & Warnow, T. PASTA: Ultra-Large Multiple Sequence Alignment. in *Research in Computational Molecular Biology* (ed. Sharan, R.) 177–191 (Springer International Publishing, 2014).
92. Chris Creevey & Nathan Weeks. *ChrisCreevey/catsequences: Version 1.3*. (Zenodo, 2021).
93. Letunic, I. & Bork, P. Interactive Tree Of Life (iTOL): an online tool for phylogenetic tree display and annotation. *Bioinformatics* **23**, 127–128 (2007).
94. Arvestad, L. Efficient methods for estimating amino acid replacement rates. *J. Mol. Evol.* **62**, 663–673 (2006).
95. Rigou, S. Scripts used in 'Past and present giant viruses diversity explored through permafrost metagenomics'. https://doi.org/10.6084/m9.figshare.20101850.v1 (2022).

Acknowledgements

We thank the PACA Bioinfo platform for computing support. Regarding the samples from Rigou et al.[14] used in this study, we would like to thank Alexander Morawitz for collecting the Kamchatka soil samples and Eugène Christo-Foroux for processing the sample and performing DNA

extraction, Dr. Jens Strauss and Dr. Guido Grosse for providing the Yukechi permafrost samples and Dr. Karine Labadie for supervising the sequencing on the Genoscope platform. We also thank François Enault and Hugo Bisio for carefully reading the manuscript. This work was supported by the CNRS Projets de Recherche Conjoints (PRC) grant (PRC1484-2018) to C.A. S.R. is supported by a doctoral fellowship obtained from Aix-Marseille University.

Author contributions

Conceptualization: M.L., C.A., and J-M.C.; Methodology: S.R. and M.L.; Software: S.R.; Validation: S.R., M.L., and S.S.; Formal analysis: S.R.; Resources: J-M.C. and S.S.; Data curation: S.R.; Writing – original draft: S.R. and M.L.; Writing – review and editing: S.R., C.A., J-M.C., and M.L.; Visualization: S.R. and M.L.; Supervision: M.L.; Project administration: M.L.; Funding acquisition: C.A. and J-M.C.

Competing interests

The authors declare no competing interests.

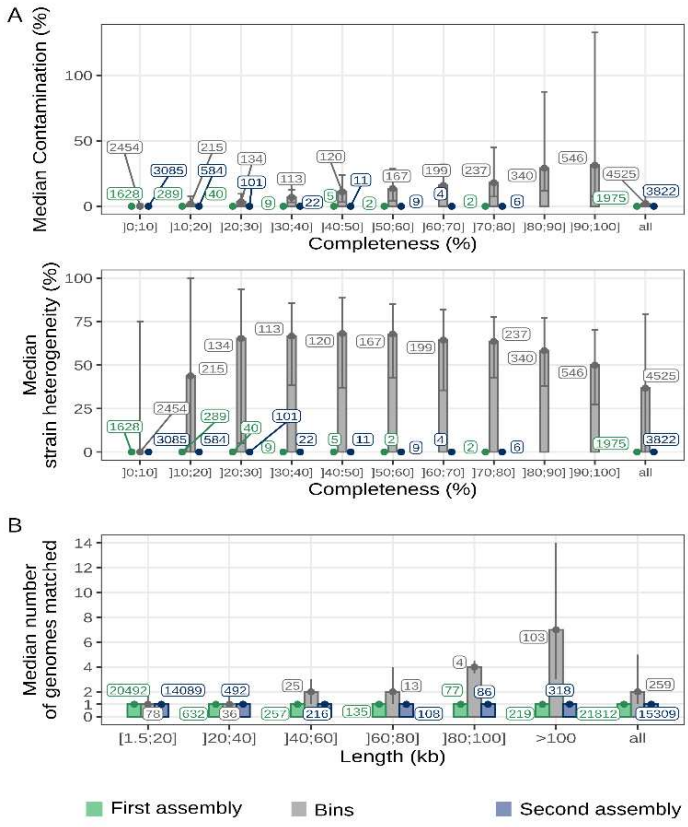
Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41467-022-33633-x.

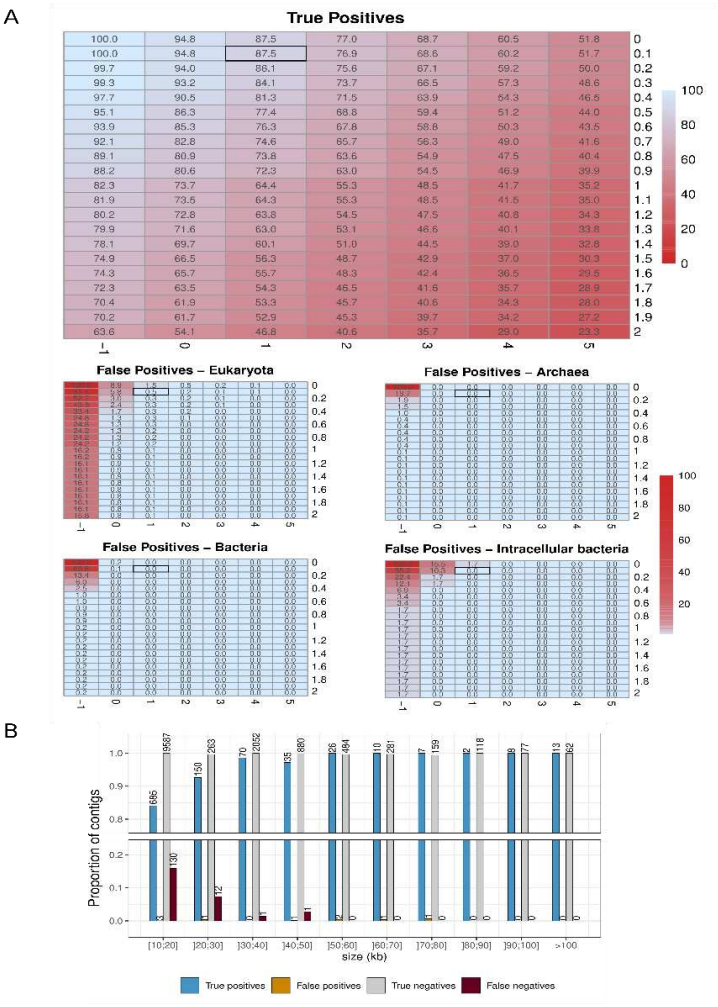
Correspondence and requests for materials should be addressed to Matthieu Legendre.

Peer review information *Nature Communications* thanks the anonymous reviewers for their contribution to the peer review of this work. Peer reviewer reports are available.

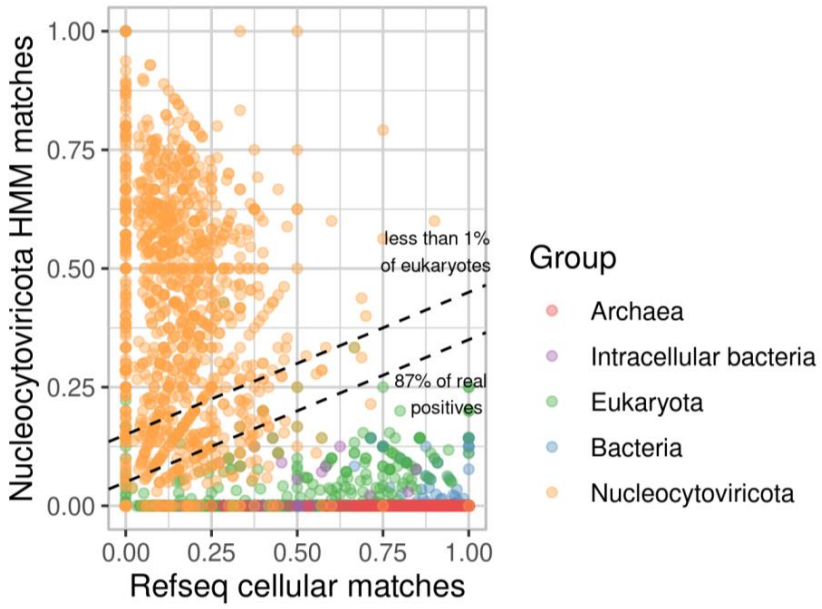
Supplementary Information file



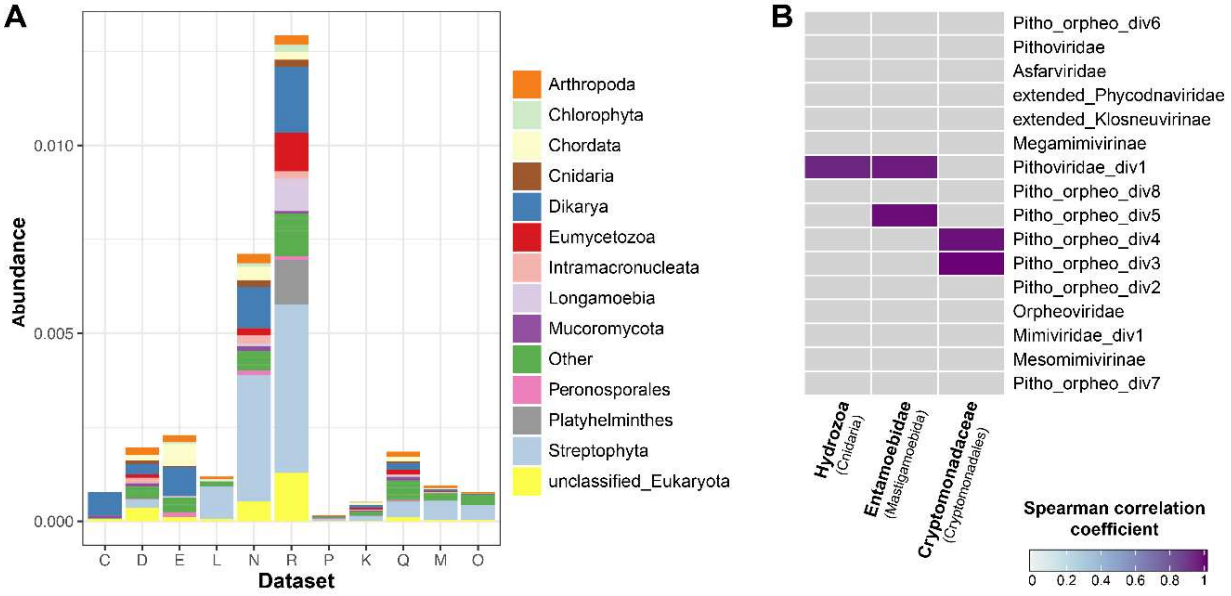
Supplementary Figure 1. Control for chimerism
CheckM was ran on the (A) 11 Russian cryosol samples to check for chimeras in contigs (first assembly), bins or scaffolds (second assembly). (B) Three MOCK complex communities were assembled using the same methodology. The resulting sequences were then aligned to their references by BLASTN to assess their level of chimerism (i.e. matching several genomes). Points indicate median values and the error bars correspond to the 25% and 75% quantiles. In the case of n < 3, no median was computed as all points correspond to zero contamination and zero strain heterogeneity. Counts are shown on top of each bar. Source data are provided as a Source Data file.



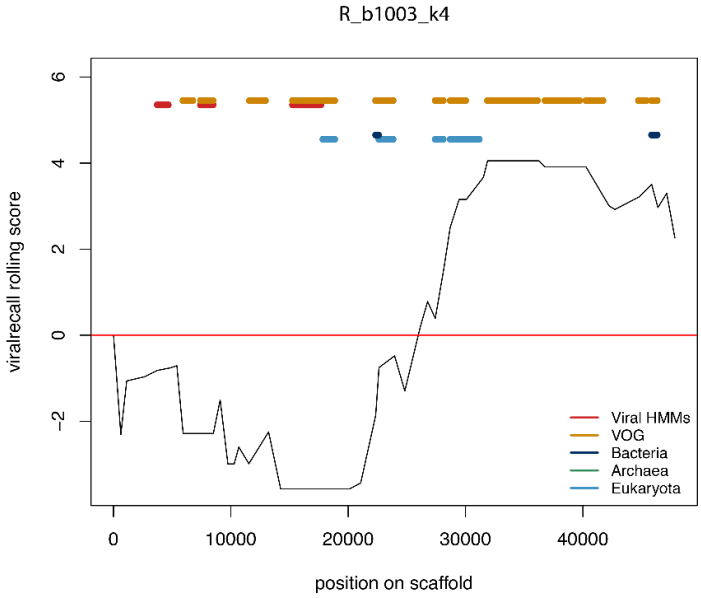
Supplementary Figure 2. Tested threshold parameters for *Nucleocytoviricota* sequence retrieval on a control dataset
(A) From a control dataset (see methods) parameters were estimated to discriminate viral and cellular contigs. The intercept parameter is shown on the x-axis and the slope on the y-axis. Notice that only eukaryotes render false positives at the optimal parameters (x=1, y=0.1). (B) Applying the threshold resulted in false/true positives and false/true negatives that were then computed for different contig sizes. The number of contigs is written above each bar.



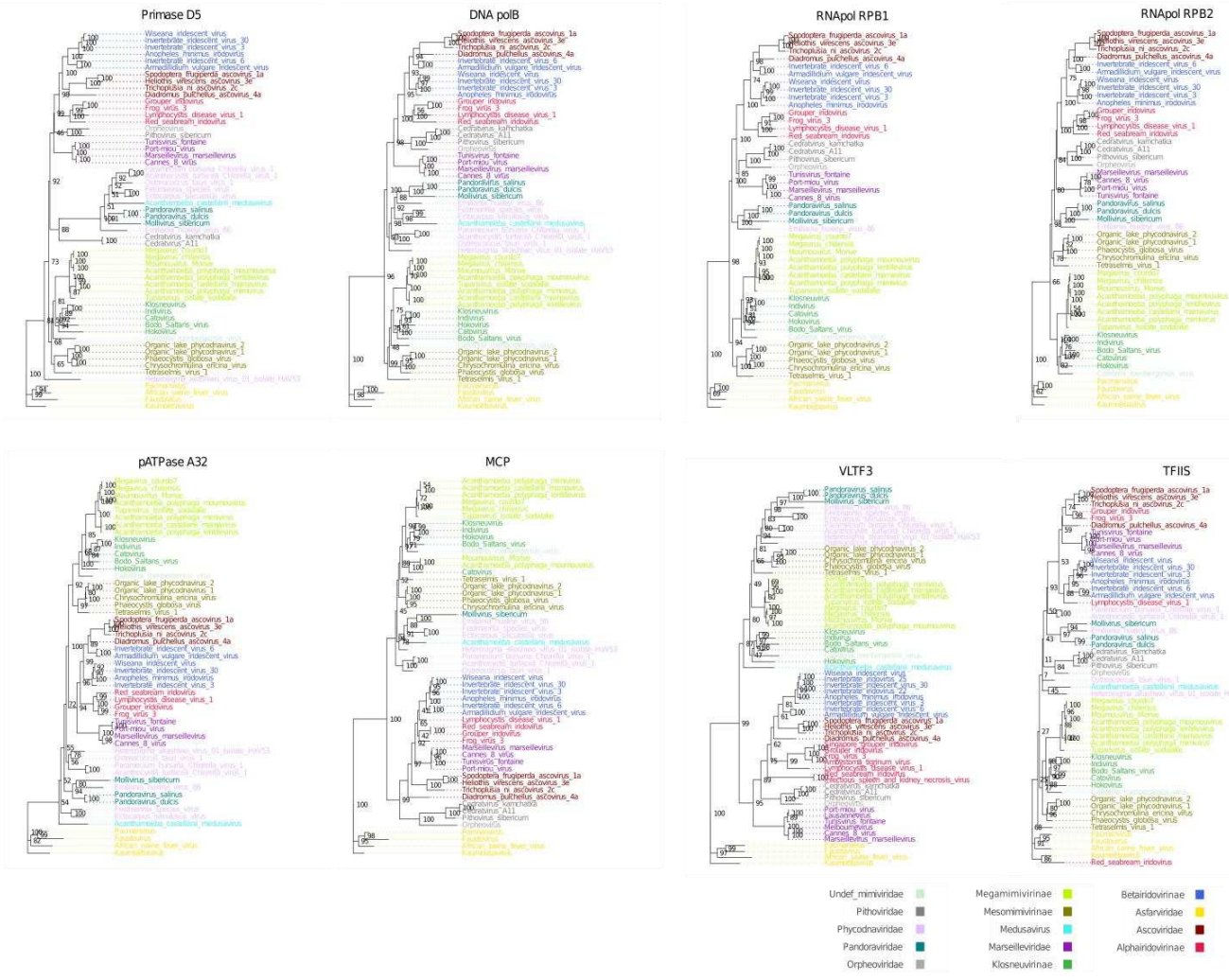
Supplementary Figure 3. Filtering of scaffolds according to their proportion of viral and cellular matches
An alternative methodology was tested in which we used the proportions instead of the number of matches. All tested parameters for determining the threshold resulted to be suboptimal compared to the previous methodology (Fig. 1 in the main text) with either more false positives (the lowest dashed line) or more false negatives (the upper dashed line). Source data are provided as a Source Data file.



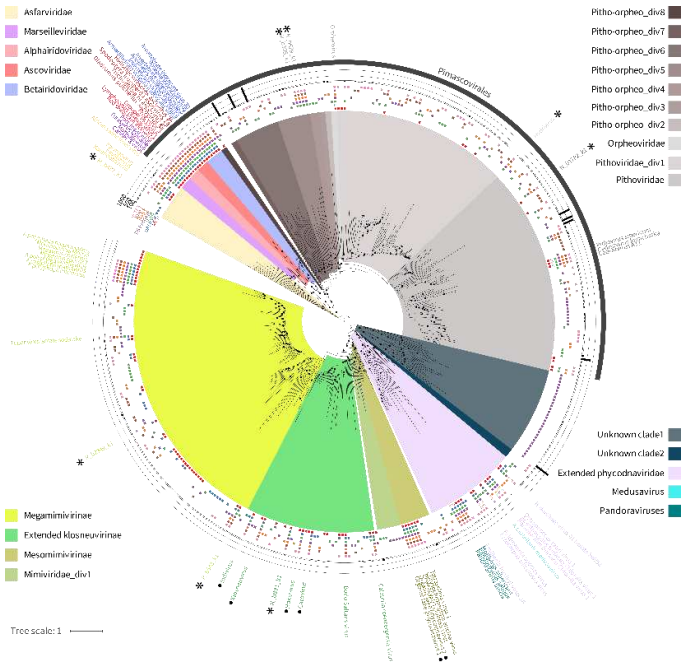
Supplementary Figure 4. Eukaryotic abundance across samples and co-occurrence with viruses
(A) The relative abundance of a given eukaryotic group is given as the sum of coverages of the sequences belonging to it divided by the total sample coverage. (B) Co-occurrence between viral and eukaryotic families. The spearman correlations were calculated from the estimated normalized abundance (sum of coverages) in each sample. Two-sided correlation test p-values were corrected for multiple testing using the Benjamini & Hochberg FDR correction. Only significantly co-occurring (corrected p-value < 0.01) Eucaryote/virus pairs are color coded. Source data are provided as a Source Data file.



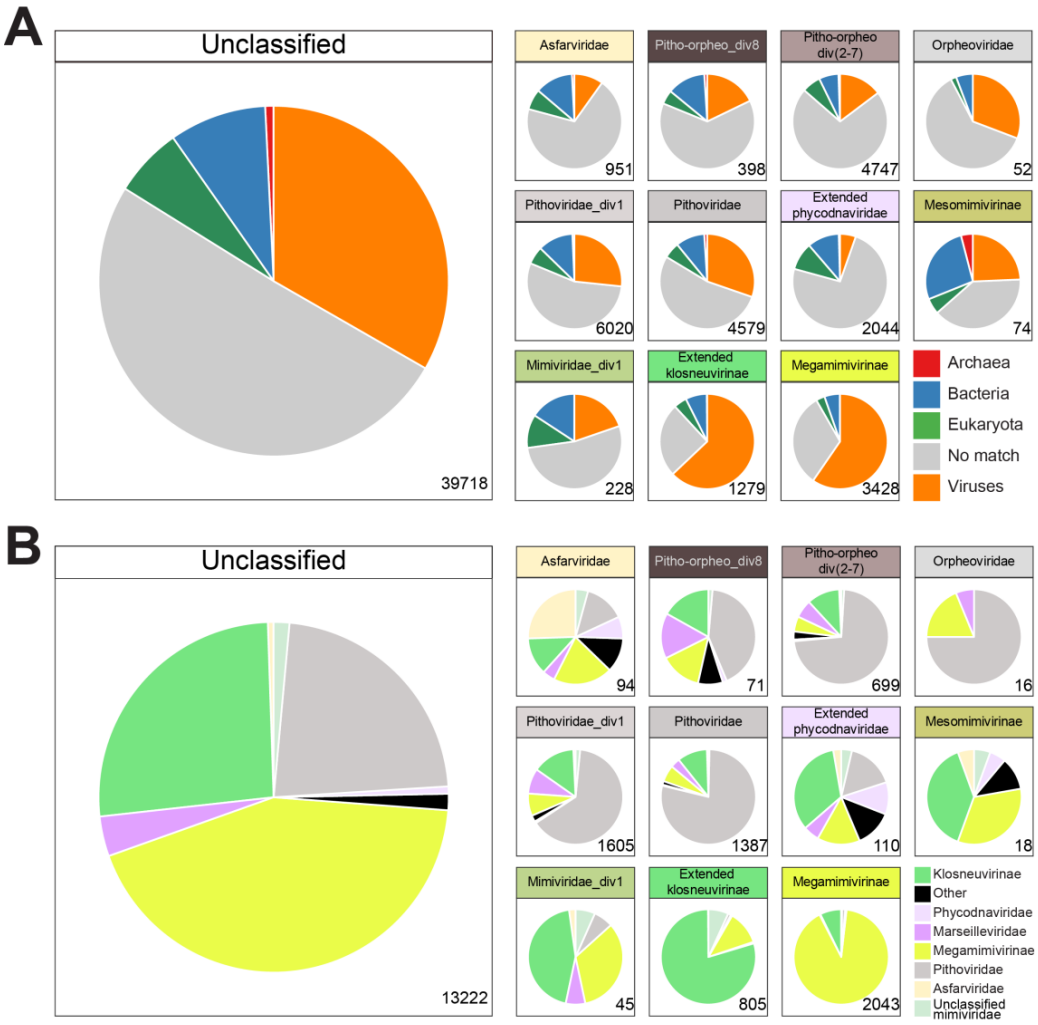
Supplementary Figure 5. Identification of potentially endogenized viruses using ViralRecall
The red line indicates the decision threshold between potentially cellular (negative) or viral (positive) genomic regions as determined by ViralRecall. In this example, the putative cellular portion does have matches in the VOG database and its DIAMOND BLASTP cellular matches are both from Eukaryotes and Bacteria.



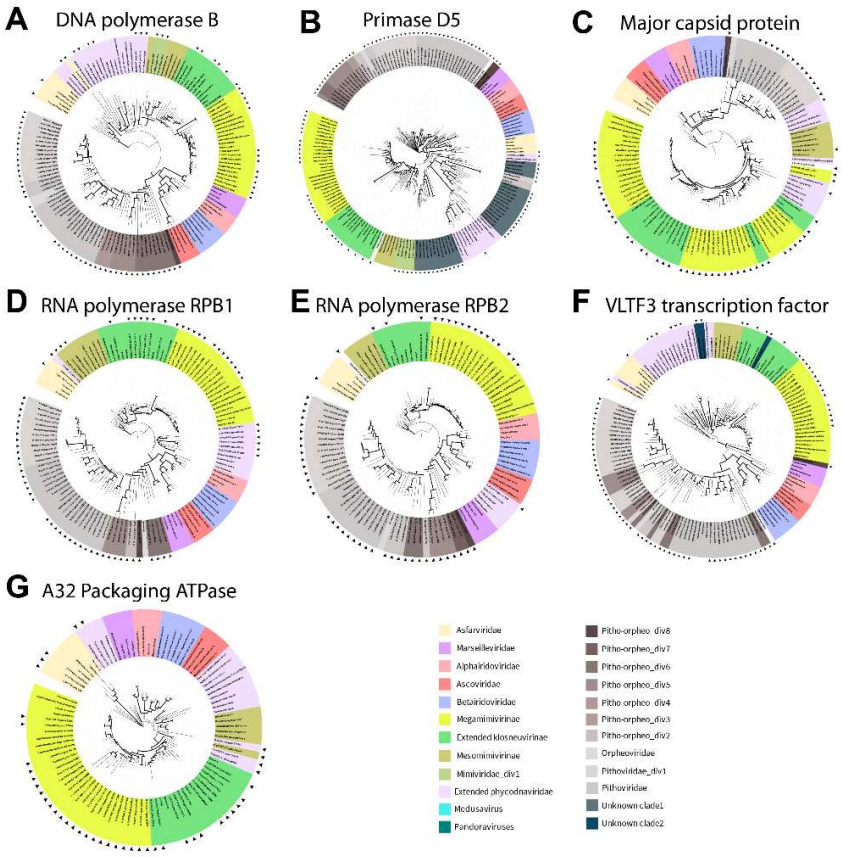
Supplementary Figure 6. Phylogenetic trees of marker genes from the reference *Nucleocytoviricota* genomes
Consensus trees (parameters –bb 1000 –bi 100) and optimal models (-m MFP) were calculated by IQ-TREE.



Supplementary Figure 7. *Nucleocytoviricota* phylogeny in all samples
Consensus of 1000 bootstrapped trees calculated by IQ-TREE through a partitioned analysis on 7 marker genes. The models used were the following: LG+F+R5 for the packaging ATPase and the MCP, LG+F+R6 for the RNA polymerase subunits RPB1 and RPB2, LG+F+R7 for the primase D5, LG+R+R10 for the DNA polymerase B and VT+F+R7 for VLTF3. Black dots represent branch bootstrap support ≥ 90. The colored labels indicate the reference genomes and the large genome fragments identified in this study (marked with a star). One should note that reference genomes coming from bins of previous metagenomic studies (marked with a black dot) are less reliable than the genomes of isolated viruses. The colored clades were manually created to be monophyletic. The marker genes used for this phylogeny are indicated as colored squares. Empty squares correspond to marker genes absent from the reference genomes. Black bars show the normalized mean coverage of the scaffolds. *Pimascovirales* are defined as the clade composed of all the *Ascoviridae*, *Iridoviridae*, *Marseilleviridae*, but also *Pithoviridae*, *Orpheoviridae* and the metagenomic intermediate clades. The Extended_phycodnaviridae group includes *Pandoraviridae* and Mollivirus. The Extended_klosneuvirinae group includes the Cafeteria roenbergensis virus.

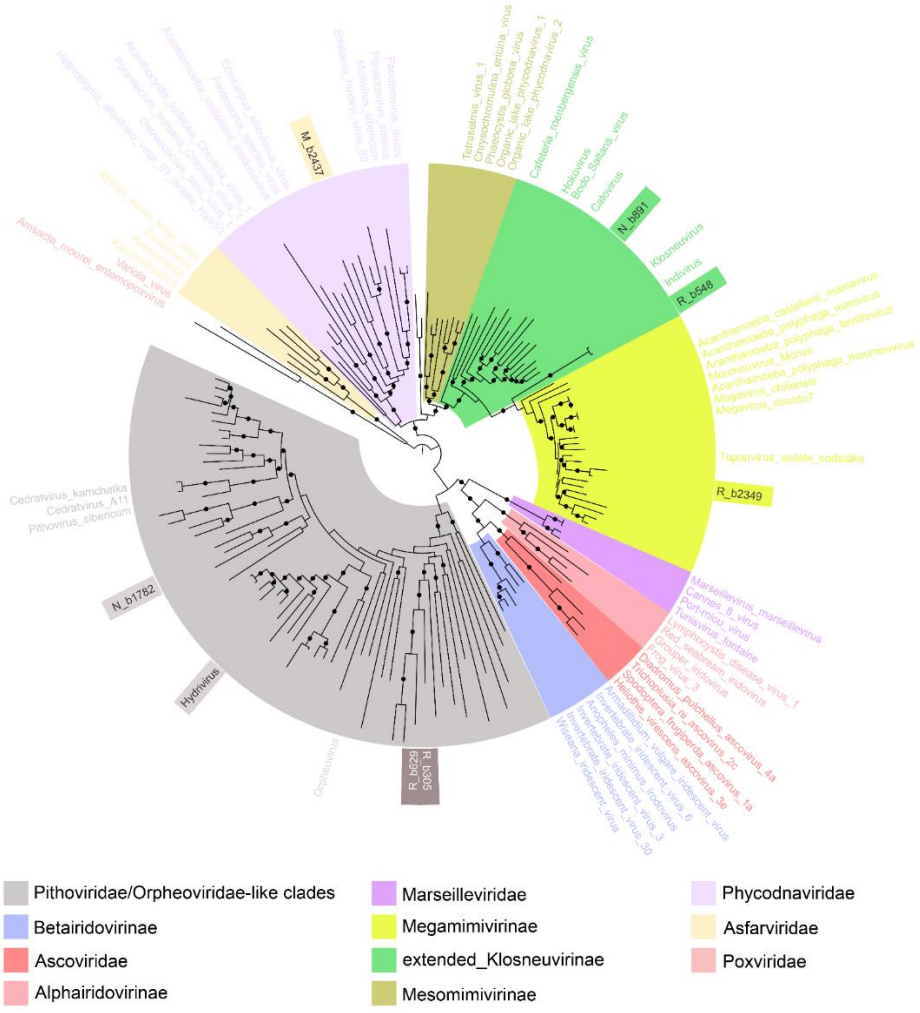


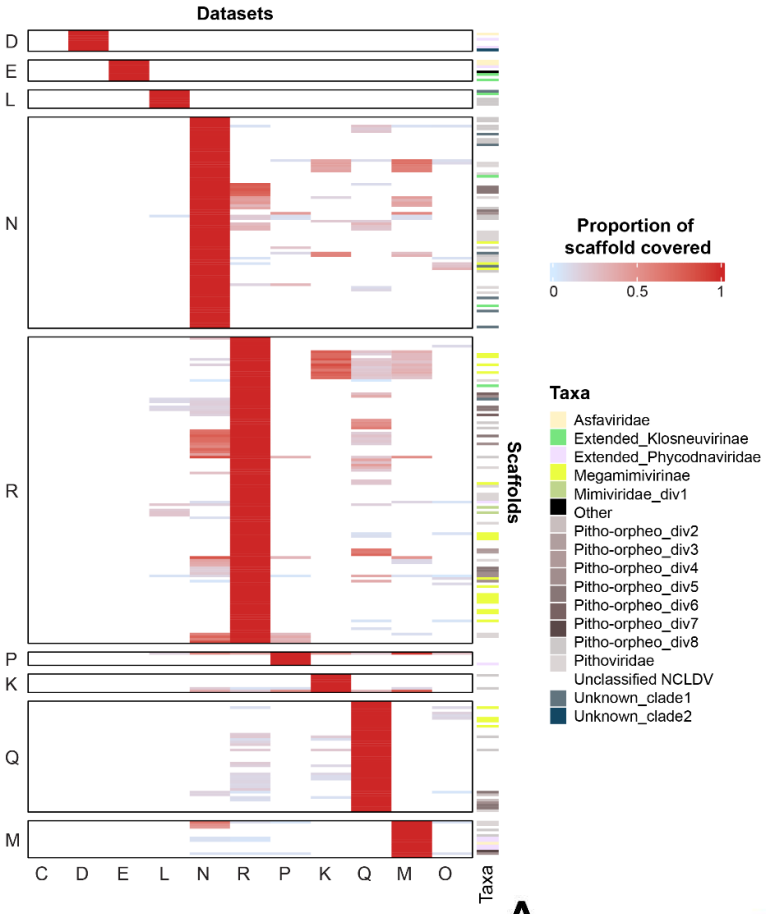
Supplementary Figure 8. Best BLASTP matches against the nr database
Pie charts of (A) best domain matches and (B) best viral matches for all permafrost sequences within phylogenetically defined groups. Total counts are shown at the bottom right corners. Source data are provided as a Source Data file.



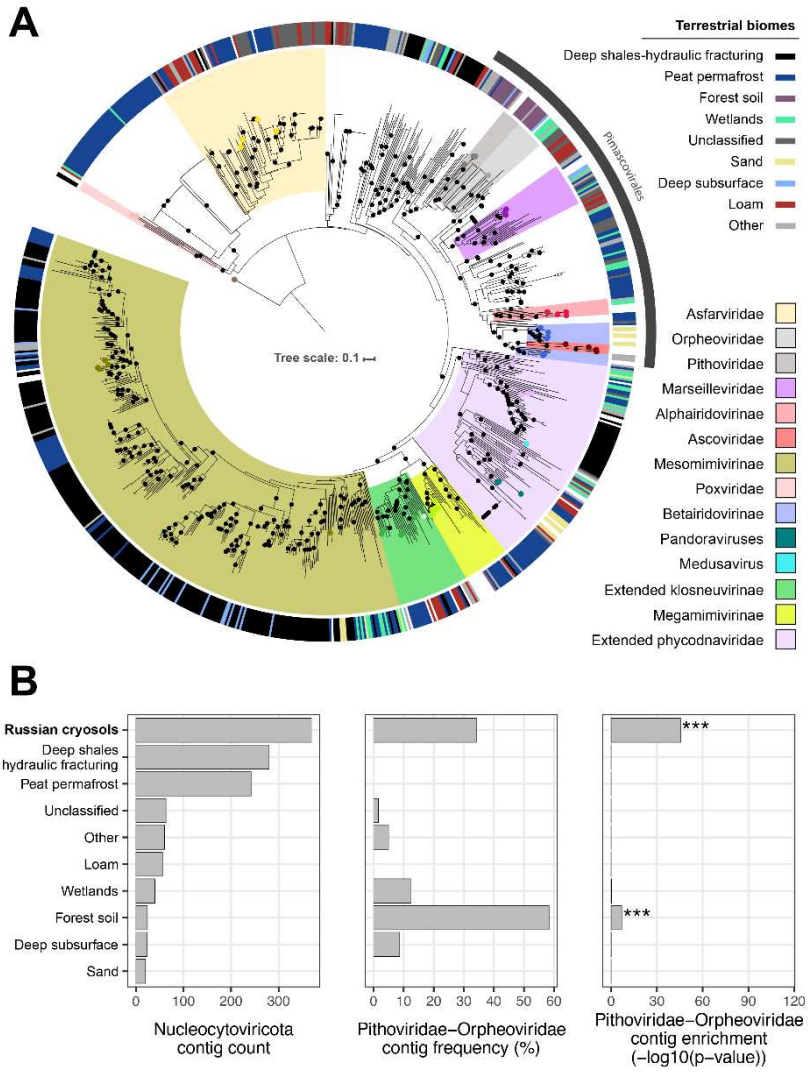
Supplementary Figure 9. Phylogenetic trees of individual *Nucleocytoviricota* marker genes
The same sequences than for the main phylogeny (Fig. 3) were used. The following models were used: LG+R+R10 for the DNA polymerase B (A), LG+F+R7 for the primase D5 (B), LG+F+R5 for the MCP (C) and the packaging ATPase (G), LG+F+R6 for the RNA polymerase subunits RPB1 (D) and RPB2 (E) and VT+F+R7 for VLTF3 (F). The wider branches indicate a support value above 80%. Permafrost sequences are marked with a black triangle.

Supplementary Figure 10. Phylogenetic tree on permafrost bins
Bins containing at least 3 marker genes were used for phylogeny. Only scaffolds that have previously passed the filter as for being a *Nucleocytoviricota* have been considered. Bins containing the large scaffolds (>500 kb) identified in this study (Fig. 4) have been labeled with a box of the color of their original scaffolds' taxonomies based on Supplementary Figure 8. Reference genomes are also labeled with white background. Of notice, the M_b2437 bin assigned to *Phycodnaviridae* contains three scaffolds with different taxonomies, indicative of potential chimerism. The black circles indicate branches that are well supported (>=90% of bootstraps).

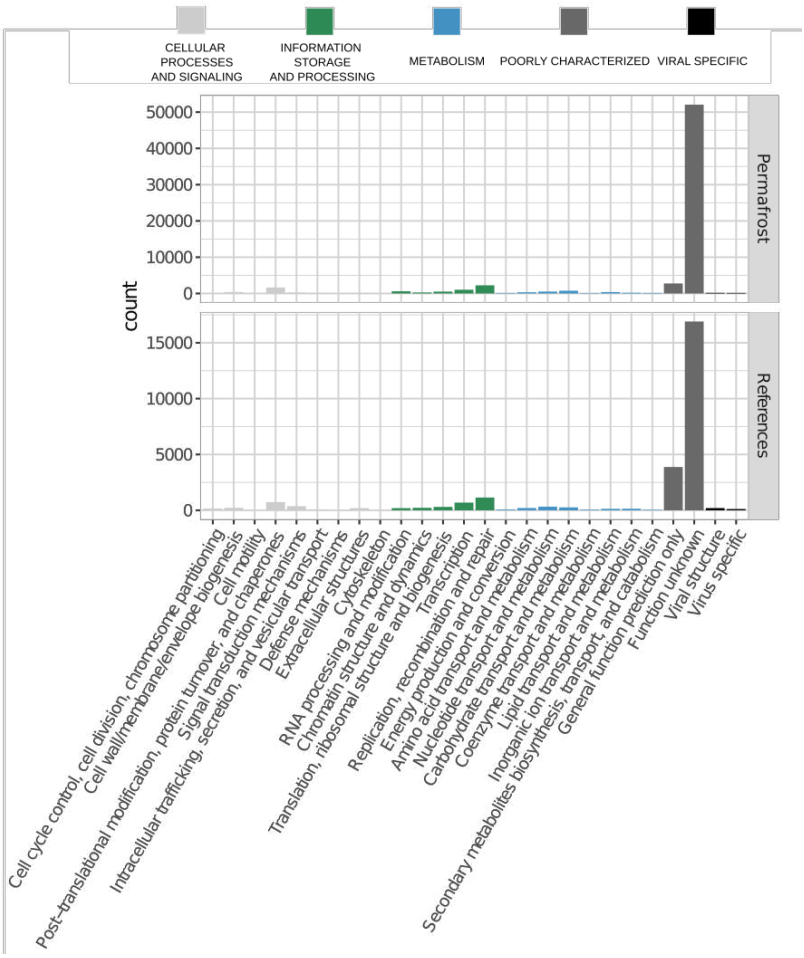




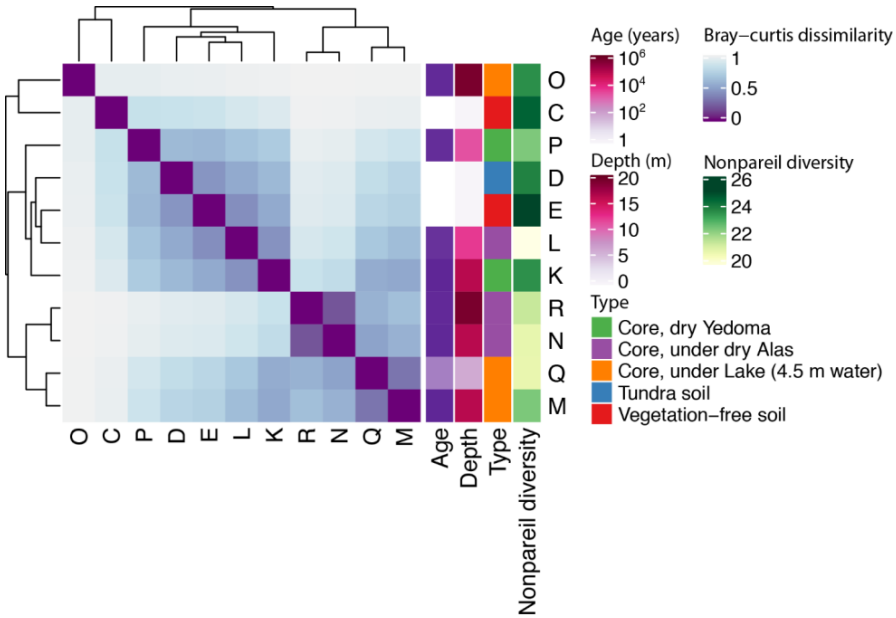
Supplementary Figure 11. Mapping of *Nucleocytoviricota* found in all samples
Reads of all samples were mapped on *Nucleocytoviricota* scaffolds over 50kb. The fraction of scaffold covered was calculated as the number of base pairs with at least one read mapped divided by the total scaffold length. White boxes indicate contigs with less than 10kb covered, considered as noise.



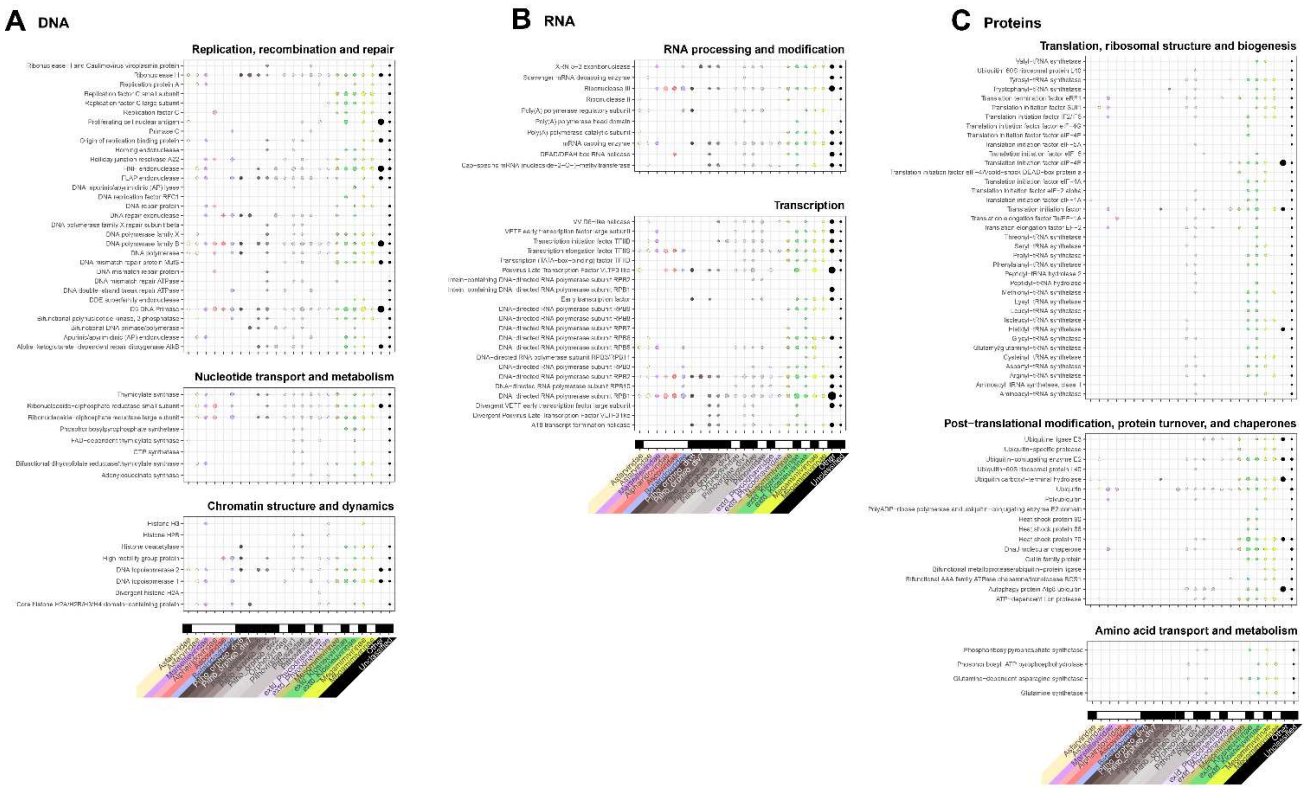
Supplementary Figure 12. *Nucleocytoviricota* recovered from terrestrial samples of the IMG/M database
(A) 804 contigs assembled from 147 terrestrial datasets of the JGI IMG/M database. Viral contigs were detected using the previously described method and placed on tree using at least one of the seven marker genes. The tree was made using Cyprinid herpesvirus 2 as outgroup. Clades containing the reference sequences were manually drawn. Colored circles at tips represent reference genomes and the outer circle shows the corresponding biome. (B) *Pithoviridae* and *Orpheoviridae* enrichment in various biomes: *Nucleocytoviricota* contig counts from our metagenomic data (Russian cryosols) and JGI samples (see Methods). The enrichment of *Pithoviridae* and *Orpheoviridae* was calculated through an asymmetrical Fisher test per biome, counting *Orpheoviridae* and *Pithoviridae* versus other *Nucleocytoviricota* among the tested biome and all other biomes.



Supplementary Figure 13. Functional categories of predicted *Nucleocytoviricota* ORFs
Viral ORFs from the Russian cryosol metagenomes were manually functionally annotated and assigned to a category. Source data are provided as a Source Data file.

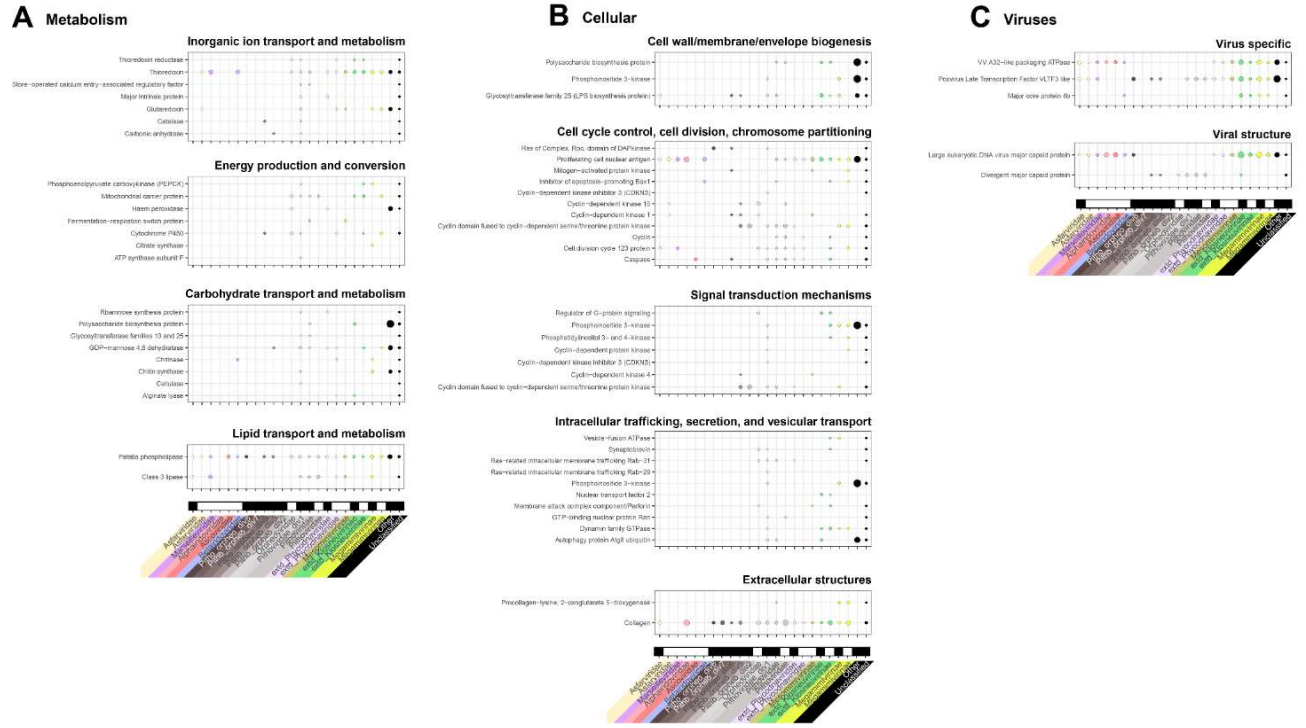


Supplementary Figure 14. Samples clustered according to the viral functional content.
The heatmap presents samples clustered based in the Bray-Curtis dissimilarity calculated on Pfam annotations occurrences. Some ecological parameters are shown for each sample.



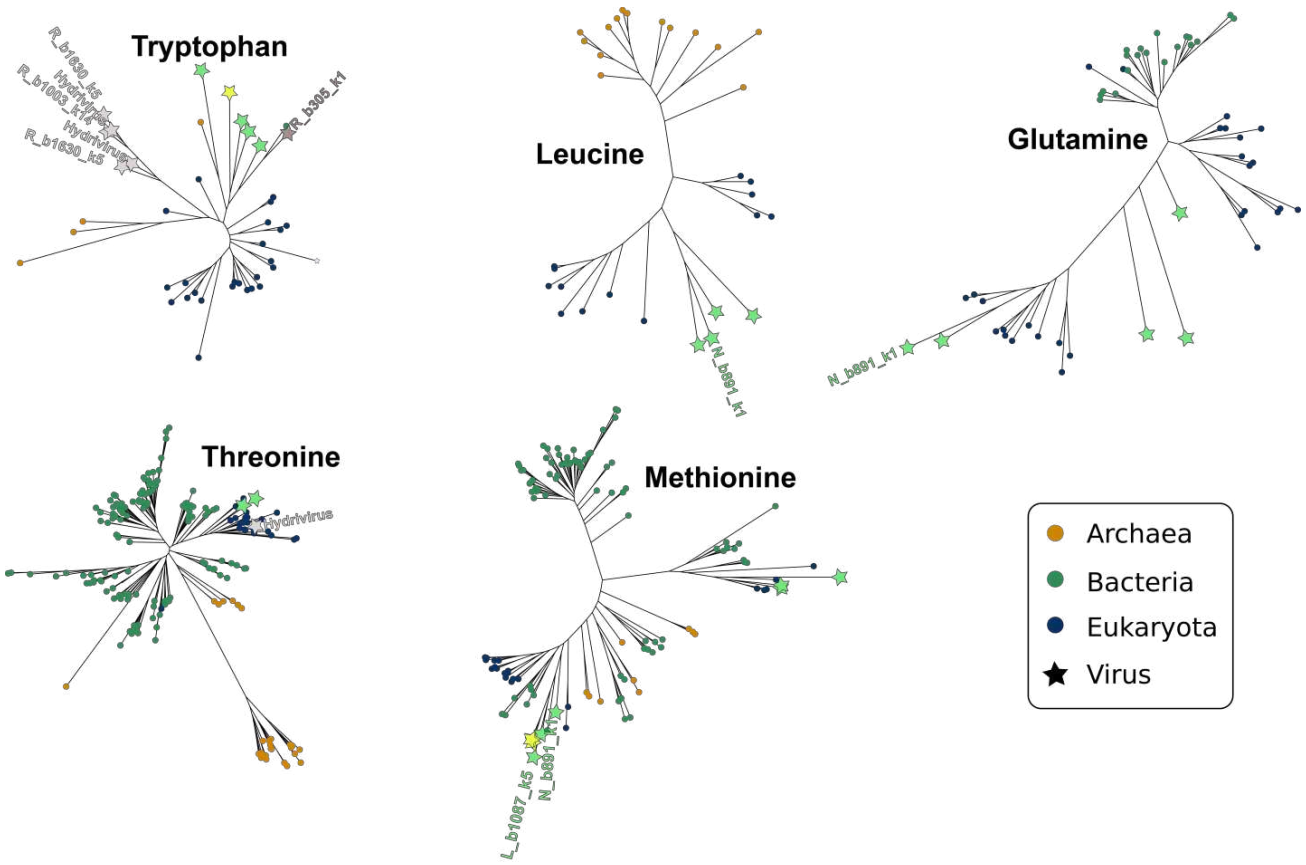
Supplementary Figure 15. Relative copy number of ORFs of selected functions within reference and permafrost *Nucleocytoviricota*

The bubbles sizes represent the relative number of copies of the functions found within a particular family. The bottom squares represent reference (white) or metagenomic (black) groups. The category "Other" includes families with less than 250 ORFs. Functions were separated according to the ones concerning DNA (A), RNA (B) or proteins (C).

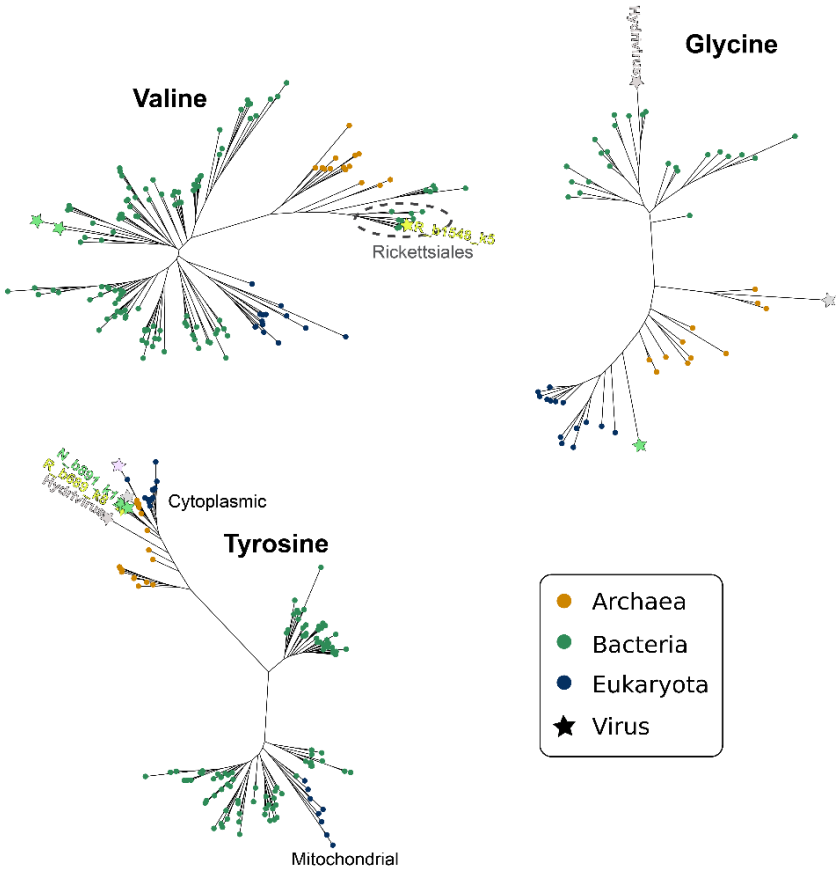


Supplementary Figure 16. Relative copy number of ORFs of selected functions within reference and permafrost *Nucleocytoviricota*

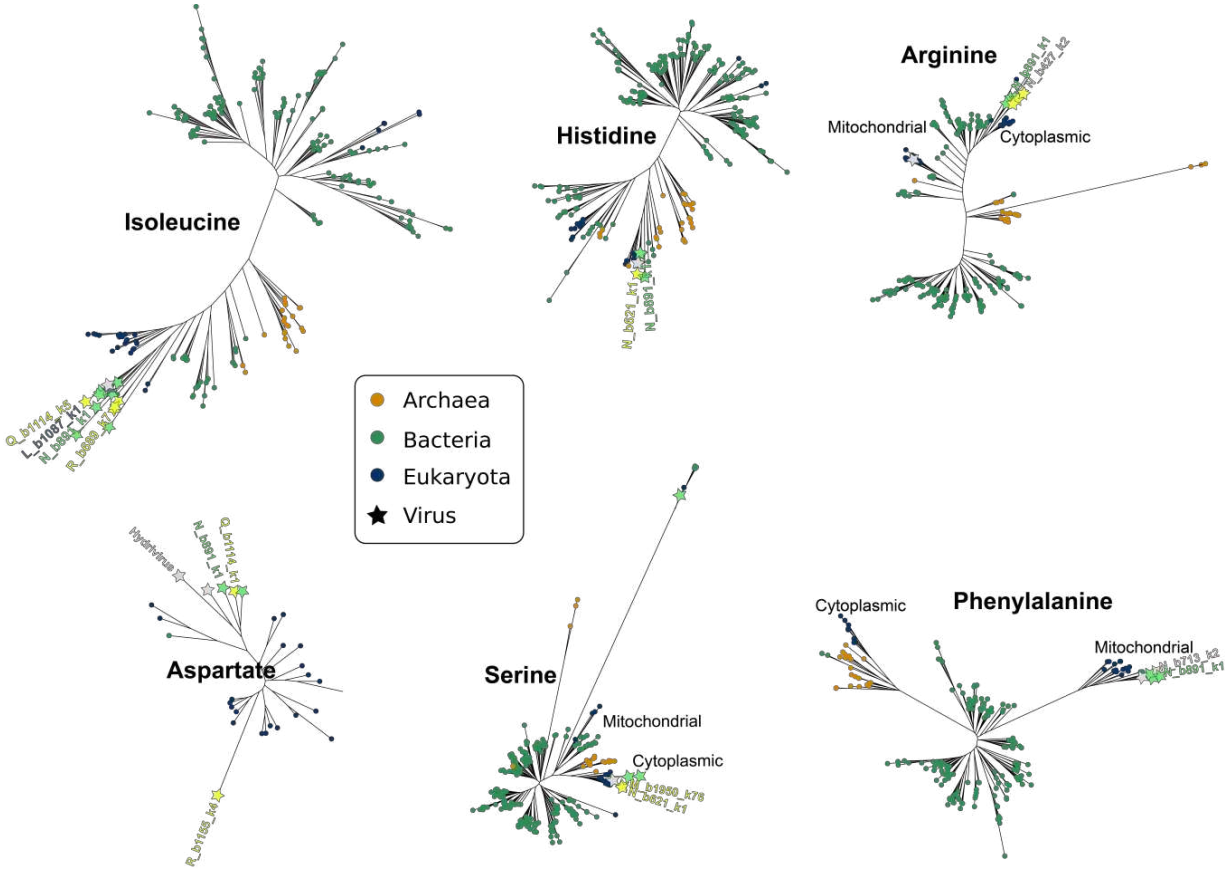
The bubbles sizes represent the relative number of copies of the functions found within a particular family. The bottom squares represent reference (white) or metagenomic (black) groups. The category "Other" includes families with less than 250 ORFs. Functions were separated into metabolism (A), cellular functions (B) and viral functions (C).



Supplementary Figure 17. Evidence of Aminoacyl-tRNA synthases exchanged between Eukaryota and *Nucleocytoviricota*
All trees are monophyletic according to their larger phylogenies (see Methods). Aminoacyl-tRNA synthetases were retrieved from IPR001412 and IPR006195 InterPro families to which were added *Nucleocytoviricota* sequence and their respective matches against the nr database.



Supplementary Figure 18. Evidence of Aminoacyl-tRNA synthases exchanged between Prokaryotes and *Nucleocytoviricota*
All trees are monophyletic according to their larger phylogenies (see Methods). Aminoacyl-tRNA synthetases were retrieved from IPR001412 and IPR006195 InterPro families to which were added *Nucleocytoviricota* sequence and their respective matches against the nr database.



Supplementary Figure 19. Potential Aminoacyl-tRNA synthase exchanges between viruses
All trees are monophyletic according to their larger phylogenies (see Methods). Aminoacyl-tRNA synthetases were retrieved from IPR001412 and IPR006195 InterPro families to which were added *Nucleocytoviricota* sequence and their respective matches against the nr database.

Supplementary Table 1. Analyzed permafrost datasets

Cryosol type	Sample code	Locality	Type	Depth (m)	Dating (103 y)	Million reads	Dataset accession
Kamchatka (surface)	C	Kronotsky river bank	Vegetation-free soil	0	0	677	ERS7649018
	D	Kizimen volcano	Tundra soil	0	0	642	ERS7649019
	E	Shapina river bank	Vegetation-free soil	0	0	685	ERS7649020
Permafrost (Yukechi Alas area)	L	Under dry Alas	Core	12	28	645	ERS7649022
	N			16	45	675	ERS7649024
	R			19	42	628	ERS7649028
	P	Dry Yedoma		11	36	616	ERS7649026
	K			16	49	569	ERS7649021
	Q	Under lake (4.5m water)		6	0.053	635	ERS7649027
	M			16	48.5	649	ERS7649023
	O			19	40	586	ERS7649025

Supplementary Table 2. Assembly statistics all datasets combined
The statistical test used to compare both assemblies in contig length is a two-sided Wilcoxon rank sum test.

	Contigs over 1 kb			Contigs over 10 kb		
	Total assembled nt	**Contigs**	**Mean (median) length**	**Total assembled nt**	**Contigs**	**Mean (median) length**
1st assembly round	1.97E-10	8,698,362	2,265 (1,473)	3.14E9	149,459	20,983 (14,660)
2nd assembly round	1.21E-10	2,720,241	4,432 (2,448)	4.51E9	185,996	24,274 (15,607)
P-value Wilcoxon test (2nd vs 1st assemblies)			<2.2E-16			<2.2E-16

Supplementary Table 3. Evalue thresholds calculated to select marker genes for phylogeny through psiblast

Marker gene	Evalue threshold
mcp	2.3875E-45
pATPase	1.72E-31
polB	1.67E-53
primase	1.4675E-36
rpo1	1.36E-68
rpo2	1.79E-74
vltF3	0.000000001065

IV. Les protéines majeures de capside métagénomique aident à comprendre le lien entre celle des *Pithoviridae* et des autres virus géants

Grâce aux séquences métagénomiques nous pouvons tenter de reconstituer l'histoire évolutive des gènes viraux. Ici nous allons nous intéresser plus spécifiquement à la MCP.

Son gène est présent chez tous les *Nucleocytoviricota* isolés à l'exception des *Pandoraviridae*. Chez les *Pithoviridae*, la protéine n'est pas retrouvée dans la capside et son gène est très divergent. Pour confirmer encore une fois que son histoire évolutive est cohérente avec le clade des *Nucleocytoviricota* et ainsi compléter l'analyse phylogénétique, nous avons procédé à l'étude de réseaux d'alignements protéiques avec BLASTP. Ainsi, nous pouvons voir le lien entre toutes les séquences de manière plus fine.

Nous avons procédé par un alignement de toutes les MCP prédites dans les séquences métagénomiques virales et de référence deux-à-deux, par BLASTP. Le résultat de ce BLAST a servi à construire des réseaux de similarité de protéines avec l'outil Cytoscape. Nous avons fait progressivement baisser le seuil de e-valeur à partir duquel un lien entre deux séquences est visible dans le réseau pour observer la dynamique de ce dernier. Les réseaux sont tous représentés selon l'algorithme « edge-weighted Spring Embedded layout » en prenant le bitscore comme poids du lien (Figure 24).

Le premier constat est que les MCP des *Pithoviridae* ne s'alignent pas avec les autres en l'absence de séquences métagénomiques en appliquant une stringence classique (Figure 24A). Ces dernières aident donc à faire le lien entre les MCP des *Pithoviridae* et celle des *Nucleocytoviricota* (Figure 24B) et ainsi reconnecter des gènes très divergents. En augmentant le seuil de stringence, les *Pithoviridae* et *Orpheoviridae* se séparent du reste du réseau jusqu'à devenir indépendant (Figure 24B et C). Cependant, une séquence reste avec les *Nucleocytoviricota* et en particulier les *Pimascovirales* après séparation avec le reste des *Pithoviridae* et *Orpheoviridae* (Figure 24C). Cette séquence appartient à un scaffold classé dans les Pitho_orpheo_div8 c'est-à-dire le clade le plus divergent. Il s'agit là d'un scaffold très probablement correctement classé puisque celui-ci possède les 6 gènes marqueurs.

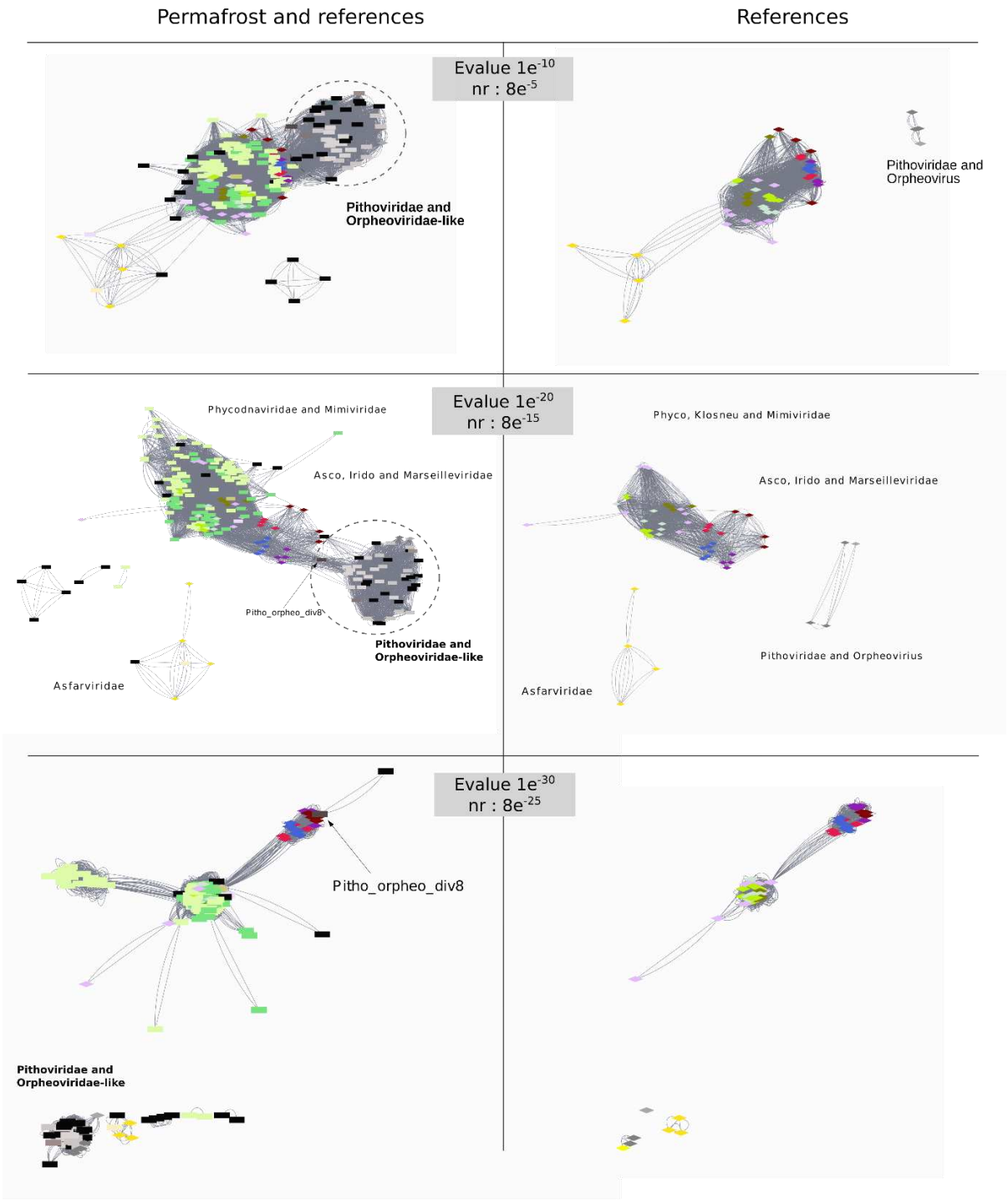


Figure 24 – Réseau des *Nucleocytoviricota* métagénomiques et de référence basé sur l'alignement de la protéine majeure de capside

Les réseaux ont été construits sur les résultats de BLASTP de toutes les séquences contre toutes en prenant le bitscore comme définissant la force du lien entre deux protéines. Trois seuils de e-valeur ont été testés. Pour chacune de ces e-valeurs, l'équivalent contre la base de données nr a été calculé en prenant en compte la différence de taille entre les deux bases de données. Les réseaux de droite ont été construits uniquement sur les MCP des génomes de référence. Les boites de couleur noir indiquent que le scaffold duquel le gène provient n'a pas été classifié. Les losanges sont pour les séquences des génomes de référence.

Face à cela nous pouvons émettre l'hypothèse que la reprogrammation fonctionnelle ou la perte de fonction de la MCP ait eu lieu après la divergence des Pitho_orpheo_div8 avec le reste du groupe puisque celle du scaffold de Pitho_orpheo_div8 a gardé une séquence plus proche de celle des autres *Nucleocytoviricota*. Nous supposons également que la MCP n'ait pas perdu sa fonction chez les *Pithoviridae* mais a changé de fonction puisque celle-ci est présente dans tous les génomes de virus isolés et dans de nombreuses séquences métagénomiques (Figure 24).

V. Présence de séquences associées aux *Pithoviridae* dans la base de données Mgnify

Comme précisé dans l'article II, de précédents travaux métagénomiques ont trouvé très peu ou pas de *Pithoviridae* dans les données analysées, pour la plupart aquatiques. Cela contraste énormément avec nos résultats, comme avec ceux de Bäckström et al. (2019). Les bases de données métagénomiques mondiales sont fortement enrichies en données provenant de milieux aquatiques par rapport aux données de métagénomique du sol. Plus un microbiome est complexe, plus les séquences assemblées provenant de ces métagénomes sont courtes et incomplètes et plus le « binning » (regroupement de séquences appartenant supposément au même organisme) est difficile. C'est pourquoi les virus géants provenant du sol se voient désavantagés dans ces grands travaux métagénomiques. Une étude sur les virus géants d'un sol de forêt applique également le binning (Schulz et al. 2018), méthode idéale pour étudier de nouveaux génomes complets mais moins bonne pour analyser une diversité générale. Pour tenter de comparer la diversité de nos échantillons avec des virus géants d'environnements terrestres, nous avons donc extrait des séquences de *Nucleocytoviricota* de la base de données IMG/M du JGI en ne téléchargeant que les données terrestres (article II, figure supplémentaire 12).

Cependant un doute subsiste ; les environnements marins sont-ils réellement pauvres en *Pithoviridae* ou est-ce une différence de méthode d'analyse qui expliquerait la diversité observée ? Pour répondre à cette question nous allons appliquer notre méthode à la base de données Mgnify de l'EBI. Pour ce faire nous avons extrait les gènes issus de données environnementales. La grande majorité des données provient de milieux aquatiques. La base de données complète téléchargeable n'est disponible qu'en séquences protéique, nous n'avons donc pas accès aux contigs. C'est pourquoi nous avons dû estimer la taille des contigs en prenant la position du dernier ORF du contig. Seul les contigs ayant une taille estimée de plus de 10 kb ont été analysés. Pour finir, nous avons appliqué la même méthode que celle

présentée dans l'article II au sujet des données du JGI pour extraire les séquences virales et calculer la phylogénie.

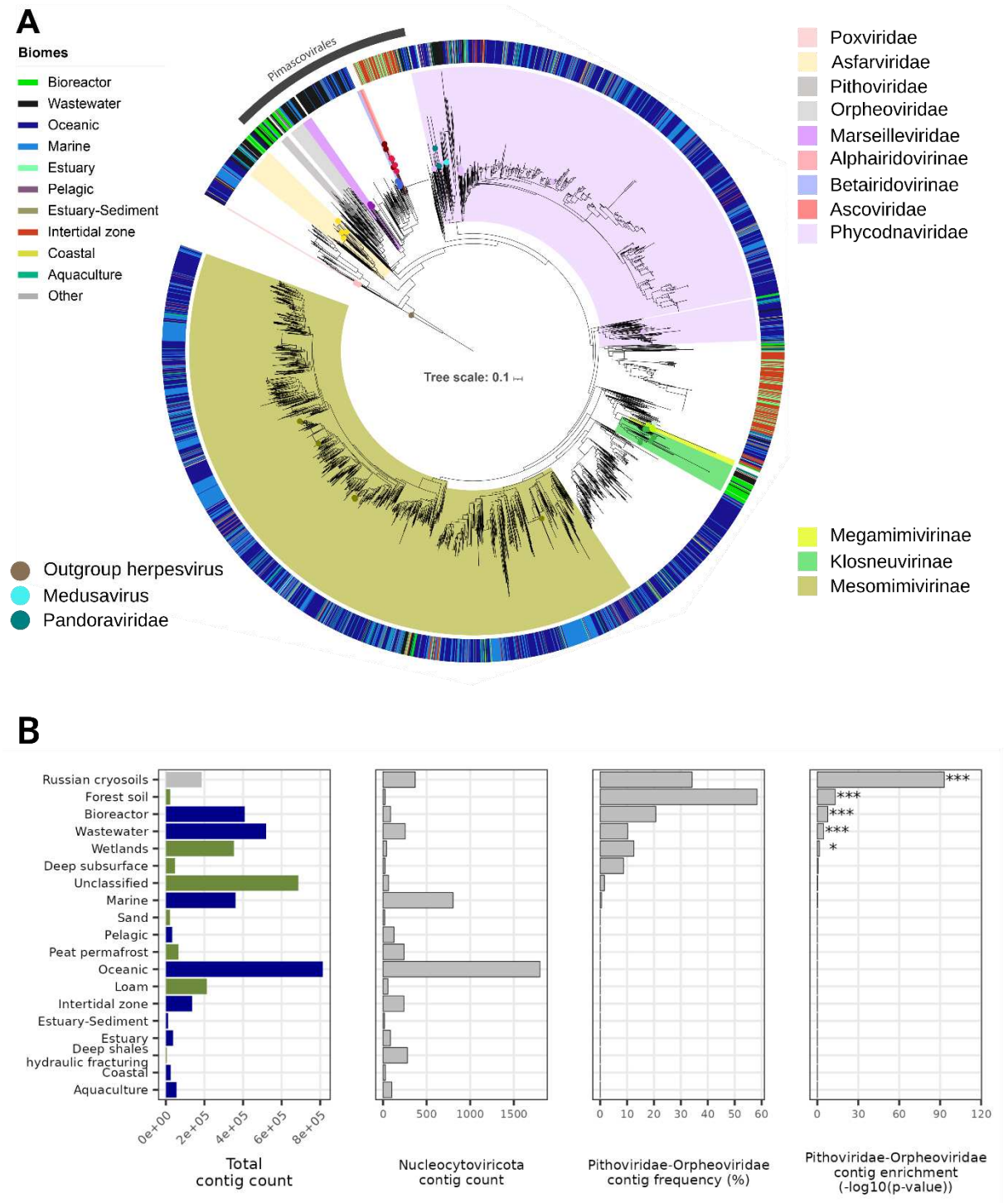


Figure 25 – *Nucleocytoviricota* dans la base de données Mgnify et les jeux de de données terrestres du JGI

(A) Pour compléter la Figure supplémentaire 11 de l'article II, une phylogénie des séquences de *Nucleocytoviricota* de la base de données Mgnify de l'EBI a été réalisée. (B) Ainsi, l'enrichissement des *Pithoviridae* au sein des différents biomes terrestres en vert a pu être comparé par des tests exacts de Fisher, également à des biomes aquatiques ou sédimentaires marins en bleu.

La Figure 25A et B, confirme que les *Pithoviridae* et *Orpheoviridae* sont presque absents des milieux marins et océaniques contrairement aux eaux usées, aux sols de forêt ou, comme conclu, au pergélisol sibérien (article II). Ainsi, les familles proches des *Pithoviridae* ont une abondance et une diversité surprenante dans les échantillons étudiés en comparaison aux autres virus géants et cela ne s'explique pas par la méthode appliquée mais par les données elles-mêmes. Il est important de noter cependant que nous ne savons pas comment l'extraction d'ADN a été faite pour générer les données téléchargées. Pour nos échantillons, du dithiothréitol (DTT) a été ajouté permettant d'ouvrir les particules virales, particulièrement récalcitrantes dans le cas des virus géants (Eugène Christo-Foroux 2020).

VI. Présence des *Pithoviridae* connus dans les échantillons de pergélisol

Nous remarquons que les séquences virales métagénomiques ne s'insèrent pas entre les *Pithoviridae* isolés dans l'arbre mais autour (article II Figure 3, article II Figure supplémentaire 11 et Figure 25), indiquant que la diversité est mal capturée par les virus isolés. Cela est surprenant sachant que Cedratvirus kamchatka provient de l'échantillon étudié D. Nous nous attendrions donc à voir au moins une séquence s'insérer entre *Cedratvirus A11* et *Pithovirus sibericum* si *Cedratvirus kamchatka* avait été assemblé dans cet échantillon.

Pour vérifier la présence de *Pithoviridae* isolés et issus de la métagénomique (Bäckström et al. 2019; Moniruzzaman et al. 2019; Schulz et al. 2020) dans le pergélisol nous avons procédé à l'alignement des lectures métagénomiques contre les génomes connus. Avant l'alignement, les régions de basse complexité ont été masquées par l'outil dustmasker. Pour l'alignement c'est l'outil bowtie2 qui a été utilisé avec l'option « --very-sensitive ». Les alignements ont ensuite été filtrés pour retirer toutes les lectures inférieures à 30 nucléotides et d'une qualité d'alignement inférieure à 30. La commande « samtools coverage » a permis de retrouver le nombre de lectures associées à chaque contig ainsi que la proportion du contig couvert.

Pithovirus est très spécifique de l'échantillon duquel il provient puisqu'il n'est retrouvé nulle part ailleurs. Les autres *Pithoviridae* sont couverts par une petite poignée de lectures dans les échantillons profonds ou de surface (Figure 26A). *Cedratvirus kamchatka* n'est pas retrouvé dans son échantillon d'origine (D). Cependant, il est surprenant d'y retrouver *Brazilian cedratvirus* avec plus de 1400 lectures alors qu'il n'a pas été retrouvé dans les assemblages métagénomiques. Ces lectures ne couvrent que 0.03% du génome soit 142 bases. Ces 142 bases sont au sein d'une région répétée (voir plus loin, article III, p. 116) qui n'est pas de basse complexité.

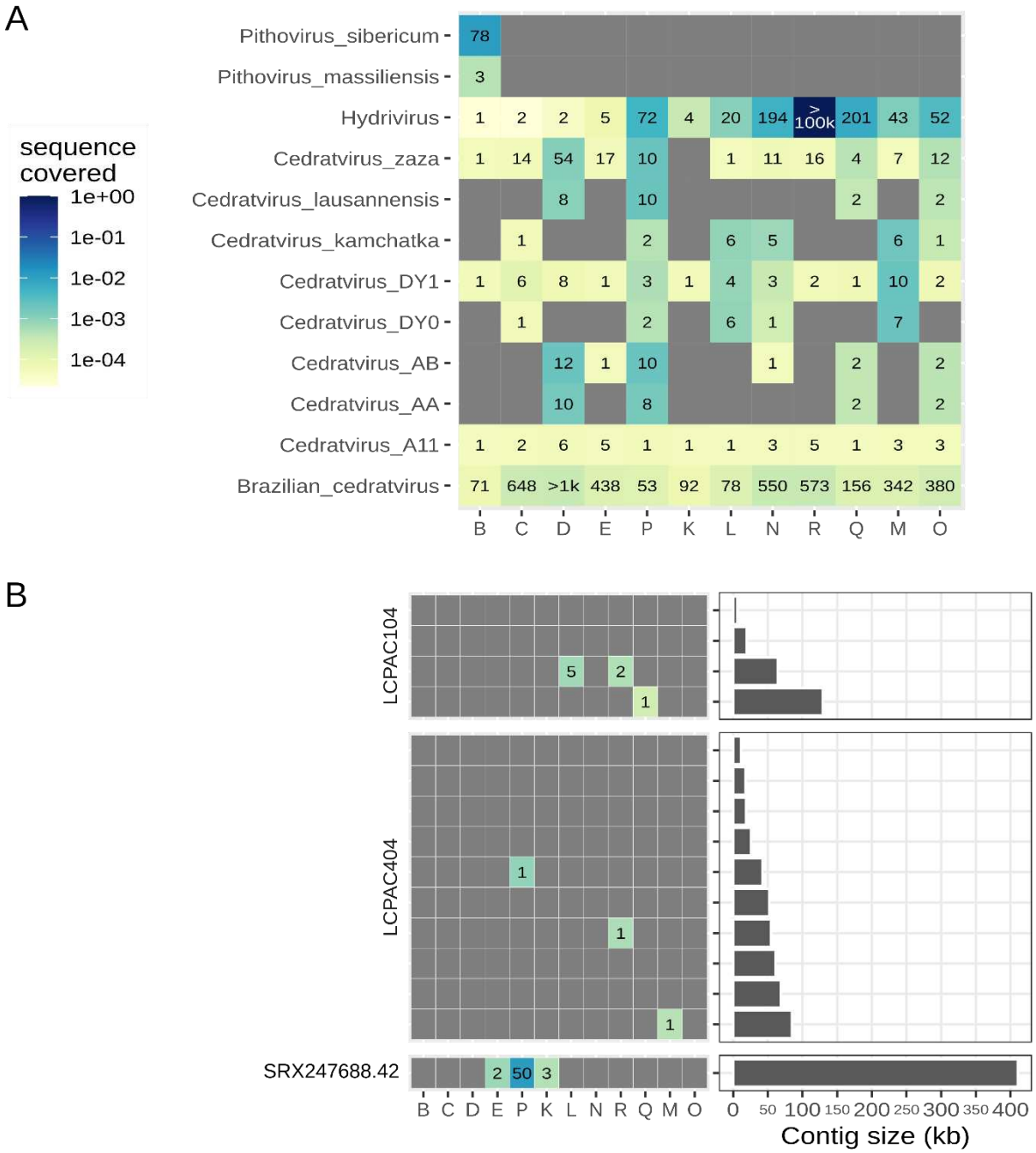


Figure 26 – Alignement des lectures métagénomiques sur les génomes des *Pithoviridae* connus

La couleur des cases des heatmaps font appel à la proportion de la séquence couverte par au moins une lecture. (A) Les lectures des jeux de données du pergélisol ont été alignées au *Pithoviridae* isolés ainsi qu'à Hydrivirus. Les nombres inscrits dans les cases indiquent le nombre de lectures associées à chaque génome. (B) Les lectures ont également été alignées aux 16 *Pithoviridae* et *Orpheoviridae* métagénomiques issus pour la plupart de l'étude de Bäckström et al. (2019), et deux proviennent des études de Schulz et al. (2020) et Moniruzzaman et al. (2019). Chaque bloc est un « bin » et chaque ligne indique un contig du « bin ». Seul 3 bins ont eu au moins 3 lectures associées et sont donc présentés. Pithovirus LCPAC404 et LCPAC104 proviennent de Bäckström et al. (2019) et SRX247688.42 est le seul *Pithovirus* trouvé par Moniruzzaman et al. (2019).

Plusieurs séquences retrouvées dans le pergélisol ont parmi leur meilleur résultat de BLAST des séquences de *Pithoviridae* divergents de sédiments océaniques profonds. Ainsi, nous sommes partis à leur recherche dans nos échantillons en incluant deux génomes de deux grandes études métagénomiques de *Nucleocytoviricota*. Sur 16 génomes, seuls 3 ont été retrouvés (Figure 26B). Le plus abondant est Pithovirus SRX247688.42 avec 50 lectures de l'échantillon P couvrant presque 5 kb. Les *Pithoviridae* divergents retrouvés dans le pergélisol sont donc différents de ceux de l'étude des sédiments marins profonds (Bäckström et al. 2019). Il faut noter ici que le terme de Pithovirus pour désigner LCDPAC01 et les autres séquences est en réalité abusif car ce ne sont pas des Pithovirus proches de ceux isolés (Figure 34 en annexe). Le manque de données sur cette famille et sur leurs clades frères est tel que l'on retrouve dans les publications le terme Pithovirus ainsi que Pitho-like pour désigner tout virus dont le plus proche parent serait un *Pithoviridae* ou un *Orpheoviridae* bien que ceux-ci soient dans les faits très divergents.

Chapitre 2. Génomique comparative des *Pithoviridae*

VII. Contexte

Les *Pithoviridae* sont très divergents par rapport aux autres *Nucleocytoviricota* isolés. Seul 32.5 % de gènes de P. sibericum avaient un homologue potentiel dans les bases de données en 2014. Cette proportion est très basse mais il n'est pas inhabituel d'avoir peu de gènes connus chez les premiers représentants d'une famille de virus géants. De la même manière, les homologues de ces protéines étaient répartis équitablement entre les virus, eucaryotes et bactéries (Legendre et al. 2014). Quand un Cedratvirus a été analysé pour la première fois il avait tout de même toujours 30.8 % de gènes sans homologue connu. Ceux ayant un homologue ressemblent surtout à P. sibericum mais il y a tout de même 108 gènes dont le meilleur alignement correspond un eucaryote (Andreani et al. 2016).

	Taille (kb)	***Groupe***	***Publication***
Pithovirus sibericum	610	Pithovirus	(Legendre et al. 2014)
Pithovirus massiliensis	686	Pithovirus	(Levasseur et al. 2016)
Pithovirus mammoth	610	Pithovirus	Non publié
Cedratvirus A11	589	Clade A	(Andreani et al. 2016)
Cedratvirus lausannensis	575		(Bertelli et al. 2017)
Cedratvirus Zaza IHUMI	561		(Rodrigues et al. 2018)
Cedratvirus CAY AA	570		Non publié
Cedratvirus CAY AB	568		Non publié
Cedratvirus Kamtchatka	473	Clade B	(Jeudy et al. 2020)
Cedratvirus DY0	468 – 469		Non publié
Cedratvirus DY1	470 – 472		Non publié
Brazilian Cedratvirus IHUMI	460	Clade C	(Rodrigues et al. 2018)
Orpheovirus IHUMI-LCC2	1474	Clade frère	(Andreani et al. 2018)
Hydrivirus	1600		Article II

Tableau 2 – Génomes étudiés dans le cadre de la génomique comparative
Les deux tailles de génomes affichées des Cedratvirus DY0 et DY1 correspondent aux tailles avant et après scaffolding

Ces dernières années le laboratoire IGS a isolé six nouveaux *Pithoviridae* : 5 Cedratvirus et un Pithovirus (Tableau 2). Ils proviennent tous d'échantillons de sol. Ces virus ont été isolés par Jean-Marie Alempic et Audrey Lartigue (IGS) sur des cultures de *A. castellanii*. Sur cette base, nous pouvons étudier les *Pithoviridae* par génomique comparative. Nous pouvons utiliser Orpheovirus pour mettre en perspective ce que nous observons chez les *Pithoviridae* mais aussi Hydrivirus issu de la métagénomique. En effet, celui-ci a été estimé complet et en un

seul morceau (article II). Enfin, Marseillevirus a été utilisé dans certains cas en tant que groupe externe.

Dans cette partie nous allons tenter de mettre en lumière les mécanismes évolutifs et l'asymétrie induite dans les génomes des *Pithoviridae*. Nous avons été largement inspirés par les travaux précédents démontrant une partie conservative et une partie créatrice dans les génomes des *Marseilleviridae* (Blanca et al. 2020) et des *Pandoraviridae* (Legendre et al. 2018; Bisio et al. 2022). Basé sur l'analyse d'orthologie, nous avons également cherché l'origine de chaque gène : duplication, transfert de gène horizontal, gène inconnu… et son niveau de conservation.

Des séquences répétées prenant une part importante du génome des Pithovirus, nous nous sommes également intéressés à ces séquences, leur structure et leurs effets sur l'évolution des régions codantes adjacentes. Pour ce faire, un outil spécifique a été créé pour permettre d'étudier précisément leur structure et délimiter ces régions sans apriori. Cet outil procède en six étapes : 1) Alignement du génome contre lui-même, 2) calcul d'un vecteur représentant un dotplot aplani, 3) délimitation des régions riches en unités répétées codé par Alain Schmitt de l'IGS, 4) délimitation des unités répétées au sein de ces régions aussi codé par Alain Schmitt, 5) regroupement de ces unités répétées et 6) regroupement des régions riches en séquences répétées.

Ensemble, ces résultats forment une première vue d'ensemble de l'évolution des *Pithoviridae*.

Avant de nous attaquer à l'étude de leurs génomes, il convient tout de même de confirmer expérimentalement la circularité génomique des *Pithoviridae* qui avait été inférée par bioinformatique. Cette première étape a été réalisée sous la supervision de Sandra Jeudy (IGS).

VIII. La circularité du génome de Cedratvirus kamchatka confirmée

Matériels et Méthodes

Cedratvirus kamchatka est un virus isolé sur une culture de *A. castellanii* à partir de l'échantillon D, du sol de toundra provenant du Kamchatka près du volcan Kizimen (Figure 23, chapitre 1). Les méthodes d'isolement, de concentration et d'extraction d'ADN ainsi que le séquençage et l'assemblage sont décrites dans l'article IV, de (Jeudy et al. 2020), en annexe (p. 180)

La circularité des *Pithoviridae* avait été démontrée bioinformatiquement mais jamais expérimentalement. Y compris la circularisation par PCR de Cedratvirus lausannensis avait échoué (Bertelli et al. 2017). Ainsi, dans le but d'obtenir une preuve formelle, nous avons réalisé une électrophorèse en champ pulsé (PFGE pour Pulse-Field Gel Electrophoresis). Le deuxième but de l'expérience était de confirmer ou non la présence d'un fragment d'ADN de quelques kilobases trouvé de manière récurrente dans plusieurs assemblages. Le choix des enzymes de restriction s'est fait avec l'outil restrict. Les enzymes sélectionnées ont été NotI avec un fragment linéaire attendu, SgrAI avec trois fragments linéaires attendus et SacII avec

deux fragments linéaires attendus. En plus de ces fragments, si l'assemblage est correct il nous faut trouver un fragment de six kb.

Pour procéder au PFGE, des bouchons (« plugs ») d'aliquots de particules virales dans de l'agarose à 2% et PBS sont formés et incubés 24h à 50°C dans du tampon PFGE (50 mM de Tris-HCl pH 8, 50 mM d'EDTA 1% (v/v) N-laurylsarcosine et 1 mg/mL de protéinase K) avec secousse. Ce tampon est renouvelé deux fois et au dernier renouvellement, 1µL de DTT est ajouté. Les bouchons sont ensuite lavés à l'eau stérile et au tampon TE (10mM de Tris-HCl pH 8, 1 mM d'EDTA). Pour digérer l'ADN, le bouchon d'intérêt a été incubé 1h au tampon TE et 1 mM de PMSF. Après rinçage, ils sont incubés dans du tampon TE puis du tampon de restriction dans un bac de glace. Après 30 min, les bouchons ont été incubés avec l'enzyme de restriction à température recommandée pour 14 heures avec renouvellement du tampon avec enzyme. Après rinçage, le bouchon fut incubé avec 500 µL de tampon PFGE contenant 1 mg/mL de protéinase K pour 2 heures à 50 °C puis lavé. Les bouchons, qu'ils contiennent de l'ADN digéré ou non ont été insérés dans les puits du gel d'agarose à 1% préparé dans du TBE à 0.5x. Du tampon TBE à 0.5x est utilisé pour faire fonctionner la machine à électrophorèse. Nous avons utilisé les programmes automatiques dont celui pour 10 kb à 500 kb à 132 mA. Le gel est ensuite révélé dans du BET.

Résultats

L'assemblage du génome de *Cedratvirus kamchatka* montre un petit génome comparé aux autres *Pithoviridae* avec environ 100 kb de moins que les *Cedratvirus* du clade A (Tableau 2). Il présente également un contig supplémentaire de 6 kb.

Dans un premier temps nous pouvons constater par un programme de PFGE adapté pour les génomes ou fragments génomiques de 1 à 750 kb qu'il n'y a pas de bande autour de 6 kb. Ce gel bien résolu sous les 20 kb n'a en fait montré aucun fragment de petite taille. Un alignement des lectures au génome assemblé montre que le deuxième contig est 78 fois moins couvert que le grand contig circulaire. De plus, les contigs supplémentaires apparaissant dans plusieurs assemblages ne sont pas strictement identiques. De tous ces éléments, nous en avons conclu que ce contig était un artefact de l'assemblage.

Dans un deuxième temps nous avons donc cherché à mieux résoudre la fraction du PFGE contenant les grandes tailles de génome. Cela a été permis par un programme pour 10 à 750 kb (Figure 27). Non digéré, l'ADN ne migre pas. Digéré sur un site de coupure unique, on obtient un fragment de génome de la taille attendue. Clivé sur trois sites, on obtient également les fragments de taille attendue, mais le plus petit fragment, sous le marqueur à 49 kb, est double. La bande est moins intense que les deux autres. Le clivage par SacII est plus surprenant. Alors que nous attendions un clivage sur deux sites, on obtient plus de fragments. Une bande faible apparaît au niveau du marqueur 485 indiquant que l'enzyme n'a pas parfaitement fonctionné sur un des deux sites de clivage. On obtient également les fragments attendus mais la bande au niveau du marqueur à 194 kb est triple. Une dernière bande faible apparaît en dessous du marqueur à 97 kb. Pour expliquer ces résultats ainsi que la double bande suite au clivage par SgrAI on peut émettre l'hypothèse d'une variabilité de souche dans l'échantillon de Cedratvirus kamchatka. Cela expliquerait que la double bande produite par le clivage de SgrAI soit moins intense que les deux autres. Pour SacII, on peut penser qu'une

souche dont le génome a été assemblé, génère deux fragments : un autour de 291 et l'autre autour de 194. La deuxième souche génèrerait trois fragments : deux autour de 194 et une autour de 97 kb. Dans les deux cas le total est de 485 kb.

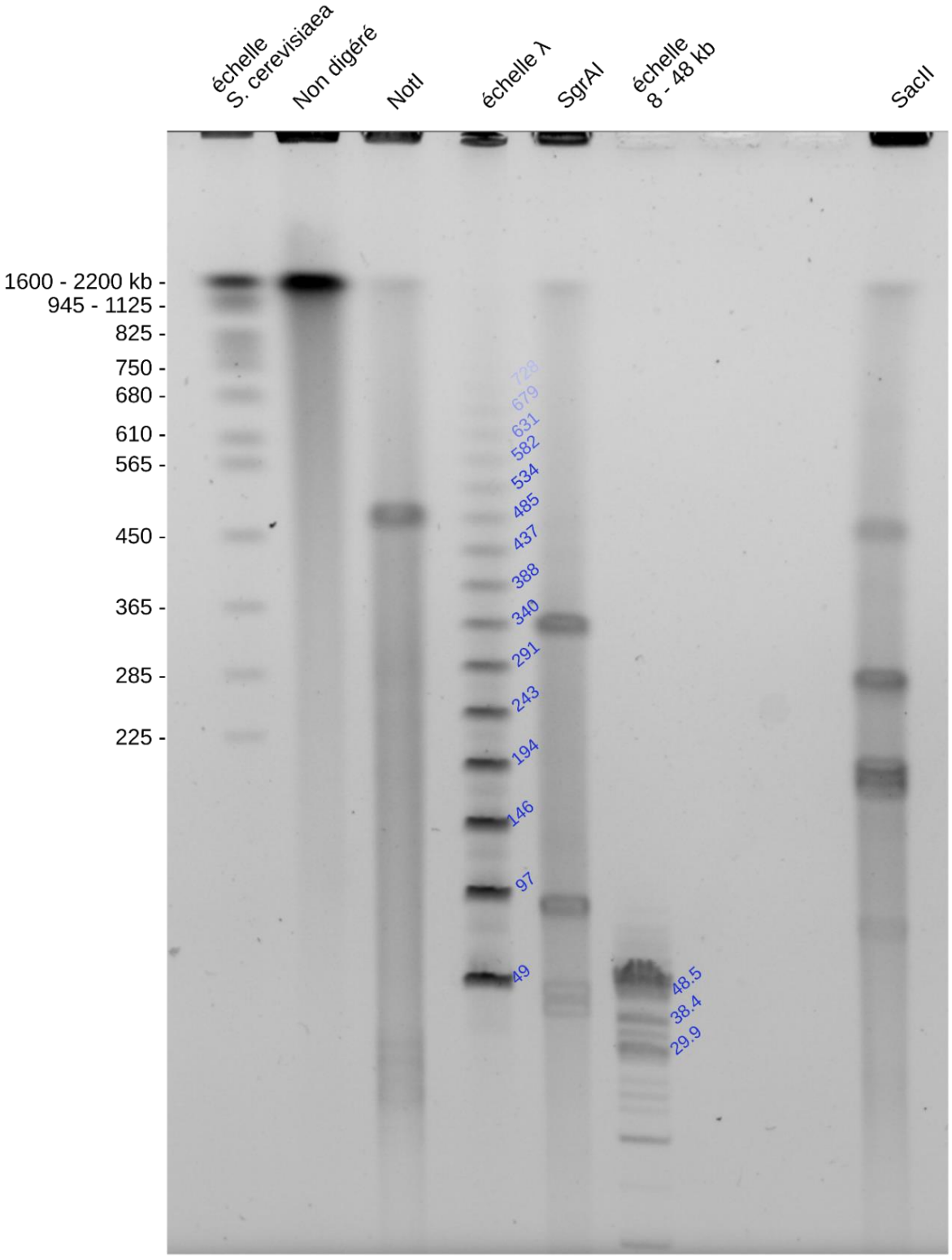


Figure 27 - Gel d'électrophorèse en champ pulsé du génome de Cedratvirus kamchatka
La migration du génome digéré ou non a eu lieu avec un programme adapté pour résoudre de 10 à 750 kb. Une image de ce gel a été publiée dans (Jeudy et al. 2020)

La bande unique après digestion par NotI et l'absence de migration de l'ADN non digéré confirment la circularité du génome de Cedratvirus kamchatka. Par extrapolation, nous avons donc la confirmation formelle de la circularité des génomes de *Pithoviridae*. Cette information sera utilisée dans l'analyse de génomique comparative qui suit.

IX. Génomique comparative, article III

L'article qui suit n'est pas finalisé. Il va donc évoluer selon le travail et recommandations des co-auteurs.

Working Draft

Evolutionary history of the drastically transposon-invaded *Pithoviridae* genomes

Sofia Rigou[1], Alain Schmitt[1], Jean-Marie Alempic[1], Audrey Lartigue[1], Peter Vendlosky[1], Eugène Christo-Forroux[1], Chantal Abergel[1], Jean-Michel Claverie[1], Matthieu Legendre[1,*]

[1]Aix–Marseille University, Centre National de la Recherche Scientifique, Information Génomique & Structurale, Unité Mixte de Recherche 7256 (Institut de Microbiologie de la Méditerranée, FR3479), 13288 Marseille Cedex 9, France

[*]Correspondence: legendre@igs.cnrs-mrs.fr

Introduction

Pithoviridae are amoeba-infecting giant viruses possessing the largest known viral particles. The prototype of the family, Pithovirus sibericum, was recovered almost 10 years ago from a 30'000-y-old permafrost sample (Legendre et al. 2014). Following this discovery, 6 additional isolates, all infecting *Acanthamoeba castellanii*, have been sequenced (Andreani et al. 2016; Levasseur et al. 2016; Bertelli et al. 2017; Rodrigues et al. 2018; Jeudy et al. 2020). Their dsDNA circular genomes range from 460 to 686 kb. The *Pithoviridae* are composed of two main clades: the pithoviruses and the cedratviruses. Both possess ovoid-shaped virions, capped by a cork-like structure at one extremity for the former and at both extremities for the latter. Orpheovirus, the closest relative to the family, infecting *Vermamoeba vermiformis*, also has an ovoid-shape virion but a much larger (1.6 Mb) genome (Andreani et al. 2018).

Pithoviridae have mostly been isolated from permafrost samples (Legendre et al. 2014; Jeudy et al. 2020), sewage samples and drinking water plants (Levasseur et al. 2016; Bertelli et al. 2017; dos Santos Silva et al. 2018). Metagenomic surveys have revealed *Pithoviridae*-like sequences in deep sea sediments (Bäckström et al. 2019) and in permafrost (Rigou et al. 2022). In addition, two such bins have been assembled from forest soil samples (Schulz et al. 2018). In every case, phylogeny of the metagenomic viral sequences showed that they are closely related to the isolated *Pithoviridae* while branching outside the clade, suggesting that the *Pithoviridae* family is highly diverse (Rigou et al. 2022).

Genomic gigantism has been observed several times in the virosphere, whether they infect prokaryotes, such as "huge" (Al-Shayeb et al. 2020) and "jumbo" phages (Yuan et Gao 2017), or eukaryotes like some members of the *Nucleocytoviricota* of which *Pithoviridae* are part of. But its origin remains a mystery as most giant viruses' ORFs have no known origin. In *Nucleocytoviricota*, massive horizontal gene transfers from their host (Moreira et Brochier-Armanet 2008) and gene duplications (Filée et Chandler 2008) have been proposed as the driving force behind their expanded genome size. Another mechanism proposed in *Pandoraviridae* is *de novo* gene creation from intergenic regions (Legendre et al. 2018). Whatever the main evolutionary process at play, different families of giant viruses exhibit an asymmetry in their genomes, by having one "creative" and one "conservative" half. This pattern is revealed by an unequal distribution of core genes, duplicated genes and genomic rearrangements, preferentially concentrated in one half of the genome (Legendre et al. 2018; Blanca et al. 2020; Christo-Foroux et al. 2020).

Another factor that might shape giant viruses' genomes are transposons. For instance different *Pandoraviridae* are known to harbor Miniature Inverted Transposable Elements (MITEs) (H.-H. Zhang et al. 2018). These are non-autonomous class II transposable elements composed of terminal inverted repeats separated by an internal sequence that lacks the transposase gene. Thus, they rely on an autonomous transposon for transposition (X. Y. Zhang et al. 2001). Their target sites are often as simple as AT dinucleotides that give rise to target site duplication (TSD) (Ge et al. 2017). In *Pandoravirus salinus*, the transposon probably associated to these MITEs has been found in the *A. castellanii* cellular host genome (Sun et al. 2015). The *Pithovirus sibericum*'s genome also contains 140 nucleotides long palindromic repeated sequences in non-coding regions that cover as much as 20 % of the whole sequence (Legendre et al. 2014). The nature of these repeated sequences has not been elucidated yet, although they have also been found in the genome of *Pithovirus massiliensis* (Levasseur et al. 2016). Strangely, Cedratviruses are completely devoid of such sequences (Andreani et al. 2016).

In this study, we present five newly isolated and sequenced viruses from the *Pithoviridae* family: Cedratvirus AA, Cedratvirus AB, Cedratvirus DY0, Cedratvirus DY1 and Pithovirus mammoth. In complement with previously published *Pithoviridae* sequences, we now have a sufficient number of sequenced genomes to get insight into the gene distribution and the evolution of the family through comparative genomics. In addition, an in-depth study of pithoviruses' repeats using a dedicated tool tells us that they correspond to highly structured MITEs that massively colonized their host genomes.

Materials & Methods

Isolation of five new *Pithoviridae*

Cedratvirus AA and Cedratvirus AB were isolated from soil samples from south of France (Parc Borély in Marseille), in February 2017. Pithovirus mammoth, Cedratvirus DY0 and Cedratvirus DY1 come from Siberian permafrost samples. The isolation and cloning of viruses from soil samples was done following the protocol as given by (Christo-Foroux et al. 2020).

Genome sequencing, assembly and annotation

All cedratviruses were sequenced using Illumina short reads and Pithovirus mammoth using a combination of Illumina short reads and Nanopore long reads. Cedratvirus DY0 was assembled after removing sequencing reads mapping to a contaminant Pandoravirus using Bowtie 2. Cedratvirus DY0 and Cedratvirus DY1 reads were assembled using SPAdes v 3.14 (Prjibelski et al. 2020) with options --careful and -k 15,17,19,21,29,33,41,55,63,71,91,101,115. The scaffolding was then performed by RaGOO (Alonge et al. 2019) using Cedratvirus kamchatka as template. Cedratvirus AA and Cedratvirus AB were assembled using SPAdes v 3.9.1 and v 3.9.0, respectively, with the –careful option. Finally, the Pithovirus mammoth sequence was assembled using Unicycler (Wick et al. 2017) v 0.4.8 with Illumina short reads and nanopore long reads larger than 40 kb.

The 3 pithoviruses and the 9 cedratviruses genomic sequences were then artificially linearized to start at the same position. The accessions of the previously published genomes used in this study can be found in Table S1A and S1B.

For functional annotation, genes were predicted using Genemark (Besemer, Lomsadze, et Borodovsky 2001) with option –virus. ORFs over 50 amino acids were kept for publication and ORFs over 100 amino acids were used for comparative genomic analysis.

ORFs were manually functionally annotated based primarily on the protein domains using InterProScan (v5.39-77.0, databases PANTHER-14.1, Pfam-32.0,ProDom-2006.1, ProSitePatterns-2019_01, ProSiteProfiles-2019_01, SMART-7.1, TIGRFAM-15.0) (Jones et al. 2014) and CDsearch (Conserved Domain Database) (Lu et al. 2020). We also searched for viral specific functions using hmmsearch on the virus orthologous groups database (https://vogdb.org/). ORFs were compared to nr and swissprot using BLASTP (Altschul et al. 1990). Finally, we also checked transmembrane domains with Phobius (Käll, Krogh, et Sonnhammer 2004).

Relative synonymous codon usage as well as other gene and genome metrics were calculated using an in-house script partly relying on Biopython (Cock et al. 2009). CAI values were calculated with the CAI python function (Lee 2018).

Orthologous gene groups computation and phylogeny

A phylogenetic viral species tree was computed by OrthoFinder (v2.5.4) (Emms et Kelly 2019) using all available *Pithoviridae* genomes in addition to the Orpheovirus (Andreani et al. 2018), Hydrivirus (Rigou et al. 2022) and Marseillevirus genomes (Table S1). The tree has then been rooted according to the distantly related Marseillevirus (Boyer et al. 2009) as outgroup. Hierarchical Orthologous Groups (hereafter defined as HOGs) were then determined by OrthoFinder (v2.5.4) using this correctly rooted tree. A final phylogeny was inferred on the concatenated alignment of single copy HOGs by IQ-TREE (Nguyen et al. 2015) with the LG+F+G4 model and options -bb 5000 -bi 200.

Selection pressure on genes was estimated by the ratios of non-synonymous substitution rates (dN) to synonymous substitution rate (dS), calculated by codeml of the PAML v4.9 package (Yang 1997). Single copy genes from HOGs at the *Pithoviridae* node (excluding Orpheovirus and Hydrivirus) were retrieved and aligned with T-Coffee (Notredame, Higgins, et Heringa 2000). Codeml was given the alignment and the species phylogenetic tree to calculate dN/dS ratios. Three models were tested: the fixed model

with all the branches following the same ratio, a two-ratios model with one ratio for pithoviruses and one ratio for cedratviruses and a four ratios model with one model for pithoviruses and one ratio per cedratvirus clade. Finally, for each alignment, the best model was chosen according to likelihood-ratio tests considering the number of parameters estimated in each model. The p-value threshold was set to 0.05. Finally, the dN and dS values of each sequence was calculated as their respective sum from the root to the branch leading to the given species.

Cedratvirus core and pangenomes estimation

Core/pangenomes sizes were calculated on HOGs (Hierarchical Orthologous Groups) at the root node. Genomes were iteratively added with all possible combinations to simulate a dataset with 1 to 9 genomes. Here we used the presence/absence matrix of HOGs instead of gene counts in the original method published by (Tettelin et al. 2005). Data was processed on R (v4.04 (R Core Team 2021)).

For comparison, the ORF predictions, orthology analyses and core/pangenome estimations were performed on other viral families: *Pandoraviridae* (Table S1C), *Marseilleviridae* (Table S1D), Ranaviruses (Table S1E), *Megavirinae* (Table S1F). The outgroups used were respectively Mollivirus sibericum, Ambystoma tigrinum virus, Red seabream iridovirus and Chrysochromulina ericina virus.

Horizontal Gene Transfers identification

Horizontal gene transfers (HGTs) were identified based phylogenetic trees of each HOG complemented with homologous sequences that were retrieved using a two steps procedure. First, the sequences of each HOG were aligned using DIAMOND BLASTP (Buchfink, Xie, et Huson 2015) against the RefSeq database (from March 2019 (O'Leary et al. 2016)) with an e-value threshold of 1e-5, keeping only matches covering more than 50 % of the query. Up to 10 matches per domain (Bacteria, Archaea, Eukaryota and Viruses) were kept for each query and Cd-hit was applied on the retrieved sequences. Secondly, the resulting sequences were queried again against the RefSeq using DIAMOND with the same e-value threshold. A maximum of two proteins per domain (whose matches covered more than 80 % of the query) were kept at this step. The HOGs and selected sequences from the first and second rounds were aligned using MAFFT v7.475 (Katoh et Standley 2013) and phylogenetic trees were built by IQ-TREE with options -bb 1000 -bi 200 -m TEST. Each resulting phylogenetic tree was rooted by mad v2.2 (Tria, Landan, et Dagan 2017). Trees were finally visually inspected and HGT events counted when one or several *Pithoviridae* were within a bacterial, eukaryotic, archaeal or different viral clade.

Genome repeats detection and classification

A pipeline was developed to retrieve repeat-rich regions and individual repeats from Pithoviruses' genomes. The steps are: (1) genome-wide alignment, (2) flattened dotplot calculation, (3) repeat-rich regions delimitation, (4) individual repeats retrieval, (5) repeat clustering.

(1) genomes were aligned against themselves by BLASTN with an e-value threshold of 1e-10.
(2) for each position of the genome, the number of times it was aligned was counted resulting in a vector (y) ; similar to a flattened dotplot.
(3) A smooth vector (y_{ss}) was first estimated by sliding mean filtering with a window size of 500 nt. A detection threshold (τ) was calculated as $\tau = \overline{ys} * sensitivity^{-1}$, with a sensitivity coefficient set to 2.5. Repeat-rich regions were detected by comparing the vector y_s to τ. Repeat-rich regions were defined as regions where y_s is above the threshold τ. Each region's start and stop are thus the positions of intersections of y_s and τ.
(4) For each previously detected region, individual repeats were extracted using a smoothed derivative of y. Smoothing was applied before and after the derivation, this time with a window size of 20 nt. Then, the absolute value was taken in order to obtain the vector $|y_s|$. Then the local maxima were considered as repeat delimitations if above a cutoff set to 10.
(5) repeats are globally aligned to each other by needle of the EMBOSS suite (Rice, Longden, et Bleasby 2000). They are then ordered according to the mean distance (100 – needle identity percentage) to

their 10 closest neighbors. The first sequence becomes a reference sequence. Then, sequences are clustered together if they are at least 70 % identical to a reference or they become themselves a reference. Finally, clusters are merged together if over half of their respective sequences are at least 70 % identical. For visual inspection to infer repeat types and similarity in-between clusters, a matrix of dotplots presenting the alignments of reference sequences is drawn.

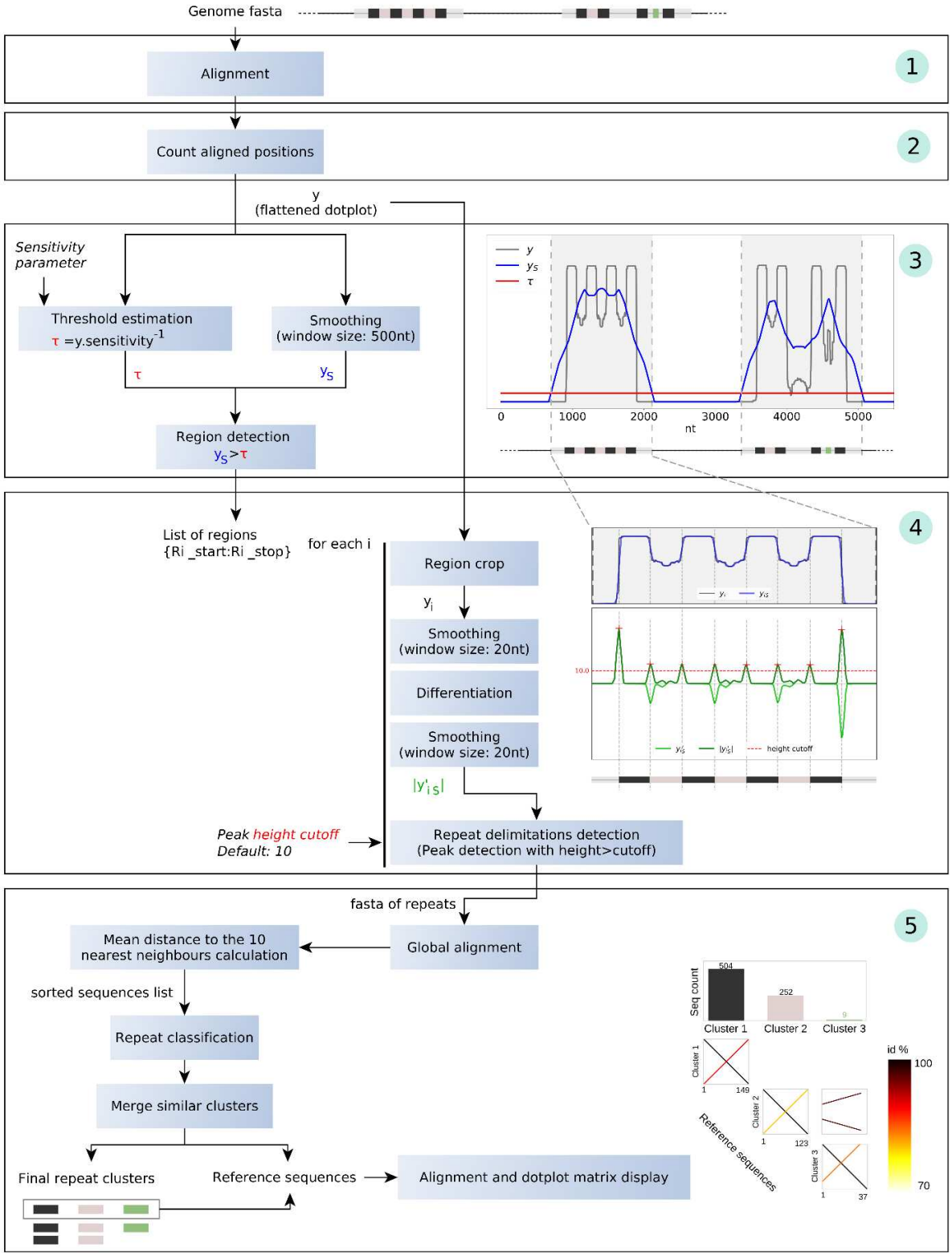


Figure 1. Workflow for repeat analysis

Steps one to five are represented within large boxes. Operations are in blue boxes while objects are shown as black text. Besides “Genome fasta” is schematized a portion of the genome containing repeats as colored boxes. The slightly grey boxed represent unclustered sequences.

For an in-depth analysis of Pithovirus' MITEs, the sequences from the cluster of repeats corresponding to the main MITE were aligned with MAFFT and sequences were trimmed were most had their terminal inverted repeat TA. The delimitations of adjacent repeats were modified accordingly. The reference sequence (see step 5) of M1 and M2 were folded by mFold (Zuker 2003). To retrieve divergent M1 and M2 clusters, the dotpots of reference sequences was inspected. Reference sequences aligned to the reference of M1 or M2 clusters were annotated as M1 or M2-like (example given by cluster 3 in step 5, Fig. 1).

MUST v2-4-002 (Ge et al. 2017) and MITE-Tracker (Crescente et al. 2018) were used to confirm the nature of the repeats.

Results

Pithoviridae viruses isolated from soil samples

We isolated five strains of *Pithoviridae* infecting *A. castellanii* from various soil environments (Table 1): one *Pithovirus* (Pithovirus mammoth) and four cedratviruses (Cedratvirus AA, Cedratvirus AB, Cedratvirus DY0 and Cedratvirus DY1) that belong to two distinct previously defined clades (Jeudy et al. 2020).

Table 1. Genome metrics of isolated *Pithoviridae* from this study compared to previously published isolates

The names of the isolated *Pithoviridae* from this study are written in italic while the names in bold represent the mean of the group considering all isolates. *Cedratvirus* clades follow the ones defined in (Jeudy et al., 2020) and are shown in Figure 2. The right part of the table shows the genome metrics after removal of the repeats identified by our pipeline (see further).

	Real genome			Without repeats		
	Length (kb)	GC%	Coding density	Length (kb)	GC%	Coding density
P. mammoth	610	35.8	0.7	469	39.5	0.9
Pithoviruses	637 ± 40.15	35.6 ± 0.13	0.6 ± 0.03	485 ± 0.04	39.5 ± 0.04	0.9 ± 0.02
C. AA	570	42.8	0.8	553	42.8	0.8
C. AB	568	42.8	0.8	552	42.8	0.9
C. Clade A	573 ± 10.49	42.8 ± 0.02	0.8 ± 0.01	556 ± 0.01	42.8 ± 0.01	0.8 ± 0.01
C. DY0	466	40.8	0.8	434	40.7	0.9
C. DY1	472	40.8	0.8	440	40.8	0.9
C. Clade B	468 ± 3.50	40.7 ± 0.10	0.8 ± 0.02	441 ± 0.00	40.7 ± 0.09	0.9 ± 0.02
C. Clade C	460	43.0	0.8	445	42.9	0.9

According to the *Pithoviridae* phylogeny (Fig. 2), the sampling locations are coherent with their sequence divergence. Pithovirus mammoth and Pithovirus sibericum (Legendre et al. 2014) have both been isolated in cryosol from Siberia. Likewise, Cedratvirus DY0 and Cedratvirus DY1 come from permafrost soil sampled in Russia as for Cedratvirus kamchatka (Jeudy et al. 2020). The two other cedratviruses (Cedratvirus AA and Cedratvirus AB) come from France as for their closest relatives Cedravirus zaza (Rodrigues et al. 2018) and Cedratvirus lausannensis (Bertelli et al. 2017). The most divergent cedratvirus, Brazilian cedratvirus, comes from the most distant sampling location (Brazil). All

included, there is now a total of 12 *Pithoviridae* sequenced genomes available to perform a comparative genomics study of the family.

Pithoviridae phylogeny

To get insight into the *Pithoviridae* family evolution, we performed a phylogenetic reconstruction of the 12 sequenced genomes in addition to the more distantly-related Orpheovirus (Andreani et al. 2018) and Hydrivirus, the only *Pithoviridae*-like metagenome assembled genome that is complete (Rigou et al. 2022). As expected, in Figure 2, Orpheovirus and Hydrivirus are the most divergent. Pithoviruses and cedratviruses split into two well established clades, and cedratviruses can be further divided into 3 previously defined clades (Jeudy et al., 2020). Although Hydrivirus and Orpheovirus cluster together in a well-supported clade, they are divergent from each other (AAI = 31%), even more than cedratviruses and pithoviruses compared together (AAI = 42.2% ± 0.2). In addition, Hydrivirus and Orpheovirus only share 140 HOGs, as compared to the more than 1400 genes they code for in their respective genomes. This suggests that the group will likely split into better defined clades as new related viruses are added.

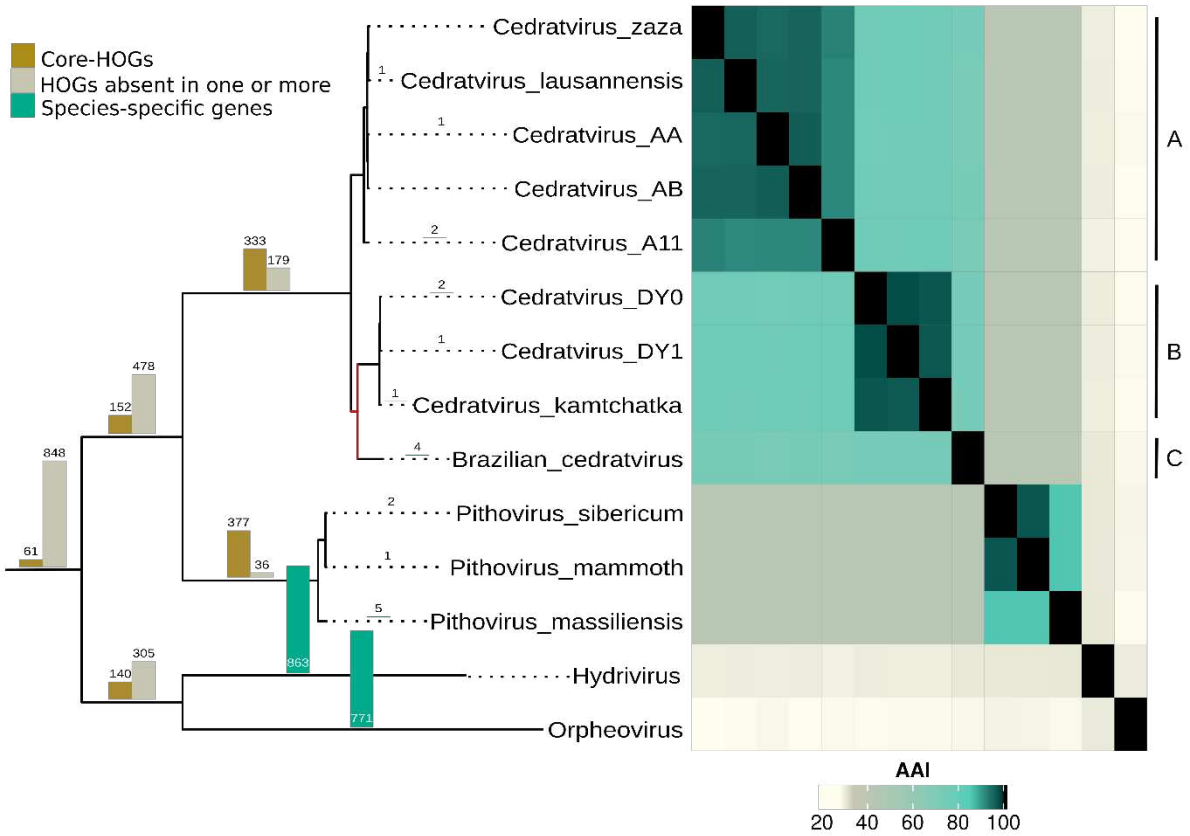


Figure 2. Phylogeny and average amino-acid identity of the *Pithoviridae* and their closest relatives

The phylogeny on the left was built from the concatenation of single copy HOGs output applying the LG+F+G4 evolutionary model. Marseillevirus was used as outgroup to root the tree. The red branch indicates a low bootstrap of 51%. All other bootstraps are of 100%. The bars on each branch represent the number of shared HOGs and other HOGs that were recomputed by OrthoFinder according to this tree. The heatmap on the right shows average amino-acid identity (AAI) between the viruses. The right-most bars (labeled A, B and C) indicate previously determined *Cedratvirus* clades (Jeudy et al. 2020).

Consistent with the phylogeny, the relative codon usage pattern shows the exact same trend, with cedratviruses tightly clustered together, as for pithoviruses, and orpheovirus being the most distant (Fig S1). This is in line with the fact that the Orpheovirus lab host is different from *Pithoviridae*.

Within cedratviruses or pithoviruses there is sequence conservation despite many rearrangements. *Pithovirus massiliensis* has one major inversion and one translocation compared to the two other pithoviruses. *Cedratvirus Kamchatka* and *Brazilian cedratvirus* have both many rearrangements compared to clade A (Fig. S2).

The closed pangenome of Cedratviruses

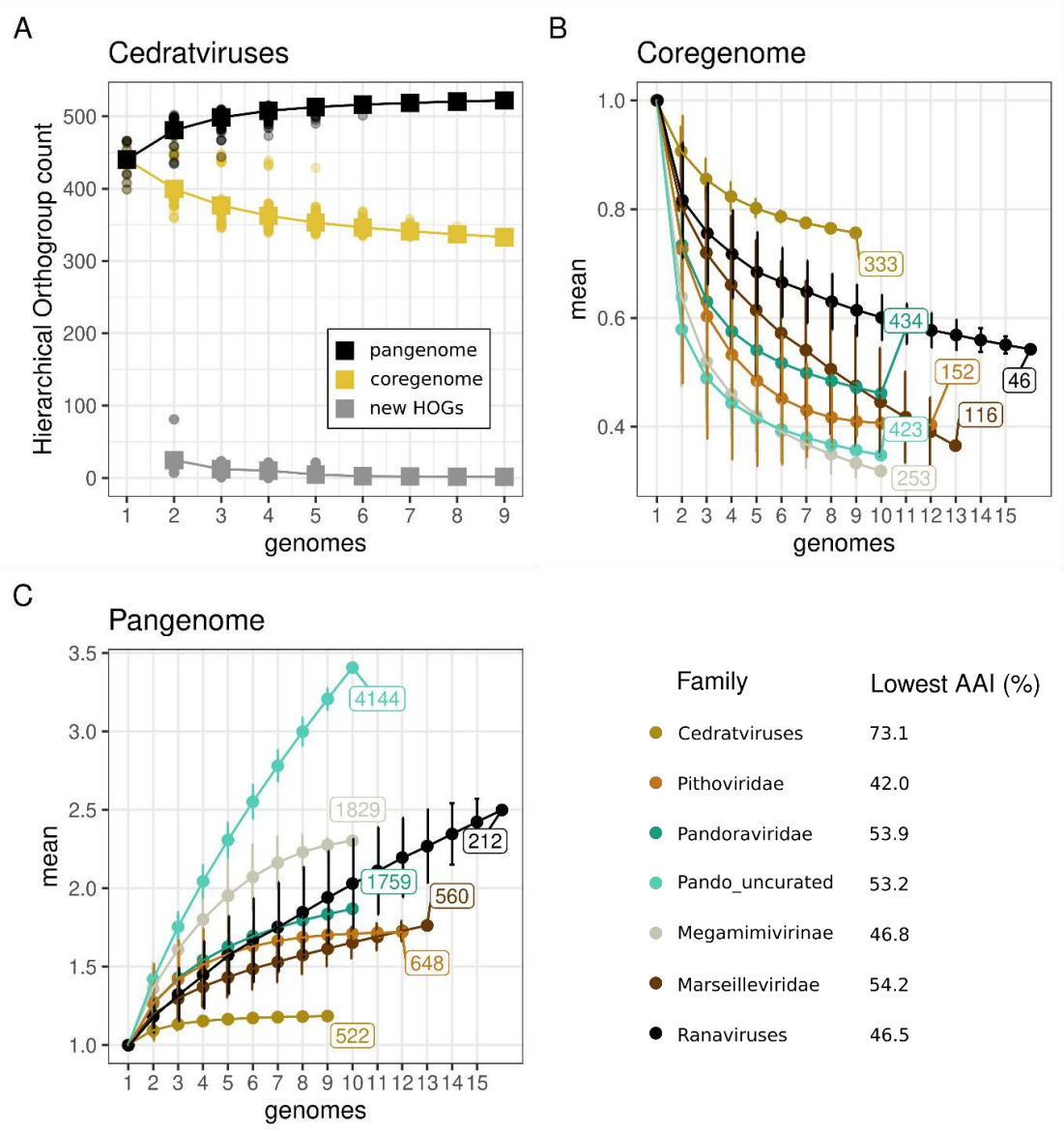


Figure 3. The core and pan-genomes of *Pithoviridae* and other *Nucleocytoviricota*
(A) Pangenome, core-genome and new HOGs have been estimated for Cedratviruses by adding new genomes to a set of previously sequenced genomes in an iterative way (Tettelin et al. 2005). On the contrary of the cited method, the three metrics have been calculated on HOGs and not genes. (B-C) For comparison, the core-genome and pangenome sizes of other *Nucleocytoviricota* have been estimated in the same iterative way. The pangenome and core-genome sizes are defined as the relative size in comparison to their initial mean size. The lowest AAI besides the legend indicates the AAI of the most distant viruses within the set of genomes used for this analysis.

As genomic novelty is an intriguing feature of giant viruses, the pangenomes and core-genomes of *Pithoviridae* were estimated in comparison to other families to get a view of the diversity of gene content. The core-genome of cedratviruses is estimated to be of 333 genes over 100 amino-acids (Figs. 2 and 3) and their minimal core-genome size is not yet known (Fig. 3A). Each new genome brings in

average less than two new HOGs to the family (Fig. 3A). The core-genome of *Pithoviridae* is twice smaller and a plateau of 153 ORFs has been reached (Fig. 3B).

The pangenomes of cedratviruses and *Pithoviridae* have both reached a plateau (Fig. 3C). Notice that the openness or closeness of pangenomes is highly influenced by the diversity within the family as exemplified by the differences in-between *Cedratvirus* and *Pithoviridae*'s pangenomes (Fig. 3). These are also highly influenced by the reliability of the dataset used; while the pangenome of *Pandoraviridae* reaches nearly 3.5 times the mean size of one genome when 10 genomes are sampled, a more realistic number would be 2 times the size of a genome as seen when using a dataset that had been curated with transcriptomic data (Fig. 3C). Thus, the openness of the pangenomes are overestimated when there is no experimental curation step. Even-so, *Pithoviridae* appear to be much more conservative (i.e. a closer pangenome) than other *Nucleocytoviricota* (Fig. 3). In the contrary of *Pandoraviridae*, the new-genes creation might not be an important strategy for *Pithoviridae*.

Furthermore, the pan-genome of *Pithoviridae* is mostly composed of core-genes within cedratviruses or pithoviruses (Fig. S3). Of these subfamily-core-genes, several functions are also present in the other subfamily core-genes. Of the sub-family core-HOGs, 84 are functionally annotated and nearly half of these have an equivalent in the other group. The most explicit example is the ribonuclease H that is separated in two HOGs; one composed of all cedratviruses and the other one with all the pithoviruses.

Most protein sequences are also constrained as seen by the dN/dS ratios calculated on single-copy genes alignments from the studied HOGs. Family ORFans have low dN/dS values but are less well conserved that other genes (Fig. S4A). It seems overall that pithoviruses are more conservative than cedratviruses in terms of protein sequence (Fig. S4B). This applies to accessory genes as well as to core-genes and ancient core-genes shared with Marseillevirus, Orpheovirus and Hydrivirus (Fig. S4B).

Duplications and HGT events are just one factor of genome gigantism

Duplications have been proposed to be an important factor of genome gigantism (Filée et Chandler 2008). There have been duplication events all along the history of *Pithoviridae* even at a short time scale as after the divergence of the closely related *Pithovirus sibericum* and *Pithovirus mammoth*. Duplications are mostly local (cedratviruses median distance = 6872 bases, mean=62557 bases; pithoviruses median=1575, mean=44062). The most duplicated genes are always of unknown function, ankyrin-repeat proteins or F-box proteins. Notably, an ankyrin-repeat protein that is part of the *Cedratvirus* core genome is present around 20 times in the clade B and C and around 50 times in clade A.

From 14 % to 28 % of the genes come from a duplication event in *Pithoviridae* (Fig. 4) with a median of 19 %. This is in line with *Marseilleviridae* (16 %), *Pandoraviridae* (15 %) and *Megamimivirinae* (14 %). In Ranaviruses only 3 % (median) of the genes appeared from a duplication event. There is an exception with Singapore grouper iridovirus that is much bigger than the other Ranaviruses and has 22% of genes coming from a duplication event. According to this analysis, duplications alone are not sufficient to explain the difference in genome sizes between *Pithoviridae* and *Marseilleviridae* or the larger *Megamimivirinae* and *Pandoraviridae* and the other groups.

On the other hand, the different gene number in cedratviruses compared to pithoviruses could be apparently explained by duplications (Fig. 4). Yet, both families share only 152 core-HOGs (Fig. 2). *Cedratvirus A11* has 364 ORFs that are apparently absent in *P. sibericum* according to OrthoFinder. The logical next step is then to check for horizontal gene transfer.

Horizontal gene transfers are even less important than duplication events in terms of genome increase. They account for 5 to 7% of the genome of *Pithoviridae* and not much more in Orpheovirus and Hydrivirus (Fig. 4). HGT events and duplication events are actually linked in *Pithoviridae* as there is significantly more duplication events of HGT-originating genes than non-HGT originating genes; on

average 27% of genes originating from an HGT event are in a duplicated HOG versus 15% of genes that are not coming from an HGT event (chi² p-value 1.39e-7). If genomes are taken separately, only Pithovirus mammoth has a significant biais for more duplications in HGT originating genes (p-value 0.03). Furthermore, HGT events recovered here gave rise to genes as constrained as the non-HGT originating genes as seen by their dN/dS values (Fig. S5A). There are still 59 % of family ORFans (considering BLASTP matches against nr, Fig. S5C) that cannot give us any indication on their origin and that tend to have a higher dN/dS value (Fig. S4A and S5A).

With no surprise, most HGT in *Pithoviridae* are coming from eukaryotes (42 % +/- 2 %) followed by Bacteria (41 % +/- 3 %), (Fig. S5B). Our analyses suggest that Hydrivirus got more ORFs through a Bacteria than through an eukaryote (54 to 40 %). Unfortunately, the HGT events coming from Eukaryota cannot easily point out to a specific host. Most often, the root of the HGT is ancient, branching before or in-between Discosea and Evosea, classes of amoebas (Fig. S5B). Orpheovirus and Hydrivirus have apparently also had exchanges with Metazoa. In *Pithoviridae*, we estimated that 10 % of the HGT events came from another virus.

Overall, the low HGT percentage is coherent with the closeness of the genomes. Duplications are partly responsible for *Pithoviridae*'s gigantism. They are more significant in Orpheovirus (Fig. 4). Yet, accumulated mutations might prevent us from recalling the history of many genes.

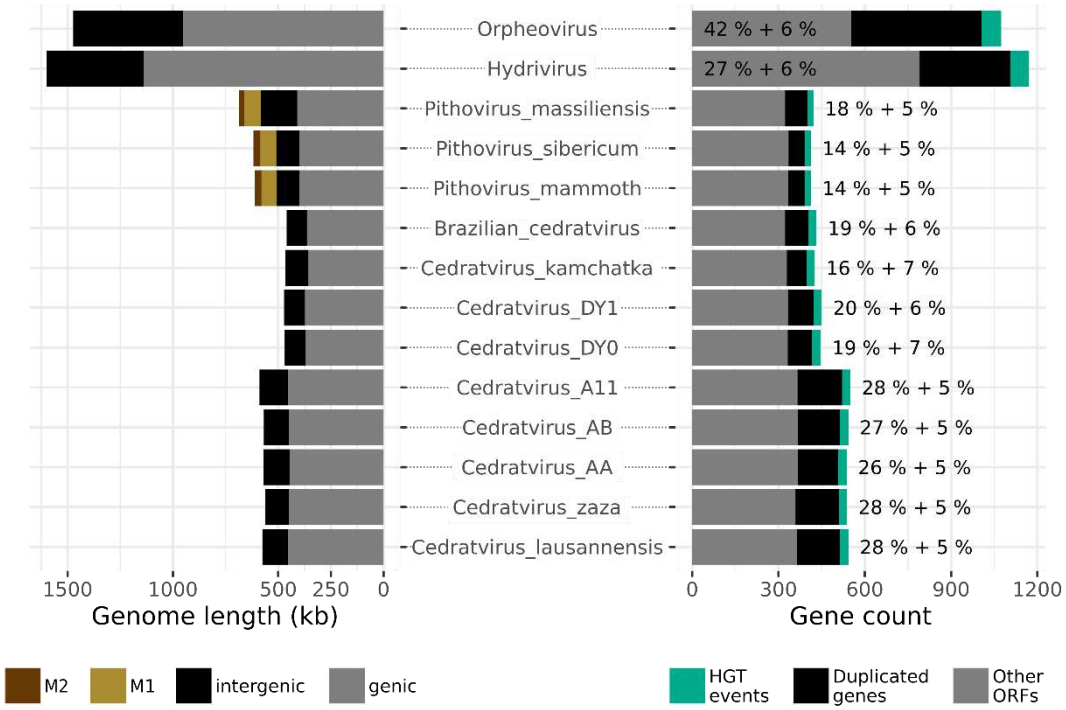


Figure 4. Summary of the genome content of *Pithoviridae* and its relatives
The left panel presents the nucleotide content of the different genomes. The right panel shows their composition in ORFs. The first number presents the percentage of genes that rose from a duplication event and the second shows the percentage of HGT events toward that genome.

Traces of asymmetry in the *Pithoviridae* genomes

The comparative genome studies of other giant virus families previously highlighted a non-uniform evolution of their genomes with one "creative" and a "conservative" half (Legendre et al. 2018; Blanca et al. 2020).

In *Pithoviridae*, duplications seem to occur in specific hotspots (Fig. 5A). These hotspots contrast with the presence of core-genes (Fig. 5B). In all three pithoviruses, the centers of their artificially

linearized genomes are very low in core-genes. In cedratviruses, core-genes are very also scarce in the mid-right portion of their genomes (Fig. 5B). No pattern was revealed by the GC skew (Fig. S6).

Although noisy, there is clearly an asymmetry also in *Pithoviridae*. Duplications and core-genes appear in hotspots that are unevenly present along their genomes. There is a slight strand preference for genes in one or the other half of the genome and a different codon usage. Overall, this suggests a non-uniform architecture of *Pithoviridae* genomes.

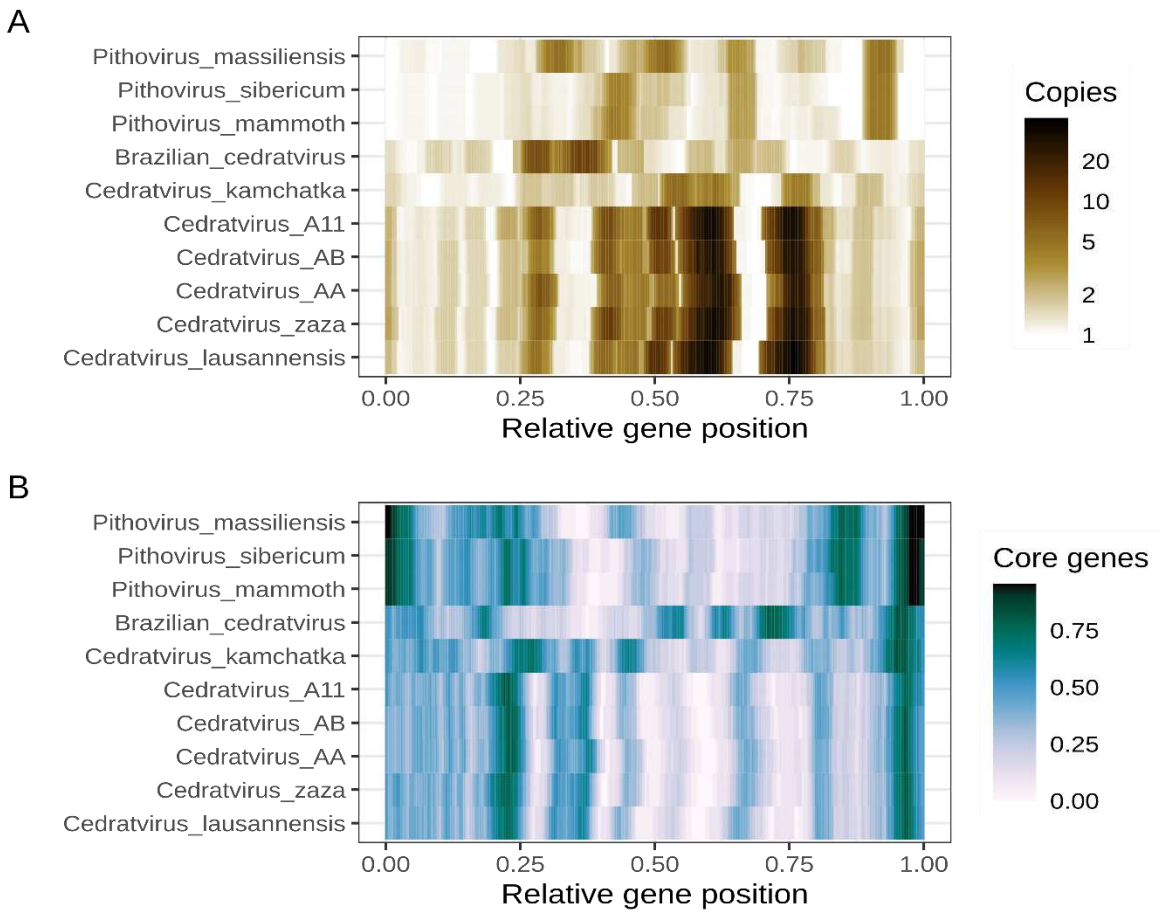


Figure 5. Genome asymmetry in gene content

(A) Mean copy number within the HOGs containing each of the genes of the trailing window. (B) Mean density of core-genes in the trailing window. Both means were calculated on training windows of 21 ORFs.

Two MITEs heavily colonized the genomes of Pithoviruses after the divergence with Cedratviruses

It is known that a regularly-interspersed palindromic sequence present in the intergenic regions is shaping the genomes of pithoviruses (Legendre et al. 2014). These repeats are present in clusters and usually separated by 140 nucleotides (median). After masking this sequence from the genomes, it turns out that other repeats where still present in the close proximity to the masked regions. In the 140 nucleotides separating the originally identified repeat resides another inverted repeat (Fig. 6C-D). By running the tools MUST, we found that both types of repeats could represent distinct Miniature Inverted Transposable Element or be a part of larger ones. With MITE-Tracker we recovered the first repeat type. Those tools helped us to determine that these two types of repeats were very probably MITEs that we will call from now on M1 and M2.

Next, a dedicated pipeline was designed to automatically retrieve these sequences and cluster them in order to get an extensive description of the structure of the repeated regions (Fig. 1). Our pipeline successfully retrieved and clustered the MITEs (Fig. 6A-D).

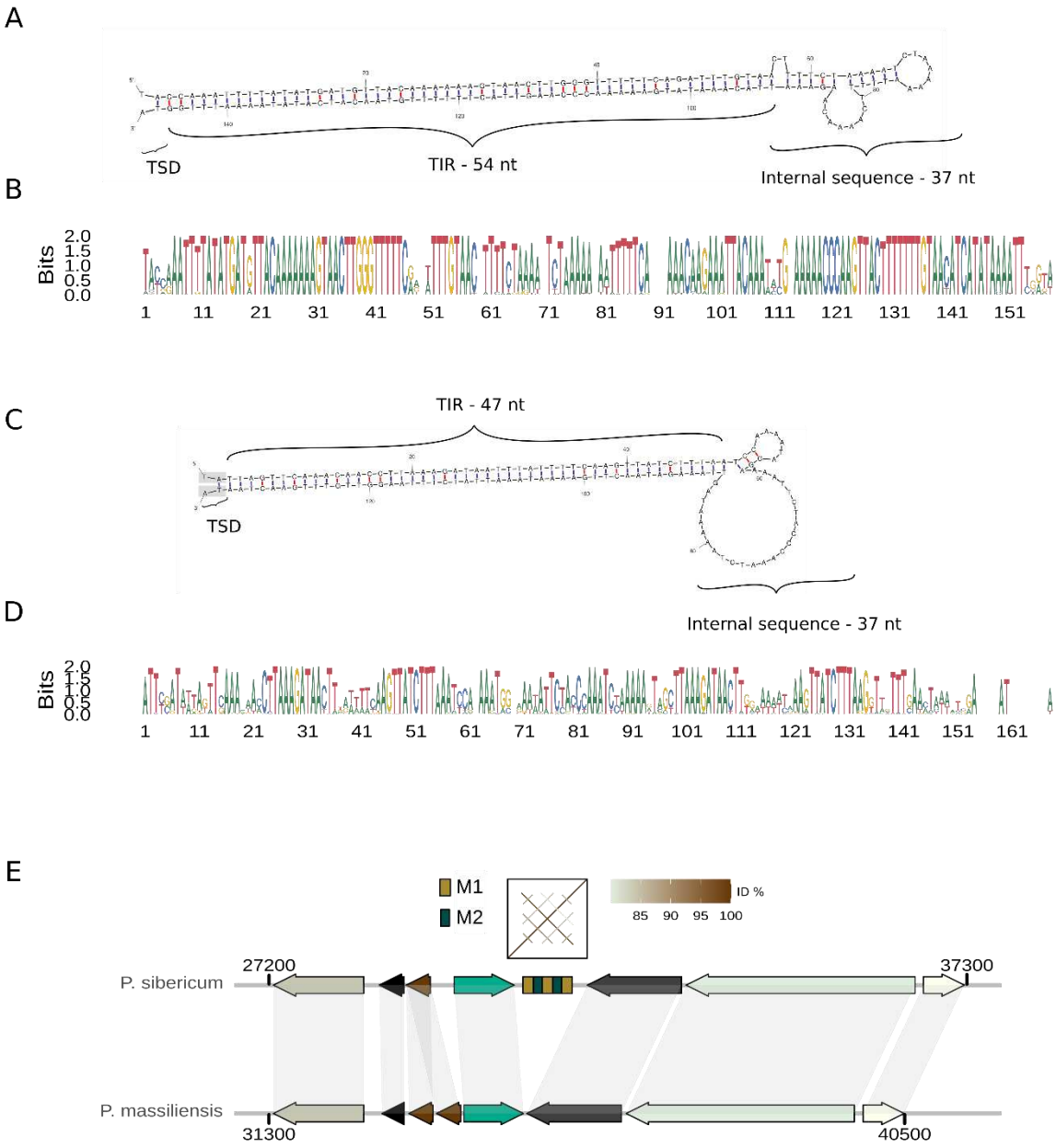


Figure 6. Structure of putative MITEs found in Pithoviruses
(A) and (C) show the DNA folding structure of the reference sequence for M1 and M2 MITEs clusters respectively. Their free energy ΔG is of -79.2 and -65.5. TSD is for Target Site Duplication and TIR for Terminal Inverted Repeat. The TSD highlighted in grey in (C) indicate that those residues are shared with the M1 MITEs next to this M2 sequence. (B) and (D) are the alignment logos of all the sequences in the clusters of M1 and M2 respectively. (E) A repeated region in *P. sibericum* from a highly synthetic portion of the genome in respect to *P. massiliensis*.

The pipeline let us establish some rules:

1) M2 can never be seen in a repeat region without M1 (Table S2)
2) M1 can be seen without M2 (Table S2)
3) When several M1 are present in a region they are always separated by a sequence of about 130 bases or more whether it is M2 or not
4) The reciprocal is true for M2

As seen from the structure of the repeated regions in the three genomes, one of the most common feature is (M1-M2){1 to 8 times}-M1. M1 is present 504 times in *Pithovirus sibericum* and M2, 252

times. If we include more degenerated forms, M1 is present 515 times and M2, 371 times. *Pithovirus mammoth* has a very similar number of regions containing M1 and M2 but the number of M1 or M2 copies per synthenic regions can often be different. Thus, differences might be explained by the extension or contraction of repeated regions rather that an insertion of a MITE in a repeat-free region. The possibility for extension is supported by the fact that repeats from the same region are more similar to each other than repeats from different regions (Fig. S7, p-values <2e-16). In *Pithovirus sibericum* on the other hand, one can observe a repeated region with three M1 and two M2 that is not present in *Pithovirus massiliensis* and could thus come from a transposition or excision of a MITE since the divergence of the two viruses (Fig. 6E). As the two regions are highly syntenic, it leaves little doubt about it. Those events might have happened several times since the divergence of the two species as *P. massiliensis* has slightly more M1-containing regions than *P. sibericum* (115 versus 110, Table S2).

M1 and M2 MITEs affect the genome metrics of Pithoviruses

The two repeats combined represent 21.4 %, 22.3 % and 24.0 % of the genomes of Pithovirus sibericum, mammoth and massiliensis respectively (Fig. 4). Those repeats highly affect the genome metrics of Pithoviruses. When artificially removing the repeats identified by our pipeline from all genomes, the GC% of *Pithoviridae* becomes homogeneous (Table 1) with a standard deviation of 1.48 much lower than the one for the real genomes that is of 3.07. Their coding density becomes homogeneous (sd: 0.02 versus 0.07). The genome size difference also lowers (sd of 53.2 kb versus 72.0 kb). The remaining differences in genome length are mainly due to the gene duplications in clade A cedratviruses (Fig. 4). Without the clade A, the standard deviation for genome sizes lowers to 27.0 kb.

While repeated sequences can potentially help duplications to occur, duplication count within repeated regions of Pithoviruses is not at all higher than the one from outside repeated regions (18% of genes are multicopy within those regions versus 20% outside). Still, those regions seem to be more prone to genetic novelty as there is a slight tendency to HGT within those regions (overall p-value: 3e-4). *Pithovirus sibericum* counts with 25 genes originating from an HGT event. 16 of those come from outside a repeat-rich region and 9 from inside. According to a chi² test, the repartition should be 20.3 and 4.7 respectively (p-value: 0.02). *Pithovirus mammoth* has 16 HGT inside a region and expects 20.5 and 10 outside but expects 5.5 (p-value: 0.03). In *Pithovirus massiliensis* alone, the tendency for HGT events in repeat-rich regions is not significant. Similarly, genes within repeated regions tend to be less ancient than other genes. This was estimated by the node in the phylogenetic tree where an ancestor had probably acquired the gene; i.e.: the position of the common ancestor of all species within the HOG (Fig. S8A, chi² test p-value : 1e-4). Finally, ORFs within repeated regions tend to have a higher dN/dS value in the three genomes, thus are less evolutionary constrained (Fig. S8B, p-value < 2e-4).

To check if this mutation rate could be the result of pseudogenization due to repeat insertion within genes we checked ORF sizes around MITEs. ORFs in-between two M1 repeats are smaller than other genes (Fig. S9). Otherwise, ORFs that have an M1 repeat only at one side (3' or 5') are not smaller than ORFs away from the repeats. In addition, the one potential insertion in *P. sibericum* since the divergence with *P. massiliensis* occurred in an intergenic region (Fig. 6). This suggests that M1 and M2 insert preferentially in the intergenic region. The differences in ORF metrics presented here might thus be the reason of a higher sequence instability within the repeat-rich regions rather than insertion inside ORFs.

M1 and M2 MITEs are unseen in other genomes

In search for a possible antonomous transposon associated to these MITEs, we screened the genomes from the non-redundant database, of *Acanthamoeba castellanii* and of reference and metagenomic *Nucleocytoviricota*, without success. The two sequences remain unseen in the databases. Still, some few *Pithoviridae*-like genomes from the environment present genomes that are highly structured by direct repeats. An example of clear satellite-arrays found in a *Pithoviridae*-like sequences is given in Fig. S8A and B. Repeats constitute 13 % of Pithovirus LCPAC302's genome according to our pipeline. In

some cases, repeats can also be regularly interspersed by another kind of repeat (Fig S10). The high abundance of repeats of all types is not constant in *Pithoviridae* as shown by the differences of the three isolates pithoviruses having 23 % (+/- 1 %) of repeats in their genomes and cedratviruses with 3 % (+/- 2 %) of repeats distributed all along the genome. The amount of diverse repeats in *Pithoviridae* and *Orpheoviridae*-like metagenomes from the permafrost (Rigou et al. 2022) is variable ranging from 0 to 8 %.

Discussion

This work presented the genomes of *Pithoviridae* as quite conservative in comparison to other groups. Their pan-genome is going towards closeness and what is not a core-gene is the most often shared either by all cedratviruses or all pithoviruses. To evaluate the level of conservation of genes, duplication events and HGT events, we largely relied on a software for orthology inference. Therefore, we chose stringent ORF size limit and the tool OrthoFinder that has shown to be more reliable than other tools in particular when adding an outgroup (Emms et Kelly 2020). There are still systematic errors even in a tool that have such good recall. For instance, HOGs with only fast-evolving genes would eventually fail to be recovered and, as a consequence, gene gain and losses are always over-estimated (Natsidis et al. 2021). How to differentiate noise from signal when studying such ancient virus family is an ever-lasting question.

HGT-originating genes in *Pithoviridae* are as constrained as putative vertically inherited genes. The former comparison of *Pithovirus massiliensis* and *Pithovirus sibericum* came to the same conclusion for those genomes although BLASTP results alone are a poor estimation of HGT events (Levasseur et al. 2016). Yet, another story might be told from what we do not see; from what is hiding behind all the family ORFans. Are they really unique or is it the higher non-synonymous mutation rate that prevents us from finding their homologues in the public databases? It is important to point out that many ORFans are expressed (Legendre et al. 2014) and not neutrally-evolving but simply less constrained comparatively to other genes.

As for other giant virus families, *Pithoviridae* have genomes divided in two parts, one more conservative than the other. The separation in two regions is less clear than other families as duplications and non-core-genes appear in hotspots rather than being evenly dispersed in the “creative” half. At the moment there is no explanation for this asymmetry but it is noticeable that very distant families, with different GC% and genome conformation share this feature (Blanca et al. 2020; Legendre et al. 2018).

A new type of MITE, highly organized, has been revealed in the genomes of Pithoviruses. Most repeat-rich regions are structured as M1{M2-M1} n times. It is not uncommon for MITEs to insert in the close proximity of another MITE such as in one Rice chromosome where such event happened multiple times (Tarchini et al. 2000). Also in Rice, it has been estimated that 11% of MITEs exist in multimers (Jiang et Wessler 2001). Still, MITE’s as organized as pithoviruses’ repeats have not been described yet. The repeat-rich regions structure indicate that M1 and M2 are probably part of the same structure and move together. Have they once been independent ? Rule 2 (“M1 can be seen without M2”) gives credit to that hypothesis. M1 is seen two times alone in *Pithovirus sibericum*. M1 is also seen many more times than M2 (Table S2). Curiously, this happens because M1 is sometimes clustered in regions where it is also separated by small sequences that can be another repeat type or a non-repeated sequence. Although M1 and M2 have not been found elsewhere, several giant viruses are repeat-rich. The repeats found in *Pithoviridae*-like metagenomes look exactly like satellite DNA arrays. As repeat proportion is highly variable from one genome to another we believe that those repeats are simply parasitic. Yet, in Pithoviruses, they seem to influence genes in close proximity. As they account for a large portion of their genomes, they may also influence the 3D structure of the latter.

Conclusion

Similarly, to other giant viruses, *Pithoviridae* have an asymmetric genome. This asymmetry in core-gene distribution and duplication occurrences is patchier than in other giant viruses' families. Cedratviruses have a rather conservative pangenome with few new genes added per new virus sequenced. In the same way, only 5 to 7 % of the genomes of *Pithoviridae* come from an HGT event according to our method. More HGT events have probably occurred as high mutation rates prevents us from recreating some gene's history. In any case, it is a sufficiently rare event so that it is not seen in the most recent *Cedratvirus* history. Pithoviruses' genomes are shaped by two mites, M1 and M2 that represent over 20 % of their genomes and affect their genes. In *Pithovirus sibericum*, we have found an insertion in a non-coding region. Why genes within repeated regions are less constrained than outside regions as well as the transposition mechanism of those repeats remains to be explored.

References

Alonge, Michael, Sebastian Soyk, Srividya Ramakrishnan, Xingang Wang, Sara Goodwin, Fritz J. Sedlazeck, Zachary B. Lippman, et Michael C. Schatz. 2019. « RaGOO: fast and accurate reference-guided scaffolding of draft genomes ». *Genome Biology* 20 (1): 224. https://doi.org/10.1186/s13059-019-1829-6.

Al-Shayeb, Basem, Rohan Sachdeva, Lin-Xing Chen, Fred Ward, Patrick Munk, Audra Devoto, Cindy J. Castelle, et al. 2020. « Clades of Huge Phages from across Earth's Ecosystems ». *Nature* 578 (7795): 425-31. https://doi.org/10.1038/s41586-020-2007-4.

Altschul, S. F., W. Gish, W. Miller, E. W. Myers, et D. J. Lipman. 1990. « Basic Local Alignment Search Tool ». *Journal of Molecular Biology* 215 (3): 403-10. https://doi.org/10.1016/S0022-2836(05)80360-2.

Andreani, Julien, Sarah Aherfi, Jacques Yaacoub Bou Khalil, Fabrizio Di Pinto, Idir Bitam, Didier Raoult, Philippe Colson, et Bernard La Scola. 2016. « Cedratvirus, a Double-Cork Structured Giant Virus, Is a Distant Relative of Pithoviruses ». *Viruses-Basel* 8 (11): 300. https://doi.org/10.3390/v8110300.

Andreani, Julien, Jacques Y. B. Khalil, Emeline Baptiste, Issam Hasni, Caroline Michelle, Didier Raoult, Anthony Levasseur, et Bernard La Scola. 2018. « Orpheovirus IHUMI-LCC2: A New Virus among the Giant Viruses ». *Frontiers in Microbiology* 8 (janvier): 2643. https://doi.org/10.3389/fmicb.2017.02643.

Bäckström, Disa, Natalya Yutin, Steffen L. Jørgensen, Jennah Dharamshi, Felix Homa, Katarzyna Zaremba-Niedwiedzka, Anja Spang, Yuri I. Wolf, Eugene V. Koonin, et Thijs J. G. Ettema. 2019. « Virus Genomes from Deep Sea Sediments Expand the Ocean Megavirome and Support Independent Origins of Viral Gigantism ». *MBio* 10 (2). https://doi.org/10.1128/mBio.02497-18.

Bertelli, Claire, Linda Mueller, Vincent Thomas, Trestan Pillonel, Nicolas Jacquier, et Gilbert Greub. 2017. « Cedratvirus Lausannensis - Digging into Pithoviridae Diversity ». *Environmental Microbiology* 19 (10): 4022-34. https://doi.org/10.1111/1462-2920.13813.

Besemer, J., A. Lomsadze, et M. Borodovsky. 2001. « GeneMarkS: A Self-Training Method for Prediction of Gene Starts in Microbial Genomes. Implications for Finding Sequence Motifs in Regulatory Regions ». *Nucleic Acids Research* 29 (12): 2607-18. https://doi.org/10.1093/nar/29.12.2607.

Blanca, Léo, Eugène Christo-Foroux, Sofia Rigou, et Matthieu Legendre. 2020. « Comparative Analysis of the Circular and Highly Asymmetrical Marseilleviridae Genomes ». *Viruses* 12 (11): 1270. https://doi.org/10.3390/v12111270.

Boyer, Mickaël, Natalya Yutin, Isabelle Pagnier, Lina Barrassi, Ghislain Fournous, Leon Espinosa, Catherine Robert, et al. 2009. « Giant Marseillevirus highlights the role of amoebae as a melting pot in emergence of chimeric microorganisms ». *Proceedings of the National Academy of Sciences* 106 (51): 21848-53. https://doi.org/10.1073/pnas.0911354106.

Buchfink, Benjamin, Chao Xie, et Daniel H. Huson. 2015. « Fast and Sensitive Protein Alignment Using DIAMOND ». *Nature Methods* 12 (1): 59-60. https://doi.org/10.1038/nmeth.3176.

Christo-Foroux, Eugene, Jean-Marie Alempic, Audrey Lartigue, Sebastien Santini, Karine Labadie, Matthieu Legendre, Chantal Abergel, et Jean-Michel Claverie. 2020. « Characterization of

Mollivirus Kamchatka, the First Modern Representative of the Proposed Molliviridae Family of Giant Viruses ». *Journal of Virology* 94 (8). https://doi.org/10.1128/JVI.01997-19.

Cock, Peter J. A., Tiago Antao, Jeffrey T. Chang, Brad A. Chapman, Cymon J. Cox, Andrew Dalke, Iddo Friedberg, et al. 2009. « Biopython: Freely Available Python Tools for Computational Molecular Biology and Bioinformatics ». *Bioinformatics (Oxford, England)* 25 (11): 1422-23. https://doi.org/10.1093/bioinformatics/btp163.

Crescente, Juan Manuel, Diego Zavallo, Marcelo Helguera, et Leonardo Sebastián Vanzetti. 2018. « MITE Tracker: an accurate approach to identify miniature inverted-repeat transposable elements in large genomes ». *BMC Bioinformatics* 19 (1): 348. https://doi.org/10.1186/s12859-018-2376-y.

Emms, David M., et Steven Kelly. 2019. « OrthoFinder: phylogenetic orthology inference for comparative genomics ». *Genome Biology* 20 (1): 238. https://doi.org/10.1186/s13059-019-1832-y.

Emms, David M, et Steven Kelly. 2020. « Benchmarking Orthogroup Inference Accuracy: Revisiting Orthobench ». *Genome Biology and Evolution* 12 (12): 2258-66. https://doi.org/10.1093/gbe/evaa211.

Filée, Jonathan, et Michael Chandler. 2008. « Convergent Mechanisms of Genome Evolution of Large and Giant DNA Viruses ». *Research in Microbiology* 159 (5): 325-31. https://doi.org/10.1016/j.resmic.2008.04.012.

Ge, Ruiquan, Guoqin Mai, Ruochi Zhang, Xundong Wu, Qing Wu, et Fengfeng Zhou. 2017. « MUSTv2: An Improved De Novo Detection Program for Recently Active Miniature Inverted Repeat Transposable Elements (MITEs). » *Journal of Integrative Bioinformatics* 14 (3): /j/jib.2017.14.issue-3/jib-2017-0029/jib-2017-0029. https://doi.org/10.1515/jib-2017-0029.

Jeudy, Sandra, Sofia Rigou, Jean-Marie Alempic, Jean-Michel Claverie, Chantal Abergel, et Matthieu Legendre. 2020. « The DNA methylation landscape of giant viruses ». *Nature Communications* 11 (1): 2657. https://doi.org/10.1038/s41467-020-16414-2.

Jiang, Ning, et Susan R. Wessler. 2001. « Insertion Preference of Maize and Rice Miniature Inverted Repeat Transposable Elements as Revealed by the Analysis of Nested Elements [W] ». *The Plant Cell* 13 (11): 2553-64. https://doi.org/10.1105/tpc.010235.

Jones, Philip, David Binns, Hsin-Yu Chang, Matthew Fraser, Weizhong Li, Craig McAnulla, Hamish McWilliam, et al. 2014. « InterProScan 5: Genome-Scale Protein Function Classification ». *Bioinformatics (Oxford, England)* 30 (9): 1236-40. https://doi.org/10.1093/bioinformatics/btu031.

Käll, Lukas, Anders Krogh, et Erik L. L. Sonnhammer. 2004. « A Combined Transmembrane Topology and Signal Peptide Prediction Method ». *Journal of Molecular Biology* 338 (5): 1027-36. https://doi.org/10.1016/j.jmb.2004.03.016.

Katoh, Kazutaka, et Daron M. Standley. 2013. « MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability ». *Molecular Biology and Evolution* 30 (4): 772-80. https://doi.org/10.1093/molbev/mst010.

Lee, Benjamin D. 2018. « Python Implementation of Codon Adaptation Index ». *Journal of Open Source Software* 3 (30): 905. https://doi.org/10.21105/joss.00905.

Legendre, Matthieu, Julia Bartoli, Lyubov Shmakova, Sandra Jeudy, Karine Labadie, Annie Adrait, Magali Lescot, et al. 2014. « Thirty-Thousand-Year-Old Distant Relative of Giant Icosahedral DNA Viruses with a Pandoravirus Morphology ». *Proceedings of the National Academy of Sciences of the United States of America* 111 (11): 4274-79. https://doi.org/10.1073/pnas.1320670111.

Legendre, Matthieu, Elisabeth Fabre, Olivier Poirot, Sandra Jeudy, Audrey Lartigue, Jean-Marie Alempic, Laure Beucher, et al. 2018. « Diversity and Evolution of the Emerging Pandoraviridae Family ». *Nature Communications* 9 (1): 2285. https://doi.org/10.1038/s41467-018-04698-4.

Levasseur, Anthony, Julien Andreani, Jeremy Delerce, Jacques Bou Khalil, Catherine Robert, Bernard La Scola, et Didier Raoult. 2016. « Comparison of a Modern and Fossil Pithovirus Reveals Its Genetic Conservation and Evolution ». *Genome Biology and Evolution* 8 (8): 2333-39. https://doi.org/10.1093/gbe/evw153.

Lu, Shennan, Jiyao Wang, Farideh Chitsaz, Myra K. Derbyshire, Renata C. Geer, Noreen R. Gonzales, Marc Gwadz, et al. 2020. « CDD/SPARCLE: The Conserved Domain Database in 2020 ». *Nucleic Acids Research* 48 (D1): D265-68. https://doi.org/10.1093/nar/gkz991.

Moreira, David, et Céline Brochier-Armanet. 2008. « Giant Viruses, Giant Chimeras: The Multiple Evolutionary Histories of Mimivirus Genes ». *BMC Evolutionary Biology* 8 (janvier): 12. https://doi.org/10.1186/1471-2148-8-12.

Natsidis, Paschalis, Paschalia Kapli, Philipp H. Schiffer, et Maximilian J. Telford. 2021. « Systematic Errors in Orthology Inference and Their Effects on Evolutionary Analyses ». *IScience* 24 (2). https://doi.org/10.1016/j.isci.2021.102110.

Nguyen, Lam-Tung, Heiko A. Schmidt, Arndt von Haeseler, et Bui Quang Minh. 2015. « IQ-TREE: A Fast and Effective Stochastic Algorithm for Estimating Maximum-Likelihood Phylogenies ». *Molecular Biology and Evolution* 32 (1): 268-74. https://doi.org/10.1093/molbev/msu300.

Notredame, C., D. G. Higgins, et J. Heringa. 2000. « T-Coffee: A Novel Method for Fast and Accurate Multiple Sequence Alignment ». *Journal of Molecular Biology* 302 (1): 205-17. https://doi.org/10.1006/jmbi.2000.4042.

O'Leary, Nuala A., Mathew W. Wright, J. Rodney Brister, Stacy Ciufo, Diana Haddad, Rich McVeigh, Bhanu Rajput, et al. 2016. « Reference Sequence (RefSeq) Database at NCBI: Current Status, Taxonomic Expansion, and Functional Annotation ». *Nucleic Acids Research* 44 (D1): D733-745. https://doi.org/10.1093/nar/gkv1189.

Prjibelski, Andrey, Dmitry Antipov, Dmitry Meleshko, Alla Lapidus, et Anton Korobeynikov. 2020. « Using SPAdes De Novo Assembler ». *Current Protocols in Bioinformatics* 70 (1): e102. https://doi.org/10.1002/cpbi.102.

R Core Team. 2021. « R: A language and environment for statistical computing ». Manual. Vienna, Austria. https://www.R-project.org/.

Rice, P., I. Longden, et A. Bleasby. 2000. « EMBOSS: The European Molecular Biology Open Software Suite ». *Trends in Genetics: TIG* 16 (6): 276-77. https://doi.org/10.1016/s0168-9525(00)02024-2.

Rigou, Sofia, Sébastien Santini, Chantal Abergel, Jean-Michel Claverie, et Matthieu Legendre. 2022. « Past and present giant viruses diversity explored through permafrost metagenomics ». *Nature Communications* 13 (1): 5853. https://doi.org/10.1038/s41467-022-33633-x.

Rodrigues, Rodrigo Araújo Lima, Julien Andreani, Ana Cláudia dos Santos Pereira Andrade, Talita Bastos Machado, Souhila Abdi, Anthony Levasseur, Jônatas Santos Abrahão, et Bernard La Scola. 2018. « Morphologic and Genomic Analyses of New Isolates Reveal a Second Lineage of Cedratviruses ». Édité par Rozanne M. Sandri-Goldin. *Journal of Virology* 92 (13): e00372-18, /jvi/92/13/e00372-18.atom. https://doi.org/10.1128/JVI.00372-18.

Santos Silva, Ludmila Karen dos, Ana Claudia dos Santos Pereira Andrade, Fabio Pio Dornas, Rodrigo Araujo Lima Rodrigues, Thalita Arantes, Erna Geessien Kroon, Claudio Antonio Bonjardim, et Jonatas Santos Abrahao. 2018. « Cedratvirus Getuliensis Replication Cycle: An in-Depth Morphological Analysis ». *Scientific Reports* 8 (mars): 4000. https://doi.org/10.1038/s41598-018-22398-3.

Schulz, Frederik, Lauren Alteio, Danielle Goudeau, Elizabeth M. Ryan, Feiqiao B. Yu, Rex R. Malmstrom, Jeffrey Blanchard, et Tanja Woyke. 2018. « Hidden Diversity of Soil Giant Viruses ». *Nature Communications* 9 (1): 1-9. https://doi.org/10.1038/s41467-018-07335-2.

Sun, Cheng, Cédric Feschotte, Zhiqiang Wu, et Rachel Lockridge Mueller. 2015. « DNA transposons have colonized the genome of the giant virus Pandoravirus salinus ». *BMC Biology* 13 (juin). https://doi.org/10.1186/s12915-015-0145-1.

Tarchini, R., P. Biddle, R. Wineland, S. Tingey, et A. Rafalski. 2000. « The Complete Sequence of 340 Kb of DNA around the Rice Adh1-Adh2 Region Reveals Interrupted Colinearity with Maize Chromosome 4. » *The Plant Cell* 12 (3): 381-91. https://doi.org/10.1105/tpc.12.3.381.

Tettelin, Hervé, Vega Masignani, Michael J. Cieslewicz, Claudio Donati, Duccio Medini, Naomi L. Ward, Samuel V. Angiuoli, et al. 2005. « Genome Analysis of Multiple Pathogenic Isolates of Streptococcus Agalactiae: Implications for the Microbial "Pan-Genome" ». *Proceedings of the National Academy of Sciences* 102 (39): 13950-55. https://doi.org/10.1073/pnas.0506758102.

Tria, Fernando Domingues Kümmel, Giddy Landan, et Tal Dagan. 2017. « Phylogenetic Rooting Using Minimal Ancestor Deviation ». *Nature Ecology & Evolution* 1 (1): 1-7. https://doi.org/10.1038/s41559-017-0193.

Wick, Ryan R., Louise M. Judd, Claire L. Gorrie, et Kathryn E. Holt. 2017. « Unicycler: Resolving Bacterial Genome Assemblies from Short and Long Sequencing Reads ». *PLOS Computational Biology* 13 (6): e1005595. https://doi.org/10.1371/journal.pcbi.1005595.

Yang, Ziheng. 1997. « PAML: a program package for phylogenetic analysis by maximum likelihood ». *Bioinformatics* 13 (5): 555-56. https://doi.org/10.1093/bioinformatics/13.5.555.

Yuan, Yihui, et Meiying Gao. 2017. « Jumbo Bacteriophages: An Overview ». *Frontiers in Microbiology* 8 (mars): 403. https://doi.org/10.3389/fmicb.2017.00403.

Zhang, Hua-Hao, Qiu-Zhong Zhou, Ping-Lan Wang, Xiao-Min Xiong, Andrea Luchetti, Didier Raoult, Anthony Levasseur, et al. 2018. « Unexpected invasion of miniature inverted-repeat transposable elements in viral genomes ». *Mobile DNA* 9 (1): 19. https://doi.org/10.1186/s13100-018-0125-4.

Zhang, X. Y., C. Feschotte, Q. Zhang, N. Jiang, W. B. Eggleston, et S. R. Wessler. 2001. « P Instability Factor: An Active Maize Transposon System Associated with the Amplification of Tourist-like MITEs and a New Superfamily of Transposases ». *Proceedings of the National Academy of Sciences of the United States of America* 98 (22): 12572-77. https://doi.org/10.1073/pnas.211442198.

Zuker, Michael. 2003. « Mfold Web Server for Nucleic Acid Folding and Hybridization Prediction ». *Nucleic Acids Research* 31 (13): 3406-15. https://doi.org/10.1093/nar/gkg595.

Supplementary material

Table S1. Assemblies used for comparative genome size analysis

A) Previously published	NCBI accessions	E) Ranaviruses	
Pithovirus sibericum	NC_023423.1	Ambystoma tigrinum virus	GC_ 000841005.1
Pithovirus massiliensis	SAMEA4074172	Bohle iridovirus	GCF_002826565.1
Cedratvirus A11	NC_032108.1	Common midwife toad virus	GCF_003033105.1
Cedratvirus lausannensis	LT907979.1	Epizootic haematopoietic necrosis virus	GCF_000897115.1
Cedratvirus zaza	LT994652.1	European catfish virus	GCF_000897115.1
Brazilian cedratvirus	LT994651.1	Frog virus 3	GCF_001717415.1
Cedratvirus kamchatka	MN873693.1	Infectious spleen and kidney necrosis virus	GCF_000848865.1
Orpheovirus (outgroup)	NC_036594.1	Lymphocystis disease virus 1	GCF_000839605.1
Hydrivirus (outgroup)	GCA_943296135.1	Lymphocystis disease virus-isolate China	GCF_000844885.1
Marseillevirus (outgroup)	NC_013756.1	Lymphocystis disease virus Sa	GCF_001974475.1
		Ranavirus maximus	GCF_001717415.1
B) New *Pithoviridae*		Largemouth bass virus	GCA_013122655.1
Cedratvirus AA	-	Scale drop disease virus	GCF_001274405.1
Cedratvirus AB	-	Short-finned eel ranavirus	GCF_001678255.2
Cedratvirus DY0	-	Singapore grouper iridovirus	GCF_000846905.1
Cedratvirus DY1	-	Grouper iridovirus	GCA_006465545.1
Pithovirus mammoth	-	Red seabream iridovirus (outgroup)	GCA_011894875.1
C) *Pandoraviridae*		**F) *Megavirinae***	
Pandoravirus braziliensis	LT972217.1	Acanthamoeba polyphaga lentillevirus	GCA_000320725.1
Pandoravirus celtis	MK174290.1	Mamavirus	GCA_002966335.1
Pandoravirus dulcis	GCA_000911655.1	Megavirus chilensis	GCF_000893915.1
Pandoravirus inopinatum	GCA_000928575.1	Megavirus courdo7	GCF_000893915.1
Pandoravirus macleodensis	GCA_003233935.1	Megavirus vitis	GCA_004156275.1
Pandoravirus massiliensis	MZ384240.1	Mimivirus	GCA_024266865.1
Pandoravirus neocaledonia	GCA_003233915.1	Moumouvirus australiensis	GCA_004156295.1
Pandoravirus pampulha	OFAJ00000000.1	Moumouvirus	GCF_000904035.1
Pandoravirus quercus	GCA_003233895.1	Tupanvirus deep ocean	GCA_002966475.2
Pandoravirus salinus	GCA_000911955.1	Tupanvirus soda lake	GCA_002966485.2
Mollivirus sibericum (outgroup)	NC_027867.1	Chrysochromulina ericina virus (outgroup)	GCF_001399245.1
D) *Marseilleviridae* as in Blanca et al., 2020 (doi: 10.3390/v12111270)			
Marseillevirus	GU071086		
Lausannevirus	HQ113105		
Cannes 8 virus	KF261120		
Insectomime virus	HG428764		
Tunisvirus	KF483846		
Brazilian marseillevirus	KT752522		
Melbournevirus	KM275475		
Port-miou virus	KT428292		
Tokyovirus	reassembled in Blanca et al., 2020		
Noumeavirus	KX066233		
Golden marseillevirus	KT835053		
Kurlavirus	KY073338		
Marseillevirus shanghai	MG827395		
Ambystoma tigrinum virus (outgroup)	MK580533.2		

s

Table S2. Pithoviruses' MITEs occurrences

A region is defined as a genomic sequence with a high density of repeats within a sliding window of 500 bp. Within each region, the number of M1 and M2 MITEs was counted. The clusters containing divergent M1 and M2 sequences were included in these results.

		P. sibericum		P. mammoth		P. massiliensis	
		M1	**M2**	**M1**	**M2**	**M1**	**M2**
Regions	**Total**	110	100	109	100	115	79
	M1 or M2	10	0	9	0	36	0
Per region	**Min count**	1	1	1	1	1	1
	Max count	11	12	13	17	13	8
	Mean	4.68	3.71	4.58	4	5.05	3.01
	Sd	2.12	2.26	2.31	3.06	2.98	1.64

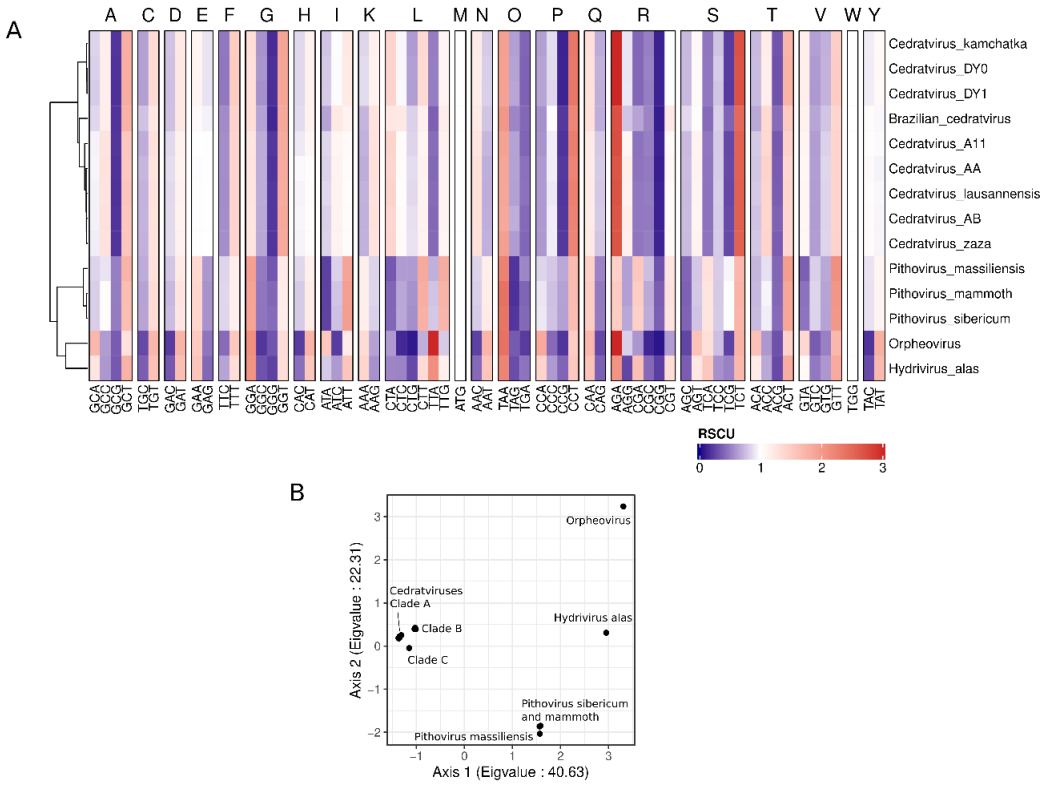


Figure S1. Relative Synonymous Codon Usage comparison

(A) The codon usage biais for each amino acid represented by the RSCU value. (B) PCOA analysis of the RSCU values without the stop codons, the tryptophan and the methionine codons.

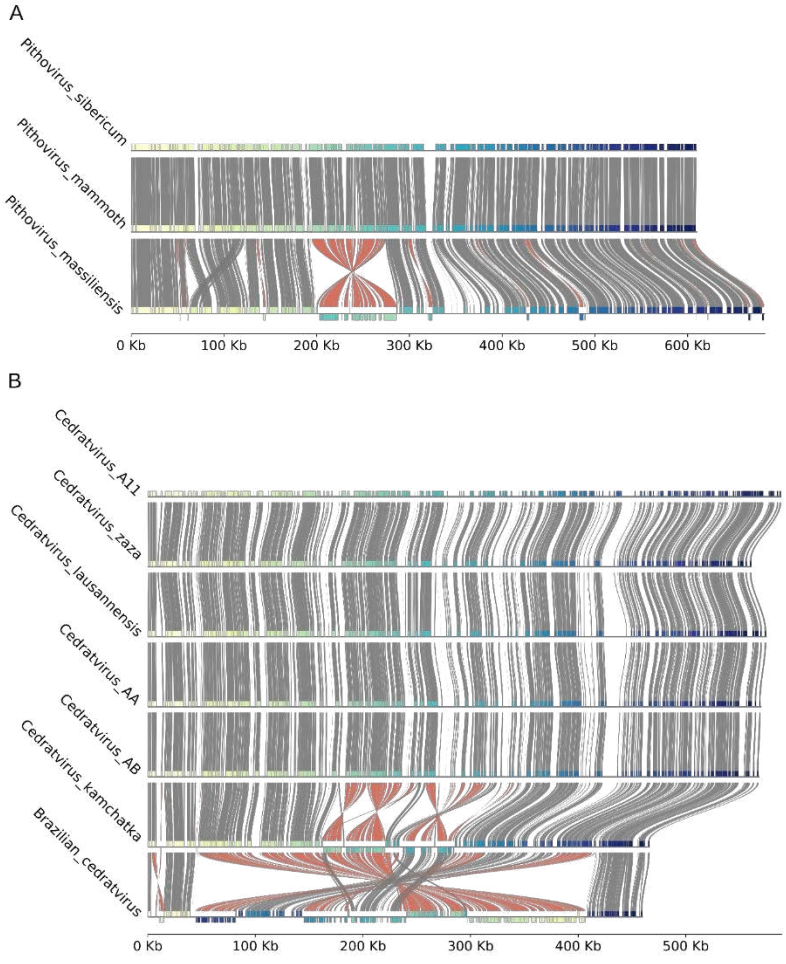


Figure S2. Genome alignment of *Pithoviridae*

Shared nucleotide sequence blocks within families were drawn based on the alignment by progressive-mauve of (A) the three pithoviruses and (B) seven cedratviruses. Cedratvirus DY0 and DY1 have been excluded since their genome alignment has been made on Cedratvirus Kamchatka as reference.

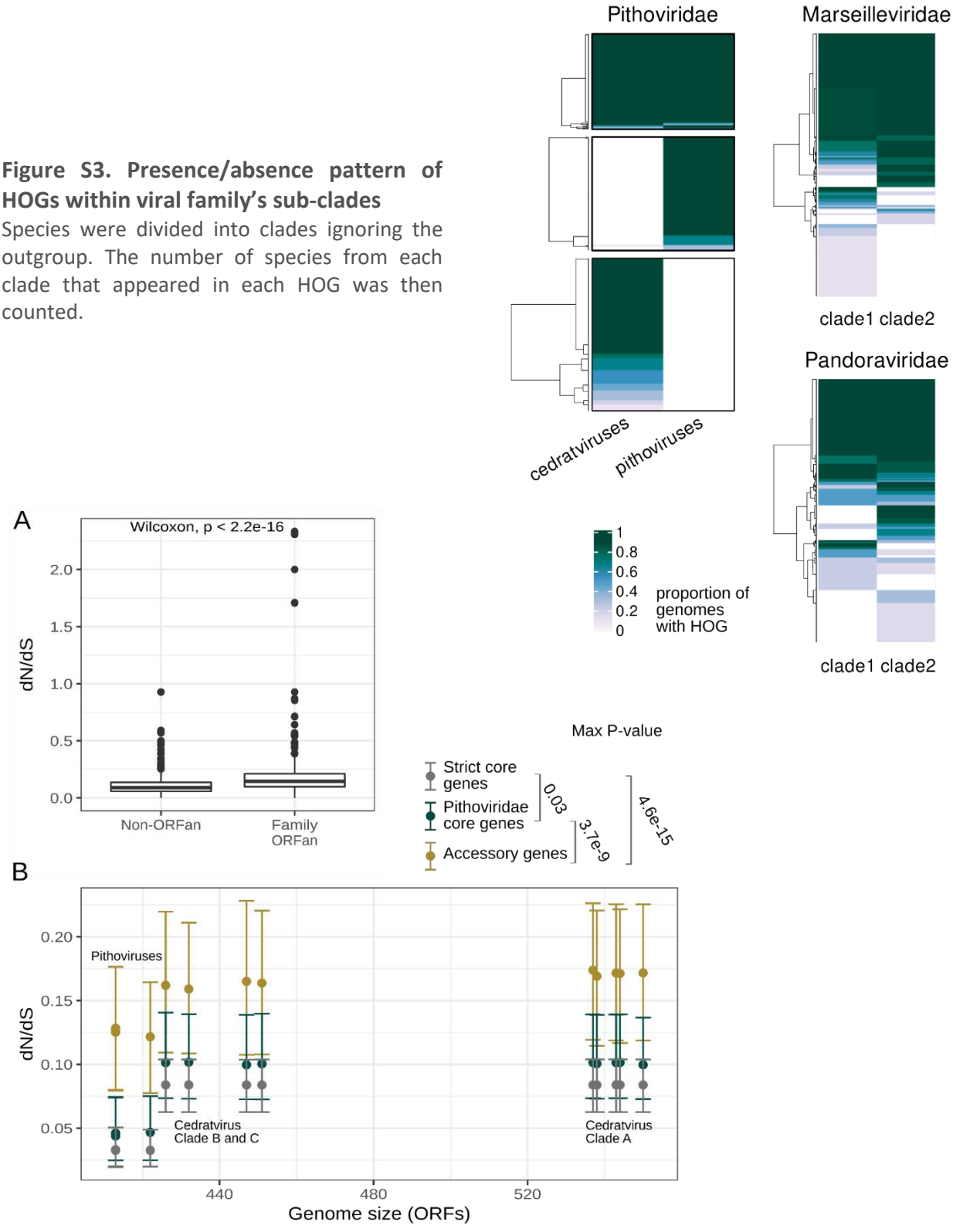


Figure S3. Presence/absence pattern of HOGs within viral family's sub-clades
Species were divided into clades ignoring the outgroup. The number of species from each clade that appeared in each HOG was then counted.

Figure S4. Selection pressure on *Pithoviridae* ORFs
(A) Family ORFans were considered if no BLASTP match was found from outside *Pithoviridae* with an e-value of $1e^{-5}$. The overall p-value from a Wilcoxon test comparing dN/dS values of both categories is given. (B) The core-genes are under stronger selective pressure that the pangenome. The strict core-genes are the ones with a homologue also in Orpheovirus and Hydrivirus. *Pithoviridae* core-genes are code-genes excluding the first category and the accessory genes represent the non-core-genes.

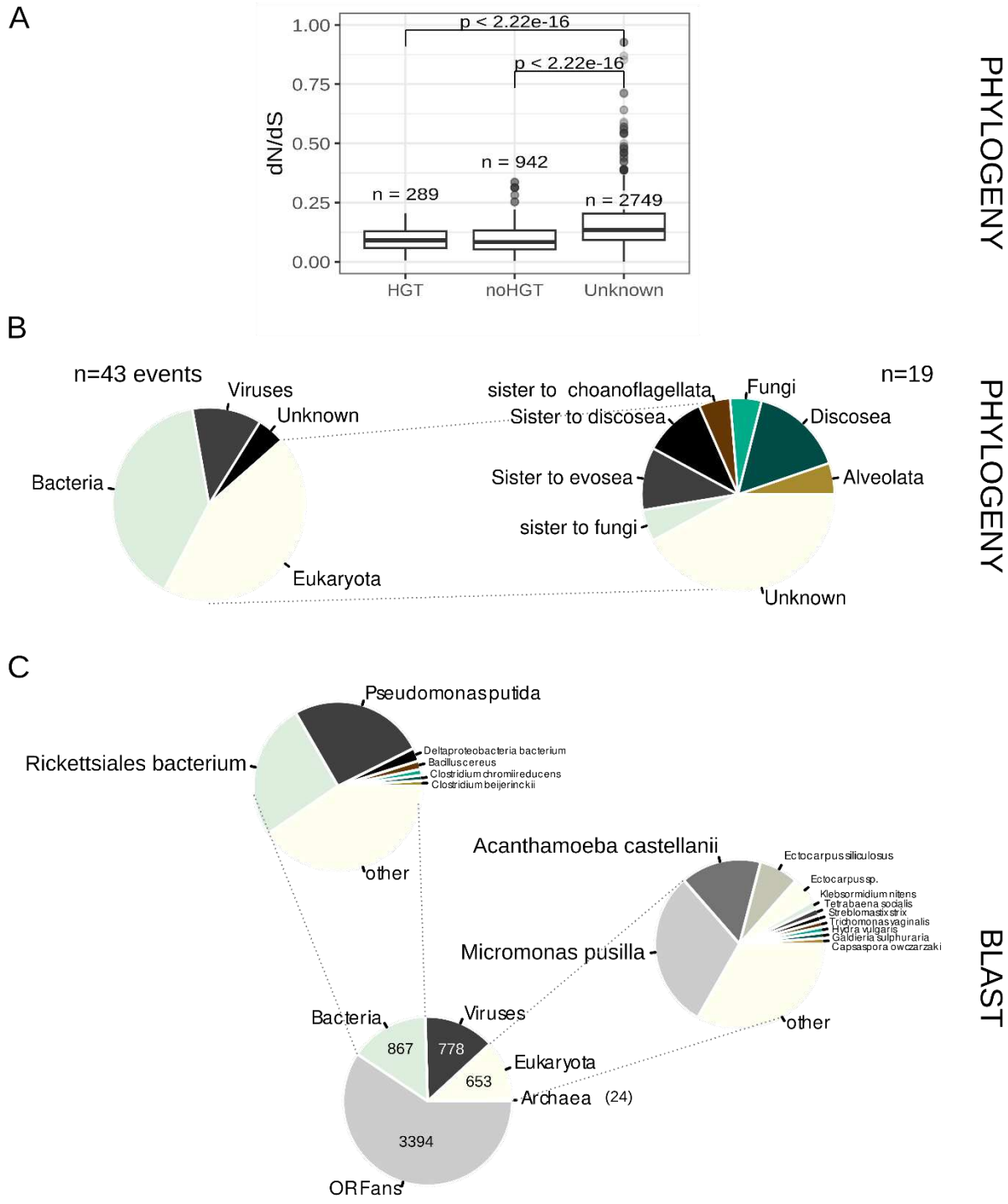


Figure S5. Horizontal Gene Transfer events in *Pithoviridae* and BLASTP control

(A) Shows the dN/dS ratios on genes resulting from an HGT event or not. The p-values were calculated by a Wilcoxon rank test and corrected according to the bonferroni method. For representation purposes, 8 outliers from the "Unknown"group with a dN/dS over 1 are not shown. (B) For each HGT event, the likely origin as estimated from the visualization of phylogenetic trees is shown. From eukaryotes, "sister to" is short for "sister group of…". (C) Best BLASTP results of the nr database free of *Pithoviridae* against *Pithoviridae* genomes. Matches with *Pithoviridae* from the public database were removed. Results under an e-value of 1e-5 were kept for analysis. Bacterial and eukaryotic species with more than 1% of matches in their respective category are shown.

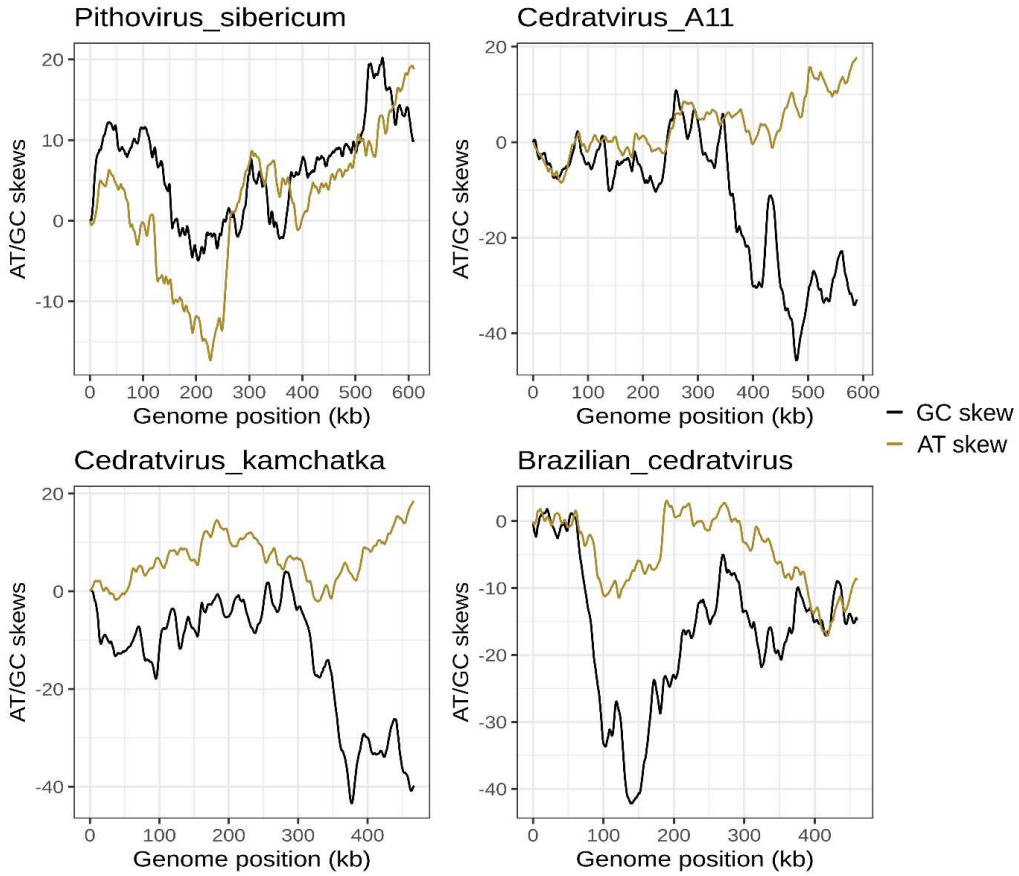


Figure S6. Cumulative GC- and AT-skews in four genomes representing four clades within *Pithoviridae*

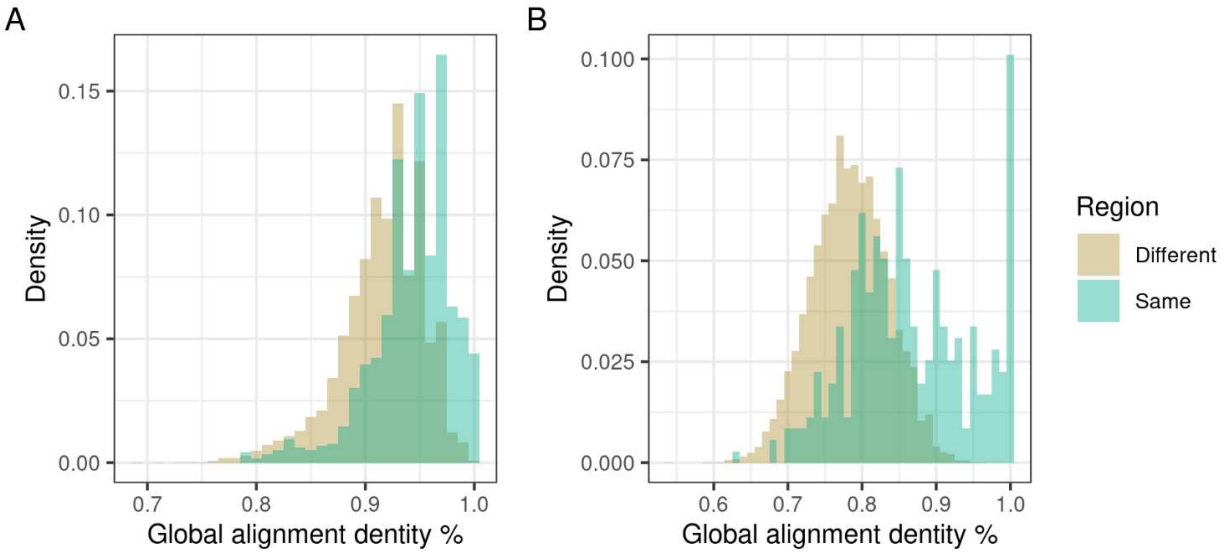


Figure S7. Pairwise identity percentage in-between repeats retrieved from the same of from distinct regions

The pairwise identity percentage were calculated by needle. (A) P1. (B) P2. Both distributions are significantly different also after sampling the data so that the number of pairs that come from different

regions matches the amount of data coming from pairs within the same region (p-value < 2e-16, wilcoxon test).

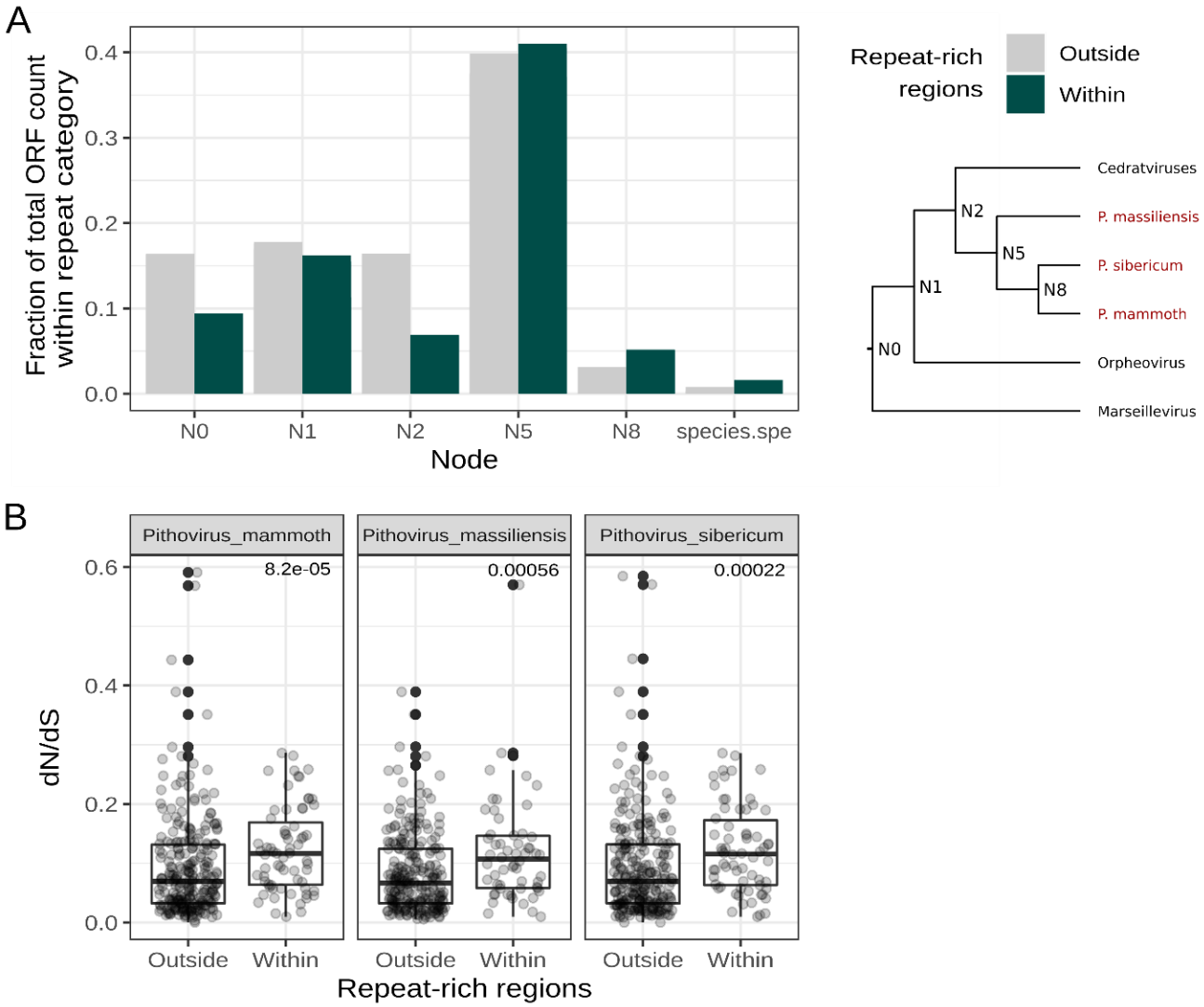


Figure S8. Evolution of ORFs within and outside of repeat-rich regions

(A) The ancestry of ORFs has been estimated by the presence of other species in the HOG. Nodes are ordered from the most ancient to the most recent as shown in the cladogram next to the plot. The number of genes was counted for each estimated ancestry and within or outside repeat-rich regions. Numbers given are relative to the number of genes in each category that are of n=357 and n=1086 within and outside regions respectively. For each node, the number of genes dN/dS values in relation to repeated regions. (B) The dN/dS values were calculated from single copy ORFs and divided into two groups respect to the repeated regions given by our repeat pipeline. The number in the top-right corners are the corrected Wilcoxon rank test p-values.

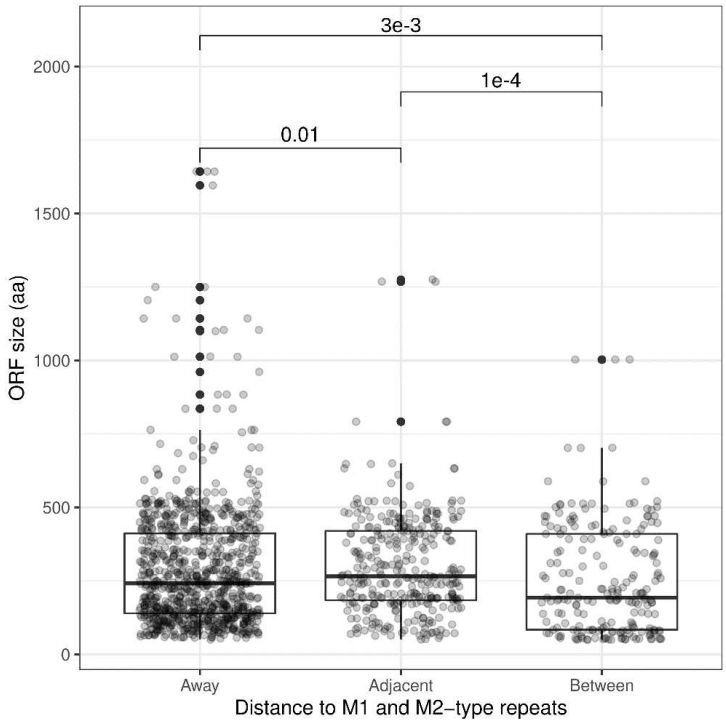


Figure S9. ORF size compared to distance from M1 and M2

ORFs were divided in three categories weather the nearest feature to them was a repeat, on both ("Between") or one ("Adjacent") side, or an ORF on both sides ("Away"). The p-value was calculated by Wilcoxon tests and corrected by the false discovery rate method.

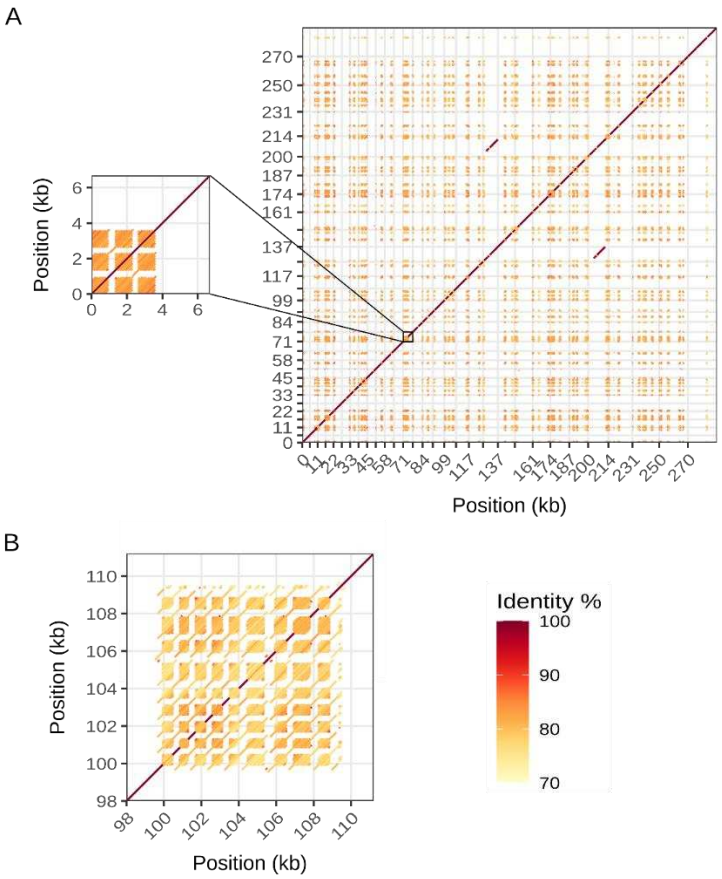


Figure S10. Repeats find in *Pithoviridae*-like metagenomes

(A) Pithovirus LCPAC302 presents numerous direct repeats. In some rare cases, these repeats are intersperced by a similar sequence as shown in the zoom. X-axis and y-axis breaks correspond to the delimitation of contigs. (B) Regularly interspersed direct repeats from a permafrost *Pithoviridae*-like metagenome (K_bin2137_k1).

X. Approche formelle visant à déterminer l'ouverture ou la fermeture du pangénome des *Cedratvirus*

Pour estimer la taille et l'ouverture du pangénome nous avions appliqué une méthode inspirée de (Tettelin et al. 2005). Les orthogroupes hiérarchiques, c'est-à-dire les orthogroupes au nœud défini dans la phylogénie, ici les *Cedratvirus*, ainsi que les gènes non classés, sont transformés en un tableau de présence/absence. Ensuite, de manière itérative de 1 à n, on essaie toutes les combinaisons possibles de génomes. Pour chacune de ces combinaisons, on compte le nombre d'orthogroupes avec au moins un représentant et le nombre d'orthogroupes partagés par tous. Pour déterminer si le pangénome et génome cœur est ouvert ou fermé, il convient de tracer la courbe des moyennes des tailles de pangénomes et génomes cœur en fonction du nombre de génomes séquencés. Si un plateau est atteint, alors nous pouvons présumer que la découverte d'un nouveau virus dans la famille ne devrait pas drastiquement changer la taille des pangénomes ou génomes cœur. On dit qu'ils sont fermés. A l'inverse, si ce n'est pas le cas on dit qu'ils sont ouverts. Un gain ou une perte de gènes fréquent donne un pangénome ouvert. Si la perte de gènes concerne également des gènes cœur, alors le génome cœur apparaît aussi comme ouvert.

Pour calculer le nombre de nouveaux gènes apportés par un nouveau génome séquencé on compte alors de 2 à n, pour chaque combinaison de n-1 et pour chaque génome i, combien d'orthogroupes sont présents dans i mais absents chez les génomes de la combinaison. Cela permet également de se faire une idée de l'ouverture de pangénome.

Dans l'article III nous avons estimé que le pangénome des *Cedratvirus* était fermé par comparaison avec d'autres groupes et par observation de l'allure de ces courbes. Cette approche peut être complétée par une analyse chiffrée. Ces données empiriques peuvent servir à nourrir un modèle pour déterminer de manière plus objective l'ouverture ou la fermeture des génomes.

Les modèles testés selon (Tettelin et al. 2005) :

(A) $$génome\ cœur = Kc \times \exp\left(\frac{-génomes}{\tau c}\right) + \Omega$$

(B) $$nouveaux\ gènes = Ks \times \exp\left(-\frac{génomes}{\tau s}\right) + \tan(\theta)$$

(C) $$pangénome = D + \tan(\theta)\,[génomes - 1] + Ks \times \exp(\tfrac{-2}{\tau s}) \times \frac{1-\exp(-\frac{génomes-1}{\tau s})}{1-\exp(-\frac{1}{\tau s})}$$

Avec génomes qui correspond au nombre de génomes échantillonnés. Kc, Ks, θ, Ω, D, Ƭc et Ƭc sont des paramètres à estimer. La tangente de l'angle θ permet d'estimer le nombre de nouveaux gènes apportés à chaque nouveau génome séquencé et la taille du génome cœur minimal est donné par Ω.

Modèles de l'outil dédié PanGP :

(D) $$génome\ cœur = A \times \exp(B \times génomes) + C$$

(E) $$nouveaux\ gènes = D \times génomes + E$$

(F) $pangénome = F \times génomes^{G} + H$

Avec A, B, C, D, E, F, G et H, des paramètres à estimer.

Enfin la loi de Heap donne le modèle :

(G) $pangénome = K \times génomes^{g}$

Avec K et g à estimer. Selon cette loi, le paramètre g permet de déterminer si le pangénome est ouvert (>1) ou fermé (<1).

Données	Équation	Paramètres	Estimations	R^2	Limite x --> Inf	Rappel de l'équation
Génome cœur	(A)	Kc	163,59	0,51		$génome\ cœur = Kc \times \exp\left(\frac{-génomes}{\tau c}\right) + \Omega$
		τc	2,369			
		Ω	331,669		Ω	
	(D)	A	210,7413	0,51		$génome\ cœur = A \times \exp(B \times génomes) + C$
		B	-0,3294			
		C	230,0637		C	
Nouveaux gènes	(B)	Ks	75,612	0,33		$nouveaux\ gènes = Ks \times \exp\left(-\frac{génomes}{\tau s}\right) + \tan(\theta)$
		τs	1,696			
		tan(θ)	0,87		tan(θ)	
	(E)	D	-2,137	0,25		$nouveaux\ gènes = D \times génomes + E$
		E	17,846		-Inf	
Pangénome	(C)			0,65		$pangénome = D + \tan(\theta)\,[génomes - 1] + Ks \times \exp(\frac{-2}{\tau s}) \times \frac{1 - \exp(-\frac{génomes - 1}{\tau s})}{1 - \exp(-\frac{1}{\tau s})}$
		D	459,2		+Inf	
	(F)	F	-100,3354	0,69		$pangénome = F \times génomes^{G} + H$
		G	-0,7917			
		H	540,229		H	
	(G)	K	452.76584	0,62		$pangénome = K \times génomes^{g}$
		g	0.07167		+Inf	

Tableau 3 – Paramètres estimés par les modèles testés sur la taille du génome cœur, le nombre de nouveaux gènes et la taille du pangénome
En rouge sont surlignées les paramètres en accord avec un génome (pan/cœur) fermé. La variable x dans la limite quand x tend vers l'infini désigne la variable génomes. Cela représente donc la taille de génome estimé en imaginant une infinité de *Cedratvirus* séquencés.

Les données calculées sur les orthogroupes ont servi à nourrir ces modèles grâce au paquet nls de R. Pour un meilleur ajustement pour chaque valeur « génomes », un poids des points a été donné, inversement proportionnel au nombre d'itérations pour le nombre de génomes échantillonnés. Cependant, les tentatives d'ajustement des différents modèles donnant des coefficients de détermination (R^2) trop faibles, il convient de prendre ces résultats avec précaution. En effet, le coefficient de détermination le plus haut est à 0,69 et le plus bas à 0,25 (Tableau 3). Cela est dû surtout à une variabilité importante d'une itération à l'autre pour un

nombre de génomes donné car certains modèles s'ajustent très bien sur la courbe des moyennes.

Le Tableau 3 nous indique qu'il y a plusieurs arguments en faveur de la fermeture du pangénome des *Cedratvirus* donnés par des équations différentes. Notamment les deux modèles, ceux de l'outil PanGP et de (Tettelin et al. 2005) montrent un pangénome fermé avec $g < 1$ et $\tan(\theta) < 1$ (Tableau 3). Nous constatons également la fermeture du génome cœur. Le modèle exponentiel (A) tend même à dire que la taille du génome cœur minimal a été atteinte puisque la valeur Ω est très proche de la valeur réelle du génome cœur actuel qui comprend 333 orthogroupes.

Ces résultats confirment l'analyse comparative fournie dans l'article III. Cependant, les faibles valeurs de R^2 nous indiquent qu'il faudrait étudier plus de génomes pour espérer améliorer l'estimation des paramètres.

Chapitre 3. Le cycle infectieux de Cedratvirus kamchatka vu par imagerie et transcriptomique

XI. Contexte

Jusqu'à aujourd'hui, seul le transcriptome d'un cycle infectieux et le protéome de la particule de Pithovirus sibericum avaient été étudiés (Legendre et al. 2014) et ce, brièvement. C'est pourquoi il nous a paru intéressant d'étudier un autre virus de la famille, Cedratvirus kamchatka, pour mettre en comparaison ces résultats. De plus, l'étude approfondie d'une cinétique de transcriptomique pourrait nous aider à expliquer certains mécanismes évolutifs mis en lumière dans l'article III. Les gènes cœur, et à faible dN/dS sont-ils plus exprimés ? Existe-t-il des motifs répétés régulant l'expression des gènes ? Est-ce que cela est bien corrélé également à l'adaptation de l'usage des codons au code de l'hôte ? Quelle est l'expression des gènes issus d'un transfert de gène horizontal ? Et enfin, pouvons-nous relier des fonctions en particulier avec une temporalité d'expression et donc potentiellement à une fonction plus précise durant le cycle ?

Avant de répondre à ces questions nous fournissons une analyse préliminaire des données de transcriptomique qui ouvrira la voie à une future recherche. Par observation directe du cycle cellulaire nous cherchons dans un premier temps à comprendre le cycle dans son ensemble et à savoir si Cedratvirus kamchatka fonctionne de manière similaire aux autres *Pithoviridae*. L'imagerie électronique a été entièrement réalisée par Lionel Bertaux. Le cycle infectieux a été suivi par Lionel Bertaux et moi-même. Le tout, évidemment, sous la supervision de Matthieu Legendre. Pour ce qui est de l'analyse du transcriptome, nous allons pour le moment nous intéresser à un gène, celui de la Protéine Majeure de Capside (très divergente chez les *Pithoviridae*) qui, chez P. sibericum, n'est pas présente dans la particule virale mais dont nous ne savons pas s'il est tout de même exprimé (Legendre et al. 2014). S'il ne l'est pas, alors nous pourrons conclure à une perte de fonction de ce gène chez Cedratvirus kamchatka.

XII. Matériels et Méthodes

Observation du cycle au microscope optique pour mise en place du plan expérimental

Pour la mise en place du projet il a d'abord fallu observer les différents points clés du cycle et en estimer la durée. Pour ce faire nous sommes partis d'une production de virus à 9.10^9 virus/mL. Les virus avaient été purifiés selon le protocole développé pour Mimivirus (Bertaux, Lartigue, et Jeudy 2020). 16 flasques de 25 cm² ont été ensemencées avec *Acanthamoeba castellanii* souche Neff donnant une confluence finale estimée à 180000 cellules/cm². Ces

cellules furent infectées par Cedratvirus kamchatka à 50 virus par cellule (MOI[29]). Après 1h d'incubation à 32°C, les flasques furent lavées deux fois avec 8 mL de milieu de culture (PPYG pour protéase peptone, extrait de levure et glucose) à 32 °C. Après le dernier lavage, 5 mL de PPYG furent ajoutés. Toutes les trente minutes jusqu'à 3 heures post-infection puis toutes les heures, une flasque a été grattée et 5 mL de culture ont été prélevés dont 500 µL ont été et centrifugés à 500 g 5 minutes. Le culot a été resuspendu avec 500 µL de tampon phosphate salin (PBS) formaldéhyde à 3.7 %. Après 15 minutes, il a été centrifugé à nouveau à 500 g durant 5 minutes puis resuspendu avec 30 à 50 µL de PBS-Hoechst. Cela permet la coloration de l'ADN. Les cellules sont déposées entre lame et lamelle et observées au microscope optique. Ceci a été réalisé jusqu'à 11 heures post-infection.

Cycle d'infection pour extraction d'ARN et observation au microscope électronique

L'étape précédente a permis de mettre en place le plan expérimental : un point de 1 heure à 3 heures toutes les 30 minutes puis toutes les deux heures jusqu'à 8 heures, fin du cycle. Nous avons également un « MOCK », c'est-à-dire un virus inactivé, qui, après 8 heures passées à 90 °C, ne génère pas d'infection. Pour qu'il fasse office de MOCK, nous avons vérifié par microscopie qu'il a bien été phagocyté (Figure 35 en annexe p. 192). Les points à 1, 3 et 6 heures et le MOCK (traité à 3h) ont été réalisés en triplicats. Pour chaque échantillon, 2 flasques de 175 cm² ont été ensemencées pour une confluence finale de 200000 cellules/cm². Les cellules furent alors infectées avec une MOI de 50. Elles sont alors incubées 45 min à 32 °C avant deux lavages en PPYG. Pour chaque point, deux flasques ont été grattées pour récupérer les cellules. 2mL ont été prélevés, centrifugés à 1000 g 5 minutes puis le culot a été gardé à -80°C pour une éventuelle analyse protéomique. Pour préparer les inclusions, 2mL de cellules avec 2 mL de PBS et 5% de glutaraldéhyde ont été refroidis 1 heure à température ambiante, centrifugés à 800 g pendant 10 minutes puis resuspendus avec 1 mL de PBS avec 2.5 % de glutaraldéhyde. Pour la transcriptomique, les 38 mL restants ont été centrifugés à 800 g durant 10 minutes et le culot fut resuspendu avec 3 mL le tampon de lyse « RLT » de QIAGEN avec du βeta-mercaptoéthanol au millième et congelé à -80 °C.

Pour l'observation des cellules au microscope électronique, les inclusions ont été réalisées en résine époxy puis les coupes ultra-fines ont été faites à l'ultramicrotome Leica UC7. Ces coupes ont été observées au microscope électronique à transmission FEI Tecnai G2 à 200 kV.

Pour chaque point, l'ARN total a été extrait avec le kit « RNeasy mini ». Le protocole du fabriquant QIAGEN a été suivi. Après cette procédure, l'ARN récupéré a été quantifié au QUBIT et au Nanodrop. Pour le qubit, des dilutions au 20ème, 100ème et 200ème ont été effectuées. Pour améliorer la pureté de la solution, un dernier traitement rigoureux à la DNase a été réalisé avec le kit « TURBO DNA free » d'invitrogen en suivant le protocole du fabricant.

Les 17 échantillons ont été séquencés par la technologie Illumina NovaSeq 6000 donnant en moyenne 44 millions de lectures de 150 nt par échantillon.

[29] Multiplicity of infection : multiplicité d'infection

Analyses préliminaires du transcriptome

Les lectures brutes et nettoyées ont été contrôlées par FastQC. Pour le nettoyage, nous avons utilisé bbduck avec l'option -ktrim=l avant tout pour retirer les adaptateurs. Les lectures ont ensuite été alignées sur les génomes de Cedratvirus kamchatka, de *Acanthamoeba castellani* souche Neff et sa mitochondrie avec l'outil STAR. Ce dernier permet un alignement en coupant les lectures permettant de prendre en compte l'épissage alternatif des ARN messager. STAR a été utilisé avec les options «--alignIntronMax 2000 --alignMatesGapMax 5000 --twopassMode Basic --quantMode TranscriptomeSAM GeneCounts » et avec un fichier d'annotation gtf. Samtools coverage a ensuite permis de compter le nombre de lectures associées à chaque contig de référence. Pour compter le pourcentage de gènes exprimés, un gène a été comptabilisé à partir de l'alignement de 3 lectures à sa séquence.

XIII. Résultats et discussion

Premier cycle observé au microscope optique

La première observation du cycle de Cedratvirus kamchatka chez *A. castellani* avec une MOI de 50 a permis d'en définir les grandes étapes.

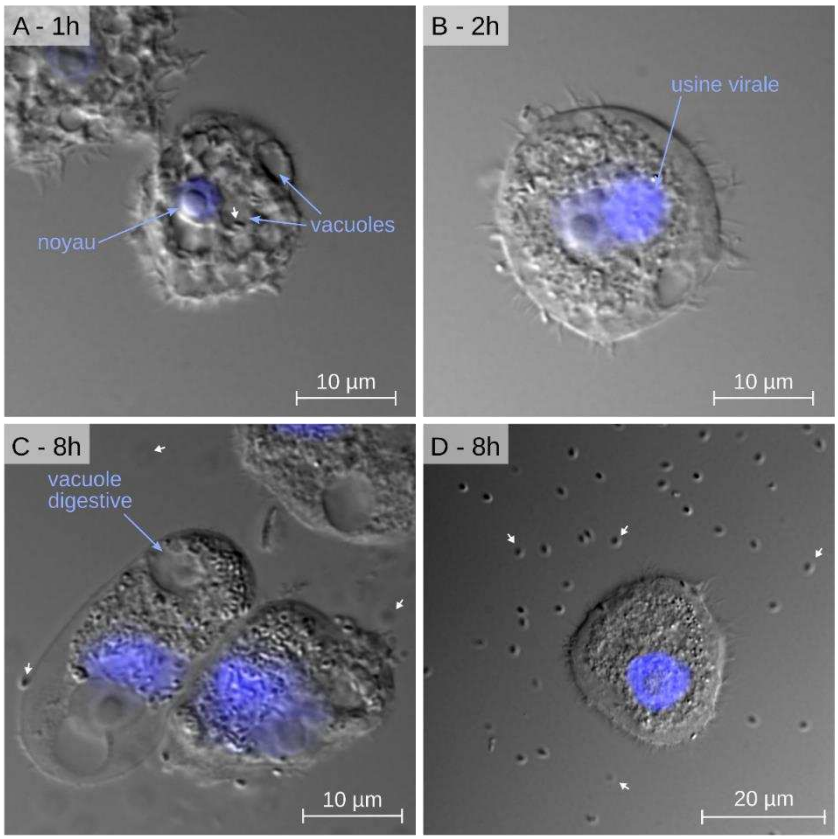


Figure 28 – Observation de cellules de *A. castellanii* infectées par Cedratvirus kamchatka Les cellules dont l'ADN est coloré au Hoechst ont été observées au microscope optique. Les petites flèches blanchent indiquent la présence de virions (non exhaustif).

On observe tout d'abord quelques virions dans les vacuoles de l'amibe dès 30 minutes et 1h post-infection (Figure 28A). Dans un premier temps seul l'ADN du noyau prend la coloration au Hoechst mais dès 1 h 30 puis, sans équivoque à 2 h post-infection, un marquage apparaît proche du noyau : c'est l'usine virale (Figure 28B). Cette dernière va grossir dans la cellule jusqu'à environ 8 heures après l'infection. Dès 3/4 h, on voit apparaitre des premiers virions néoformés dans la cellule, à l'extérieur de l'usine virale. Le nombre de ces virions ne fera qu'augmenter au moins jusqu'à 11 heures post-infection. Au temps d'infection tardifs, on peut voir l'espace dans la cellule presque entièrement occupé par les nouveaux virions, l'usine virale, le noyau et la vacuole digestive (Figure 28C). A 7 heures et surtout à 8 heures post-infection, apparaissent au microscope optique, des particules virales dans le milieu de culture (Figure 28C et D). Nous avons donc considéré le premier cycle comme terminé à 8h pour séquençage ARN.

Infection par Cedratvirus kamchatka observée au microscope électronique

L'observation du cycle infectieux au microscope électronique a permis de confirmer les observations faites au microscope optique et d'observer les modifications morphologiques de la cellule et des virions de manière détaillée. Le cycle de Cedratvirus kamchatka comporte les mêmes grandes étapes que celles observées pour d'autres *Pithoviridae* (dos Santos Silva et al. 2018; Legendre et al. 2014). Le virus est d'abord phagocyté puis un bouchon apical va être expulsé pour permettre la fusion des membranes du virion et du phagosome (Figure 29A). A 1 heure, comme à 30 minutes, des vacuoles contiennent des particules vides. C'est le signe que le contenu interne du virion a été relargué dans le cytoplasme. A 2h et 2h30, une zone éclaircie, sans organelles, peu définie mais légèrement moins granuleuse que le cytoplasme, s'est mise en place : c'est l'usine virale (Figure 29C). Dès 3 heures, des virions à plusieurs stades de formation sont visibles dans l'usine virale (Figure 29D). Tôt dans la morphogénèse, on aperçoit un début de membrane externe formée en demi-cercle ou en angles droits. Très rapidement le bouchon et la structure à sa racine sont visibles. Les membranes vides sont ensuite remplies, puis se rajoute le tégument, très dense en électrons, qui sera par la suite épaissi. Cela est cohérent avec l'analyse de la morphogénèse de Cedratvirus getuliensis (dos Santos Silva et al. 2018). A 6 heures post-infection nous voyons déjà qu'il y a beaucoup de virions matures sortis de l'usine virale (Figure 29E). Ces virions matures peuvent être entourés ou non de membrane (Figure 29F). Cela nous indique que probablement la membrane vient entourer les virions matures après leur sortie de l'usine virale. On aperçoit très distinctement une zone claire, vide autour des particules matures (Figure 29F). Comme chez les autres *Pithoviridae*, des temps très tardifs dans l'infection voient apparaître des « blobs », grands virions malformés, indiquant une accumulation excessive de matériel. Nous avons cependant fait une observation qui n'a pas été relevée auparavant ; avant la formation de l'usine virale, à 1h/1h30 post-infection, il semblerait que du matériel ressemblant à des membranes de l'appareil de golgi se rassemblent en un point proche du noyau (Figure 29B). Ce phénomène est observé de manière récurrente dans de nombreuses cellules mais n'ayant pas été remarqué par d'autres auteurs chez les *Pithoviridae*, il faudrait plus d'observations pour le confirmer.

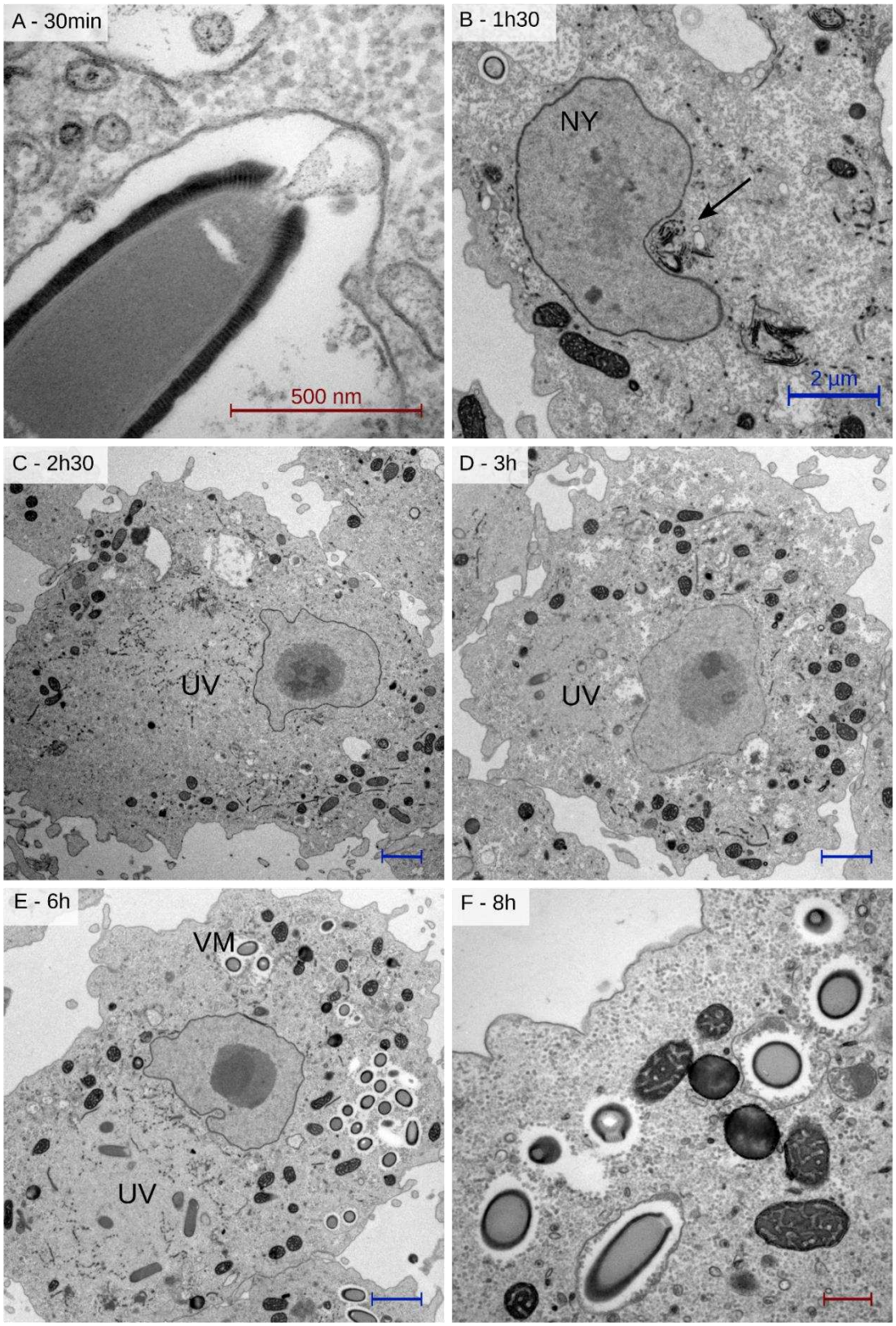


Figure 29 – Observation de cellules de *A. castallanii* infectées par Cedratvirus kamchatka vues au microscope électronique à transmission
Les échelles de taille bleues indiquent l'équivalent de 2 µm et les rouges, 500 nm. NY : noyau, UV : usine virale, VM : virions matures. (B) la flèche indique la structure accolée au noyau potentiellement formée par le virus

Il a été proposé à partir d'observations de Cedratvirus getuliensis qu'il y avait recrutement des mitochondries à la périphérie de l'usine virale (dos Santos Silva et al. 2018). En effet, nous remarquons également quelques usines virales avec de nombreuses mitochondries autour. Cependant, cette usine virale étant grande, il est envisageable que des mitochondries se trouvent en périphérie de cette zone qui « pousse » simplement les organelles vers l'extérieur. Dans de très rares cas, il nous a tout de même été donné d'observer des mitochondries au sein même des usines virales (annexe Figure 36).

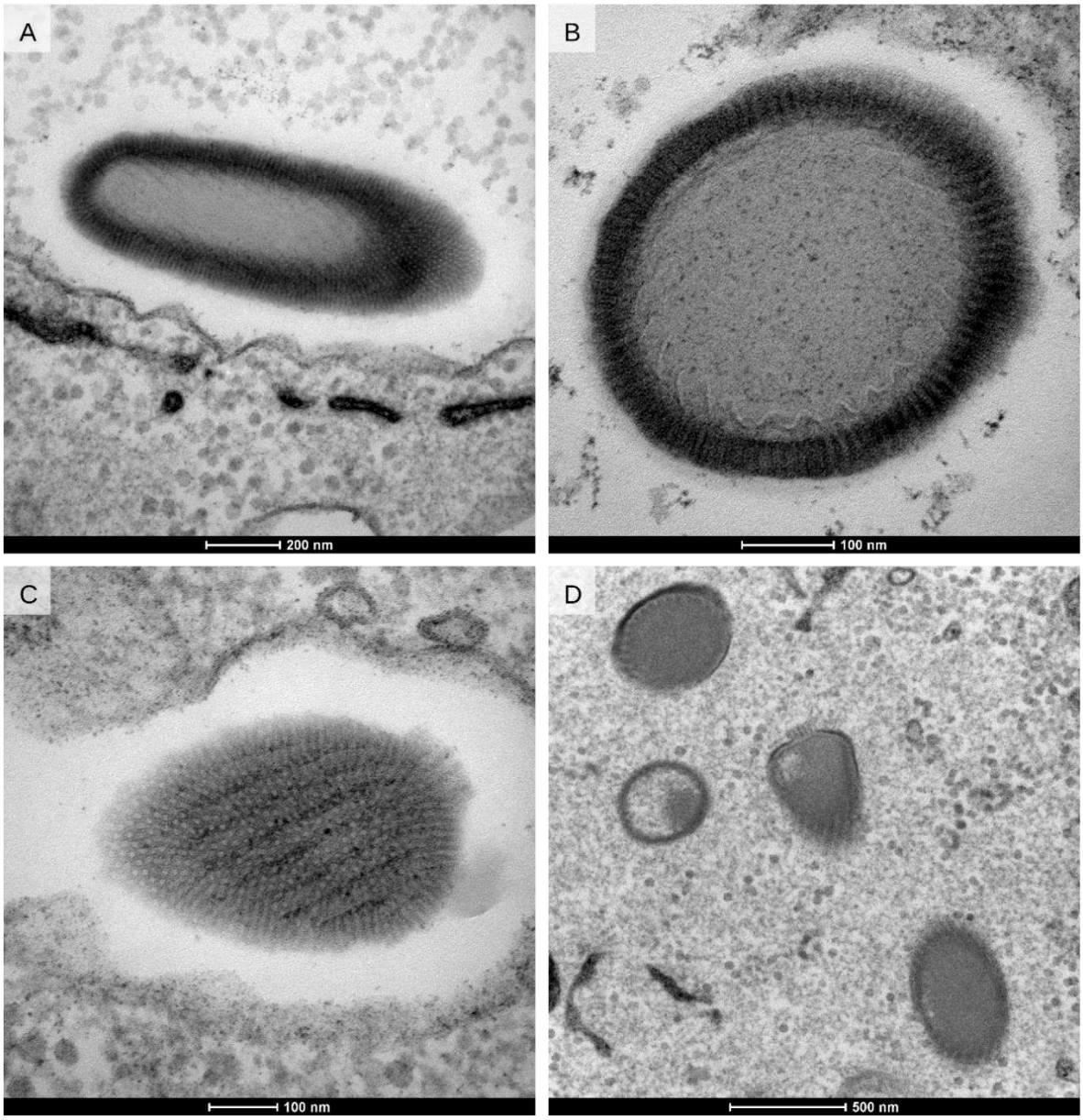


Figure 30 – Morphologie de la particule virale de C. kamchatka
La membrane interne et possiblement le tégument semblent former une structure hélicoïdale visible à 30 minutes (A), 1 heure (B), 1h30 (C) et chez les virions néoformés, 7h après l'infection (D).

Les images obtenues nous ont permis d'observer une structure encore passée inaperçue chez cette famille virale : la membrane interne et peut-être le tégument présenteraient un motif répété. Ce motif, sous forme de vaguelettes selon une coupe transversale (Figure 30B et D) et de strilles selon une coupe longitudinale (Figure 30A et C), a une période d'environ 40 nm. Nous l'apercevons chez des virions phagocytés par l'amibe (Figure 30A et B), sur le tégument de virions matures (Figure 30C) et aussi en formation (Figure 30D). Cette structure ressemble aux « ribs », les côtes, du virus de chétognathe récemment découvert, Meelsvirus (Shinn et Bullard 2018). La taxonomie de ce virus géants n'est pas encore connue.

La cinétique transcriptomique d'une infection par Cedratvirus kamchatka – résultats préliminaires

Le séquençage transcriptomique de 17 échantillons obtenus lors d'une cinétique d'infection a engendré des données homogènes en qualité. L'alignement par STAR a montré qu'en moyenne 94.5 % (+/- 0.36) étaient correctement alignés en un seul endroit. Ensuite, 3.38 % des lectures sont alignées à plusieurs endroits des génomes et 2%, trop courtes, n'ont pas pu être alignées.

Le transcriptome total est largement dominé par l'expression du génome nucléaire de l'hôte qui ne passe pas en dessous des 97% de lectures alignées. La courbe est inversement proportionnelle à celle de Cedratvirus kamchatka puisque le génome mitochondrial est très peu exprimé en comparaison. L'expression relative du virus est déjà plus élevée à 1h que chez le MOCK. L'expérience MOCK n'a d'ailleurs révélé aucune lecture virale indiquant qu'elle a fonctionné (Figure 31A). L'expression virale arrive à un maximum de 2.8% à 2h30 pour ensuite redescendre vers un plateau à 0.4% à partir de 4h. La faible expression du génome de la mitochondrie semble suivre chaotiquement celle du virus. Elle augmente jusqu'à 1h30/2h30 puis baisse lentement. Son expression relative est plus élevée avec le virus qu'avec le MOCK et ce, dès 1h, bien que les barres d'erreur se chevauchent (Figure 31A). Une profondeur de séquençage plus importante permettrait peut-être de voir une courbe de l'expression mitochondriale plus nette.

Si l'on normatlise à présent par la taille du génome transcrit, l'expression relative du virus parait à présent beaucoup plus importante. En effet, la somme de la longueur des gènes est 117 fois plus importante chez l'amibe que chez le virus. On voit à présent qu'à 1h30, l'expression virale relative équivaut à l'expression de l'hôte. A 2h et 2h30, l'expression relative est 3 fois plus importante chez le virus que chez le noyau de l'amibe. L'expression virale baisse et au plateau, c'est l'expression relative amibienne qui reprend le dessus par un facteur 2 (Figure 31B).

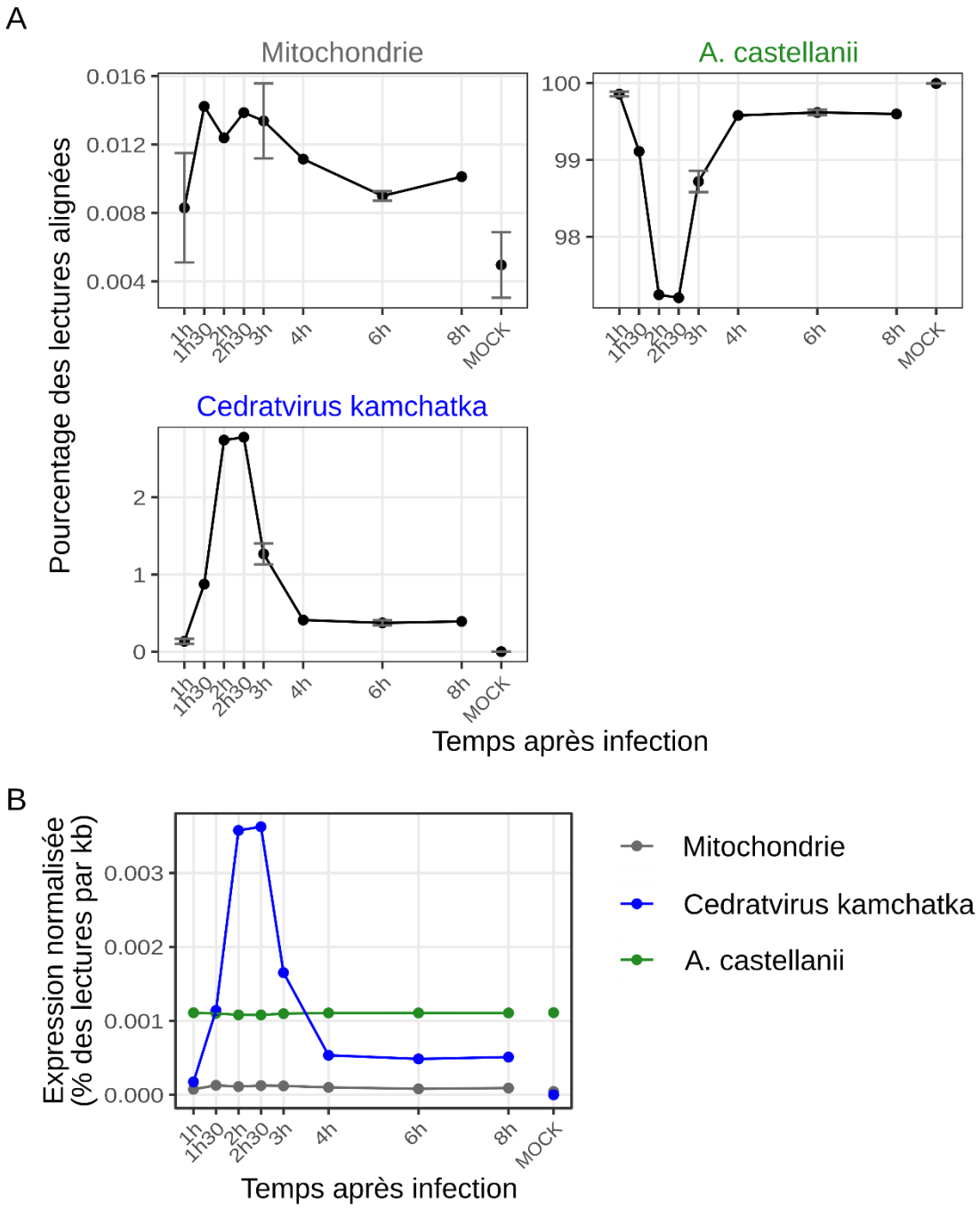


Figure 31 – Pourcentage de lectures alignées au génome viral et de l'hôte au court de la cinétique d'infection
(A) Pour chaque génome et chaque temps, le nombre relatif de lectures alignées a été compté. Les points à 1h, 3h, 6h et le MOCK sont en triplicats. (B) Expression normalisée par la taille du génome : les pourcentages obtenus en (A) ont été normalisés par la somme de la taille des transcrits potentiels en kilobases de chacun des trois génomes.

Le pourcentage de gènes viraux exprimés est aussi le plus élevé à 2h et 2h30 avec 96,5 % des gènes exprimés. Le génome mitochondrial arrive à son maximum à 1h30 avec 38.6%. L'amibe exprime 87.3% de ses gènes dans la condition MOCK et 90.2 à 1h30 après l'infection.

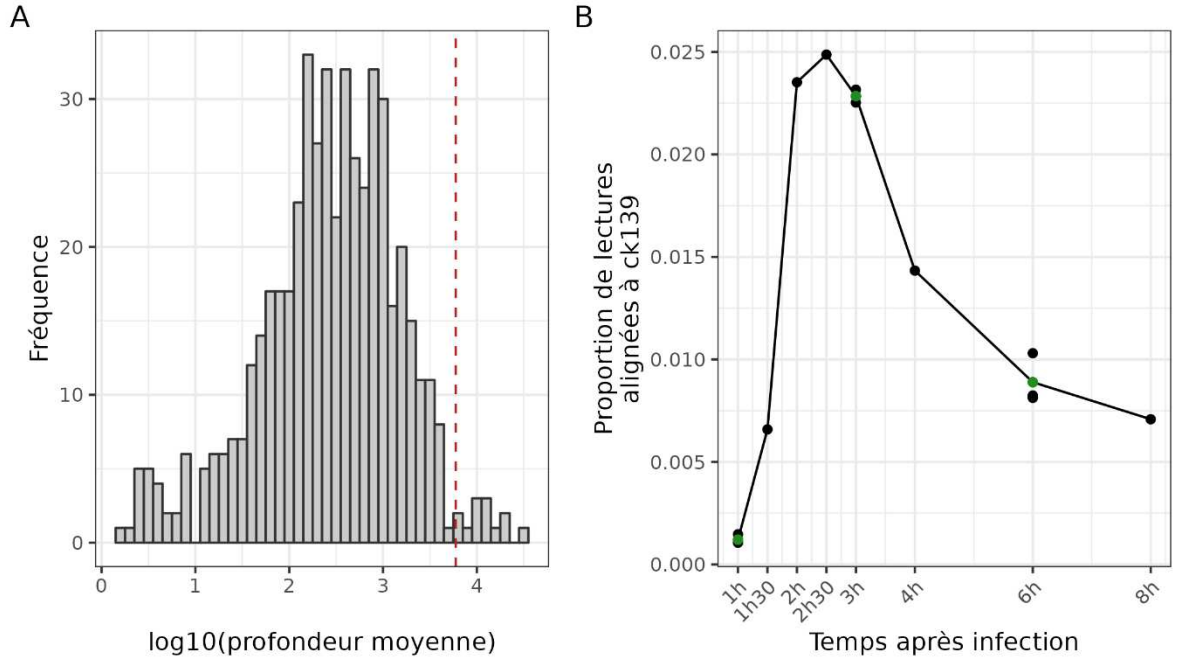


Figure 32 – Expression du gène de la MCP de Cedratvirus kamchatka
(A) Histogramme de l'expression des gènes de C. kamchatka. Les moyennes des points en triplicat ont été calculées avant de sommer la profondeur de séquençage des gènes pour tous les temps d'infection. La ligne verticale indique la profondeur de séquençage du gène de à la MCP. (B) Expression relative de la MCP. Le nombre de lectures alignées au gène de la MCP de C. kamchatka a été normalisé par le nombre total de lectures alignées au génome viral à un temps t. Les points verts à 1h, 3h et 6h indiquent la moyenne calculée sur les trois réplicats.

L'expression des gènes viraux est cependant très hétérogène comprenant cinq ordres de grandeur dans la profondeur de séquençage moyenne (Figure 32A). Parmi les gènes viraux, un retient notre attention dans cette analyse préliminaire. Il s'agit de la MCP. Nous avions fait l'hypothèse précédemment d'une reprogrammation fonctionnelle de ce gène, absent de la particule de P. sibericum (Legendre et al. 2014) mais présent chez tous les virus isolés et beaucoup de séquences métagénomiques (

Les protéines majeures de capside métagénomique aident à comprendre le lien entre celle des *Pithoviridae* et des autres virus géants, p. 105). En effet, le gène de la MCP, ck139, fait partie des quelques gènes très fortement exprimés chez Cedratvirus kamchatka (Figure 32A). Son expression normalisée par l'expression virale totale n'est pas non plus constante. Elle est maximale à 2h30 après l'infection, moment où la transcription virale est aussi la plus importante, et diminue progressivement (Figure 32B). Ce gène semble a priori régulé. Sachant que la MCP n'est plus une protéine constituante du virion (Legendre et al. 2014), son profil d'expression chez Cedratvirus kamchatka est le signe d'une reprogrammation fonctionnelle ou, moins probablement, d'un vestige de son utilité passée. Cette dernière explication est moins probable puisque les gènes de la MCP sont généralement exprimés tardivement comme chez Marseillevirus (Rodrigues et al. 2020).

Discussion générale et perspectives

I. Des virus proches de *Pithoviridae* et *Orpheoviridae* peuvent être très abondants et divers dans le pergélisol russe

Les échantillons étudiés en Sibérie et au Kamchatka ont montré une abondance virale très hétérogène. L'étude de cette diversité montre une surprenante abondance de séquences apparentés avec les *Pithoviridae* et *Orpheoviridae*. Après estimation de l'abondance par la somme des couvertures associées aux différents groupes, on obtient des *Pithoviridae* et *Orpheoviridae* non seulement divers mais aussi abondants. Ce n'est donc pas un hasard si parmi les huit scaffolds supérieures à 500 kb, quatre font partie de ces groupes. Nous avons vu aussi que certaines séquences virales étaient présentes dans plusieurs échantillons d'âges différents. Ces virus ont donc perduré dans le temps.

La diversité présentée ici rappelle celle observée dans des sédiments profonds à 15 km de la source hydrothermale de Loki's Castle (Introduction, p. 50). Cependant, l'alignement des lectures à ces génomes montre que ce ne sont pas les mêmes virus. Cela contraste avec la diversité trouvée dans d'autres études métagénomiques globales (Moniruzzaman et al. 2019; Schulz et al. 2020) qui ont d'avantage mis en avant la grande diversité des *Phycodnaviridae* et des *Mesomimivirinae*.

Nous avons fait l'hypothèse que ces différences s'expliquent par deux facteurs. Le premier étant la méthode appliquée : plus on donne de l'importance aux gènes de la MCP et de l'ATPase d'empaquetement pour reconstruire la phylogénie ou sélectionner des bins viraux, moins on a de chances de mettre en avant les *Pithoviridae*. Pour étudier cette hypothèse nous pouvons comparer la récupération de scaffolds en utilisant les profils métagénomiques de l'étude de (Schulz et al. 2020) avec ou sans les profils spécifiques de *Pithoviridae*. Ces derniers ont permis de récupérer des séquences de groupes proches des *Pithoviridae* mais n'ont augmenté que de 5% le nombre de séquences virales totales. Cette explication ne suffit donc pas. La deuxième hypothèse que nous avons fait est que les différences taxonomiques s'expliquent par le milieu étudié. En effet, bien que plusieurs études métagénomiques des virus géants se veuillent globales, les bases de données sont souvent plus riches en données marines et aquatiques (et par des microbiotes humains) que terrestres. Cela s'explique par la difficulté d'assembler des jeux de données complexes que sont les données de métagénomique du sol. Ainsi, nous avons comparé la diversité taxonomique de *Nucleocytoviricota* dans divers types d'environnements. Les données terrestres se sont avérées plus compliquées à analyser mais nous avons tout de même pu conclure que les *Pithoviridae* et *Orpheoviridae* sont moins présents dans les données de métagénomique marine et plus abondants dans les échantillons étudiés du pergélisol, dans des eaux usées et dans du sol de forêt. Notre étude ne prend cependant pas en compte d'éventuelles différences de rendement dans l'extraction d'ADN. Cette procédure est généralement beaucoup plus aisée pour un virus de la famille des *Mimiviridae* qu'un virus de type

Pithoviridae et donc influe grandement sur la diversité estimée. En l'absence de données quantifiées, nous ne pouvons pas estimer l'influence de ce biais.

Tout d'abord, la diversité des séquences virales dans les échantillons indique que leur abondance n'est pas due à une espèce virale unique qui aurait pu se multiplier fortement à un moment donné. A l'inverse, l'abondance importante de virus géants dans certains échantillons est due à une diversité complexe. De plus, certains virus se sont maintenus dans des échantillons ayant jusqu'à 13 000 ans d'écart ou plus. Ces virus sont donc potentiellement des acteurs importants de leurs milieux et leur rôle dans les sols devrait par conséquent être plus précisément investigué. Pour cela il faudrait commencer par connaitre l'hôte des *Pithoviridae* divergents. Nous pouvons faire l'hypothèse qu'au moins certains d'entre eux sont capables d'infecter *Acanthamoeba castellanii* et *Vermamoeba vermiformis*, deux espèces présentes dans l'eau douce et dans les sols et, dans une moindre mesure, dans l'eau de mer. Ceci pourrait expliquer pourquoi les *Pithoviridae* ne sont pas retrouvés dans des échantillons marins. L'alternative serait que ce ne soit pas l'hôte mais les virus qui ne résistent pas à la salinité. Cependant, cela est étonnant pour des virus qui résistent à 30 000 ans dans le pergélisol (Legendre et al. 2014), à l'éthanol, au peroxyde d'hydrogène et à une température de 55°C (Bertelli et al. 2017). Une dernière possibilité est que ces virus sédimentent trop vite de par leur forme, leur densité et l'absence de fibrilles (cheveux) présents chez les *Mimiviridae*.

Ces interrogations ne trouveront probablement pas de réponse dans les jeux de données publiques de Mgnify ou du JGI car les métadonnées associées sont insuffisantes pour avoir un recul et une compréhension écologique des échantillons. C'est également ce qui limite le projet Permagenomics présenté ici. Le sol et surtout le pergélisol étant très hétérogènes, une grande variabilité a été observée dans la composition virale des échantillons. Cela rend plus difficile la comparaison et l'extrapolation. De plus, les connaissances actuelles sur les protozoaires du pergélisol sont surtout basées sur la culture de ces derniers. Les données métagénomiques étudiées ici ont révélé peu d'eucaryotes. Cela est dû à des biais bioinformatiques mais également probablement à l'enkystement des amibes du pergélisol qui rend l'extraction d'ADN impossible en conditions normales. Au vu de ces résultats il serait intéressant d'entreprendre un projet pour mieux comprendre spécifiquement les interactions eucaryotes-virus qui ont lieu dans le pergélisol. Une étude comparative à l'interface de la couche active et du pergélisol pourrait nous aider à comprendre pourquoi les virus géants et en particulier *Pithoviridae* et *Orpheoviridae* divergents, forment dans certains échantillons de Yukechi une part considérable du microbiome. Peut-il en être ainsi dans la couche active également grâce aux dynamiques hôte-pathogène ou cela traduit-il une meilleure résistance aux conditions du pergélisol ? La métagénomique peut être complétée par une étude des eucaryotes via le métabarcoding des gènes de la sous unité 18S de l'ARN ribosomique et la région de l'espaceur interne transcrit (ITS) de l'ADN ribosomique. L'ADN des cystes amibiens pour être extrait nécessite que les échantillons soient chauffés à très haute température (Laummaunwai, Ruangjirachuporn, et Boonmars 2012). Des observations directes au microscope, bien que fastidieuses pour les échantillons de sol, ainsi que de la cytométrie en flux pourrait venir confirmer ou non notre constat d'une plus grande abondance de virus géants que d'eucaryotes dans les échantillons. Enfin, pour tenter de trouver de manière fiable l'hôte des virus environnementaux, il faudrait procéder à un séquençage de cellules uniques d'organismes d'intérêt pour espérer y trouver des interactions avec des virus. D'autres milieux candidats à explorer pour étudier les *Pithoviridae* et *Orpheoviridae* sont d'après nos analyses

les eaux usées et les sols de forêt. Soulignons que deux bins de *Pithoviridae*/*Orpheoviridae* divergents ont été étudiés dans le cadre d'une étude métagénomique du sol de forêt d'Harvard (Schulz et al. 2018).

II. Métagénomique et isolement ne s'accordent pas

Face à l'analyse des scaffold métagénomiques nous constatons l'absence de séquences des virus isolés à partir de ces mêmes échantillons. Les virus suffisamment abondants pour avoir été séquencés et assemblés en contigs ne sont pas les virus cultivés. A l'échelle de la lecture on retrouve cependant Pithovirus sibericum dans son échantillon mais pas Cedratvirus kamchatka. La réciproque de cette observation est que nous n'arrivons pas à isoler les virus importants observés en métagénomique comme Hydrivirus alors même que l'IGS a développé une expertise dans le domaine.

Une faible abondance dans le métagénome de l'échantillon à partir duquel Mollivirus kamchatka avait été isolé a également été observé avec seulement 14 lectures alignées (Eugene Christo-Foroux et al. 2020). *Mollivirus sibericum* et *Pithovirus sibericum* ont en revanche 336 et 125 lectures de leurs échantillons correctement alignées respectivement (Legendre et al. 2015). Cela n'est pas suffisant pour l'assemblage. Une étude de divers échantillons de microbiotes humains (hors intestinal) a réussi à cultiver 61 espèces bactériennes et fongiques. 5 de ces espèces n'ont pas été retrouvées par métagénomique. Les auteurs expliquent cela par l'erreur humaine (l'espèce identifiée visuellement sur la plaque de culture n'est peut-être pas la bonne), par de la contamination ou par une abondance bactérienne trop faible qui empêcherait des espèces en particulier d'être séquencées (Abayasekara et al. 2017). Cette dernière explication est la plus probable dans le cas de nos virus. L'abondance est suffisante pour un faible séquençage (en général) mais pas pour l'assemblage. Réciproquement, ne pas réussir à cultiver des organismes prédits en métagénomique est classique et a ses racines dans ce qui a été appelé « La grande anomalie du dénombrement »[30]. Dans une étude du microbiote de serpent, seul 2% des genres bactériens identifiés par un séquençage 16S en métabarcoding ont pu être cultivés (Mao et al. 2021). L'incapacité à isoler les virus abondants vus en métagénomique vient simplement du fait que nous ne connaissions pas l'hôte exacte du virus. C'est probablement ce biais d'hôte qui explique que les *Pithoviridae* isolés sur *A. castellanii* se retrouvent, en phylogénie, dans une branche monophylétique, séparée des séquences métagénomiques. Une explication supplémentaire possible est que les virus puissent être inactivés après des dizaines de milliers d'années passées dans le pergélisol. Enfin, l'hétérogénéité du sol pourrait également donner des abondances de virus très variables à l'échelle du centimètre comme c'est le cas pour la biodiversité bactérienne (O'Brien et al. 2016). Pour contrer cet effet et pour avoir suffisamment de matériel génétique, de grandes quantités de pergélisol ont été utilisées pour extraire l'ADN. Ce n'a pas été le cas pour les échantillons de surface, plus divers (article I) et riches en ADN où seul 0,25 g de sol a été utilisé, ce qui a réduit drastiquement les chances de retrouver *Cedratvirus kamchatka*. Comme pour la section précédente, un séquençage de

[30] Traduction de « The great plate count anomaly »

cellules uniques pourrait nous réorienter sur de nouveaux hôtes pour isoler plus efficacement, si l'hôte est cultivable, de nouveaux virus de familles encore sans représentant isolé.

III. La mosaïque de fonctions et les échanges de gènes entre virus questionnent la monophylie des *Nucleocytoviricota*

Dans l'article II nous avons fait l'analyse des fonctions encodées dans les séquences de virus géants assemblées à partir de l'ADN total d'échantillons de pergélisol. L'étude des gènes cœur présents dans ces séquences et dans les génomes viraux de référence montrent qu'il peut y avoir une cohérence taxonomique dans les fonctions encodées. C'est l'exemple de l'ATPase d'empaquetement, gène présent chez tous les *Nucleocytoviricota* sauf chez les *Pithoviridae* et *Orpheoviridae* isolés et métagénomiques. En revanche, cette cohérence semble perdue quand on regarde la distribution des gènes accessoires[31]. Des fonctions typiquement présentes chez les *Mimiviridae* comme les aminoacyl-ARNt transférases, les ARN de transfert ou les gènes de protéines de choc thermique sont présents chez Hydrivirus ou d'autres virus de ce type ainsi que chez Orpheovirus alors qu'ils sont absents chez tous les autres *Pimascovirales* isolés. Les phylogénies de six des aminoacyl-ARNt transférases (article II, Fig. S18) ainsi que celle de la sous-unité H2A des histones et de l'hémoglobine tronquée (annexe III, p. 193) ont montré un échange très probable entre virus impliquant des *Mimiviridae*, *Pimascovirales* et/ou *Phycodnaviridae*.

Ce n'est pas la première fois que des auteurs s'étonnent de fonctions présentes chez les virus géants métagénomiques. C'est le cas par exemple de gènes de la photosynthèse et aussi de gènes de la rhodopsine, abondement présents dans les données de (Schulz et al. 2020). Des gènes de la glycolyse seraient également présents chez les virus géants (Moniruzzaman et al. 2019). On peut constater une plasticité génomique importante chez les virus géants (introduction p. 25). Cela avec l'échange de gènes entre virus questionne la méthodologie de l'ICTV de baser la classification de tous les virus sur une protéine unique ; la protéine majeure de capside. Puisque le partage d'hôte encourage l'échange de gènes (T.-W. Sun et Ku 2021; Mönttinen et al. 2021) et que certains *Nucleocytoviricota* partagent un hôte nous sommes en droit de questionner la monophilie de ce groupe. Ce phylum comprend des virus dont la morphologie des virions est extrêmement diverse et dont le cycle réplicatif peut être variable également (certains sont strictement cytoplasmiques et d'autres ont une phase nucléaire). Le schéma de présence/absence de fonctions est aux premiers abords chaotique en comparaison avec la phylogénie des *Nucleocytoviricota* (article II). Il en va de même pour la présence/absence d'orthogroupes (Koonin et Yutin 2018). Pourtant, la phylogénie estimée à partir de ce schéma de présence/absence peut être cohérente avec l'arbre des *Nucleocytoviricota* tel qu'il est proposé aujourd'hui (Bäckström et al. 2019). A l'origine, ce phylum avait d'ailleurs été défini selon ce type de méthode, par cladistique, sur les *Poxviridae*, les *Phycodnaviridae*, Iridovirus et African Swine Fever Virus (Iyer, Aravind, et Koonin 2001).

[31] Utilisé ici pour désigner les gènes présents sporadiquement, dont les fonctions ne font pas partie des plus partagées par toutes les familles

Mais déjà, les auteurs notaient que la phylogénie ne rétablissait pas cette monophylie. Ils justifient cela par une méthode peu adaptée aux familles virales anciennes dont les mutations plus fréquentes que dans le monde cellulaire peuvent changer l'histoire évolutive apparente. Depuis certains auteurs, Eugène Koonin (chercheur au NCBI) notamment, grand défenseur de cette famille virale, ont étudié et questionné sa monophyllie directement. Un article en particulier décrit une analyse phylogénétique complète des orthogroupes probablement ancestraux de cette famille et leurs homologues cellulaires et viraux (Yutin et Koonin 2012). On constate que certaines phylogénies ne sont pas en accord avec la monophylie de la famille et qu'il faut contraindre l'arbre (forcer une certaine topologie) pour la retrouver. La phylogénie de l'ADN polymérase non contrainte, sépare les *Asfarviridae* et *Poxviridae* du reste de la famille (Kazlauskas et al. 2020), cependant l'arbre contraint avec tous les *Nucleocytoviricota* et les *Baculoviridae* est plus vraisemblable (meilleur « Log-likelighood ») (Yutin et Koonin 2012). Certaines phylogénies contraintes, comme celle de la primase D5 ou l'ARN ligase, ont en revanche été statistiquement rejetées. Malgré cela, la balance penche plutôt vers la monophylie de la famille dans leur analyse. Une autre équipe de recherche a questionné cette phylogénie via l'analyse des orthogroupes. Il en ressort qu'un seul gène est partagé par tous et une dizaine sont présents dans la majorité des génomes (Mönttinen et al. 2021). Dans cette même étude, quatre des cinq phylogénies sur les gènes les plus partagés dans la famille sont en accord avec la monophylie. Cependant, au vu de la plasticité génomique et de l'échange de gènes possible, dont l'ATPase d'empaquetement de type *Nucleocytoviricota* retrouvée chez Yaravirus, un virus probablement externe au phylum, les auteurs concluent que le phylum n'est peut-être pas monophylétique. Pourtant, la plasticité génomique ne contredit pas la monophylie. Au contraire, elle expliquerait la difficulté à la retracer. De plus, nous constatons une constance dans l'histoire retracée par les gènes cœur de la famille quand on ne prend qu'eux dans la phylogénie (Guglielmini et al. 2018). Cette histoire évolutive se reflète également dans les résultats de BLASTP de la MCP (entre autres) de ces virus (Chapitre 2, p. 105). L'histoire évolutive retracée par ces gènes cœur aux fonctions essentielles est un argument fort en faveur du phylum. L'analyse phylogénétique de gènes accessoires impliquant également des séquences métagénomiques peut aussi être cohérente avec une monophylie des *Mimiviridae* et *Phycodnaviridae* ; deux familles très diverses, sans aucune contrainte (Moniruzzaman et al. 2019).

En prenant en compte tous ces résultats et en considérant également que la polyphylie signifie que ces virus auraient dû émerger plusieurs fois dans l'histoire, ce qui n'est pas non plus très parcimonieux, nous avons fait le choix d'admettre la famille des *Nucleocytoviricota*. De plus, la répartition des meilleurs résultats de BLASTP du grand virus métagénomique inconnu M_b2437_k1 (article II, chapitre 1) comprenant presque toutes les grandes familles de *Nucleocytoviricota* (mais pas les *Poxviridae*) renforce l'idée selon laquelle cette famille est en grande partie réelle. Bien que cette monophylie semble probable, il pourrait être intéressant que d'autres équipes reproduisent l'étude de (Yutin et Koonin 2012) également en faisant une analyse statistique comparant des arbres contraints et non contraints selon plusieurs hypothèses. Il se pourrait que les nouveaux virus isolés ou séquencés depuis renforcent l'analyse statistique ou à l'inverse, fassent tomber ce phylum. Il serait nécessaire de

requestionner le lien avec les *Asfarviridae* et *Poxviridae* puisqu'ils semblent les plus lointains et donc, compliqués à évaluer.

IV. Les génomes des *Cedratvirus* sont conservatifs

Les génomes des *Pithoviridae* sont petits comparés à Orpheovirus et aux autres virus géants d'amibe, exceptés les *Marseilleviridae*. L'analyse des orthogroupes a montré que le pangénome des Cedratvirus était fermé. Peu de nouveaux orthogroupes sont ajoutés à chaque nouveau Cedratvirus séquencé. De plus, la fraction des gènes qui ne sont pas partagés par tous les *Pithoviridae* est souvent partagée par toute la sous-famille : *Pithovirus* ou *Cedratvirus*. Tous ces résultats semblent montrer des gènes bien conservés malgré une asymmétrie génomique comprenant une ou des régions plus variables.

Ce résultat contraste avec l'importante plasticité génomique chez les virus géants (introduction : gigantisme génomique, p. 25). C'est ce que nous montrons par la figure 2 de l'article III en appliquant la même méthode à tous les groupes analysés : les *Pithoviridae* sont plus conservatifs que les *Pandoraviridae*, les Mega*mimiviridae* et même les *Ranavirus*. Déjà, la comparaison des gènes deux à deux chez *Pithovirus sibericum* et *massiliensis* avait donné une valeur de dN/dS moyenne comparable avec celle retrouvée chez des bactéries du genre Neisseria et plus basse (donc mieux conservé) que chez les *Helicobacter* analysés (Levasseur et al. 2016).

Aux vus des expériences d'évolution contrôlées menées sur un *Marseilleviridae* (Mueller et al. 2017) et le virus de la vaccine (Elde et al. 2012) nous savons qu'un environnement stable dans lequel le virus est bien adapté peut mener à une réduction génomique. Nous savons également que les *Pithoviridae* sont spécialistes des Acanthamoeba du groupe II (Bertelli et al. 2017; Hoffmann et al. 1997). Ce dernier point néanmoins est peu certain car un nombre très restreint d'hôtes a été testé. Cependant, avec ces informations nous pouvons supposer que les *Pithoviridae* sont des virus spécialistes et ainsi, ont connu des périodes de réduction génomique. Cela expliquerait que leur génome soit bien conservé puisqu'il est adapté à un milieu particulier. Au contraire, l'importante duplication chez les *Cedratvirus* du gène contenant des répétitions de type Ankyrine (36.3 copies en moyenne) sont la trace d'une période d'expansion génomique. Cela pourrait être le signe d'une évolution en accordéon comme ce qui a été montré chez le *Poxviridae* (Elde et al. 2012).

Une limite à ce raisonnement est que peu de *Cedratvirus* et très peu de *Pithovirus* ont été séquencés. Peut-être sont-ils particulièrement proches. Deux nouveaux génomes de *Cedratvirus* devraient bientôt apparaître dans les bases de données (dos Santos Silva et al. 2018; Kördel et al. 2021) et certainement d'autres encore dans les années qui viennent. Il conviendra alors de refaire l'analyse des gènes orthologues pour améliorer le modèle concluant à la fermeture du pangénome. Il serait intéressant de reproduire des expériences d'évolution accélérées en conditions de laboratoire avec (Mueller et al. 2017) ou sans (Boyer et al. 2011) compétiteurs bactériens ou viraux et avec ou sans bactérie symbiotique. Cette dernière condition n'a jamais été testée sur un virus géant. Cela pourrait nous indiquer

éventuellement si le génome est capable de réduire fortement sa taille et si certains gènes en particulier sont nécessaires pour l'adaptation à une condition spécifique. Enfin, nous avons vu que les génomes des *Cedratvirus* semblent bien conservés entre-eux. Cependant, les génomes ont beaucoup divergé depuis le dernier ancêtre commun des *Pithovirius* et *Cedratvirus* puisque les deux familles ont environ une moitié de leurs orthogroupes absents chez l'autre groupe. Il serait intéressant de chercher à définir, par d'autres moyens qu'avec Orthofinder, l'origine de ces orthogroupes non partagés. Sachant que peu d'entre eux proviennent d'un transfert horizontal peut-être sont-ils des homologues lointains ayant muté rapidement ? Pour vérifier cette hypothèse nous pourrions procéder à un alignement des structures de protéines calculées par AlphaFold. En effet, le consensus scientifique dit que la structure protéique est mieux conservée que sa séquence (Illergård, Ardell, et Elofsson 2009). Des auteurs s'en servent déjà pour rétablir des homologies lointaines en commençant par la recherche des meilleurs alignements structuraux réciproques (Monzon et al. 2022).

V. Les séquences répétées chez Pithovirus couvrent une part importante du génome et ressemblent à des MITE organisés

23% des génomes des Pithovirus sont constitués de séquences répétées. Ces séquences palindromiques sont rassemblées dans des régions donnant un aspect de motif écossais aux « dotplots », représentations graphiques de la matrice d'alignement de ces génomes. Les deux sous-unités principales de ces régions ressemblent fortement à des MITE (Miniature Interted-repeat Transposable Element) ce pourquoi nous les avons nommées M1 et M2. Ces séquences sont uniques, à l'exception de la séquence d'insertion, TA, typique des MITE qui est retrouvée. Nous nous sommes donc basés sur leur architecture plutôt que sur leur séquence pour les classer. Les régions sont structurées par une succession de M1 et de M2 tel que (M1-M2){1 à 8 fois}-M1.

Lors de la découverte de *Pithovirus sibericum*, l'importance des séquences répétées dans le génome sautait déjà aux yeux avec une estimation de 21.2% du génome (Legendre et al. 2014). Les MITE ne sont pas communs dans les génomes viraux mais ils existent. Les *Pandoraviridae* possèdent de nombreux MITEs de la famille Submariner dont la transposase associée a été identifiée chez l'hôte, *A. castellanii* (C. Sun et al. 2015). *Pandoravirus salinus*, riche en MITE, n'a pourtant qu'environ 38 000 bases dans son génome soit environ 1.5%, représentées par ces répétitions. Les auteurs trouvent également des MITE chez deux bracovirus, un ascovirus et un ichnovirus (Zhang et al. 2018). Une forte proportion de répétitions est en général une caractéristique des génomes eucaryotes. Le génome humain a entre 50% (selon RepeatMasker) et 69% (selon P-Clouds) de séquences répétées dans son génome (Koning et al. 2011). Chez les eucaryotes, se sont majoritairement les séquences répétées qui expliquent les différences de taille de génome après les duplications de chromosomes entiers. De plus, les duplications de larges fractions génomiques sont souvent permises par la présence de microsatellites (Richard, Kerrest, et Dujon 2008). Contrairement aux répétitions en tandem, les MITE font partie de ce qui est appelé les répétitions dispersées. Cependant, il a déjà été vu que des MITE se suivent dans le génome. Dans le génome du riz ce

sont 11% des MITE qui sont présents en multimères (Jiang et Wessler 2001). Chez le Maïs, il semblerait également que certaines familles de MITE aient des préférences d'insertion précises dans d'autres MITE (Jiang et Wessler 2001). Pourtant, une structure impliquant des MITE aussi organisés que chez les Pithovirus n'a pas été décrite auparavant. L'ADN satellite, constitué de répétitions directes dont l'unité de base fait de quelques paires de bases à quelques centaines de paires de bases, est en revanche souvent plus organisé quand observé sur un dotplot. Les répétitions directes peuvent être concentrées comme chez la cuscute d'Europe où l'on retrouve des blocs d'ADN satellite séparés par d'autres répétitions dont des rétrotransposons (Vondrak et al. 2021), Figure 33). Les centromères sont typiquement des régions riches en ADN satellite et en transposons (Hartley et O'Neill 2019).

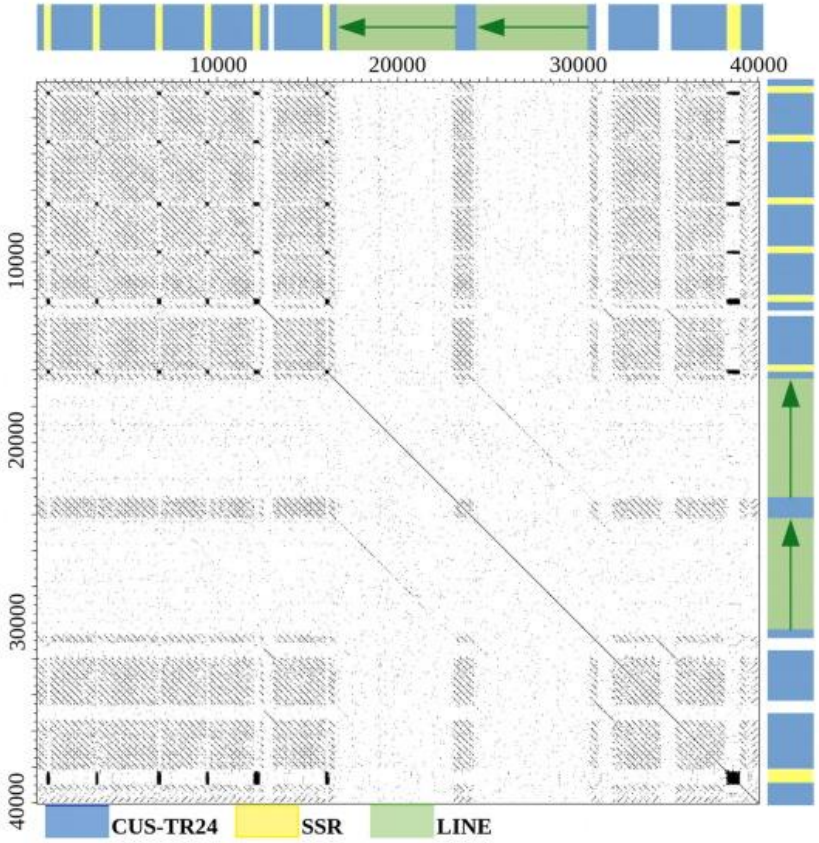


Figure 33 – Matrice d'alignement d'une lecture Nanopore contenant le locus CUS-TR24 chez *Cuscuta europaea* (Vondrak et al. 2021)
L'alignement de 40 kb d'une lecture de 98 kb est montré ici. La séquence est annotée par les rectangles colorés. Le bleu indique les blocs d'ADN satellite, le jaune ; des séquences répétées, le vert ; les rétrotransposons de type LINE.

Nous avons donc chez les Pithovirus des unités répétées qui ont la structure de MITE et dont l'organisation rappelle celle de l'ADN satellite. Les *Pithoviridae* métagénomiques divergents, eux, peuvent contenir de grandes portions d'ADN satellite. Les mécanismes de multiplication des séquences répétées ne sont pas très bien compris. Nous savons cependant que l'ADN satellite s'étend par glissement de l'ADN polymérase (Leclercq, Rivals, et Jarne 2010). Les MITE, eux, peuvent changer de position par transposition ou se multiplier selon le mode de transposition « copier-coller » à l'aide d'un transposon actif. Un MITE est souvent isolé mais

peut aussi s'insérer au sein d'un autre type de MITE ou au sein d'un MITE de la même famille (Jiang et Wessler 2001). La séquence interne du MITE détermine si celui-ci est encore actif, c'est-à-dire s'il est capable de lier la transposase et de se multiplier (Yang et al. 2009). Il est proposé que le MITE et le transposon autonome associé aient une origine commune et donc, ont des répétitions terminales (TIR) qui se ressemblent.

Le mécanisme de transposition des séquences répétées chez Pithovirus ne semble pas être celui de MITE classiques puisque nous avons vu qu'il y avait probablement multiplication par extension au sein des régions. Il y a aussi transposition puisqu'une région répétée M1-M2-M1-M2-M1 est présente chez *P. sibericum* et absente chez *P. massiliensis*. Nous sommes donc face à une répétition qui peut-être se multiplie par extension comme les séquences satellite et qui se transpose comme les MITE.

Nous ne savons pas non plus si ces répétitions sont parasitaires ou ont évolué vers une fonction bénéfique pour le virus. Dans le cas où la séquence serait parasitaire : Comment les Pithovirus arrivent-ils à rester compétitifs ? Un élément de réponse est que ces répétitions se trouvent dans les régions non transcrites (Legendre et al. 2014). Peut-être que ces MITE ont évolué pour se transposer à côté d'un autre MITE proche pour éviter de se transposer dans un gène et alors disparaître avec le virus. Une autre question se pose si la séquence est parasitaire : la taille du génome est-elle un facteur limitant dans la valeur sélective[32] du virus ? Si c'est le cas, alors ces séquences rallongeant le temps et les ressources pour la réplication mais n'apportant pas de bénéfices, pourraient mener les Pithovirus à leur perte.

Une première approche pour tenter de savoir si ces répétitions peuvent être bénéfiques est de savoir si elles lient des protéines et lesquelles. Une méthode de choix pour cela en l'absence de candidat est l'immunoprécipitation inversée de la chromatine (R-ChIp) suivie de spectrométrie de masse pour déterminer les protéines liées à la séquence d'intérêt. Si nous posons l'hypothèse d'un rôle structurel des régions répétées des *Pithoviridae,* comme c'est le cas pour les régions répétées chez les eucaryotes (Cournac, Koszul, et Mozziconacci 2016; Winter et al. 2018), nous pouvons également étudier la structure 3D du génome. Cela peut se faire à travers le séquençage Hi-C, donnant une carte de contacts de l'ADN. Pour ce qui est du mécanisme de transposition, nous sommes face à une énigme. Avant de chercher à l'identifier il faut que nous sachions s'il est reproductible en laboratoire dans nos conditions. Pour commencer, une expérience d'évolution accélérée avec de nombreux cycles puis séquençage en longues lectures pourrait nous indiquer si les MITEs se transposent ou s'étendent avant de chercher à en étudier le mécanisme.

[32] Traduction de « fitness », concept en écologie permettant d'estimer la sélection naturelle via la quantification du succès reproducteur d'un organisme

VI. Le cycle infectieux de Cedratvirus kamchatka produit de grandes usines virales mais le transcriptome n'est jamais dominé par le virus

Comme nous l'avons constaté par microscopie électronique, les limites de l'usine virale ne sont pas bien définies mais une grande surface moins granuleuse et sans organelle est visible dès les débuts du cycle, vers 2 heures après l'infection (Figure 29, p. 149). Une fois installée elle continue à produire des virions, même défectueux, jusqu'à la lyse cellulaire. L'analyse préliminaire du séquençage du transcriptome de Cedratvirus kamchatka et de son hôte à différents temps d'infection a donné un maximum relatif de transcription virale à 2h30 après infection. Alors, elle ne représente que 2.8% des lectures séquencées et alignées. Cette valeur normalisée par la somme de la taille du génome transcrit montre néanmoins que l'activité transcriptionnelle par kb du virus est trois fois plus importante que celle de l'hôte à 2h30 post-infection pour baisser par la suite.

Il est étonnant de voir que le transcriptome viral arrive à un pic d'abondance relative pour ensuite baisser. Il est plus commun de voir l'infection s'installer et l'expression virale prendre complètement le dessus sur la cellule comme chez Marseillevirus (Rodrigues et al. 2020). Chez une infection par SARS-CoV-2, 20% des lectures sont associées au virus 7 h après infection puis 60% 24 h après contre environ 10% associées à l'hôte (J. Sun et al. 2020). Mimivirus atteint 20% de l'expression 1h30 après infection. Cette proportion va augmenter jusqu'à complètement dominer le transcriptome et que seulement 10% des lectures soient alignées au génome de l'hôte de 6 à 12h après infection (Legendre et al. 2010). L'expression relative de Cedratvirus kamchatka est presque négligeable en comparaison.

La première chose à laquelle nous pouvons penser est que les cellules n'ont pas toutes été infectées. Cette explication peut être écartée car une MOI de 50 a servi à la première infection et qu'à l'imagerie nous avons constaté que toutes les cellules ont bien été infectées. Nous savons également par observation par imagerie que le cycle est relativement rapide avec des virions matures visibles à 6h post-infection et une cellule remplie de virions à 8h post-infection (Figure 28, Chapitre 3). Les éléments nécessaires à la morphogénèse sont donc bien présents aux temps d'infection auxquels nous avons séquencé l'ARN. D'autres pistes sont à explorer pour comprendre cette faible expression virale relative :

Un défaut majeur de l'étude transcriptomique par RNA-seq est qu'il s'agit d'une mesure relative. Ceci est d'autant plus vrai ici puisqu'il existe plusieurs partenaires (le virus et l'hôte) qui interagissent entre eux au cours du cycle. A titre d'exemple, nous ne savons pas si l'expression des gènes viraux est réellement maximale à 2h30 ou si ce sont les gènes de l'amibe qui sont le plus réprimés à ce temps-là. Une façon de déjouer ce problème aurait été d'utiliser un ARN « spike-in », en quantité connue servant à calibrer les différents transcriptomes. Sans cette calibration, nous pouvons tout de même évaluer s'il y a régulation des gènes en comparant séparemment les gènes viraux et amibiens.

Peut-être que le matériel viral prend une part plus importante dans le protéome que le transcriptome. Pour évaluer cette hypothèse nous pouvons procéder à une analyse de

protéome à différents temps à travers la spectrométrie de masse. Ainsi nous pourrons voir si la traduction compense l'apparente faible quantité d'ARN messagers du virus.

Enfin, une hypothèse à investiguer est que le virus ait besoin d'une cellule en bon fonctionnement et une transcription du génome de l'hôte importante pour son cycle et/ou pour maintenir l'exocytose. L'exocytose est un mécanisme de sortie pour les *Pithoviridae* en plus de la lyse cellulaire (Legendre et al. 2014; dos Santos Silva et al. 2018). La proportion de l'un ou l'autre des mécanismes n'est pas clairement identifiée. Si l'exocytose est un mécanisme progressif et important pour la sortie, alors on comprend que la cellule soit maintenue en vie pour pouvoir relaguer des virions. Si la lyse est le mécanisme de sortie le plus important comme le suggèrent (dos Santos Silva et al. 2018), alors il faudra chercher une autre explication. Nous pouvons penser également que le virus utilise à son avantage des gènes particuliers de l'hôte. C'est ce que nous nous proposons d'étudier en nous intéressant en particulier aux échantillons en triplicat (1h, 3h, 6h et MOCK). Ainsi nous pourrons voir si statistiquement, il y a un impact sur l'expression de certains gènes amibiens. Il est possible que l'expression de l'hôte et de certaines fonctions cellulaires particulières soit utile à l'infection et vienne appuyer la production grâce à une régulation virale. Une fois identifiés des gènes amibiens d'intérêt, il serait intéressant d'observer la localisation des protéines correspondantes. La question est de savoir si ces protéines entrent dans l'usine virale pour aider la production. Pour ce faire il est possible d'exprimer des protéines couplées à la GFP. L'imagerie fluorescente avec marqueur de l'ADN nous dira si ces protéines se localisent vers l'usine virale ou non.

Conclusion

Ce travail constitue un premier pas vers la compréhension des mécanismes évolutifs à l'œuvre chez les *Pithoviridae* et de leur répartition. Les *Pithoviridae* environnementaux sont phylogénétiquement divergents des *Pithoviridae* isolés. C'est le signe que la diversité des virus isolés ne représente qu'une petite partie de la diversité réelle de cette famille virale. Des *Pithoviridae* divergents sont présents et parfois très abondants dans le pergélisol sibérien de Yakoutie. Les environnements échantillonnés se sont avérés être de très bons candidats pour retrouver des *Pithoviridae* comparativement à des jeux de données métagénomiques publiques. Cette comparaison a également permis de montrer que les *Pithoviridae* sont presque absents dans les océans (hors sédiments) mais plus communs dans les environnements terrestres ou d'eau douce.

L'analyse métagénomique d'un échantillon de pergélisol de Yakoutie de 42000 ans a permis d'assembler un virus complet de 1,6 Mb, Hydrivirus, qui partage un ancêtre commun avec Orpheovirus, plus proche parent des *Pithoviridae*. Avec Orpheovirus, ils constituent les plus grands génomes de *Pimascovirales*. Les deux virus ont de très nombreux gènes uniques, leur gigantisme serait donc en partie indépendant. A l'inverse, les *Cedratvirus* présentent très peu de nouveauté génomique ; plus de 75% de leurs gènes sont cœur et très peu de nouveaux gènes sont ajoutés à chaque nouveau génome séquencé.

Contrairement aux *Cedratvirus*, les *Pithovirus* ont plus de 20% de leurs génomes couverts par des séquences répétées. Ces séquences se structurent comme une suite de deux types de MITE (Element Transposable Miniature à répétition Inversée). Ces MITE ont été transposés au moins une fois depuis le dernier ancêtre commun entre *Pithovirus sibericum* et *Pithovirus massiliensis*.

Ces études comparatives ont été permises par la découverte de nouveaux virus depuis 2014, date de la publication du génome de *Pithovirus sibericum*. Nous pouvons espérer que de nouveaux *Pithoviridae* viendront soutenir ces résultats dans le futur. Malgré le petit nombre de génomes, leur comparaison a pu montrer des traits significatifs. Cela ouvre de nombreuses pistes à explorer : comment les *Pithoviridae* interagissent- ils avec leur hôte dans l'environnement ? Est-ce que les séquences répétées des génomes des Pithovirus leur sont délétères ? Quelle est la fonction des gènes orphelins des *Pithoviridae* ? Des premières pistes à cette dernière question pourront être données par l'analyse de la cinétique de transcription et de traduction de ces gènes. Nous avons choisi *Cedratvirus kamchatka* comme organisme modèle pour étudier un cycle infectieux. Les analyses préliminaires ont montré que jusqu'à 97% du génome viral était exprimé et que la protéine majeure de capside, qui n'est pas un constituant de la particule virale, est pourtant très exprimée.

Références

Abayasekara, Lalanika M., Jennifer Perera, Vishvanath Chandrasekharan, Vaz S. Gnanam, Nisala A. Udunuwara, Dileepa S. Liyanage, Nuwani E. Bulathsinhala, et al. 2017. « Detection of Bacterial Pathogens from Clinical Specimens Using Conventional Microbial Culture and 16S Metagenomics: A Comparative Study ». *BMC Infectious Diseases* 17 (1): 1-11. https://doi.org/10.1186/s12879-017-2727-8.

Abergel, Chantal, Matthieu Legendre, et Jean-Michel Claverie. 2015. « The Rapidly Expanding Universe of Giant Viruses: Mimivirus, Pandoravirus, Pithovirus and Mollivirus ». *FEMS Microbiology Reviews* 39 (6): 779-96. https://doi.org/10.1093/femsre/fuv037.

Abramov, Andrey, Tatiana Vishnivetskaya, et Elizaveta Rivkina. 2021. « Are permafrost microorganisms as old as permafrost? » *FEMS Microbiology Ecology* 97 (2). https://doi.org/10.1093/femsec/fiaa260.

Agarkova, Irina V., Leslie C. Lane, David D. Dunigan, Cristian F. Quispe, Garry A. Duncan, Elad Milrot, Abraham Minsky, Ahmed Esmael, Jayadri S. Ghosh, et James L. Van Etten. 2021. « Identification of a Chlorovirus PBCV-1 Protein Involved in Degrading the Host Cell Wall during Virus Infection ». *Viruses* 13 (5): 782. https://doi.org/10.3390/v13050782.

Andreani, Julien, Sarah Aherfi, Jacques Yaacoub Bou Khalil, Fabrizio Di Pinto, Idir Bitam, Didier Raoult, Philippe Colson, et Bernard La Scola. 2016. « Cedratvirus, a Double-Cork Structured Giant Virus, Is a Distant Relative of Pithoviruses ». *Viruses-Basel* 8 (11): 300. https://doi.org/10.3390/v8110300.

Andreani, Julien, Jacques Y. B. Khalil, Emeline Baptiste, Issam Hasni, Caroline Michelle, Didier Raoult, Anthony Levasseur, et Bernard La Scola. 2018. « Orpheovirus IHUMI-LCC2: A New Virus among the Giant Viruses ». *Frontiers in Microbiology* 8 (janvier): 2643. https://doi.org/10.3389/fmicb.2017.02643.

Angly, Florent E., Ben Felts, Mya Breitbart, Peter Salamon, Robert A. Edwards, Craig Carlson, Amy M. Chan, et al. 2006. « The Marine Viromes of Four Oceanic Regions ». *PLOS Biology* 4 (11): e368. https://doi.org/10.1371/journal.pbio.0040368.

Arantes, Thalita Souza, Rodrigo Araujo Lima Rodrigues, Ludmila Karen dos Santos Silva, Graziele Pereira Oliveira, Helton Luis de Souza, Jacques Y. B. Khalil, Danilo Bretas de Oliveira, et al. 2016. « The Large Marseillevirus Explores Different Entry Pathways by Forming Giant Infectious Vesicles ». *Journal of Virology* 90 (11): 5246-55. https://doi.org/10.1128/JVI.00177-16.

Asgari, Sassan, Dennis K. Bideshi, Yves Bigot, et Brian A. Federici. 2021. « Ascoviruses (Ascoviridae) ». In *Encyclopedia of Virology*, 724-31. Elsevier. https://doi.org/10.1016/B978-0-12-809633-8.21548-3.

Aszalós, Júlia Margit, Attila Szabó, Melinda Megyes, Dóra Anda, Balázs Nagy, et Andrea K. Borsodi. 2020. « Bacterial Diversity of a High-Altitude Permafrost Thaw Pond Located on Ojos del Salado (Dry Andes, Altiplano-Atacama Region) ». *Astrobiology* 20 (6): 754-65. https://doi.org/10.1089/ast.2018.2012.

Axelsson, Per. 2009. « “Do Not Eat Those Apples; They’ve Been on the Ground!”: Polio Epidemics and Preventive Measures, Sweden 1880s-1940s ». *Asclepio; Archivo Iberoamericano De Historia De La Medicina Y Antropologia Medica* 61 (1): 23-38. https://doi.org/10.3989/asclepio.2009.v61.i1.270.

Bäckström, Disa, Natalya Yutin, Steffen L. Jørgensen, Jennah Dharamshi, Felix Homa, Katarzyna Zaremba-Niedwiedzka, Anja Spang, Yuri I. Wolf, Eugene V. Koonin, et Thijs J. G. Ettema. 2019. « Virus Genomes from Deep Sea Sediments Expand the Ocean Megavirome and Support Independent Origins of Viral Gigantism ». *MBio* 10 (2). https://doi.org/10.1128/mBio.02497-18.

Baltimore, D. 1971. « Expression of animal virus genomes ». *Bacteriological Reviews* 35 (3): 235-41. https://doi.org/10.1128/br.35.3.235-241.1971.

Bertaux, Lionel, Audrey Lartigue, et Sandra Jeudy. 2020. « Giant Mimiviridae CsCl Purification Protocol ». *Bio-protocol* 10 (22): e3827-e3827.

Bertelli, Claire, Linda Mueller, Vincent Thomas, Trestan Pillonel, Nicolas Jacquier, et Gilbert Greub. 2017. « Cedratvirus Lausannensis - Digging into Pithoviridae Diversity ». *Environmental Microbiology* 19 (10): 4022-34. https://doi.org/10.1111/1462-2920.13813.

Bideshi, Dennis K., Marie-Véronique Demattei, Florence Rouleux-Bonnin, Karine Stasiak, Yeping Tan, Sylvie Bigot, Yves Bigot, et Brian A. Federici. 2006. « Genomic Sequence of Spodoptera frugiperda Ascovirus 1a, an Enveloped, Double-Stranded DNA Insect Virus That Manipulates Apoptosis for Viral Reproduction ». *Journal of Virology* 80 (23): 11791-805. https://doi.org/10.1128/JVI.01639-06.

Bisio, Hugo, Matthieu Legendre, Claire Giry, Nadege Philippe, Jean-Marie Alempic, Sandra Jeudy, et Chantal Abergel. 2022. « Evolution of giant pandoravirus from small icosahedral viruses revealed by CRISPR/Cas9 ». *bioRxiv*, janvier, 2022.08.18.504477. https://doi.org/10.1101/2022.08.18.504477.

Biskaborn, Boris K., Sharon L. Smith, Jeannette Noetzli, Heidrun Matthes, Gonçalo Vieira, Dmitry A. Streletskiy, Philippe Schoeneich, et al. 2019. « Permafrost Is Warming at a Global Scale ». *Nature Communications* 10 (1): 264. https://doi.org/10.1038/s41467-018-08240-4.

Blanca, Léo, Eugène Christo-Foroux, Sofia Rigou, et Matthieu Legendre. 2020. « Comparative Analysis of the Circular and Highly Asymmetrical Marseilleviridae Genomes ». *Viruses* 12 (11): 1270. https://doi.org/10.3390/v12111270.

Bowers, B., et E. D. Korn. 1968. « The Fine Structure of Acanthamoeba Castellanii. I. The Trophozoite ». *The Journal of Cell Biology* 39 (1): 95-111. https://doi.org/10.1083/jcb.39.1.95.

Boycott, A. E. 1928. « The Transition from Live to Dead: The Nature of Filtrable Viruses ». *Proceedings of the Royal Society of Medicine* 22 (1): 55-69. https://doi.org/10.1177/003591572802200121.

Boyer, Mickaël, Saïd Azza, Lina Barrassi, Thomas Klose, Angélique Campocasso, Isabelle Pagnier, Ghislain Fournous, et al. 2011. « Mimivirus shows dramatic genome reduction after intraamoebal culture ». *Proceedings of the National Academy of Sciences* 108 (25): 10296-301. https://doi.org/10.1073/pnas.1101118108.

Brahim Belhaouari, Djamal, Jean-Pierre Baudoin, Franck Gnankou, Fabrizio Di Pinto, Philippe Colson, Sarah Aherfi, et Bernard La Scola. 2019. « Evidence of a Cellulosic Layer in Pandoravirus massiliensis Tegument and the Mystery of the Genetic Support of Its Biosynthesis ». *Frontiers in Microbiology* 10. https://www.frontiersin.org/articles/10.3389/fmicb.2019.02932.

Bratbak, G, Jk Egge, et M Heldal. 1993. « Viral Mortality of the Marine Alga Emiliania Huxleyi (Haptophyceae) and Termination of Algal Blooms ». *Marine Ecology Progress Series* 93: 39-48. https://doi.org/10.3354/meps093039.

Breitbart, Mya, Ian Hewson, Ben Felts, Joseph M. Mahaffy, James Nulton, Peter Salamon, et Forest Rohwer. 2003. « Metagenomic Analyses of an Uncultured Viral Community from Human Feces ». *Journal of Bacteriology* 185 (20): 6220-23. https://doi.org/10.1128/JB.185.20.6220-6223.2003.

Breitbart, Mya, Peter Salamon, Bjarne Andresen, Joseph M. Mahaffy, Anca M. Segall, David Mead, Farooq Azam, et Forest Rohwer. 2002. « Genomic analysis of uncultured marine viral communities ». *Proceedings of the National Academy of Sciences* 99 (22): 14250-55. https://doi.org/10.1073/pnas.202488399.

Brum, Jennifer R., J. Cesar Ignacio-Espinoza, Simon Roux, Guilhem Doulcier, Silvia G. Acinas, Adriana Alberti, Samuel Chaffron, et al. 2015. « Patterns and ecological drivers of ocean viral communities ». *Science* 348 (6237): 1261498. https://doi.org/10.1126/science.1261498.

Burkert, Alexander, Thomas A. Douglas, Mark P. Waldrop, et Rachel Mackelprang. 2019. « Changes in the Active, Dead, and Dormant Microbial Community Structure across a Pleistocene Permafrost Chronosequence ». *Applied and Environmental Microbiology* 85 (7): e02646-18. https://doi.org/10.1128/AEM.02646-18.

Byers, Thomas J. 1986. « Molecular Biology of DNA InAcanthamoeba, Amoeba, Entamoeba, and Naegleria ». In *International Review of Cytology*, édité par G. H. Bourne, J. F. Danielli, et K. W. Jeon, 99:311-41. Molecular Approaches to the Study of Protozoan Cells. Academic Press. https://doi.org/10.1016/S0074-7696(08)61430-8.

Byrne, Deborah, Renata Grzela, Audrey Lartigue, Stéphane Audic, Sabine Chenivesse, Stéphanie Encinas, Jean-Michel Claverie, et Chantal Abergel. 2009. « The polyadenylation site of Mimivirus transcripts obeys a stringent 'hairpin rule' ». *Genome Research* 19 (7): 1233-42. https://doi.org/10.1101/gr.091561.109.

Carlson, Colin J., Wayne M. Getz, Kyrre L. Kausrud, Carrie A. Cizauskas, Jason K. Blackburn, Fausto A. Bustos Carrillo, Rita Colwell, et al. 2018. « Spores and Soil from Six Sides: Interdisciplinarity and the Environmental Biology of Anthrax (Bacillus Anthracis) ». *Biological Reviews* 93 (4): 1813-31. https://doi.org/10.1111/brv.12420.

Cashdollar, J.l., et L. Wymer. 2013. « Methods for Primary Concentration of Viruses from Water Samples: A Review and Meta-Analysis of Recent Studies ». *Journal of Applied Microbiology* 115 (1): 1-11. https://doi.org/10.1111/jam.12143.

Chaudhari, Harshali V., Mandar M. Inamdar, et Kiran Kondabagil. 2021. « Scaling Relation between Genome Length and Particle Size of Viruses Provides Insights into Viral Life History ». *IScience* 24 (5): 102452. https://doi.org/10.1016/j.isci.2021.102452.

Chinchar, V. Gregory, Kwang H. Yu, et James K. Jancovich. 2011. « The Molecular Biology of Frog Virus 3 and Other Iridoviruses Infecting Cold-Blooded Vertebrates ». *Viruses* 3 (10): 1959-85. https://doi.org/10.3390/v3101959.

Christo-Foroux, Eugène. 2020. « Etude de virus réactivés à partir d'échantillons de cryosol ». These de doctorat, Aix-Marseille. http://www.theses.fr/2020AIXM0347.

Christo-Foroux, Eugene, Jean-Marie Alempic, Audrey Lartigue, Sebastien Santini, Karine Labadie, Matthieu Legendre, Chantal Abergel, et Jean-Michel Claverie. 2020. « Characterization of Mollivirus Kamchatka, the First Modern Representative of the Proposed Molliviridae Family of Giant Viruses ». *Journal of Virology* 94 (8). https://doi.org/10.1128/JVI.01997-19.

Claverie, Jean-Michel. 2006. « Viruses Take Center Stage in Cellular Evolution. » *Genome Biology* 7 (6): 110. https://doi.org/10.1186/gb-2006-7-6-110.

Claverie, Jean-Michel, Hiroyuki Ogata, Stéphane Audic, Chantal Abergel, Pierre-Edouard Fournier, et Karsten Suhre. 2005. « Mimivirus and the emerging concept of "giant" virus ». https://hal.archives-ouvertes.fr/hal-00005228.

Cournac, Axel, Romain Koszul, et Julien Mozziconacci. 2016. « The 3D folding of metazoan genomes correlates with the association of similar repetitive elements ». *Nucleic Acids Research* 44 (1): 245-55. https://doi.org/10.1093/nar/gkv1292.

Decker, E R, et G J Bucher. 1977. « Geothermal studies in Antarctica ». *Antarct. J. U. S.; (United States)* 12:4 (octobre). https://www.osti.gov/biblio/6112190.

Eagles, G. Hardy. 1933. « The in Vitro Cultivation of Filterable Viruses ». *Biological Reviews* 8 (3): 335-44. https://doi.org/10.1111/j.1469-185X.1933.tb01092.x.

Efimenko, T. A., O. V. Efremenkova, E. V. Demkina, M. A. Petrova, I. G. Sumarukova, B. F. Vasilyeva, et G. I. El'-Registan. 2018. « Bacteria Isolated from Antarctic Permafrost Are Efficient Antibiotic Producers ». *Microbiology* 87 (5): 692-98. https://doi.org/10.1134/S0026261718050089.

Elde, Nels C., Stephanie J. Child, Michael T. Eickbush, Jacob O. Kitzman, Kelsey S. Rogers, Jay Shendure, Adam P. Geballe, et Harmit S. Malik. 2012. « Poxviruses Deploy Genomic Accordions to Adapt Rapidly against Host Antiviral Defenses ». *Cell* 150 (4): 831-41. https://doi.org/10.1016/j.cell.2012.05.049.

Emerson, Joanne B., Simon Roux, Jennifer R. Brum, Benjamin Bolduc, Ben J. Woodcroft, Ho Bin Jang, Caitlin M. Singleton, et al. 2018. « Host-Linked Soil Viral Ecology along a Permafrost Thaw Gradient ». *Nature Microbiology* 3 (8): 870-80. https://doi.org/10.1038/s41564-018-0190-y.

Endo, Hisashi, Romain Blanc-Mathieu, Yanze Li, Guillem Salazar, Nicolas Henry, Karine Labadie, Colomban de Vargas, et al. 2020. « Biogeography of Marine Giant Viruses Reveals Their

Interplay with Eukaryotes and Ecological Functions ». *Nature Ecology & Evolution*, septembre, 1-11. https://doi.org/10.1038/s41559-020-01288-w.

Etten, James L. Van, et Russel H. Meints. 1999. « Giant Viruses Infecting Algae ». *Annual Review of Microbiology* 53: 447-94. https://doi.org/10.1146/annurev.micro.53.1.447.

Fabre, Elisabeth, Sandra Jeudy, Sébastien Santini, Matthieu Legendre, Mathieu Trauchessec, Yohann Couté, Jean-Michel Claverie, et Chantal Abergel. 2017. « Noumeavirus Replication Relies on a Transient Remote Control of the Host Nucleus ». *Nature Communications* 8 (1): 15087. https://doi.org/10.1038/ncomms15087.

Federici, Brian A., Just M. Vlak, et John J.YR 1990 Hamm. 1990. « Comparative Study of Virion Structure, Protein Composition and Genomic DNA of Three Ascovirus Isolates ». *Journal of General Virology* 71 (8): 1661-68. https://doi.org/10.1099/0022-1317-71-8-1661.

Filée, Jonathan, et Michael Chandler. 2008. « Convergent Mechanisms of Genome Evolution of Large and Giant DNA Viruses ». *Research in Microbiology* 159 (5): 325-31. https://doi.org/10.1016/j.resmic.2008.04.012.

Frauenfeld, Oliver W., Tingjun Zhang, Roger G. Barry, et David Gilichinsky. 2004. « Interdecadal Changes in Seasonal Freeze and Thaw Depths in Russia ». *Journal of Geophysical Research: Atmospheres* 109 (D5). https://doi.org/10.1029/2003JD004245.

Galperin, Charles. 1987. « Le bactériophage, la lysogénie et son déterminisme génétique ». *History and Philosophy of the Life Sciences* 9 (2): 175-224.

Gauthier, Jeff, Antony T Vincent, Steve J Charette, et Nicolas Derome. 2019. « A brief history of bioinformatics ». *Briefings in Bioinformatics* 20 (6): 1981-96. https://doi.org/10.1093/bib/bby063.

Geballa-Koukoulas, Khalil, Julien Andreani, Bernard La Scola, et Guillaume Blanc. 2021. « The Kaumoebavirus LCC10 Genome Reveals a Unique Gene Strand Bias among "Extended Asfarviridae" ». *Viruses* 13 (2). https://doi.org/10.3390/v13020148.

Ghedin, Elodie, et Jean-Michel Claverie. 2005. « Mimivirus Relatives in the Sargasso Sea ». *Virology Journal* 2 (août): 62. https://doi.org/10.1186/1743-422X-2-62.

Gilichinsky, D. A., G. S. Wilson, E. I. Friedmann, C. P. McKay, R. S. Sletten, E. M. Rivkina, T. A. Vishnivetskaya, et al. 2007. « Microbial Populations in Antarctic Permafrost: Biodiversity, State, Age, and Implication for Astrobiology. » *Astrobiology* 7 (2): 275-311. https://doi.org/10.1089/ast.2006.0012.

Gregory, Ann C., Ahmed A. Zayed, Nádia Conceição-Neto, Ben Temperton, Ben Bolduc, Adriana Alberti, Mathieu Ardyna, et al. 2019. « Marine DNA Viral Macro- and Microdiversity from Pole to Pole ». *Cell* 177 (5): 1109-1123.e14. https://doi.org/10.1016/j.cell.2019.03.040.

Greub, Gilbert, et Didier Raoult. 2004. « Microorganisms Resistant to Free-Living Amoebae ». *Clinical Microbiology Reviews* 17 (2): 413-33. https://doi.org/10.1128/CMR.17.2.413-433.2004.

Grosse, G., B. Jones, et C. Arp. 2013. « 8.21 Thermokarst Lakes, Drainage, and Drained Basins ». In *Treatise on Geomorphology*, édité par John F. Shroder, 325-53. San Diego: Academic Press. https://doi.org/10.1016/B978-0-12-374739-6.00216-5.

Guglielmini, Julien, Anthony Woo, Mart Krupovic, Patrick Forterre, et Morgan Gaia. 2018. « Diversification of Giant and Large Eukaryotic DsDNA Viruses Predated the Origin of Modern Eukaryotes ». *BioRxiv*, octobre, 455816. https://doi.org/10.1101/455816.

Ha, Anh D., Mohammad Moniruzzaman, et Frank O. Aylward. 2021. « High Transcriptional Activity and Diverse Functional Repertoires of Hundreds of Giant Viruses in a Coastal Marine System ». *MSystems* 6 (4): e0029321. https://doi.org/10.1128/mSystems.00293-21.

Hartley, Gabrielle, et Rachel J. O'Neill. 2019. « Centromere Repeats: Hidden Gems of the Genome ». *Genes* 10 (3): 223. https://doi.org/10.3390/genes10030223.

Hingamp, Pascal, Nigel Grimsley, Silvia G. Acinas, Camille Clerissi, Lucie Subirana, Julie Poulain, Isabel Ferrera, et al. 2013. « Exploring Nucleo-Cytoplasmic Large DNA Viruses in Tara Oceans Microbial Metagenomes ». *The ISME Journal* 7 (9): 1678-95. https://doi.org/10.1038/ismej.2013.59.

Hoffmann, Ralf L., Rolf Michel, Karl-Dieter Müller, et E.N. Schmid. 1997. « Archaea like endocytobiotic organisms isolated from Acanthamoeba sp (gr II) - JPortal ». *Endocytobiosis & Cell Res.*, n° 12 (septembre): 185-88.

Hua, Jianfei, Alexis Huet, Carlos A. Lopez, Katerina Toropova, Welkin H. Pope, Robert L. Duda, Roger W. Hendrix, et James F. Conway. 2017. « Capsids and Genomes of Jumbo-Sized Bacteriophages Reveal the Evolutionary Reach of the HK97 Fold ». *mBio* 8 (5): e01579-17. https://doi.org/10.1128/mBio.01579-17.

Illergård, Kristoffer, David H. Ardell, et Arne Elofsson. 2009. « Structure Is Three to Ten Times More Conserved than Sequence--a Study of Structural Response in Protein Cores ». *Proteins* 77 (3): 499-508. https://doi.org/10.1002/prot.22458.

İnce, İkbal Agah, Orhan Özcan, Ayca Zeynep Ilter-Akulke, Erin D. Scully, et Arzu Özgen. 2018. « Invertebrate Iridoviruses: A Glance over the Last Decade ». *Viruses* 10 (4): 161. https://doi.org/10.3390/v10040161.

Irwin, Nicholas A. T., Alexandros A. Pittis, Thomas A. Richards, et Patrick J. Keeling. 2022. « Systematic Evaluation of Horizontal Gene Transfer between Eukaryotes and Viruses ». *Nature Microbiology* 7 (2): 327-36. https://doi.org/10.1038/s41564-021-01026-3.

Iyer, Lakshminarayan M., L. Aravind, et Eugene V. Koonin. 2001. « Common Origin of Four Diverse Families of Large Eukaryotic DNA Viruses ». *Journal of Virology* 75 (23): 11720-34. https://doi.org/10.1128/JVI.75.23.11720-11734.2001.

Jacquet, Stéphan, Mikal Heldal, Debora Iglesias-Rodriguez, Aud Larsen, William Wilson, et Gunnar Bratbak. 2002. « Flow Cytometric Analysis of an Emiliana Huxleyi Bloom Terminated by Viral Infection ». *Aquatic Microbial Ecology* 27 (2): 111-24. https://doi.org/10.3354/ame027111.

Jansson, Janet K., et Neslihan Taş. 2014. « The Microbial Ecology of Permafrost ». *Nature Reviews. Microbiology* 12 (6): 414-25. https://doi.org/10.1038/nrmicro3262.

Jeudy, Sandra, Chantal Abergel, Jean-Michel Claverie, et Matthieu Legendre. 2012. « Translation in Giant Viruses: A Unique Mixture of Bacterial and Eukaryotic Termination Schemes ». *PLOS Genetics* 8 (12): e1003122. https://doi.org/10.1371/journal.pgen.1003122.

Jeudy, Sandra, Sofia Rigou, Jean-Marie Alempic, Jean-Michel Claverie, Chantal Abergel, et Matthieu Legendre. 2020. « The DNA methylation landscape of giant viruses ». *Nature Communications* 11 (1): 2657. https://doi.org/10.1038/s41467-020-16414-2.

Jiang, Ning, et Susan R. Wessler. 2001. « Insertion Preference of Maize and Rice Miniature Inverted Repeat Transposable Elements as Revealed by the Analysis of Nested Elements [W] ». *The Plant Cell* 13 (11): 2553-64. https://doi.org/10.1105/tpc.010235.

Johnson, Sarah Stewart, Martin B. Hebsgaard, Torben R. Christensen, Mikhail Mastepanov, Rasmus Nielsen, Kasper Munch, Tina Brand, et al. 2007. « Ancient Bacteria Show Evidence of DNA Repair ». *Proceedings of the National Academy of Sciences* 104 (36): 14401-5. https://doi.org/10.1073/pnas.0706787104.

Jongejans, Loeka L., Susanne Liebner, Christian Knoblauch, Kai Mangelsdorf, Mathias Ulrich, Guido Grosse, George Tanski, et al. 2021. « Greenhouse Gas Production and Lipid Biomarker Distribution in Yedoma and Alas Thermokarst Lake Sediments in Eastern Siberia ». *Global Change Biology* 27 (12): 2822-39. https://doi.org/10.1111/gcb.15566.

Kazlauskas, Darius, Mart Krupovic, Julien Guglielmini, Patrick Forterre, et Česlovas Venclovas. 2020. « Diversity and evolution of B-family DNA polymerases ». *Nucleic Acids Research* 48 (18): 10142-56. https://doi.org/10.1093/nar/gkaa760.

Khan, Naveed Ahmed. 2006. « Acanthamoeba : biology and increasing importance in human health ». *FEMS Microbiology Reviews* 30 (4): 564-95. https://doi.org/10.1111/j.1574-6976.2006.00023.x.

Koning, A. P. Jason de, Wanjun Gu, Todd A. Castoe, Mark A. Batzer, et David D. Pollock. 2011. « Repetitive Elements May Comprise Over Two-Thirds of the Human Genome ». *PLOS Genetics* 7 (12): e1002384. https://doi.org/10.1371/journal.pgen.1002384.

Koonin, Eugene V., et Natalya Yutin. 2018. « Multiple Evolutionary Origins of Giant Viruses ». *F1000Research* 7: F1000 Faculty Rev-1840. https://doi.org/10.12688/f1000research.16248.1.

Kördel, Mikael, Martin Svenda, Hemanth K. N. Reddy, Emelie Fogelqvist, Komang G. Y. Arsana, Bejan Hamawandi, Muhammet S. Toprak, Hans M. Hertz, et Jonas A. Sellberg. 2021. « Quantitative conversion of biomass in giant DNA virus infection ». *Scientific Reports* 11 (mars). https://doi.org/10.1038/s41598-021-83547-9.

Kostyrka, Gladys. 2018. « La place des virus dans le monde vivant », 845.

La Scola, Bernard, Stéphane Audic, Catherine Robert, Liang Jungang, Xavier de Lamballerie, Michel Drancourt, Richard Birtles, Jean-Michel Claverie, et Didier Raoult. 2003. « A Giant Virus in Amoebae ». *Science (New York, N.Y.)* 299 (5615): 2033. https://doi.org/10.1126/science.1081867.

Laummaunwai, Porntip, Wipaporn Ruangjirachuporn, et Thidarut Boonmars. 2012. « A Simple PCR Condition for Detection of a Single Cyst of Acanthamoeba Species ». *Parasitology Research* 110 (4): 1569-72. https://doi.org/10.1007/s00436-011-2662-3.

Leclercq, Sébastien, Eric Rivals, et Philippe Jarne. 2010. « DNA Slippage Occurs at Microsatellite Loci without Minimal Threshold Length in Humans: A Comparative Genomic Approach ». *Genome Biology and Evolution* 2: 325-35. https://doi.org/10.1093/gbe/evq023.

Legendre, Matthieu, Stéphane Audic, Olivier Poirot, Pascal Hingamp, Virginie Seltzer, Deborah Byrne, Audrey Lartigue, et al. 2010. « MRNA Deep Sequencing Reveals 75 New Genes and a Complex Transcriptional Landscape in Mimivirus ». *Genome Research* 20 (5): 664-74. https://doi.org/10.1101/gr.102582.109.

Legendre, Matthieu, Julia Bartoli, Lyubov Shmakova, Sandra Jeudy, Karine Labadie, Annie Adrait, Magali Lescot, et al. 2014. « Thirty-Thousand-Year-Old Distant Relative of Giant Icosahedral DNA Viruses with a Pandoravirus Morphology ». *Proceedings of the National Academy of Sciences of the United States of America* 111 (11): 4274-79. https://doi.org/10.1073/pnas.1320670111.

Legendre, Matthieu, Elisabeth Fabre, Olivier Poirot, Sandra Jeudy, Audrey Lartigue, Jean-Marie Alempic, Laure Beucher, et al. 2018. « Diversity and Evolution of the Emerging Pandoraviridae Family ». *Nature Communications* 9 (1): 2285. https://doi.org/10.1038/s41467-018-04698-4.

Legendre, Matthieu, Audrey Lartigue, Lionel Bertaux, Sandra Jeudy, Julia Bartoli, Magali Lescot, Jean-Marie Alempic, et al. 2015. « In-Depth Study of Mollivirus Sibericum, a New 30,000-y-Old Giant Virus Infecting Acanthamoeba ». *Proceedings of the National Academy of Sciences* 112 (38): E5327-35. https://doi.org/10.1073/pnas.1510795112.

Levasseur, Anthony, Julien Andreani, Jeremy Delerce, Jacques Bou Khalil, Catherine Robert, Bernard La Scola, et Didier Raoult. 2016. « Comparison of a Modern and Fossil Pithovirus Reveals Its Genetic Conservation and Evolution ». *Genome Biology and Evolution* 8 (8): 2333-39. https://doi.org/10.1093/gbe/evw153.

Liskova, Elena A., Irina Y. Egorova, Yuri O. Selyaninov, Irina V. Razheva, Nadezhda A. Gladkova, Nadezhda N. Toropova, Olga I. Zakharova, et al. 2021. « Reindeer Anthrax in the Russian Arctic, 2016: Climatic Determinants of the Outbreak and Vaccination Effectiveness ». *Frontiers in Veterinary Science* 8. https://www.frontiersin.org/articles/10.3389/fvets.2021.668420.

Liu, Yang, Hugo Bisio, Chelsea Marie Toner, Sandra Jeudy, Nadege Philippe, Keda Zhou, Samuel Bowerman, et al. 2021. « Virus-Encoded Histone Doublets Are Essential and Form Nucleosome-like Structures ». *Cell* 184 (16): 4237-4250.e19. https://doi.org/10.1016/j.cell.2021.06.032.

LoGiudice, Kathleen, Richard S. Ostfeld, Kenneth A. Schmidt, et Felicia Keesing. 2003. « The ecology of infectious disease: Effects of host diversity and community composition on Lyme disease risk ». *Proceedings of the National Academy of Sciences of the United States of America* 100 (2): 567-71. https://doi.org/10.1073/pnas.0233733100.

Luisi, Pier Luigi, et Pasquale Stano. 2011. *The Minimal Cell The Biophysics of Cell Compartment and the Origin of Cell Functionality*. 1st ed. 2011. Dordrecht: Springer Netherlands. https://doi.org/10.1007/978-90-481-9944-0.

Luria, Salvador Edward. 1957. « Letter from Salvador E. Luria to Andre Lwoff - Digital Collections - National Library of Medicine ». 27 novembre 1957. https://collections.nlm.nih.gov/catalog/nlm:nlmuid-101584611X97-doc.

Lwoff, André. 1953. « LYSOGENY1 ». *Bacteriological Reviews* 17 (4): 269-337.

———. 1957. « The Concept of Virus ». *Microbiology* 17 (2): 239-53. https://doi.org/10.1099/00221287-17-2-239.

Mackenzie, J. S., et D. T. Williams. 2009. « The Zoonotic Flaviviruses of Southern, South-Eastern and Eastern Asia, and Australasia: The Potential for Emergent Viruses ». *Zoonoses and Public Health* 56 (6-7): 338-56. https://doi.org/10.1111/j.1863-2378.2008.01208.x.

Majji, S., V. Thodima, R. Sample, D. Whitley, Y. Deng, J. Mao, et V. G. Chinchar. 2009. « Transcriptome Analysis of Frog Virus 3, the Type Species of the Genus Ranavirus, Family Iridoviridae. » *Virology* 391 (2): 293-303. https://doi.org/10.1016/j.virol.2009.06.022.

Malavin, Stas, Lyubov Shmakova, Jean-Michel Claverie, et Elizaveta Rivkina. 2020. « Frozen Zoo: A Collection of Permafrost Samples Containing Viable Protists and Their Viruses ». *Biodiversity Data Journal* 8 (juillet): e51586. https://doi.org/10.3897/BDJ.8.e51586.

Mao, Yan-Chiao, Han-Ni Chuang, Chien-Hung Shih, Han-Hsueh Hsieh, Yu-Han Jiang, Liao-Chun Chiang, Wen-Loung Lin, Tzu-Hung Hsiao, et Po-Yu Liu. 2021. « An Investigation of Conventional Microbial Culture for the Naja Atra Bite Wound, and the Comparison between Culture-Based 16S Sanger Sequencing and 16S Metagenomics of the Snake Oropharyngeal Bacterial Microbiota ». *PLOS Neglected Tropical Diseases* 15 (4): e0009331. https://doi.org/10.1371/journal.pntd.0009331.

Martínez-Puchol, Sandra, Marta Itarte, Marta Rusiñol, Eva Forés, Cristina Mejías-Molina, Cristina Andrés, Andrés Antón, et al. 2021. « Exploring the diversity of coronavirus in sewage during COVID-19 pandemic: Don’t miss the forest for the trees ». *Science of The Total Environment* 800 (décembre): 149562. https://doi.org/10.1016/j.scitotenv.2021.149562.

Matthey-Doret, Cyril, Morgan Colp, Pedro Escoll, Agnès Thierry, Bruce Curtis, Matt Sarrasin, Michael Gray, et al. 2021. « Chromosome-scale assemblies of Acanthamoeba castellanii genomes provide insights into Legionella pneumophila infection-related chromatin re-organizatio ». https://doi.org/10.1101/2021.10.26.465878.

Maumus, Florian, et Guillaume Blanc. 2016. « Study of Gene Trafficking between Acanthamoeba and Giant Viruses Suggests an Undiscovered Family of Amoeba-Infecting Viruses ». *Genome Biology and Evolution* 8 (11): 3351-63. https://doi.org/10.1093/gbe/evw260.

Metcalf, T. G., J. L. Melnick, et M. K. Estes. 1995. « Environmental Virology: From Detection of Virus in Sewage and Water by Isolation to Identification by Molecular Biology--a Trip of over 50 Years ». *Annual Review of Microbiology* 49: 461-87. https://doi.org/10.1146/annurev.mi.49.100195.002333.

Michel, R., K.-D. Müller, E. N. Schmid, L. Zöller, et R. Hoffmann. 2003. « Endocytobiont KC5/2 Induces Transformation into Sol-like Cytoplasm of Its Host Acanthamoeba Sp. as Substrate for Its Own Development ». *Parasitology Research* 90 (1): 52-56. https://doi.org/10.1007/s00436-002-0710-8.

Michel, R., E. N. Schmid, R. Hoffmann, et K.-D. Müller. 2003. « Endoparasite KC5/2 Encloses Large Areas of Sol-like Cytoplasm within Acanthamoebae. Normal Behavior or Aberration? » *Parasitology Research* 91 (4): 265-66. https://doi.org/10.1007/s00436-003-0944-0.

Mihara, Tomoko, Hitoshi Koyano, Pascal Hingamp, Nigel Grimsley, Susumu Goto, et Hiroyuki Ogata. 2018. « Taxon Richness of “Megaviridae” Exceeds Those of Bacteria and Archaea in the Ocean ». *Microbes and Environments* 33 (2): 162-71. https://doi.org/10.1264/jsme2.ME17203.

Mizuno, Carolina M., Charlotte Guyomar, Simon Roux, Régis Lavigne, Francisco Rodriguez-Valera, Matthew B. Sullivan, Reynald Gillet, Patrick Forterre, et Mart Krupovic. 2019. « Numerous Cultivated and Uncultivated Viruses Encode Ribosomal Proteins ». *Nature Communications* 10 (1): 752. https://doi.org/10.1038/s41467-019-08672-6.

Moniruzzaman, Mohammad, Carolina A. Martinez-Gutierrez, Alaina R. Weinheimer, et Frank O. Aylward. 2019. « Dynamic Genome Evolution and Blueprint of Complex Virocell Metabolism in Globally-Distributed Giant Viruses ». Preprint. Microbiology. https://doi.org/10.1101/836445.

Moniruzzaman, Mohammad, Alaina R. Weinheimer, Carolina A. Martinez-Gutierrez, et Frank O. Aylward. 2020. « Widespread Endogenization of Giant Viruses Shapes Genomes of Green Algae ». *Nature* 588 (7836): 141-45. https://doi.org/10.1038/s41586-020-2924-2.

Moniruzzaman, Mohammad, Louie L. Wurch, Harriet Alexander, Sonya T. Dyhrman, Christopher J. Gobler, et Steven W. Wilhelm. 2017. « Virus-Host Relationships of Marine Single-Celled Eukaryotes Resolved from Metatranscriptomics ». *Nature Communications* 8 (1): 16054. https://doi.org/10.1038/ncomms16054.

Mönttinen, Heli A. M., Cedric Bicep, Tom A. Williams, et Robert P. Hirt. 2021. « The genomes of nucleocytoplasmic large DNA viruses: viral evolution writ large ». *Microbial Genomics* 7 (9): 000649. https://doi.org/10.1099/mgen.0.000649.

Monzon, Vivian, Typhaine Paysan-Lafosse, Valerie Wood, et Alex Bateman. 2022. « Reciprocal Best Structure Hits: Using AlphaFold Models to Discover Distant Homologues ». bioRxiv. https://doi.org/10.1101/2022.07.04.498216.

Moran, Nancy A, et Gordon R Plague. 2004. « Genomic changes following host restriction in bacteria ». *Current Opinion in Genetics & Development* 14 (6): 627-33. https://doi.org/10.1016/j.gde.2004.09.003.

Mueller, Linda, Claire Bertelli, Trestan Pillonel, Nicolas Salamin, et Gilbert Greub. 2017. « One Year Genome Evolution of Lausannevirus in Allopatric versus Sympatric Conditions ». *Genome Biology and Evolution* 9 (6): 1432-49. https://doi.org/10.1093/gbe/evx074.

Mykytczuk, Nadia C. S., Simon J. Foote, Chris R. Omelon, Gordon Southam, Charles W. Greer, et Lyle G. Whyte. 2013. « Bacterial Growth at −15 °C; Molecular Insights from the Permafrost Bacterium Planococcus Halocryophilus Or1 ». *The ISME Journal* 7 (6): 1211-26. https://doi.org/10.1038/ismej.2013.8.

Nazir, Amina, Azam Ali, Hong Qing, et Yigang Tong. 2021. « Emerging Aspects of Jumbo Bacteriophages ». *Infection and Drug Resistance* 14 (novembre): 5041-55. https://doi.org/10.2147/IDR.S330560.

Needham, David M., Susumu Yoshizawa, Toshiaki Hosaka, Camille Poirier, Chang Jae Choi, Elisabeth Hehenberger, Nicholas A. T. Irwin, et al. 2019. « A Distinct Lineage of Giant Viruses Brings a Rhodopsin Photosystem to Unicellular Marine Predators ». *Proceedings of the National Academy of Sciences* 116 (41): 20574-83. https://doi.org/10.1073/pnas.1907517116.

Norrby, Erling. 1983. « 2 - The Morphology of Virus Particles. Classification of Viruses ». In *Textbook of Medical Virology*, édité par Erik Lycke et Erling Norrby, 4-16. Butterworth-Heinemann. https://doi.org/10.1016/B978-0-407-00253-1.50007-4.

———. 2008. « Nobel Prizes and the Emerging Virus Concept ». *Archives of Virology* 153 (6): 1109-23. https://doi.org/10.1007/s00705-008-0088-8.

Ober, W. B., et N. Aloush. 1982. « The plague at Granada, 1348-1349: Ibn Al-Khatib and ideas of contagion. » *Bulletin of the New York Academy of Medicine* 58 (4): 418-24.

O'Brien, Sarah L., Sean M. Gibbons, Sarah M. Owens, Jarrad Hampton-Marcell, Eric R. Johnston, Julie D. Jastrow, Jack A. Gilbert, Folker Meyer, et Dionysios A. Antonopoulos. 2016. « Spatial Scale Drives Patterns in Soil Bacterial Diversity ». *Environmental Microbiology* 18 (6): 2039-51. https://doi.org/10.1111/1462-2920.13231.

Obu, J. 2021. « How Much of the Earth's Surface Is Underlain by Permafrost? » *Journal of Geophysical Research: Earth Surface* 126 (5): e2021JF006123. https://doi.org/10.1029/2021JF006123.

Okamoto, Kenta, Naoyuki Miyazaki, Hemanth K. N. Reddy, Max F. Hantke, Filipe R. N. C. Maia, Daniel S. D. Larsson, Chantal Abergel, et al. 2018. « Cryo-EM Structure of a Marseilleviridae Virus Particle Reveals a Large Internal Microassembly ». *Virology* 516 (mars): 239-45. https://doi.org/10.1016/j.virol.2018.01.021.

Okamoto, Kenta, Naoyuki Miyazaki, Chihong Song, Filipe R. N. C. Maia, Hemanth K. N. Reddy, Chantal Abergel, Jean-Michel Claverie, Janos Hajdu, Martin Svenda, et Kazuyoshi Murata. 2017.

« Structural Variability and Complexity of the Giant Pithovirus Sibericum Particle Revealed by High-Voltage Electron Cryo-Tomography and Energy-Filtered Electron Cryo-Microscopy ». *Scientific Reports* 7 (1): 13291. https://doi.org/10.1038/s41598-017-13390-4.

Pace, Norman R., David A. Stahl, David J. Lane, et Gary J. Olsen. 1986. « The Analysis of Natural Microbial Populations by Ribosomal RNA Sequences ». In *Advances in Microbial Ecology*, édité par K. C. Marshall, 1-55. Advances in Microbial Ecology. Boston, MA: Springer US. https://doi.org/10.1007/978-1-4757-0611-6_1.

Pagarete, António, Gildas Le Corguillé, Bela Tiwari, Hiroyuki Ogata, Colomban de Vargas, William H. Wilson, et Michael J. Allen. 2011. « Unveiling the transcriptional features associated with coccolithovirus infection of natural Emiliania huxleyi blooms ». *FEMS Microbiology Ecology* 78 (3): 555-64. https://doi.org/10.1111/j.1574-6941.2011.01191.x.

Peterson, A. Townsend, Jorge Soberón, Janine Ramsey, et Luis Osorio-Olvera. 2020. « Co-occurrence Networks do not Support Identification of Biotic Interactions ». *Biodiversity Informatics* 15 (1): 1-10. https://doi.org/10.17161/bi.v15i1.9798.

Philippe, Nadège, Matthieu Legendre, Gabriel Doutre, Yohann Couté, Olivier Poirot, Magali Lescot, Defne Arslan, et al. 2013. « Pandoraviruses: Amoeba Viruses with Genomes Up to 2.5 Mb Reaching That of Parasitic Eukaryotes ». *Science* 341 (6143): 281-86. https://doi.org/10.1126/science.1239181.

Pound, Helena L., Eric R. Gann, Xiangming Tang, Lauren E. Krausfeldt, Matthew Huff, Margaret E. Staton, David Talmy, et Steven W. Wilhelm. 2020. « The “Neglected Viruses” of Taihu: Abundant Transcripts for Viruses Infecting Eukaryotes and Their Potential Role in Phytoplankton Succession ». *Frontiers in Microbiology* 11. https://www.frontiersin.org/article/10.3389/fmicb.2020.00338.

Prodinger, Florian, Hisashi Endo, Yoshihito Takano, Yanze Li, Kento Tominaga, Tatsuhiro Isozaki, Romain Blanc-Mathieu, et al. 2022. « Year-Round Dynamics of Amplicon Sequence Variant Communities Differ among Eukaryotes, Imitervirales and Prokaryotes in a Coastal Ecosystem ». *FEMS Microbiology Ecology* 97 (12): fiab167. https://doi.org/10.1093/femsec/fiab167.

R., R. 1877. « ESSAIS ET NOTICES. BACTÈRIDIES ET VIBRIONS ». *Revue des Deux Mondes (1829-1971)* 22 (4): 948-58.

Raoult, Didier, Stéphane Audic, Catherine Robert, Chantal Abergel, Patricia Renesto, Hiroyuki Ogata, Bernard La Scola, Marie Suzan, et Jean-Michel Claverie. 2004. « The 1.2-Megabase Genome Sequence of Mimivirus ». *Science* 306 (5700): 1344-50. https://doi.org/10.1126/science.1101485.

Raoult, Didier, et Patrick Forterre. 2008. « Redefining Viruses: Lessons from Mimivirus ». *Nature Reviews Microbiology* 6 (4): 315-19. https://doi.org/10.1038/nrmicro1858.

Reynolds, Mary G., et Inger K. Damon. 2012. « Outbreaks of Human Monkeypox after Cessation of Smallpox Vaccination ». *Trends in Microbiology* 20 (2): 80-87. https://doi.org/10.1016/j.tim.2011.12.001.

Richard, Guy-Franck, Alix Kerrest, et Bernard Dujon. 2008. « Comparative Genomics and Molecular Dynamics of DNA Repeats in Eukaryotes ». *Microbiology and Molecular Biology Reviews* 72 (4): 686-727. https://doi.org/10.1128/MMBR.00011-08.

Richards, Gary P. 1985. « Outbreaks of Shellfish-Associated Enteric Virus Illness in the United States: Requisite for Development of Viral Guidelines ». *Journal of Food Protection* 48 (9): 815-23. https://doi.org/10.4315/0362-028X-48.9.815.

Rivers, T. M. 1927. « Filterable viruses a critical review ». *Journal of Bacteriology* 14 (4): 217-58. https://doi.org/10.1128/jb.14.4.217-258.1927.

Rodrigues, Rodrigo Araújo Lima, Jônatas Santos Abrahão, Betânia Paiva Drumond, et Erna Geessien Kroon. 2016. « Giants among Larges: How Gigantism Impacts Giant Virus Entry into Amoebae ». *Current Opinion in Microbiology* 31 (juin): 88-93. https://doi.org/10.1016/j.mib.2016.03.009.

Rodrigues, Rodrigo Araújo Lima, Julien Andreani, Ana Cláudia dos Santos Pereira Andrade, Talita Bastos Machado, Souhila Abdi, Anthony Levasseur, Jônatas Santos Abrahão, et Bernard La Scola.

2018. « Morphologic and Genomic Analyses of New Isolates Reveal a Second Lineage of Cedratviruses ». Édité par Rozanne M. Sandri-Goldin. *Journal of Virology* 92 (13): e00372-18, /jvi/92/13/e00372-18.atom. https://doi.org/10.1128/JVI.00372-18.

Rodrigues, Rodrigo Araújo Lima, Amina Cherif Louazani, Agnello Picorelli, Graziele Pereira Oliveira, Francisco Pereira Lobo, Philippe Colson, Bernard La Scola, et Jônatas Santos Abrahão. 2020. « Analysis of a Marseillevirus Transcriptome Reveals Temporal Gene Expression Profile and Host Transcriptional Shift ». *Frontiers in Microbiology* 11. https://doi.org/10.3389/fmicb.2020.00651.

Rodríguez-Zaragoza, S. 1994. « Ecology of Free-Living Amoebae ». *Critical Reviews in Microbiology* 20 (3): 225-41. https://doi.org/10.3109/10408419409114556.

Santos Silva, Ludmila Karen dos, Ana Claudia dos Santos Pereira Andrade, Fabio Pio Dornas, Rodrigo Araujo Lima Rodrigues, Thalita Arantes, Erna Geessien Kroon, Claudio Antonio Bonjardim, et Jonatas Santos Abrahao. 2018. « Cedratvirus Getuliensis Replication Cycle: An in-Depth Morphological Analysis ». *Scientific Reports* 8 (mars): 4000. https://doi.org/10.1038/s41598-018-22398-3.

Schulz, Frederik, Lauren Alteio, Danielle Goudeau, Elizabeth M. Ryan, Feiqiao B. Yu, Rex R. Malmstrom, Jeffrey Blanchard, et Tanja Woyke. 2018. « Hidden Diversity of Soil Giant Viruses ». *Nature Communications* 9 (1): 1-9. https://doi.org/10.1038/s41467-018-07335-2.

Schulz, Frederik, Simon Roux, David Paez-Espino, Sean Jungbluth, David A. Walsh, Vincent J. Denef, Katherine D. McMahon, et al. 2020. « Giant Virus Diversity and Host Interactions through Global Metagenomics ». *Nature* 578 (7795): 432-36. https://doi.org/10.1038/s41586-020-1957-x.

Shinn, George L., et Brianna L. Bullard. 2018. « Ultrastructure of Meelsvirus: A Nuclear Virus of Arrow Worms (Phylum Chaetognatha) Producing Giant "Tailed" Virions ». *PLOS ONE* 13 (9): e0203282. https://doi.org/10.1371/journal.pone.0203282.

Souza, Fernanda, Rodrigo Rodrigues, Erik Reis, Maurício Lima, Bernard La Scola, et Jônatas Abrahão. 2019. « In-Depth Analysis of the Replication Cycle of Orpheovirus ». *Virology Journal* 16 (1): 158. https://doi.org/10.1186/s12985-019-1268-8.

Speciale, Immacolata, Anna Notaro, Chantal Abergel, Rosa Lanzetta, Todd L. Lowary, Antonio Molinaro, Michela Tonetti, James L. Van Etten, et Cristina De Castro. 2022. « The Astounding World of Glycans from Giant Viruses ». *Chemical Reviews*, juillet. https://doi.org/10.1021/acs.chemrev.2c00118.

Stanley, W. M. 1935. « ISOLATION OF A CRYSTALLINE PROTEIN POSSESSING THE PROPERTIES OF TOBACCO-MOSAIC VIRUS ». *Science (New York, N.Y.)* 81 (2113): 644-45. https://doi.org/10.1126/science.81.2113.644.

Sun, Cheng, Cédric Feschotte, Zhiqiang Wu, et Rachel Lockridge Mueller. 2015. « DNA transposons have colonized the genome of the giant virus Pandoravirus salinus ». *BMC Biology* 13 (juin). https://doi.org/10.1186/s12915-015-0145-1.

Sun, Jiya, Fei Ye, Aiping Wu, Ren Yang, Mei Pan, Jie Sheng, Wenjie Zhu, et al. 2020. « Comparative Transcriptome Analysis Reveals the Intensive Early Stage Responses of Host Cells to SARS-CoV-2 Infection ». *Frontiers in Microbiology* 11. https://www.frontiersin.org/articles/10.3389/fmicb.2020.593857.

Sun, Tsu-Wang, et Chuan Ku. 2021. « Unraveling gene content variation across eukaryotic giant viruses based on network analyses and host associations ». *Virus Evolution* 7 (2): veab081. https://doi.org/10.1093/ve/veab081.

Sundberg, Lotta-Riina, et Katja Pulkkinen. 2015. « Genome Size Evolution in Macroparasites ». *International Journal for Parasitology* 45 (5): 285-88. https://doi.org/10.1016/j.ijpara.2014.12.007.

Tettelin, Hervé, Vega Masignani, Michael J. Cieslewicz, Claudio Donati, Duccio Medini, Naomi L. Ward, Samuel V. Angiuoli, et al. 2005. « Genome Analysis of Multiple Pathogenic Isolates of Streptococcus Agalactiae: Implications for the Microbial "Pan-Genome" ». *Proceedings of the National Academy of Sciences* 102 (39): 13950-55. https://doi.org/10.1073/pnas.0506758102.

Thomas, Vincent, Gerald McDonnell, Stephen P. Denyer, et Jean-Yves Maillard. 2010. « Free-Living Amoebae and Their Intracellular Pathogenic Microorganisms: Risks for Water Quality ». *FEMS Microbiology Reviews* 34 (3): 231-59. https://doi.org/10.1111/j.1574-6976.2009.00190.x.

Toenshoff, Elena R., Peter D. Fields, Yann X. Bourgeois, et Dieter Ebert. 2018. « The End of a 60-year Riddle: Identification and Genomic Characterization of an Iridovirus, the Causative Agent of White Fat Cell Disease in Zooplankton ». *G3: Genes|Genomes|Genetics* 8 (4): 1259-72. https://doi.org/10.1534/g3.117.300429.

Toft, Christina, et Siv G. E. Andersson. 2010. « Evolutionary microbial genomics: insights into bacterial host adaptation ». *Nature Reviews Genetics* 11 (7): 465-75. https://doi.org/10.1038/nrg2798.

Tsai, Chih-Tung, Jing-Wen Ting, Ming-Hsien Wu, Ming-Feng Wu, Ing-Cherng Guo, et Chi-Yao Chang. 2005. « Complete Genome Sequence of the Grouper Iridovirus and Comparison of Genomic Organization with Those of Other Iridoviruses. » *Journal of Virology* 79 (4): 2010-23. https://doi.org/10.1128/JVI.79.4.2010-2023.2005.

Turner, Neil A., Giancarlo A. Biagini, et David Lloyd. 1997. « Anaerobiosis-induced differentiation of Acanthamoeba castellanii ». *FEMS Microbiology Letters* 157 (1): 149-53. https://doi.org/10.1111/j.1574-6968.1997.tb12766.x.

Vasiliev, Gennadii G., Anton A. Dzhaljabov, et Igor A. Leonovich. 2021. « Analysis of the causes of engineering structures deformations at gas industry facilities in the permafrost zone ». *Journal of Mining Institute* 249 (septembre): 377-85. https://doi.org/10.31897/PMI.2021.3.6.

Venter, J. Craig, Karin Remington, John F. Heidelberg, Aaron L. Halpern, Doug Rusch, Jonathan A. Eisen, Dongying Wu, et al. 2004. « Environmental Genome Shotgun Sequencing of the Sargasso Sea ». *Science (New York, N.Y.)* 304 (5667): 66-74. https://doi.org/10.1126/science.1093857.

Vishnivetskaya, T. A., E. V. Spirina, A. V. Shatilovich, L. G. Erokhina, E. A. Vorobyova, et D. A. Gilichinsky. 2003. « The Resistance of Viable Permafrost Algae to Simulated Environmental Stresses: Implications for Astrobiology ». *International Journal of Astrobiology* 2 (3): 171-77. https://doi.org/10.1017/S1473550403001575.

Vondrak, Tihana, Ludmila Oliveira, Petr Novák, Andrea Koblížková, Pavel Neumann, et Jiří Macas. 2021. « Complex Sequence Organization of Heterochromatin in the Holocentric Plant Cuscuta Europaea Elucidated by the Computational Analysis of Nanopore Reads ». *Computational and Structural Biotechnology Journal* 19 (janvier): 2179-89. https://doi.org/10.1016/j.csbj.2021.04.011.

Wernegreen, Jennifer J. 2005. « For Better or Worse: Genomic Consequences of Intracellular Mutualism and Parasitism ». *Current Opinion in Genetics & Development*, Genomes and evolution, 15 (6): 572-83. https://doi.org/10.1016/j.gde.2005.09.013.

Wilhelm, Steven W., et Curtis A. Suttle. 1999. « Viruses and Nutrient Cycles in the Sea: Viruses play critical roles in the structure and function of aquatic food webs ». *BioScience* 49 (10): 781-88. https://doi.org/10.2307/1313569.

Williamson, Kurt E., Jeffry J. Fuhrmann, K. Eric Wommack, et Mark Radosevich. 2017. « Viruses in Soil Ecosystems: An Unknown Quantity Within an Unexplored Territory ». *Annual Review of Virology* 4 (1): 201-19. https://doi.org/10.1146/annurev-virology-101416-041639.

Williamson, Shannon J., Douglas B. Rusch, Shibu Yooseph, Aaron L. Halpern, Karla B. Heidelberg, John I. Glass, Cynthia Andrews-Pfannkoch, et al. 2008. « The Sorcerer II Global Ocean Sampling Expedition: Metagenomic Characterization of Viruses within Aquatic Microbial Samples ». *PLOS ONE* 3 (1): e1456. https://doi.org/10.1371/journal.pone.0001456.

Windirsch, Torben, Guido Grosse, Mathias Ulrich, Lutz Schirrmeister, Alexander N. Fedorov, Pavel Y. Konstantinov, Matthias Fuchs, et al. 2020. « Organic Carbon Characteristics in Ice-Rich Permafrost in Alas and Yedoma Deposits, Central Yakutia, Siberia ». *Biogeosciences* 17 (14): 3797-3814. https://doi.org/10.5194/bg-17-3797-2020.

Winter, David J., Austen R. D. Ganley, Carolyn A. Young, Ivan Liachko, Christopher L. Schardl, Pierre-Yves Dupont, Daniel Berry, Arvina Ram, Barry Scott, et Murray P. Cox. 2018. « Repeat elements organise 3D genome structure and mediate transcription in the filamentous fungus Epichloë festucae ». *PLoS Genetics* 14 (10): e1007467. https://doi.org/10.1371/journal.pgen.1007467.

Woolhouse, Mark E. J., et Sonya Gowtage-Sequeria. 2005. « Host Range and Emerging and Reemerging Pathogens ». *Emerging Infectious Diseases* 11 (12): 1842-47. https://doi.org/10.3201/eid1112.050997.

Wu, Xiaofen, Archana Chauhan, Alice C. Layton, Maggie C. Y. Lau Vetter, Brandon T. Stackhouse, Daniel E. Williams, Lyle Whyte, Susan M. Pfiffner, Tullis C. Onstott, et Tatiana A. Vishnivetskaya. 2021. « Comparative Metagenomics of the Active Layer and Permafrost from Low-Carbon Soil in the Canadian High Arctic ». *Environmental Science & Technology* 55 (18): 12683-93. https://doi.org/10.1021/acs.est.1c00802.

Xue, Yaxin, Inge Jonassen, Lise Øvreås, et Neslihan Taş. 2020. « Metagenome-assembled genome distribution and key functionality highlight importance of aerobic metabolism in Svalbard permafrost ». *FEMS Microbiology Ecology* 96 (5): fiaa057. https://doi.org/10.1093/femsec/fiaa057.

Yang, Guojun, Dawn Holligan Nagel, Cédric Feschotte, C. Nathan Hancock, et Susan R. Wessler. 2009. « Tuned for Transposition: Molecular Determinants Underlying the Hyperactivity of a Stowaway MITE ». *Science (New York, N.Y.)* 325 (5946): 1391-94. https://doi.org/10.1126/science.1175688.

Yashina, Svetlana, Stanislav Gubin, Stanislav Maksimovich, Alexandra Yashina, Edith Gakhova, et David Gilichinsky. 2012. « Regeneration of whole fertile plants from 30,000-y-old fruit tissue buried in Siberian permafrost ». *Proceedings of the National Academy of Sciences of the United States of America* 109 (10): 4008-13. https://doi.org/10.1073/pnas.1118386109.

Yutin, Natalya, et Eugene V. Koonin. 2012. « Hidden evolutionary complexity of Nucleo-Cytoplasmic Large DNA viruses of eukaryotes ». *Virology Journal* 9 (1): 161. https://doi.org/10.1186/1743-422X-9-161.

———. 2013. « Pandoraviruses are highly derived phycodnaviruses ». *Biology Direct* 8 (1): 25. https://doi.org/10.1186/1745-6150-8-25.

Zhang, Hua-Hao, Qiu-Zhong Zhou, Ping-Lan Wang, Xiao-Min Xiong, Andrea Luchetti, Didier Raoult, Anthony Levasseur, et al. 2018. « Unexpected invasion of miniature inverted-repeat transposable elements in viral genomes ». *Mobile DNA* 9 (1): 19. https://doi.org/10.1186/s13100-018-0125-4.

Zimmer, Carl. 2017. *Planète de virus*. Humensis.

Annexes

I. Phylogénie des Pithoviridae isolés et métagénomiques

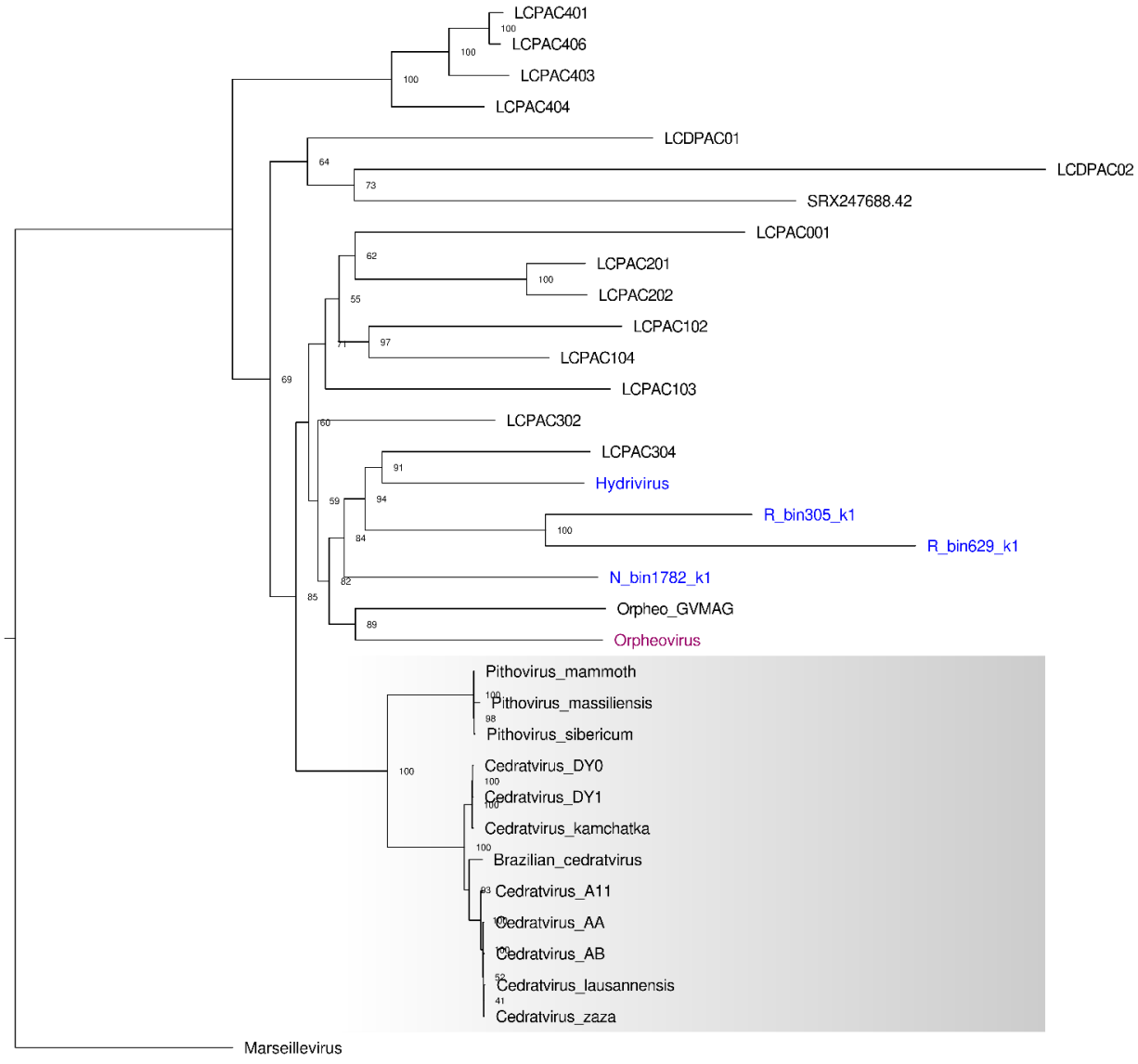


Figure 34 – Phylogénie des Pithoviridae isolés et métagénomiques
La phylogénie a été calculée par IQtree sur les alignements de l'ADN polymérase, de VLTF3, des deux sous-unités principales de l'ARN polymérase, des sous-unités 5 et 10 de l'ARN polymérase et de l'enzyme coiffante. Les valeurs de bootstraps ont été calculés sur 5000 itérations en éliminant les 200 premières. Le clade surligné correspond aux génomes de virus isolés. Les noms en bleu sont les grands virus alignés par l'article II (p. 80). Orpheovirus est en mauve. Tous les autres séquences sont issues de la métagénomique (Bäckström et al. 2019; Schulz et al. 2020; Moniruzzaman et al. 2019).

nature COMMUNICATIONS

ARTICLE

https://doi.org/10.1038/s41467-020-16414-2 OPEN

The DNA methylation landscape of giant viruses

Sandra Jeudy[1], Sofia Rigou[1], Jean-Marie Alempic[1], Jean-Michel Claverie[1], Chantal Abergel[1] & Matthieu Legendre[1✉]

DNA methylation is an important epigenetic mark that contributes to various regulations in all domains of life. Giant viruses are widespread dsDNA viruses with gene contents overlapping the cellular world that also encode DNA methyltransferases. Yet, virtually nothing is known about the methylation of their DNA. Here, we use single-molecule real-time sequencing to study the complete methylome of a large spectrum of giant viruses. We show that DNA methylation is widespread, affecting 2/3 of the tested families, although unevenly distributed. We also identify the corresponding viral methyltransferases and show that they are subject to intricate gene transfers between bacteria, viruses and their eukaryotic host. Most methyltransferases are conserved, functional and under purifying selection, suggesting that they increase the viruses' fitness. Some virally encoded methyltransferases are also paired with restriction endonucleases forming Restriction-Modification systems. Our data suggest that giant viruses' methyltransferases are involved in diverse forms of virus-pathogens interactions during coinfections.

[1] Aix Marseille Univ., CNRS, IGS, Information Génomique & Structurale (UMR7256), Institut de Microbiologie de la Méditerranée (FR 3489), Marseille, France. ✉email: legendre@igs.cnrs-mrs.fr

Methylation of DNA is an important class of epigenetic modification observed in the genomes of all domains of life. In eukaryotes, it is involved in biological processes as diverse as gene expression regulation, transposon silencing, genomic imprinting, or development[1–4]. In prokaryotes, DNA methylation often results from the targeted activity of methyltransferases (MTases) that are components of the restriction-modification (R-M) systems, which involve methylation and restriction activity. Within these systems, restriction enzymes (REases) cleave the DNA only if the shared recognized motifs are unmethylated[5]. This provides prokaryotes with a powerful weapon against foreign DNA, such as the one of infecting viruses[6]. Besides R-M systems, prokaryotic DNA MTases may occur without cognate REases, in which case they are coined orphan. Prokaryotic orphan MTases are involved in the regulation of gene expression[7,8], DNA replication[9], DNA repair[10], and cell cycle regulation[11].

Outside of the cellular world, some DNA viruses exploit DNA methylation as a mechanism to regulate their replication cycle. For instance, the transition from latent to lytic infection in Epstein–Barr Virus is mediated by the expression of genes that are silenced or transcribed according to the methylation status of their promoter[12]. Iridoviruses and ascoviruses sometimes exhibit heavily methylated genomes and encode their own MTases[12]. *Phycodnaviridae* members also encode functional MTases able to methylate their own DNA[13], and among them some chloroviruses encode complete R-M systems with their associated REases[14]. These endonucleases, packaged in the virions, contribute to the degradation of host DNA either to allow for the recycling of deoxynucleotides, or to inhibit the expression of host genes by shifting the transcription from host to viral DNA[14].

Over the last 15 years several viruses whose particles are large enough to be seen by light microscopy were discovered[15–23]. These so-called giant viruses exhibit DNA genomes as large and complex as prokaryotes[16], or even parasitic eukaryotes[22]. A growing body of metagenomics surveys shows that they are widespread on the planet in diverse environments[24,25]. The first family of giant viruses to be discovered, the *Mimiviridae*, have megabase-sized AT-rich linear genomes packaged in icosahedral capsids[16–18]. Intriguingly some *Mimiviridae* members are infected by smaller 20-kb-dsDNA viruses, dubbed virophages[26,27] and sometimes found in association with 7-kb-DNA episomes called transpovirons[28,29]. In contrast to *Mimiviridae*, the pandoraviruses have GC-rich linear genomes, twice as big with up to 2.5 Mb, packaged in amphora-shaped capsids[22,30]. Again different, pithoviruses[20] and cedratviruses[21,31] have smaller circular AT-rich genomes, ranging from 400 to 700 Kb, and packaged in the largest known amphora-shaped capsids. Thus, although these different giant viruses infect the same hosts (amoebas of the *Acanthamoeba* genus), they exhibit different morphological features, replication cycles, gene contents, and potential epigenomic modifications. To date virtually nothing is known about the epigenomes of these giant viruses. In particular, the methylation status of their DNA is unknown despite the presence of predicted MTases in their genomes. Pandoravirus dulcis for instance encodes up to five different DNA MTases (UniprotKB IDs: [S4VR68], [S4VTY0], [S4VS49], [S4VUD3], and [S4VQ82])[22]. Yet, it remains to be assessed whether any of these enzymes methylate the viral DNA.

Most of the genome-wide studies of eukaryotic DNA methylation have been performed using bisulfite sequencing techniques[32]. These approaches only detect 5-methyl-cytosine modifications and are thus not well suited for the analysis of prokaryotic-like epigenomic modifications, mostly composed of N^6-methyl-adenines and N^4-methyl-cytosines. However, the recently developed single-molecule real-time (SMRT) sequencing method overcomes this limitation[33]. Briefly this approach analyses the kinetics of incorporation of modified nucleotides by the polymerase compared to the non-modified ones. The Inter-Pulse Duration ratio (IPDr) metric can then be computed for each genomic position, and makes it possible to map all modified nucleotides and methylated motifs along the genome. This approach is now extensively used to study the methylation landscapes of isolated bacteria[34], archaea[35], and even prokaryotic metagenomes[36].

Here, we use SMRT sequencing to survey the complete methylome of a large spectrum of giant viruses. We analyze two distinct *Mimiviridae* members and their associated transpovirons, as well as a virophage[28]. We also survey a *Marseilleviridae* member[37], five distinct pandoraviruses[22,30,38], a mollivirus[23] and a pithovirus[20]. Finally, we isolate a new cedratvirus (cedratvirus kamchatka) that we sequence using SMRT sequencing to assess its methylome. Furthermore, we thoroughly annotate MTases and REases contained in all these genomes and analyze their phylogenetic histories. Our findings reveal that DNA methylation is widespread among giant viruses and open new avenues of research on its role in their population dynamics.

Results

Methylome and MTase gene contents of giant viruses. We gathered PacBio SMRT data of diverse families from previously published genomes sequenced by our group to analyze the DNA methylation profile of a wide range of giant viruses. SMRT genomic data were collected for the following viruses: the *Mimiviridae* member moumouvirus australiensis and its associated transpoviron[28], the *Marseilleviridae* member melbournevirus[37] and five pandoraviruses (pandoravirus celtis[38], pandoravirus dulcis[30], pandoravirus neocaledonia[30], pandoravirus quercus[30], and pandoravirus salinus[30]). In addition, we resequenced on the PacBio platform the complete genomes of mollivirus sibericum, pithovirus sibericum, the *Lavidaviridae* member zamilon vitis and the megavirus vitis *Mimiviridae* member together with its associated transpoviron. Finally, we sequenced a newly isolated strain of cedratvirus (cedratvirus kamchatka). The datasets are listed in Supplementary Table 1. The sequence data obtained from the whole collection corresponds to an average coverage of 192-fold.

We then aligned the SMRT reads to their corresponding reference genomes (see Supplementary Table 1) and computed IPDr at each genomic position (see Methods). These genome-wide profiles were used to identify overrepresented sequence motifs at positions with high IPDr values. All the identified motifs were palindromic and prone to either N^4-methyl-cytosine or N^6-methyl-adenine methylations (Fig. 1). As a control, we applied the same procedure to DNA samples of mollivirus sibericum and pithovirus sibericum subjected to whole genome amplifications (WGA), which in principle erase methylation marks[33]. As expected, no overrepresented methylated motif was detected in these controls, and the median IPDr of the motifs previously detected in the wild-type datasets were basal here (see Supplementary Fig. 1 and Fig. 1).

In parallel, we analyzed the DNA MTases encoded in these genomes and predicted their target sequences based on their homology with characterized MTases (see Methods). All the MTases for which a target site could be predicted were putative type II MTases. In 15 out of the 19 cases (79%) where a MTase target could be predicted or a methylated motif detected, we found an agreement between the two (Fig. 1). It is worth mentioning that this result also highlights the reliability of MTases' targets predictions based on protein homology.

***Marseilleviridae* members encode complete R-M systems**. Melbournevirus encodes a DNA MTase (mel_016) predicted to

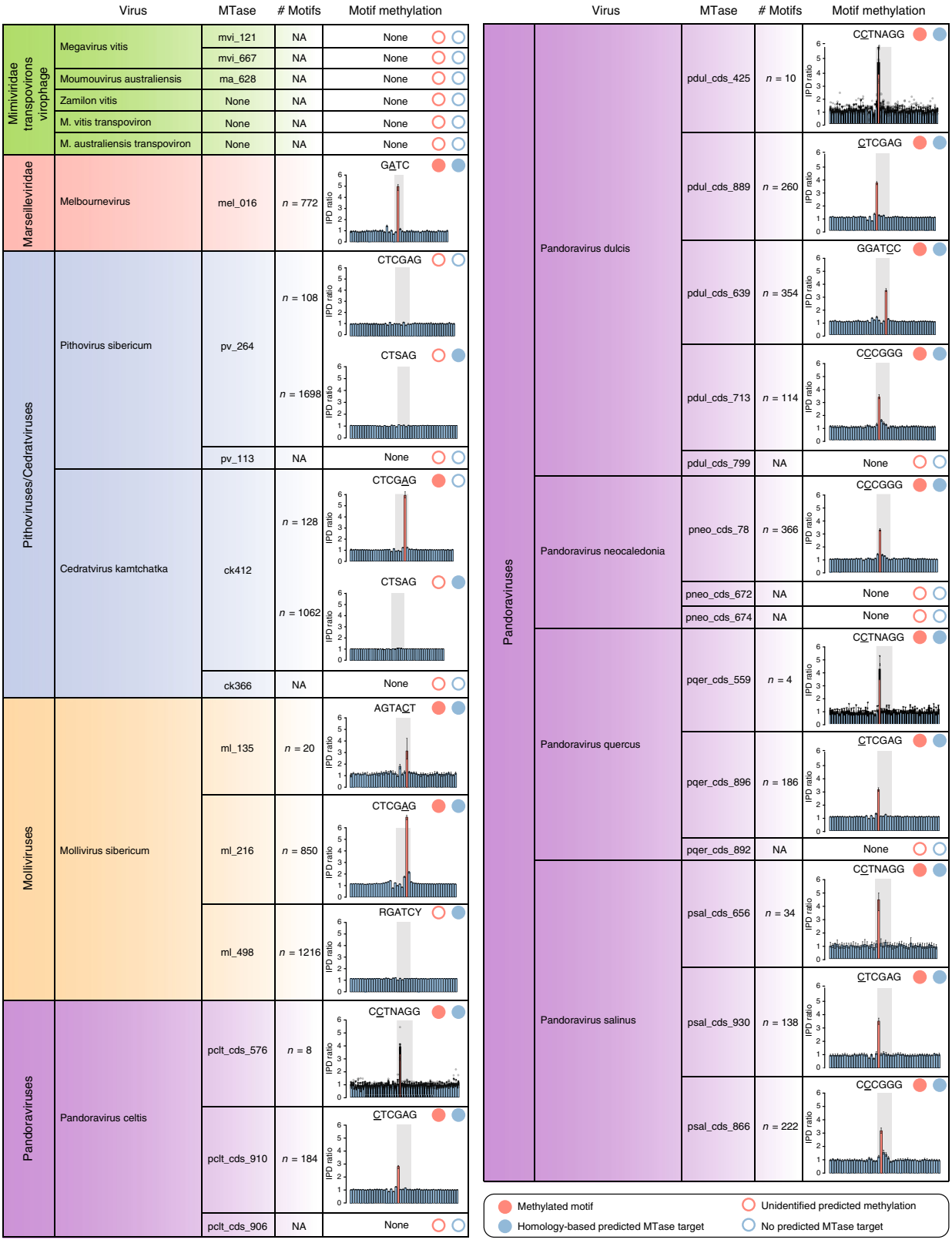


Fig. 1 Encoded MTases and targeted methylated motifs in the giant viruses' genomes. The encoded DNA MTases of each virus are shown along with the number of occurrences of the predicted targets (if any) on both strands of the cognate genomic sequence. Modified nucleotides within the motifs are underlined. Red circles indicate methylated motifs experimentally verified from SMRT data (filled circles) or with unidentified predicted methylation (empty circles). Likewise, predicted (filled circle) and not predicted (empty circle) targets based on sequence homology of the encoded MTase are shown using blue circles. Bar graphs correspond to the median IPDr profiles of the motifs (gray region) and the surrounding 20 nucleotides on each side. Each bar displays the median IPDr value and a 95% confidence interval (error bars) based on 1000 bootstraps. These statistics are derived from the number of occurrences (n) of the motifs in each genome. Red bars correspond to positions with significantly high IPDr values. Individual data points are displayed for viruses with $n \leq 10$.

target GATC sites. Our data confirm that GATC motifs were modified (underlined characters indicate methylated bases) with N^6-methyl-adenines (Fig. 1). We then searched for the possible cognate REase in the genomic vicinity of mel_016 and identified the neighboring mel_015 gene as a candidate. Although the encoded protein does not exhibit a recognizable motif using standard domain search tools[39,40], a search against REbase, the database dedicated to R-M systems[41], identified it as a probable GATC-targeting REase. Moreover, the mel_015 protein is similar (blast *E*-value $= 2 \times 10^{-33}$) to the *Paramecium bursaria* Chlorella virus 1 (PBCV-1) CviAI REase known to target GATC sites. Melbournevirus thus encodes a complete R-M system.

The N^6-methyl-adenine modification is typical of prokaryotic MTases. Since melbournevirus is a eukaryotic virus, our finding immediately questioned the evolutionary history of its encoded R-M system. We thus reconstructed the phylogeny of the complete system, including the MTase and the REase (see Methods). The mel_016 MTase strikingly branches within the prokaryotes along with other viruses, mostly chloroviruses and members of the *Marseilleviridae* and *Mimiviridae* of the proposed mesomimivirinae subfamily[42] (Supplementary Fig. 2A). In agreement with its enzymatic activity, this phylogeny suggests that the melbournevirus MTase was acquired from a prokaryote. Likewise, the phylogeny of the mel_015 REase suggests its prokaryotic origin (Supplementary Fig. 2B). Altogether, these results support a relatively ancient acquisition from prokaryotes as the origin of the complete marseilleviruses R-M system.

Surprisingly, we did not identify orthologues of the melbournevirus R-M system in all *Marseilleviridae* members. As shown in Fig. 2, only 5 out of the 13 marseilleviruses genomes contain both a MTase and a REase, always encoded next to each other. The other marseilleviruses encode neither the MTase nor the REase. The *Marseilleviridae* phylogeny based on core genes (see Methods) clearly coincides with its dichotomous distribution (Fig. 2). All the clade A members of the *Marseilleviridae* encode a complete R-M system, while the others do not. This suggests that the marseilleviruses R-M system was acquired by the clade A ancestor. It is worth noticing that once acquired, the R-M system was maintained, with none of the two enzymes undergoing pseudogenization. This suggests that the encoded R-M system has a functional role in these viruses.

Activity of the *Marseilleviridae* members R-M system on non-self DNA. Once methylated by the mel_016-encoded enzyme, we expect the viral DNA to be protected from its digestion by the cognate mel_015 REase. If not, the melbournevirus genome would be theoretically fragmented into 387 fragments of 954 nt on average. We verified this prediction by conducting DNA restriction experiments using two endonucleases targeting GATC sites: DpnI and DpnII. The former only cleaves DNA at modified GATC sites containing a N^6-methyl-adenine, while the latter conversely only cleaves unmethylated GATC sites. Figure 3 demonstrates that melbournevirus DNA is digested by DpnI but not by DpnII. Thus, assuming that mel_015 REase is functional, we can infer that melbournevirus is able to protect its own genome from its encoded R-M system digestion. As a control, we reproduced the above experiment with the DNA of noumeavirus[43], a *Marseilleviridae* member belonging to the clade B that do not encode a R-M system (Fig. 2). As expected, the noumeavirus genome is digested by DpnII but not by DpnI (Fig. 3). This demonstrates that noumeavirus DNA is not methylated at GATC sites and is thus susceptible to degradation by a co-infecting marseillevirus bearing a functional R-M system.

To verify whether the *Acanthamoeba castellanii* host genome was sensitive to DNA degradation at GATC sites we performed digestion assays using the same enzymes. According to its sequence (GenBank accession AEYA00000000), the *A. castellanii* genome should be fragmented into 171,298 pieces of 273 nt on average. Surprisingly, the *A. castellanii* genome is cleaved by both enzymes (Fig. 3). This indicates that the host DNA contains a mixture of methylated and unmethylated GATC sites. However, the restriction profiles show that unmethylated positions are in larger proportion than methylated ones.

Since the host DNA is (at least partially) unprotected from the marseilleviruses encoded R-M systems we next assessed its potential degradation during a melbournevirus infection. As shown in Supplementary Fig. 3, A. *castellanii* DNA is not degraded during the infection. As expected, the infection of *A. castellanii* with noumeavirus, which do not encode the GATC R-M system, do not alter its DNA. Hence, marseilleviruses encoded R-M systems do not contribute to host DNA degradation.

To further investigate the role of the marseilleviruses R-M systems we analyzed the timing of expression of both enzymes (the MTase and the REase) during a melbournevirus infection. Figure 4 shows that the mel_015 REase gene is first transcribed between 30 min and 45 min post infection, followed by the mel_016 MTase gene expressed between 45 min and 1 h post infection. In addition, proteomic data of the melbournevirus virion from[43] confirm that the mel_016 MTase protein is packaged in the particle whereas the REase is not.

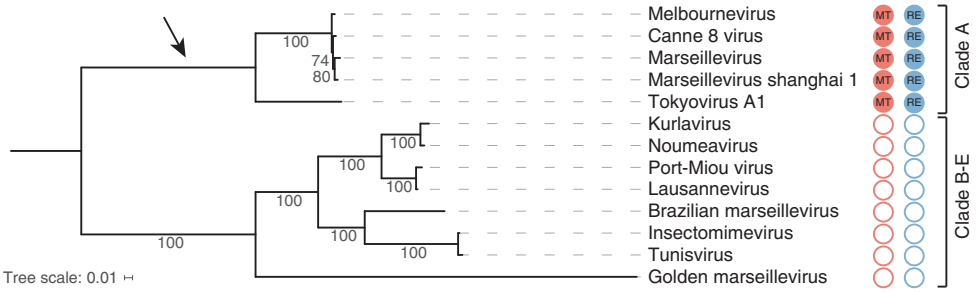


Fig. 2 Presence/absence of R-M systems in the *Marseilleviridae* family. Phylogeny of the *Marseilleviridae* completely sequenced viruses with the following GenBank accessions: melbournevirus (KM275475.1)[37], cannes 8 virus (KF261120.1)[88], marseillevirus (GU071086.1)[89], marseillevirus shanghai 1 (MG827395.1), tokyovirus A1 (AP017398.1)[90], kurlavirus (KY073338.1)[91], noumeavirus (KX066233.1)[43], port-Miou virus (KT428292.1)[92], lausannevirus (HQ113105.1)[48], brazilian marseillevirus (KT752522.1)[93], insectomimevirus (KF527888.1)[94], tunisvirus (KF483846.1)[95] and golden marseillevirus (KT835053.1)[96]. The phylogeny was based on protein sequence alignments of the 115 strictly conserved single copy orthologues (see Supplementary Data 1). The tree was calculated using the best model of each partitioned alignment as determined by IQtree[82]. Bootstrap values were computed using the UFBoot[84] method. The red and blue filled circles highlight the presence of encoded marseilleviruses R-M systems MTases and REases respectively. The empty circles highlight the absence of the MTase (in red) and the REase (in blue). The arrow points to the most parsimonious acquisition of a R-M system in the *Marseilleviridae* family.

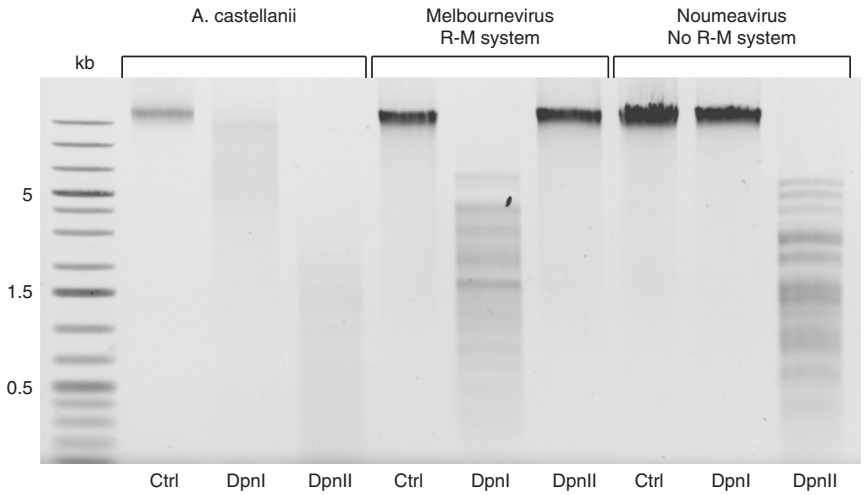


Fig. 3 Host and marseilleviruses DNA protection against GATC-targeting REases. Agarose gel electrophoresis analysis of *A. castellanii*, melbournevirus and noumeavirus DNA digested with GATC-targeting restriction enzymes. Restriction patterns using DpnI and DpnII enzymes are presented with control DNA. DpnI cleaves DNA at GATC sites containing N^6-methyl-adenines and DpnII at GATC sites containing unmethylated adenines. This experiment was repeated twice with similar results.

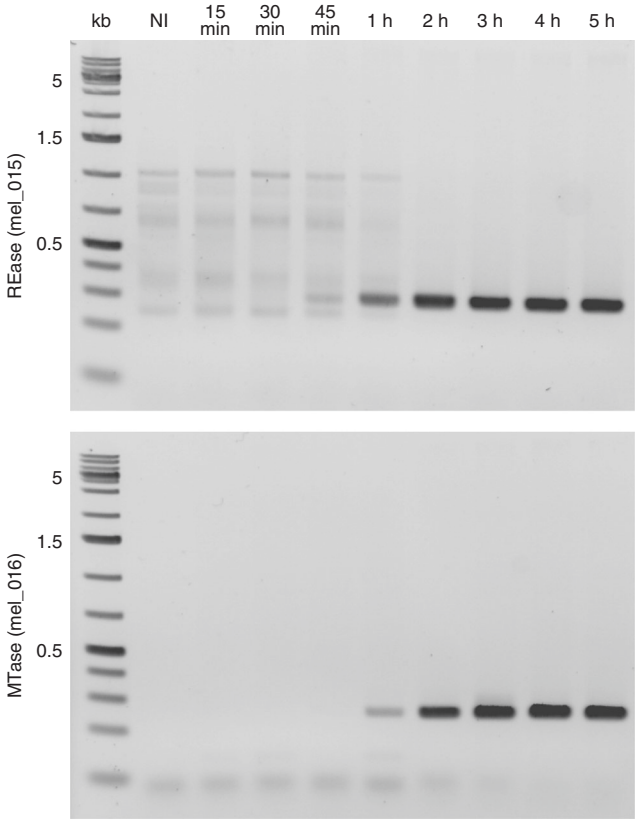


Fig. 4 Expression timing of the melbournevirus R-M system MTase and REase. Shown are the RT-PCRs of the transcripts corresponding to the mel_015 and mel_016 genes during a melbournevirus infection. Times (post infection) are listed on the top of the figure. NI corresponds to non-infected. This experiment was repeated twice with similar results.

The newly isolated cedratvirus kamchatka. Cedratviruses are giant viruses morphologically and to some extant genetically related to pithoviruses[31]. Four completely sequenced genomes are available today: cedratvirus A11[21], cedratvirus zaza[44], cedratvirus lausannensis[31], and brazilian cedratvirus[44]. We isolated a new strain of cedratvirus (named cedratvirus kamchatka) from a muddy grit soil sample collected near a lake at kizimen volcano, Kamchatka (Russian Federation N 55°05'50 E 160°20'58) (see Methods). SMRT sequencing was used to characterize both its genome and methylome. The cedratvirus kamchatka genome was assembled into a circular 466,767 bp DNA molecule (41% G + C), predicted to encode 545 protein-coding genes The genome size and topology was confirmed by pulsed field gel electrophoresis (PFGE) (Supplementary Fig. 4A).

The phylogenetic tree computed from the pithoviruses and cedratviruses core genes shows that they cluster in well-separated groups (Supplementary Fig. 4B). Their orthologous proteins share an average of 46% identical residues. The available cedratviruses appear to split into three distinct clades: clade A contains cedratvirus A11, cedratvirus zaza and cedratvirus lausannensis, clade B contains cedratvirus kamchatka, and clade C contains brazilian cedratvirus (Supplementary Fig. 4B). This classification might be challenged as new strains will be characterized. We found that 51 of the 545 cedratvirus kamchatka genes were unique to this strain compared to the other cedratviruses. According to the presence/absence of pseudogenes in the other strains, as detected using tblastn, we designed a putative evolutionary scenario for each of these genes (see Supplementary Table 2). As previously discussed for pandoraviruses[30,38], the process of de novo gene creation seems to participate to the shaping of cedratviruses genomes. Interestingly, among the 51 genes unique to cedratvirus kamchatka only one has a clear predicted function: the ck412 DNA MTase.

DNA methylation is widespread but unevenly distributed among giant viruses genomes. The SMRT sequencing of cedratvirus kamchatka and the other methylome datasets show that the various giant virus families exhibit distinct methylation features. Whereas the genomes of pithoviruses and *Mimiviridae* members are devoid of DNA modifications, those of molliviruses, pandoraviruses and cedratviruses clearly contain methylated nucleotides (Fig. 1).

More specifically, cedratvirus kamchatka DNA is methylated at CTCGAG motifs (Fig. 1). Although the ck412 DNA MTase has a slightly different predicted target (CTSAG), it is probably responsible for the CTCGAG methylation as CTSAG motifs are

not methylated (Fig. 1). Cedratvirus kamchatka ck366 gene encodes an additional predicted FkbM domain-containing MTase for which the predicted specific target, if any, is unknown. Importantly, we found no REase associated with the cedratvirus kamchatka predicted MTases.

Pithovirus sibericum encodes two DNA MTases: pv_264 predicted to target the CTSAG motif and pv_113, an FkbM domain-containing MTase targeting an unknown site. Yet, pithovirus sibericum DNA exhibit no methylated sites (including CTSAG and CTCGAG) (Fig. 1). Surprisingly, the RNA-seq data from[20] shows that both transcripts are significantly expressed all along the replication cycle (see Supplementary Tables 3 and 4). Finally, none of the genes surrounding the MTases are predicted to encode a functional REase. Thus, according to our SMRT-seq data pithovirus sibericum MTase-like proteins do not methylate the viral DNA even though they are expressed.

Although megavirus vitis encodes a type 11 domain MTase (mvi_121) and an FkbM domain-containing MTase (mvi_667) shared with moumouvirus australiensis (ma_628), we cannot infer their DNA target specificities from sequence homology or experimental evidences since none of the two genomes appear to be methylated (Fig. 1). In addition to these MTase-like candidates, megavirus vitis and moumouvirus australiensis encode a 6-O-methylguanine-DNA methyltransferase (mvi_228/ma_196) probably involved in DNA repair. We also surveyed the DNA methylation of the *Mimiviridae* members' mobilome, namely the zamilon vitis virophage, the megavirus vitis transpoviron and the moumouvirus australiensis transpoviron. These genomes that do not encode DNA MTases are not methylated (Fig. 1). Collectively this data show that the putative DNA MTases encoded by the *Mimiviridae* members infecting *Acanthamoeba* do not methylate the viral or mobilome DNA.

In contrast, all the surveyed pandoraviruses' genomes are methylated (Fig. 1). Unexpectedly they exhibit N^4-methyl-cytosines instead of the N^6-methyl-adenines found in *Marseilleviridae* members and cedratviruses. The number of distinct methylated motifs is also quite variable: a single one in pandoravirus neocaledonia, but two in pandoravirus celtis and pandoravirus quercus, three in pandoravirus salinus, and up to four in pandoravirus dulcis. We successfully assigned each of all methylated motifs to their cognate encoded MTases. However, none appeared to be associated with a REase. In addition, we found a MTase type 25 domain-containing gene in pandoravirus neocaledonia (pneo_cds_672), as well as a MTase type 11 domain gene (pneo_cds_674) with orthologs in pandoravirus dulcis, pandoravirus quercus, and pandoravirus celtis (pdul_cds_799, pqer_cds_892 and pclt_cds_906). None of them had predicted DNA targets.

Finally, the mollivirus sibericum genome is prone to both types of modification: N^4-methyl-cytosines and N^6-methyl-adenines. It is methylated at the AGCACT sites by the ml_135 encoded MTase and at the CTCGAG sites by the ml_216 MTase (Fig. 1). A third MTase encoded by this genome (ml_498) is predicted to recognize the RGATCY sites but our methylome data clearly show that they are not methylated (Fig. 1). We first suspected that the ml_498 gene was not transcribed but transcriptomic data from[23] clearly show that ml_498 is expressed, mostly in the early phase of the infection (see Supplementary Table 5). However, the analysis of the gene structure shows a long 5'UTR, suggesting a N-terminal truncation of the protein (Supplementary Fig. 5A) compared to its homologs (Supplementary Fig. 5B). As a single frameshift is sufficient to restore the N-terminal part of the protein, ml_498 probably underwent a recent pseudogenization.

The activity of ml_216 was further confirmed by restriction experiments showing that mollivirus sibericum DNA is cleaved after WGA at CTCGAG sites but not in wild-type conditions (Supplementary Fig. 6A). By contrast, the RGATCY sites are not protected, as expected from the ml_498 loss of function. Interestingly while the RGATCY, AGTACT and CTCGAG sites are unmethylated in host DNA, the CCCGGG motifs are protected against degradation (Supplementary Fig. 6A, B). This probably corresponds to CpG methylation of the host DNA.

None of the mollivirus sibericum MTases appear to be associated with a corresponding REase. This lack of nuclease activity was confirmed by the absence of host DNA degradation during the infection cycle (Supplementary Fig. 7). One might expect that the host DNA is protected against putative CTCGAG and AGTACT targeting viral REases by endogenously encoded MTases. We exclude this possibility since host DNA is sensitive to degradation at those sites in uninfected conditions (Supplementary Fig. 6B).

The complex evolutionary history of giant viruses' MTases. In the *Marseilleviridae* family, we observed methylation patterns typical of prokaryotic MTases. This raised the question of their evolutionary histories. In Fig. 5 we now present a global phylogenetic analysis of all giant viruses' MTases analyzed in this work (Fig. 1). They clearly do not share a common origin, and appear partitioned in five main groups, interspersed among bacterial and occasional amoebal homologs.

First, one observes that most viral MTases are either embedded within clusters of prokaryotic sequences (green, and red groups in Fig. 5) or constitute sister groups of prokaryotes (orange, purple and blue groups). Thus, as for *Marseilleviridae* members these MTases are most likely of bacterial origin.

Of special interest, the tree also exhibits MTases encoded by different *Acanthamoeba* species (see GenBank accessions in Fig. 5), the main known hosts of giant viruses. For instance, the closest ml_135 mollivirus MTase homolog is found in *A. polyphaga*, suggesting a recent exchange between virus and host. The direction of the gene transfer cannot be determined from these data. However, in the red and orange groups (Fig. 5) other acanthameoba homologs appear well nested within pandoraviruses MTases. This supports transfers occurring from the giant viruses to the host genome. We also noticed divergent MTases attributed to various *Acanthamoeba* species in the purple group. However, a closer inspection of the taxonomic assignment of the corresponding contigs indicates that they are bacterial sequences, probably from the *Bradyrhizobiaceae* family (see Supplementary Fig. 8). These bacteria are amoeba resistant intracellular microorganisms[45] that probably contaminated the eukaryotic host sequencing project.

The purple group also contains three orthologous MTases targeting CTCGAG sites: ml_216 from mollivirus sibericum, pv_264 from pithovirus sibericum and ck412 from cedratvirus kamchatka. These viruses belong to two distinct viral families but infect similar hosts. It is thus likely that these MTases were recently exchanged between these viruses. In the green group there are also three orthologous MTases from distinct viral families: pdul_cds_639 and ppam_cds_578 from pandoravirus dulcis and pandoravirus pampulha respectively, as well as ml_498 from mollivirus sibericum. The two pandoraviruses are closely related with an average of 83% sequence identity between shared proteins. As this MTase is not found in other pandoraviruses (Supplementary Fig. 9), a gene exchange might have occurred between the pandoravirus pampulha-pandoravirus dulcis ancestor and mollivirus sibericum.

Selection pressure acting on giant viruses' MTases. Following the above phylogenetic analysis of the giant viruses MTases, we investigated the selection pressure acting on them. We first

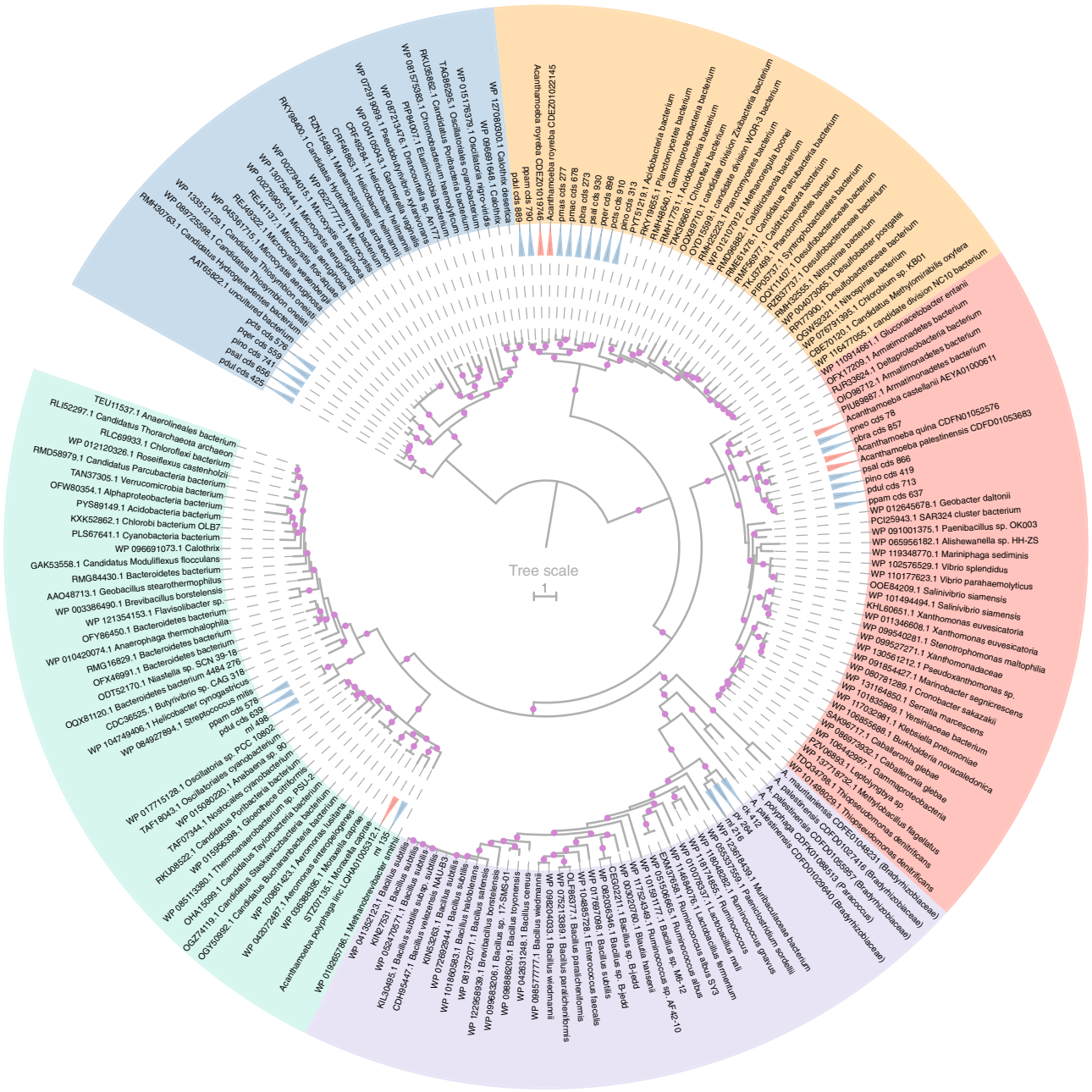


Fig. 5 Phylogenetic tree of the giant viruses MTases. Phylogenetic tree of the giant viruses' MTases along with prokaryotic and eukaryotic homologs. The blue triangles mark viral genes, the red ones eukaryotic genes and the unmarked genes are prokaryotic. The tree was computed using the LG + R6 model from a multiple alignment of 678 informative sites. Bootstrap values were computed using the UFBoot[84] method from IQtree[82]. All branches with support value > 80 are highlighted using purple circles. The GenBank accessions and taxonomic assignations extracted from GenBank entries are shown. The tree was rooted using the midpoint rooting method. The tree was split into five subgroups highlighted using different colors (blue, orange, red, purple, and green).

noticed that some MTases were conserved for long periods of time in various viral families. The CTCGAG and CCCGGG targeting MTases, most likely gained by a Pandoravirus ancestor, remain present in most of the extant Pandoraviruses (Supplementary Fig. 9). Likewise, the marseilleviruses R-M MTase and the CCTNAGG pandoravirus targeting MTase were kept in almost all members of their respective clades (Fig. 2 and Supplementary Fig. 9). By contrast, the ml_498 mollivirus sibericum MTase was found to be recently pseudogenized (Supplementary Fig. 5). In addition, the pino_cds_419 gene from pandoravirus inopinatum and the ppam_cds_578 from pandoravirus pampulha are most likely truncated pseudogenes, even though we do not have SMRT data to confirm their loss of function.

We then computed the ratios (ω) of non-synonymous (dN) to synonymous (dS) substitution rates to quantify the selection pressure acting on the MTases. The ω of MTases with predicted targets were calculated using Codeml[46] according to three different models (see Methods). We then selected the best fitted models using likelihood ratio tests (LRT) to determine whether ω were significantly different from one. As shown in Supplementary Table 6, the majority (11/20) of the MTases had a ω significantly smaller than one and the rest could not be statistically

distinguished from neutral evolution. This indicates that most giant viruses MTases are under purifying selection.

Discussion

Following the initial description of mimivirus[16], the last decade has seen an acceleration in the pace of discovery of giant viruses, now distributed in multiple different families[15,18,20–23,47], both thanks to the physical isolation of new specimens and to the rapid accumulation of metagenomics data[25]. Although the number of genomic sequences steadily increased during this period, the epigenomic status of giant viruses remained virtually unknown. Yet, the presence of numerous predicted DNA modification functions in their gene contents, as well as histone homologs in some of them[48], suggest that epigenetic may have a general impact on giant viruses's fitness, most likely through virus-virus and host-virus interactions. Here, we presented the first investigation of the DNA methylome of a large diversity of giant viruses using SMRT sequencing. Our analyses reveal that DNA methylation is widespread as it was detected in four of the 6 distinct giant viruses' families tested (cedratviruses, molliviruses, pandoraviruses and marseilleviruses). The recent advances in SMRT sequencing of metagenomes[36] will probably soon enable the survey of cultivation-independent giant viruses[25,49] and provide further evidence to test this hypothesis.

Our detailed investigation of the DNA MTase gene contents first confirmed the ubiquity of DNA methylation in giant viruses. We identified homologs of these enzymes in all analyzed viruses, with the exception of the *Mimiviridae* members' mobilome. Although widely present in giant viruses, the number of encoded MTases (and targeted sites) is unevenly distributed. It ranges from a single one in melbournevirus and moumouvirus australiensis, up to five in pandoravirus dulcis. Even within the same family, such as the pandoraviruses, the number of encoded DNA MTases is variable. Furthermore, the number of occurrences of each methylated sites per genome is highly variable, from 4 (CCTNAGG in pandoravirus quercus) to 850 (CTCGAG in mollivirus sibericum) (Fig. 1). The range of relative frequency of these sites is even larger from 10^{-6} for CCTNAGG in pandoravirus quercus up to 10^{-3} for the GATC motif in melbournevirus. Therefore, as already noticed in prokaryotes[34–36], DNA methylation is widespread but has a patchy distribution in giant viruses.

The non-uniformity of DNA methylation in giant viruses is partially explained by the loss-of-function of some encoded MTases. For instance, the mollivirus sibericum ml_498 MTase lacks a conserved (D/N/S)PP(Y/F) motif, involved in the formation of a hydrophobic pocket that binds the targeted nucleotide[50]. This was probably caused by a recent frameshift mutation in the 5′ region of the gene. Another case is the pithovirus sibericum pv_264 MTase also unable to methylate viral DNA, although the protein does not seem to be truncated. Here, uncharacterized mutations in a critical part of the protein might be at play and explain the loss of function. Alternatively, although less likely, the lack of methylation could be the result of transient methylation where methylation marks of viral DNA are eventually erased by an unknown N^6-methyl adenine demethylase. Finally, one cannot exclude that SMRT-seq is not sensitive enough to detect cryptic methylation of pithovirus DNA.

As expected from their enzymatic specificities, giant viruses' MTases are all of bacterial origins (Fig. 5). More surprisingly, we found that some of them were transferred from giant viruses (mostly pandoraviruses) to *Acanthamoeba* genomes. It remains to be determined whether this is an evolutionary dead end or if the transferred enzymes are still active in the host. Although host-to-virus gene exchanges are traditionally deemed more frequent than virus-to-host transfers[51], we previously noticed that this might not be true in pandoraviruses[30]. The picture is even more complex concerning MTases transferred between viruses, as illustrated by the orthologous MTases identified in cedratvirus kamchatka and mollivirus sibericum, two viruses from different families. If we previously noticed that some genes might be swapped between strains of pandoraviruses[38], the present case involves an exchange between viruses from totally different families only sharing the *Acanthamoeba* host. In the recently discovered mollivirus kamchatka[52], a MTase without ortholog in Mollivirus sibericum was probably acquired from a pandoravirus. In prokaryotes, the analysis of the co-occurrence of R-M systems and genetic fluxes between bacteria revealed that genetic exchanges are favored between genomes that share the same R-M systems, regardless of their evolutionary distance[53]. A similar phenomenon might be at play between molliviruses and pandoraviruses, and partially explains their shared gene content[52]. Previous analyses have also suggested that amoeba act as a genetic melting pot between intracellular bacteria[54], a concept that should now be extended to include amoeba-infecting viruses.

Our global survey of the DNA methylome of giant viruses revealed unexpected features. Besides the N^6-methyl-adenine modifications identified in the melbournevirus, cedratvirus kamchatka and mollivirus sibericum, we unveiled unexpected N^4-methyl-cytosines in the genomes of mollivirus sibericum and all tested pandoraviruses. To our knowledge, these are the first of such modifications reported for eukaryotic viruses. Some chloroviruses contain large amounts of N^6-methyl-adenines and 5-methyl-cytosines[55] also found in other eukaryotic viruses such as herpesviruses[12,56], and members of the *Iridoviridae*[12,57] and *Adenoviridae*[12,58] families. However, N^4-methyl-cytosine modifications were until now thought to be restricted to the prokaryotes and their viruses.

Another unexpected finding is the discovery of different types of modifications of the same CTCGAG site in giant viruses. If cedratvirus kamchatka and mollivirus sibericum exhibit CTCGAG motifs (with N^6-methyl-adenines), the pandoraviruses exhibit CTCGAG with N^4-methyl-cytosines (Fig. 1). The corresponding MTases belong to two distinct phylogenetic groups (green and orange groups in Fig. 5) and were acquired from distinct prokaryotes. Structural studies will be needed to elucidate how these MTases differently methylate the same DNA motif.

It was recently discovered that the AGCT tetramer is specifically eliminated from the pandoraviruses' genomes, in particular the ones belonging to the A-clade[59]. The evolutionary mechanism causing the elimination of this motif is still mysterious. One of the rejected hypothesis was that a R-M system targeting AGCT sites could be involved[59]. Our methylome data consistently confirm that the AGCT motif is not methylated in the pandoraviruses of neither clade A nor B (Supplementary Fig. 10).

Our digestion experiments revealed the presence of N^6-methyl-adenine-modified GATC sites in the genome of *A. castellanii* (Fig. 3). N^6-methyl-adenines were long thought to be restricted to prokaryotes, until several studies showed that some eukaryotes are also subject to these modifications[60,61]. In Chlamydomonas for instance, the N^6-methyl-adenines preferentially localize at the vicinity of transcription start sites, in the nucleosome free regions, to mark transcriptionally active genes[60]. Since the *A. castellanii* genome contains a mixture of methylated and unmethylated GATC sites we expect that similar biased modification patterns could be revealed by SMRT sequencing.

Most of the MTases (16 over 18) that had a testable (i.e., predicted) target site were found to be functional (Fig. 1). This either suggests that they were recently acquired, or that they were conserved because they increased the recipient viruses' fitness. Several evidences favor the second hypothesis. First, we found several of them conserved in entire clades (Fig. 2 and

Supplementary Fig. 9), indicating that they were retained throughout the family's radiation. Secondly, most of them are under purifying selection (Supplementary Table 6).

We observed that the complete R-M systems found in the *Marseilleviridae* members were phylogenetically related to that of the chloroviruses, in which it functions as a host DNA recycling mechanism[14]. There was thus a possibility that the marseilleviruses R-M system could play a similar role. However, our data clearly refute this hypothesis. Even though we found that the host DNA remains vulnerable at the GATC site targeted by the marseilleviruses REases (Fig. 3), the actual infection did not induce its degradation (Supplementary Fig. 3). This result is coherent with what we know about the replication cycle of these viruses[43,48]. Even if melbournevirus temporarily requires nucleus functions to initiate its replication cycle[43], most of it then proceeds in the cytoplasm, without contact with the host DNA[43]. This supports the non-involvement of marseilleviruses R-M systems, as well as other encoded REases, in the recycling of the host DNA.

By contrast, our results revealed that REases corresponding to the marseilleviruses R-M system could degrade the DNA of other marseilleviruses devoid of the same system. As previously proposed for chloroviruses[14,62] this suggests that the marseilleviruses R-M systems are involved in the exclusion of other viruses in cases of multiple infections. Indeed, *Acanthamoeba* can be infected by a wide variety of giant viruses[15,63]. In this context, a R-M system becomes an efficient way for a virus to fight against competitors and increase its fitness. In addition, *Acanthamoeba* feed on bacteria and are the reservoir of many intracellular bacteria, some of them remain as cytoplasmic endosymbionts[45,64]. The marseilleviruses R-M systems could thus be involved in recycling the DNA of these intracellular parasites.

In line with such putative role in pathogen exclusion, we found a congruent pattern of expression of the melbournevirus R-M system REase and MTase. The REase is first transcribed in the early phase of the infection cycle, where it could degrade the DNA of eventual co-infecting bacteria or viruses. The MTase is then transcribed 15 min later and the enzyme finally packaged in the virion, where it could protect the viral DNA from the REase, pending the next infection.

Besides *Marseilleviridae* family members, giant viruses for which we identified MTases and observed DNA methylation do not seem to encode cognate REases. Such so-called orphan MTases are common in bacteria where they regulate various biological processes, such as replication initiation, mismatch repair or gene expression[7–11]. Accordingly, the targeted genomic positions are not uniformly distributed, with hotspots and cold-spots of fully-, hemi- and unmethylated sites. Likewise, DNA modifications in phages have epigenetic roles beyond R-M systems. For instance the P1 phage Dmt-encoded MTase is involved in the control of DNA concatemers cleavage at the initiation of DNA packaging[65–67]. Again the methylated sites are clustered in the so-called pac regions. One could hypothesize a similar replication-related epigenetic role of orphan giant viruses MTases. However, giant viruses' genomes exhibit unimodal distributions of IPDr values and the corresponding motifs are globally uniformly distributed (Supplementary Fig. 11). In the case of bacterial orphan MTases involved in gene regulations the methylated sites tend to be located in the upstream non-coding regions of the regulated genes[34]. This is not true for giant viruses where methylated motifs are not enriched in intergenic regions (empirical two sided p-values > 0.1, see Methods) and thus not favoring such an epigenetic role.

How could we then interpret the presence of orphan MTases in giant viruses? Their role could be to protect the viral genome from its digestion by cognate REases from other *Acanthamoeba*-co-infecting bacteria or viruses harboring complete R-M systems. This would explain the tendency for some viruses, such as pandoravirus dulcis, to accumulate functional MTases in their genomes. In absence of the corresponding REases in the environment, the selection pressure would be relaxed on less solicited MTases, leading to their pseudogenization (Supplementary Table 6).

Giant viruses of the *Mimiviridae* family are involved in complex networks of interactions with the cellular host, the virophages, and the transpovirons[28,29,68]. It has been proposed that R-M systems could be used as an anti-virophage agent in these multipartite systems[69]. Accordingly, we investigated the role of DNA methylation in these cross talks. Our analysis of *Acanthamoeba*-infecting *Mimiviridae* members do not currently support this view, as DNA methylation does not seem to be a key player in this network (Fig. 1). However, future studies might reveal DNA methylation of the *Mimiviridae*-virophage system using different methylation detection techniques or by analyzing different strains/viruses of this family. Systems involving other hosts, such as the *Cafeteria roenbergensis*-croV-mavirus trio[19,27,68], might depend on DNA methylation to regulate their intricate interactions. DNA methylation is also a key factor in the switch between latent and integrated forms of some viruses[12]. In the *Cafeteria roenbergensis*-croV-mavirus context, one might wonder about the role DNA methylation could play in the maintenance of the host integrated mavirus provirophage, or in its awakening in the presence of croV infections.

R-M systems provide the carrying bacteria an immediate protection against the most lethal bacteriophages present it its environment[6]. Our work suggests that DNA methylation is equally important in giant viruses and involved in several types of interactions depending on the presence or absence of REase activity, and on the strictly cytoplasmic or nucleus dependency of their replication cycle. In chloroviruses, R-M systems offer a way to attack host DNA and exploit its nucleotide pool[14]. Our work on marseilleviruses now suggests that they act as an offensive weapon against competing pathogens. By contrast, the many orphan MTases found in giant viruses are potential self-defense weapons against other pathogens bearing active R-M systems with similar targets. Therefore, DNA methylation might allow giant viruses to face the fierce battles taking place in their amoebal hosts. In that context, it seems odd that the most frequent giant viruses in the environment that we studied, the *Mimiviridae* members, are apparently devoid of DNA methylation. However, one might remember that bacteria developed different lethal weapons to survive phage invasion: R-M and CRISPR systems. The many other REases found in the giant viruses' genomes could thus be involved in other competition/resistance processes such as the Mimivire system suggested to be directed against the parasitic virophage competing with some *Mimiviridae* members for its replication in *Acanthamoeba*[70]. Alongside toxin-antitoxin systems[69,71], Mimivire or other yet to be discovered CRISPR-like systems, DNA methylation might be part of the giant viruses' arsenal to cope with their numerous competitors.

Methods

Cedratvirus kamchatka characterization. Cedratvirus kamchatka was isolated from muddy grit soil collected near a lake at kizimen volcano, Kamchatka (Russian Federation N 55° 05′ 50 E 160° 20′ 58). The sample was resuspended in phosphate buffer saline containing ampicillin (100 µg mL^{-1}), chloramphrenicol (30 µg mL^{-1}) and kanamycin (25 µg mL^{-1}), and an aliquot was incubated with *A. castellanii Neff* (ATCC 30010^{TM}) cells (2000 cells per cm^2) adapted to 2.5 µg mL^{-1} of Amphotericin B (Fungizone), in protease-peptone-yeast-extract-glucose (PPYG) medium. Cultures exhibiting infectious phenotypes were recovered, centrifuged 5 min at 500 × g to eliminate the cells debris and the supernatant was centrifuged for 1 h at 16,000 × g at room temperature. T75 flasks were seeded with 60,000 cells per cm^2 and infected with the resuspended viral pellet. After a succession of passages, viral particles produced in sufficient quantity were recovered and purified. The viral

pellet was resuspended in PBS and loaded on a 1.2 to 1.5 density cesium gradient. After 16 h of centrifugation at 200,000 × *g*, the viral disk was washed three times in PBS and stored at 4 °C.

The genomic DNA of cedratvirus kamchatka was recovered from 2×10^{10} purified particles resuspended in 300 µL of water incubated with 500 µL of a buffer containing 100 mM Tris-HCl pH 8, 1.4 M NaCl, 20 mM Na2EDTA, 2% (w/v) CTAB (cetyltrimethylammonium bromide), 6 mM DTT and 1 mg mL^{-1} proteinase K at 65 °C for 90 min. After treatment with 0.5 mg mL^{-1} RNase A for 10 min, 500 µL of chloroform was added and the sample was centrifuged at 16,000 × *g* for 10 min at 4 °C. One volume of chloroform was added to the supernatant and centrifuged at 16,000 × *g* for 5 min at 4 °C. The aqueous phase was incubated with 2 volumes of precipitation buffer (5 g L^{-1} CTAB, 40 mM NaCl, pH 8) for 1 h at room temperature and centrifuged for 5 min at 16,000 × *g*. The pellet was then resuspended in 350 µL of 1.2 M NaCl and 350 µL of chloroform was added and centrifuged at 16,000 × *g* for 10 min at 4 °C. The aqueous phase was mixed to 0.6 volume of isopropanol, centrifuged for 10 min at room temperature and the pellet was washed with 500 µL 70% ethanol, centrifuged again and resuspended with nuclease-free water.

The DNA was sequenced using the PacBio SMRT technology, resulting in 868 Mb of sequence data (76,825 reads). SMRT reads were filtered using the SMRTanalysis package version 2.3.0. We then used Flye[72] version 2.4.2 to perform the de novo genome assembly with the pacbio_raw and $g = 450{,}000$ parameters. The assembly resulted in two distinct contigs (one of 466,767 nt and a second of 6049 nt). A pulsed field gel electrophoresis (PFGE) confirmed the genome size and its circular structure (Supplementary Fig. 4A). The smaller contig, not seen on the PFGE, potentially corresponds to an assembly artifact. Finally the assembly was subsequently polished using the Quiver tool from SMRTanalysis.

The cedratvirus kamchatka gene annotation was performed using GeneMarkS[73] with the virus option. Only genes predicted to encode proteins of at least 50 amino acids were kept for functional annotation. These proteins were aligned against the NR and Swissprot databases using BlastP[74] (with an *E*-value cutoff of 10^{-5}) and submitted to CD search[39], to InterProScan[75] with the Pfam, PANTHER, TIGRFAM, SMART, ProDom, ProSiteProfiles, ProSitePatterns and Hamapts databases, and to Phobius[76]. The genome was then manually curated according to these data.

SMRT resequencing. Pithovirus sibericum, mollivirus sibericum and megavirus vitis/zamilon vitis/megavirus vitis transpoviron genomic DNAs were extracted from 2×10^{10} purified particles using the PureLink Genomic DNA Extraction Mini Kit (Thermo Scientific) according to the manufacturer protocol. For pithovirus sibericum, we performed two successive purifications, and added 10 mM DTT in the lysis buffer for the first one.

The sequencing of pithovirus sibericum, mollivirus sibericum and megavirus vitis/zamilon vitis/megavirus vitis transpoviron performed using PacBio SMRT technology resulted in, respectively, 779 Mb (143,675 reads), 371 Mb (66,453 reads), and 997 Mb (66,464 reads) of sequence data. A second SMRT sequencing was performed on pithovirus sibericum and mollivirus sibericum viral DNA after WGA amplification using the Illustra GenomiPhi V2 DNA Amplification kit (GE Healthcare) according to the manufacturer instructions. This resulted in 632 Mb (78,685 reads) and 346 Mb (55,797 reads) of sequences for pithovirus sibericum and mollivirus sibericum WGA amplified DNA.

Identification of methylated motifs. Methylated motifs were identified using the modification and motif analysis module from the SMRTanalysis package using the datasets described in Supplementary Table 1 and the corresponding reference genome sequences. In addition we used the per-base resolution file of IPD ratios calculated by SMRTanalysis to compute the global IPDr of detected motifs and predicted MTases targets.

Annotation of MTases, REases, and target prediction. We analyzed the MTases and REases encoded in giant viruses genomes based on published annotations when available, as well as a combination of CD search and Interproscan protein domain search tools analyses[39,40]. In addition, we performed blast alignments against the REbase database[41] to support the annotations and to predict their targets when possible.

Phylogenies. All the phylogenies were calculated from the protein multiple alignments of homologous sequences computed using Expresso[77], Mafft[78], Mcoffee[79], and Clustal Omega[80]. We next selected the best multiple alignment using TrimAl[81].

For phylogenies based on single genes we identified homologs using BlastP against the NR database with an *E*-value cutoff of 10^{-10}. IQtree[82] was used to compute the tree using the best model as defined by ModelFinder[83]. Bootstrap values were computed using the UFBoot method[84].

Phylogenies of viral families (Fig. 2 and Supplementary Fig. 9) were based on the multiple alignments of strictly conserved single copy genes as defined by the OrthoFinder algorithm[85]. Next we used IQtree[82] with the -p option to compute the tree using the best model of each partitioned alignment. Each orthogroup (i.e., cluster of single copy orthologues) and the corresponding best models found by IQtree[83,86] are listed in Supplementary Data 1.

Selection pressure measurements. We performed codon-based multiple alignments of each subgroup of giant viruses MTases (Fig. 5) using protein multiple alignments (see Phylogenies Methods) and nucleotide sequences. The ω were computed for each gene of interest using Codeml[46] through the ETE framework[87]. The M0 model, which considers a unique ω for the whole tree, was first computed. Genes that were too divergent, i.e., where multiple substitutions might have occurred ($dS > 1.5$), were excluded. We then calculated the ω of the remaining genes using the B_free model, which assigns two distinct ω (one for the branch of interest and one for the rest of the tree), and the B_neut model with a fixed $\omega = 1$ for the gene of interest. LTR tests with a *p*-value cutoff of 0.05 were then performed to select the best model and to decide whether ω were significantly different from one.

DNA restriction experiments. Mollivirus sibericum, melbournevirus and noumeavirus genomic DNA were extracted using the PureLink Genomic DNA Extraction Mini Kit (Thermo Scientific) according to the manufacturer protocol. *A. castellanii* genomic DNA was extracted using the Wizard® Genomic DNA Purification Kit (Promega). The DNA were digested with 10 units of the appropriate restriction enzymes (New England Biolabs) for 1 h at 37 °C and loaded on a 1% agarose gel.

Mollivirus sibericum-infected *A. castellanii* cells DNA extraction. *A. castellanii* overexpressing the ml_216-GFP fusion and wild-type cells were grown in T25 flasks. They were infected with mollivirus sibericum at MOI 100. After 1 h of infection at 32 °C, cells were washed three times with PPYG to eliminate the excess of viruses. For each infection time (1 h to 6 h), a T25 flask was recovered and cells were centrifuged for 5 min at 1000 × *g*. DNA was extracted using the Wizard® Genomic DNA Purification Kit (Promega) according to the manufacturer's protocol and loaded on a 0.8% agarose gel.

Melbournevirus R-M system MTase and REase expression timing. *A. castellanii* cells were grown in T25 flasks and infected with melbournevirus at MOI 50. After 15 min of infection at 32 °C, cells were washed 3 times with PPYG to eliminate the excess of viruses. For each infection time (15 min to 5 h), a T25 flask was recovered and cells were centrifuged for 5 min at 1000 × *g*. RNA was extracted using the RNeasy Mini kit (QIAGEN) according to the manufacturer's protocol. Briefly, cells were resuspended in the provided buffer and disrupted by −80 °C freezing and thawing, and shaken vigorously. Total RNA was eluted with 50 µL of RNase free water. Total RNA was quantified on the nanodrop spectrophotometer (Thermo Scientific). Poly(A) enrichment was performed (Life Technologies, Dynabeads oligodT_{25}) and first-strand complementary DNA (cDNA) poly(A) synthesis was performed with the SmartScribe Reverse Transcriptase (Clontech Laboratories) using an oligo(d$T)_{24}$ primer and then treated with RNase H (New England Biolabs). For each time point, PCR reactions were performed using mel_015 and mel_016 genes specific primers and one unit of Phusion DNA Polymerase (Thermo Scientific) in a 50 µL final volume.

Motif enrichment in intergenic regions. Analysis of motif enrichment in intergenic regions was done by computing the number of motifs identified in these regions and in coding regions. We next shuffled coding coordinates 1000 times and calculated the same values. *Z*-scores were calculated from these randomizations and transformed in empirical *p*-values. We excluded the CCTNAGG motifs from this calculation as there were not enough occurrences of this motif (Fig. 1) to compute accurate *p*-values.

Reporting summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

Data supporting the findings of this work are available within the paper and its Supplementary Information files. The raw SMRT sequence datasets generated and analyzed in the current study were deposited in the Sequence Read Archive database under the following accession PRJNA612691. In addition individual datasets accessions are all reported in the Supplementary Table 1. The assembled cedratvirus kamchatka genome has been deposited to the GenBank database under the following accession MN873693. MTases annotations were performed using the REbase database [http://rebase.neb.com/rebase/rebase.html] and the cedratvirus kamchatka gene annotations using the Blast NCBI NR (GenBank CDS translations+PDB + SwissProt+PIR + PRF) database and the Uniprot-Swissprot [https://www.uniprot.org/uniprot/] database. The source data underlying Fig. 1, Supplementary Figs. 1 and 10, Supplementary Tables 3–5 are provided as source data file.

Received: 15 January 2020; Accepted: 3 May 2020;
Published online: 27 May 2020

References

1. Jones, P. A. Functions of DNA methylation: islands, start sites, gene bodies and beyond. *Nat. Rev. Genet.* **13**, 484–492 (2012).
2. Arand, J. et al. In vivo control of CpG and non-CpG DNA methylation by DNA methyltransferases. *PLoS Genet.* **8**, e1002750 (2012).
3. Li, E., Beard, C. & Jaenisch, R. Role for DNA methylation in genomic imprinting. *Nature* **366**, 362–365 (1993).
4. Greenberg, M. V. C. & Bourc'his, D. The diverse roles of DNA methylation in mammalian development and disease. *Nat. Rev. Mol. Cell Biol.* **20**, 590–607 (2019).
5. Loenen, W. A. M., Dryden, D. T. F., Raleigh, E. A., Wilson, G. G. & Murray, N. E. Highlights of the DNA cutters: a short history of the restriction enzymes. *Nucleic Acids Res.* **42**, 3–19 (2014).
6. Murray, N. E. Immigration control of DNA in bacteria: self versus non-self. *Microbiology* **148**, 3–20 (2002).
7. Adhikari, S. & Curtis, P. D. DNA methyltransferases and epigenetic regulation in bacteria. *FEMS Microbiol. Rev.* **40**, 575–591 (2016).
8. Casselli, T. et al. DNA Methylation by Restriction modification systems affects the global transcriptome profile in Borrelia burgdorferi. *J. Bacteriol.* **200**, pii: e00395-1 (2018).
9. Messer, W., Bellekes, U. & Lother, H. Effect of dam methylation on the activity of the E. coli replication origin, oriC. *EMBO J.* **4**, 1327–1332 (1985).
10. Putnam, C. D. Evolution of the methyl directed mismatch repair system in Escherichia coli. *DNA Repair* **38**, 32–41 (2016).
11. Domian, I. J., Reisenauer, A. & Shapiro, L. Feedback control of a master bacterial cell-cycle regulator. *Proc. Natl Acad. Sci. U.S.A.* **96**, 6648–6653 (1999).
12. Hoelzer, K., Shackelton, L. A. & Parrish, C. R. Presence and role of cytosine methylation in DNA viruses of animals. *Nucleic Acids Res.* **36**, 2825–2837 (2008).
13. Wilson, W. H., Van Etten, J. L. & Allen, M. J. The Phycodnaviridae: the story of how tiny giants rule the world. *Curr. Top. Microbiol. Immunol.* **328**, 1–42 (2009).
14. Agarkova, I. V., Dunigan, D. D. & Van Etten, J. L. Virion-associated restriction endonucleases of chloroviruses. *J. Virol.* **80**, 8114–8123 (2006).
15. 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).
16. Raoult, D. et al. The 1.2-megabase genome sequence of Mimivirus. *Science* **306**, 1344–1350 (2004).
17. 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).
18. 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).
19. Fischer, M. G., Allen, M. J., Wilson, W. H. & Suttle, C. A. Giant virus with a remarkable complement of genes infects marine zooplankton. *Proc. Natl Acad. Sci. U.S.A.* **107**, 19508–19513 (2010).
20. Legendre, M. et al. 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 (2014).
21. Andreani, J. et al. Cedratvirus, a double-cork structured giant virus, is a distant relative of pithoviruses. *Viruses* **8**, 300 (2016).
22. Philippe, N. et al. Pandoraviruses: amoeba viruses with genomes up to 2.5 Mb reaching that of parasitic eukaryotes. *Science* **341**, 281–286 (2013).
23. 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).
24. Bäckström, D. et al. Virus genomes from deep sea sediments expand the ocean megavirome and support independent origins of viral gigantism. *mBio* **10**, e02497–18 (2019).
25. Schulz, F. et al. Giant virus diversity and host interactions through global metagenomics. *Nature* **578**, 432–436 (2020).
26. La Scola, B. et al. The virophage as a unique parasite of the giant mimivirus. *Nature* **455**, 100–104 (2008).
27. Fischer, M. G. & Suttle, C. A. A virophage at the origin of large DNA transposons. *Science* **332**, 231–234 (2011).
28. Jeudy, S. et al. Exploration of the propagation of transpovirons within Mimiviridae reveals a unique example of commensalism in the viral world. *ISME J.* 1–13 (2019) https://doi.org/10.1038/s41396-019-0565-y.
29. Desnues, C. et al. Provirophages and transpovirons as the diverse mobilome of giant viruses. *Proc. Natl Acad. Sci. U.S.A.* **109**, 18078–18083 (2012).
30. Legendre, M. et al. Diversity and evolution of the emerging Pandoraviridae family. *Nat. Commun.* **9**, 2285 (2018).
31. Bertelli, C. et al. Cedratvirus lausannensis-digging into Pithoviridae diversity. *Environ. Microbiol.* **19**, 4022–4034 (2017).
32. Barros-Silva, D., Marques, C. J., Henrique, R. & Jerónimo, C. Profiling DNA methylation based on next-generation sequencing approaches: new insights and clinical applications. *Genes* **9**, pii: E429 (2018).
33. Flusberg, B. A. et al. Direct detection of DNA methylation during single-molecule, real-time sequencing. *Nat. Methods* **7**, 461–465 (2010).
34. Blow, M. J. et al. The epigenomic landscape of prokaryotes. *PLoS Genet.* **12**, e1005854 (2016).
35. Fullmer, M. S., Ouellette, M., Louyakis, A. S., Papke, R. T. & Gogarten, J. P. The patchy distribution of restriction–modification system genes and the conservation of orphan methyltransferases in halobacteria. *Genes* **10**, 233 (2019).
36. Hiraoka, S. et al. Metaepigenomic analysis reveals the unexplored diversity of DNA methylation in an environmental prokaryotic community. *Nat. Commun.* **10**, 1–10 (2019).
37. Doutre, G., Philippe, N., Abergel, C. & Claverie, J. -M. Genome analysis of the first Marseilleviridae representative from Australia indicates that most of its genes contribute to virus fitness. *J. Virol.* **88**, 14340–14349 (2014).
38. Legendre, M. et al. Pandoravirus celtis illustrates the microevolution processes at work in the giant *Pandoraviridae* genomes. *Front. Microbiol.* **10**, 430 (2019).
39. Marchler-Bauer, A. et al. CDD/SPARCLE: functional classification of proteins via subfamily domain architectures. *Nucleic Acids Res.* **45**, D200–D203 (2017).
40. Mitchell, A. L. et al. InterPro in 2019: improving coverage, classification and access to protein sequence annotations. *Nucleic Acids Res.* **47**, D351–D360 (2019).
41. Roberts, R. J., Vincze, T., Posfai, J. & Macelis, D. REBASE-a database for DNA restriction and modification: enzymes, genes and genomes. *Nucleic Acids Res.* **43**, D298–299 (2015).
42. Claverie, J. -M. & Abergel, C. Mimiviridae: an expanding family of highly diverse large dsDNA viruses infecting a wide phylogenetic range of aquatic eukaryotes. *Viruses* **10**, 506 (2018).
43. Fabre, E. et al. Noumeavirus replication relies on a transient remote control of the host nucleus. *Nat. Commun.* **8**, 15087 (2017).
44. Rodrigues, R. A. L. et al. Morphologic and genomic analyses of new isolates reveal a second lineage of Cedratviruses. *J. Virol.* **92**, e00372–18 (2018).
45. Greub, G. & Raoult, D. Microorganisms resistant to free-living amoebae. *Clin. Microbiol. Rev.* **17**, 413–433 (2004).
46. Yang, Z. PAML 4: phylogenetic analysis by maximum likelihood. *Mol. Biol. Evol.* **24**, 1586–1591 (2007).
47. Yoshikawa, G. et al. Medusavirus, a novel large DNA virus discovered from hot spring water. *J. Virol.* **93**, e02130–18 (2019).
48. Thomas, V. et al. Lausannevirus, a giant amoebal virus encoding histone doublets. *Environ. Microbiol.* **13**, 1454–1466 (2011).
49. Needham, D. M. et al. A distinct lineage of giant viruses brings a rhodopsin photosystem to unicellular marine predators. *Proc. Natl Acad. Sci. U.S.A* **116**, 20574–20583 (2019).
50. Jeltsch, A. Beyond Watson and Crick: DNA methylation and molecular enzymology of DNA methyltransferases. *Chembiochem. Eur. J. Chem. Biol.* **3**, 274–293 (2002).
51. Moreira, D. & Brochier-Armanet, C. Giant viruses, giant chimeras: the multiple evolutionary histories of Mimivirus genes. *BMC Evol. Biol.* **8**, 12 (2008).
52. Christo-Foroux, E. et al. Characterization of Mollivirus kamchatka, the first modern representative of the proposed molliviridae family of giant viruses. *J. Virol.* **94**, e01997–19 (2020).
53. Oliveira, P. H., Touchon, M. & Rocha, E. P. C. Regulation of genetic flux between bacteria by restriction-modification systems. *Proc. Natl Acad. Sci. U. S.A.* **113**, 5658–5663 (2016).
54. Gimenez, G. et al. Insight into cross-talk between intra-amoebal pathogens. *BMC Genomics* **12**, 542 (2011).
55. Nelson, M., Burbank, D. E. & Van Etten, J. L. Chlorella viruses encode multiple DNA methyltransferases. *Biol. Chem.* **379**, 423–428 (1998).
56. Davison, A. J., Cunningham, C., Sauerbier, W. & McKinnell, R. G. Genome sequences of two frog herpesviruses. *J. Gen. Virol.* **87**, 3509–3514 (2006).
57. Willis, D. B. & Granoff, A. Frog virus 3 DNA is heavily methylated at CpG sequences. *Virology* **107**, 250–257 (1980).
58. Gunthert, U., Schweiger, M., Stupp, M. & Doerfler, W. DNA methylation in adenovirus, adenovirus-transformed cells, and host cells. *Proc. Natl Acad. Sci. U.S.A.* **73**, 3923–3927 (1976).
59. Poirot, O., Jeudy, S., Abergel, C. & Claverie, J.- M. A puzzling anomaly in the 4-Mer composition of the giant pandoravirus genomes reveals a stringent new evolutionary selection process. *J. Virol.* **93**, e01206–19 (2019).
60. Fu, Y. et al. N6-methyldeoxyadenosine marks active transcription start sites in Chlamydomonas. *Cell* **161**, 879–892 (2015).
61. Luo, G. -Z., Blanco, M. A., Greer, E. L., He, C. & Shi, Y. DNA N(6)-methyladenine: a new epigenetic mark in eukaryotes? *Nat. Rev. Mol. Cell Biol.* **16**, 705–710 (2015).

62. Chase, T. E., Nelson, J. A., Burbank, D. E. & Van Etten, J. L. Mutual exclusion occurs in a Chlorella-like green alga inoculated with two viruses. *J. Gen. Virol.* **70**(Pt 7), 1829–1836 (1989).
63. Rolland, C. et al. Discovery and further studies on giant viruses at the IHU Mediterranee infection that modified the perception of the virosphere. *Viruses* **11**, pii: E312 (2019).
64. Marciano-Cabral, F. Introductory remarks: bacterial endosymbionts or pathogens of free-living amebae1. *J. Eukaryot. Microbiol.* **51**, 497–501 (2004).
65. Sternberg, N. & Coulby, J. Cleavage of the bacteriophage P1 packaging site (pac) is regulated by adenine methylation. *Proc. Natl Acad. Sci. U.S.A.* **87**, 8070–8074 (1990).
66. Iyer, L. M., Zhang, D., Burroughs, A. M. & Aravind, L. Computational identification of novel biochemical systems involved in oxidation, glycosylation and other complex modifications of bases in DNA. *Nucleic Acids Res.* **41**, 7635–7655 (2013).
67. Murphy, J., Mahony, J., Ainsworth, S., Nauta, A. & van Sinderen, D. Bacteriophage orphan DNA methyltransferases: insights from their bacterial origin, function, and occurrence. *Appl. Environ. Microbiol.* **79**, 7547–7555 (2013).
68. Fischer, M. G. & Hackl, T. Host genome integration and giant virus-induced reactivation of the virophage mavirus. *Nature* **540**, 288–291 (2016).
69. Filée, J. Giant viruses and their mobile genetic elements: the molecular symbiosis hypothesis. *Curr. Opin. Virol.* **33**, 81–88 (2018).
70. Levasseur, A. et al. MIMIVIRE is a defence system in mimivirus that confers resistance to virophage. *Nature* **531**, 249–252 (2016).
71. 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**, pii: e33014 (2018).
72. Kolmogorov, M., Yuan, J., Lin, Y. & Pevzner, P. A. Assembly of long, error-prone reads using repeat graphs. *Nat. Biotechnol.* **37**, 540–546 (2019).
73. Besemer, J., Lomsadze, A. & Borodovsky, M. GeneMarkS: a self-training method for prediction of gene starts in microbial genomes. Implications for finding sequence motifs in regulatory regions. *Nucleic Acids Res.* **29**, 2607–2618 (2001).
74. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *J. Mol. Biol.* **215**, 403–410 (1990).
75. Jones, P. et al. InterProScan 5: genome-scale protein function classification. *Bioinformatics* **30**, 1236–1240 (2014).
76. Käll, L., Krogh, A. & Sonnhammer, E. L. L. A combined transmembrane topology and signal peptide prediction method. *J. Mol. Biol.* **338**, 1027–1036 (2004).
77. Armougom, F. et al. Expresso: automatic incorporation of structural information in multiple sequence alignments using 3D-Coffee. *Nucleic Acids Res.* **34**, W604–608 (2006).
78. Katoh, K. & Standley, D. M. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol. Biol. Evol.* **30**, 772–780 (2013).
79. Wallace, I. M., O'Sullivan, O., Higgins, D. G. & Notredame, C. M-Coffee: combining multiple sequence alignment methods with T-Coffee. *Nucleic Acids Res.* **34**, 1692–1699 (2006).
80. Sievers, F. & Higgins, D. G. Clustal Omega, accurate alignment of very large numbers of sequences. *Methods Mol. Biol. Clifton NJ* **1079**, 105–116 (2014).
81. Capella-Gutiérrez, S., Silla-Martínez, J. M. & Gabaldón, T. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinforma. Oxf. Engl.* **25**, 1972–1973 (2009).
82. Nguyen, L. -T., Schmidt, H. A., von Haeseler, A. & Minh, B. Q. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol. Biol. Evol.* **32**, 268–274 (2015).
83. Kalyaanamoorthy, S., Minh, B. Q., Wong, T. K. F., von Haeseler, A. & Jermiin, L. S. ModelFinder: fast model selection for accurate phylogenetic estimates. *Nat. Methods* **14**, 587–589 (2017).
84. Minh, B. Q., Nguyen, M. A. T. & von Haeseler, A. Ultrafast approximation for phylogenetic bootstrap. *Mol. Biol. Evol.* **30**, 1188–1195 (2013).
85. Emms, D. M. & Kelly, S. OrthoFinder: phylogenetic orthology inference for comparative genomics. *Genome Biol.* **20**, 238 (2019).
86. Chernomor, O., von Haeseler, A. & Minh, B. Q. Terrace aware data structure for phylogenomic inference from supermatrices. *Syst. Biol.* **65**, 997–1008 (2016).
87. Huerta-Cepas, J., Serra, F. & Bork, P. ETE 3: Reconstruction, analysis, and visualization of phylogenomic data. *Mol. Biol. Evol.* **33**, 1635–1638 (2016).
88. Aherfi, S. et al. Complete genome sequence of Cannes 8 virus, a new member of the proposed family 'Marseilleviridae'. *Virus Genes* **47**, 550–555 (2013).
89. Boyer, M. et al. Giant Marseillevirus highlights the role of amoebae as a melting pot in emergence of chimeric microorganisms. *Proc. Natl Acad. Sci. U. S.A.* **106**, 21848–21853 (2009).
90. Takemura, M. Draft genome sequence of tokyovirus, a member of the family marseilleviridae isolated from the Arakawa river of Tokyo, Japan. *Genome Announc.* **4**, e00429–16 (2016).
91. Chatterjee, A. & Kondabagil, K. Complete genome sequence of Kurlavirus, a novel member of the family Marseilleviridae isolated in Mumbai, India. *Arch. Virol.* **162**, 3243–3245 (2017).
92. Doutre, G. et al. Complete genome sequence of a new member of the marseilleviridae recovered from the Brackish Submarine Spring in the Cassis Port-Miou Calanque, France. *Genome Announc.* **3**, pii: e01148–15 (2015).
93. Dornas, F. P. et al. A Brazilian Marseillevirus Is the founding member of a lineage in Family Marseilleviridae. *Viruses* **8**, 76 (2016).
94. Boughalmi, M. et al. First isolation of a Marseillevirus in the Diptera Syrphidae Eristalis tenax. *Intervirology* **56**, 386–394 (2013).
95. Aherfi, S. et al. Complete genome sequence of Tunisvirus, a new member of the proposed family Marseilleviridae. *Arch. Virol.* **159**, 2349–2358 (2014).
96. Dos Santos, R. N. et al. A new marseillevirus isolated in Southern Brazil from Limnoperna fortunei. *Sci. Rep.* **6**, 35237 (2016).

Acknowledgements

We are deeply indebted to our volunteer collaborator Alexander Morawitz for collecting the Kamchatka soil samples. We thank the PACA Bioinfo platform for computing support. This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation program (grant agreement No 832601), from the FRM prize Lucien Tartois and from CNRS (PRC1484-2018) to C. Abergel. The funding bodies had no role in the design of the study, analysis, and interpretation of data and in writing the manuscript.

Author contributions

S.J., C.A. and M.L. designed research. S.J. performed most of the experiments. S.R. assembled, annotated, and analyzed the cedratvirus kamchatka genome. J.M.A. isolated cedratvirus kamchatka and provided research assistance. S.J., C.A. and J.M.C. contributed to manuscript writing. M.L. directed the research, carried most of the data analysis, and wrote the paper.

Competing interests

The authors declare no competing interests.

Additional information

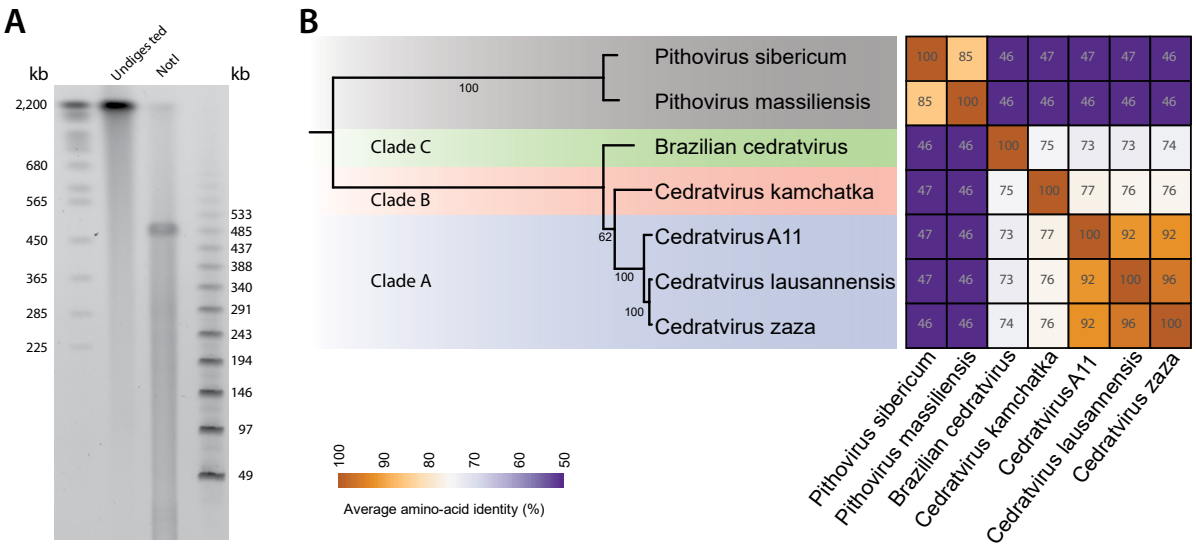
Supplementary information is available for this paper at https://doi.org/10.1038/s41467-020-16414-2.

Correspondence and requests for materials should be addressed to M.L.

Peer review information *Nature Communications* thanks Frank Aylward and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Peer reviewer reports are available.

Reprints and permission information is available at http://www.nature.com/reprints

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



Supplementary Figure 4. Characterization of the cedratviruse kamchatka genome

A) Pulse Field Gel Electrophoresis of the cedratvirus kamchatka genome. The first and last lanes correspond to DNA ladders. Undigested and NotI digested DNA are shown. As expected from its circular structure the undigested DNA migrates higher than the assembled genomic sequence length. The NotI digested DNA, predicted to be cleaved at a single position migrates at the expected position. This experiment was repeated twice with similar results. **B**) Phylogenetic tree of the completely sequenced genomes of cedratviruses and pithoviruses based on the protein sequences alignments of 134 strictly conserved single copy orthologues. The tree was calculated using the best model of each partitioned alignment as determined by IQtree [2]. Bootstrap values (n = 5,000) were computed using the UFBoot [1] method from IQtree [2]. The average amino acid identity matrix was computed using CompareM [5] on the shared ORFs.

II. Observations de cellules infectées par Cedratvirus kamchatka au microscope électronique

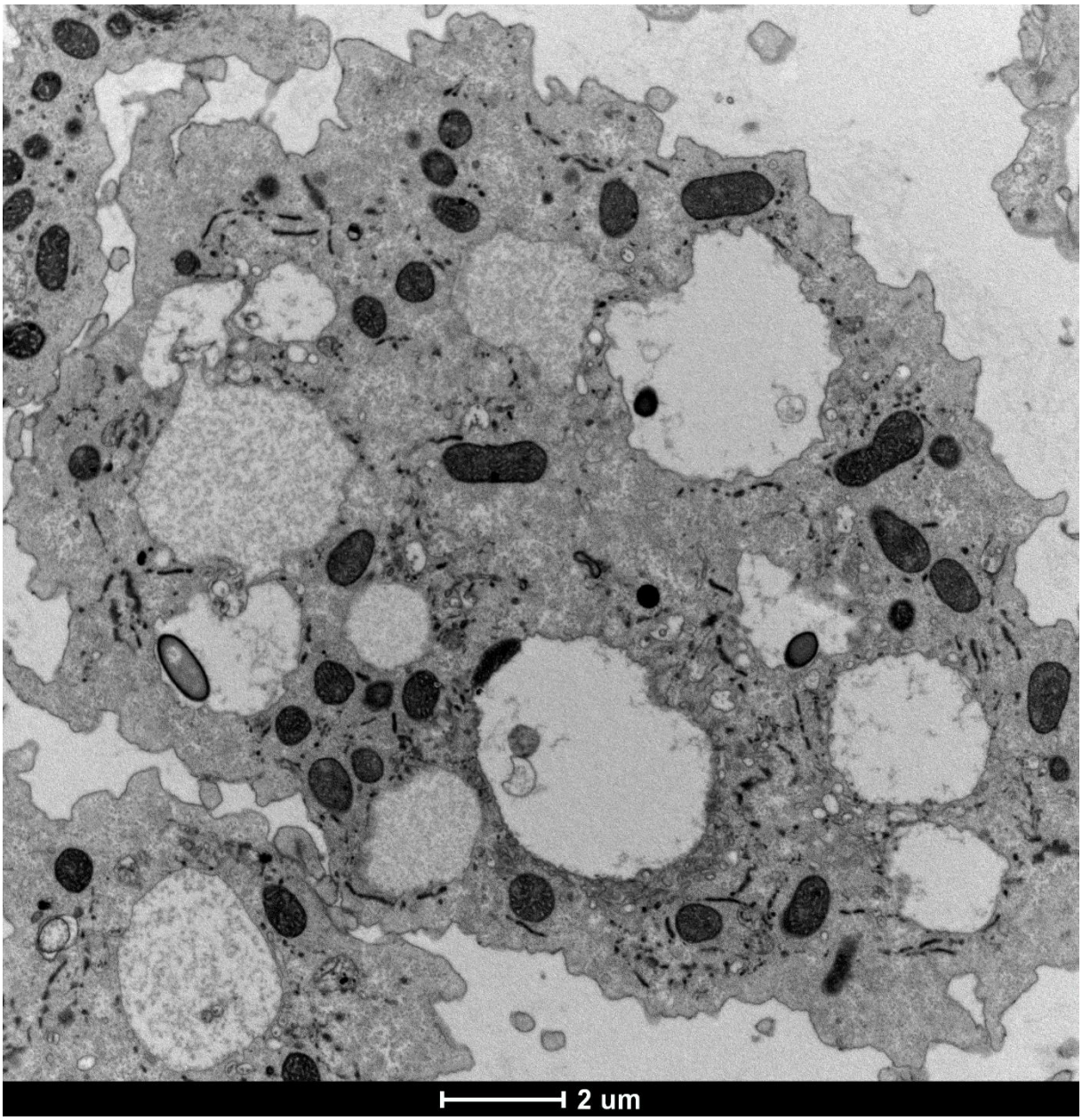


Figure 35 – Observation de cellules de *A. castellanii* au contact de particules virales MOCK de Cedratvirus kamchatka
L'image provient d'un microscope électronique à transmission. Une particule virale désactivée est visible dans une vacuole. La cellule a été fixée 3 heures après infection.

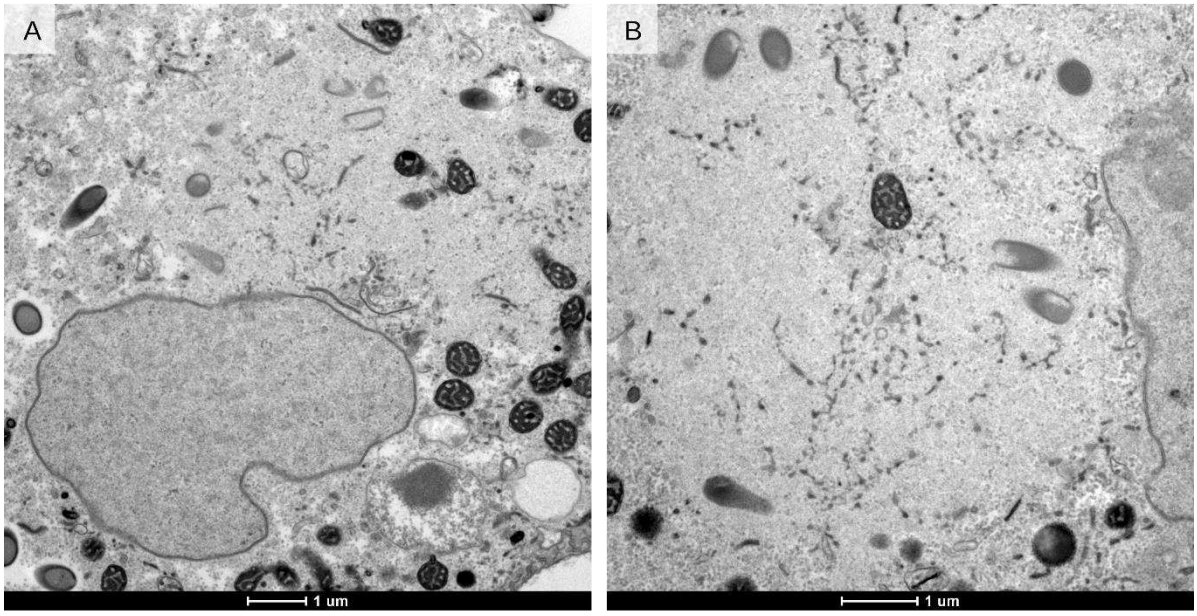


Figure 36 – Usines à virions de C. kamchatka vues avec une ou deux mitochondries en son sein
Les images proviennent d'une infection de *A. castellanii*, 4 heures après, observée au microscope électronique à transmission.

III. Phylogénies de gènes particuliers trouvés dans les séquences virales de métagénomique du pergélisol

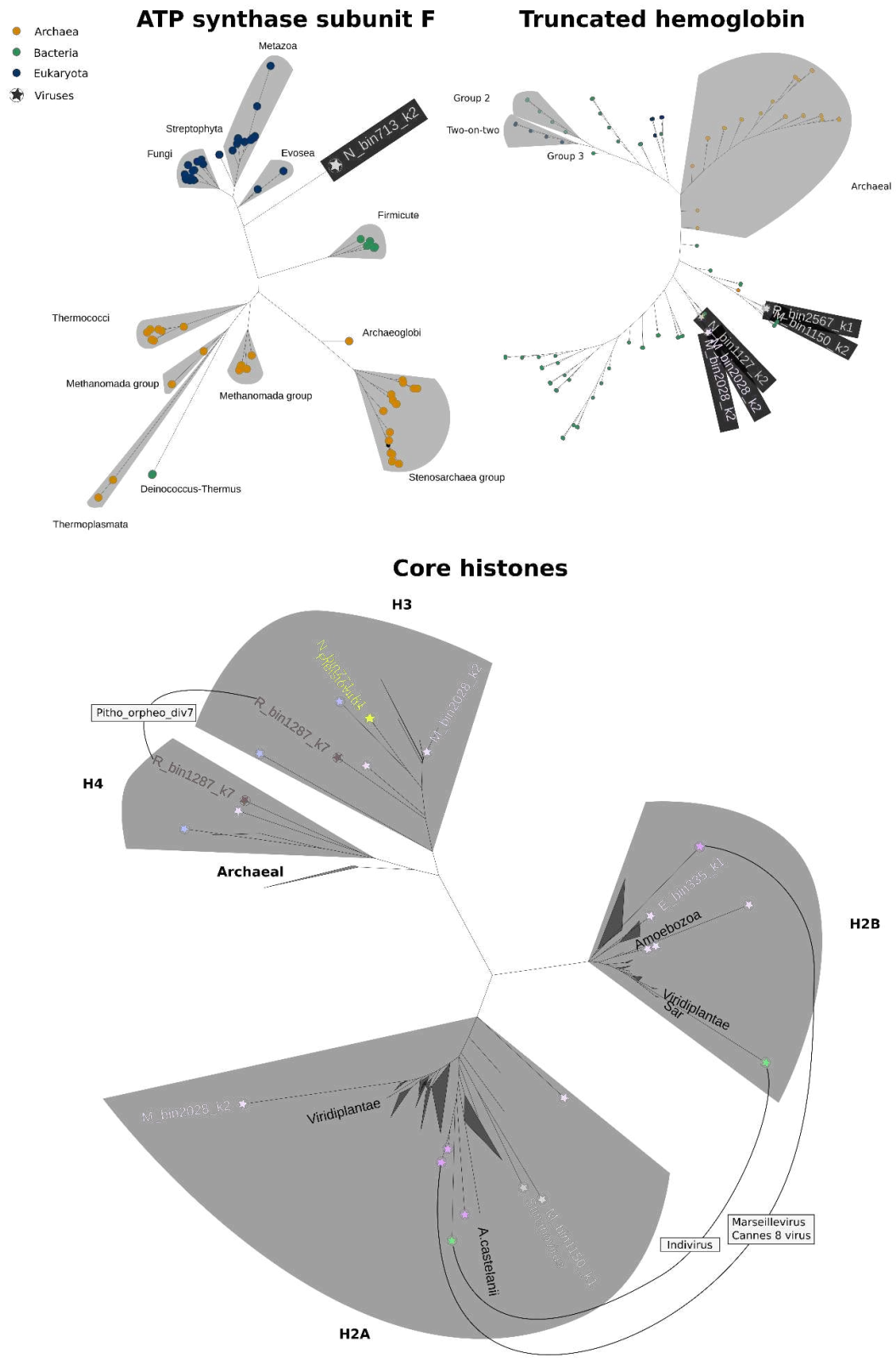


Figure 37 – Phylogénie de gènes nouveaux chez les *Pithoviridae* et *Orpheoviridae*
Une phylogénie de trois fonctions inattendues retrouvées dans des séquences métagénomiques a été retracée en intégrant des gènes de la base de données InterPro et HystoneDB et les résultats de BLASTP des protéines d'intérêt. Les gènes viraux sont indiqués par une étoile. Seul les gènes d'Indivirus et de *Marseilleviridae* dans la phylogénie des Histones cœur proviennent d'un virus de référence.

viruses

MDPI

Article

Comparative Analysis of the Circular and Highly Asymmetrical *Marseilleviridae* Genomes

Léo Blanca, Eugène Christo-Foroux, Sofia Rigou and Matthieu Legendre *

CNRS, IGS, Information Génomique & Structurale (UMR7256), Institut de Microbiologie de la Méditerranée (FR 3489), Aix Marseille Univ., 13288 Marseille, France; leoblanca18@gmail.com (L.B.); eugene.christo-foroux@igs.cnrs-mrs.fr (E.C.-F.); rigou@igs.cnrs-mrs.fr (S.R.)

* Correspondence: legendre@igs.cnrs-mrs.fr

Received: 7 October 2020; Accepted: 5 November 2020; Published: 7 November 2020

Abstract: *Marseilleviridae* members are large dsDNA viruses with icosahedral particles 250 nm in diameter infecting *Acanthamoeba*. Their 340 to 390 kb genomes encode 450 to 550 protein-coding genes. Since the discovery of marseillevirus (the prototype of the family) in 2009, several strains were isolated from various locations, among which 13 are now fully sequenced. This allows the organization of their genomes to be deciphered through comparative genomics. Here, we first experimentally demonstrate that the *Marseilleviridae* genomes are circular. We then acknowledge a strong bias in sequence conservation, revealing two distinct genomic regions. One gathers most *Marseilleviridae* paralogs and has undergone genomic rearrangements, while the other, enriched in core genes, exhibits the opposite pattern. Most of the genes whose protein products compose the viral particles are located in the conserved region. They are also strongly biased toward a late gene expression pattern. We finally discuss the potential advantages of *Marseilleviridae* having a circular genome, and the possible link between the biased distribution of their genes and the transcription as well as DNA replication mechanisms that remain to be characterized.

Keywords: comparative genomics; large DNA viruses; marseillevirus; genome evolution

1. Introduction

Marseilleviridae is an expanding family of large double-stranded DNA viruses infecting free-living amoeba of the *Acanthamoeba* genus. Their icosahedral capsids of 250 nm diameters enclose a 340 to 390 kb genome predicted to encode an average of 500 protein-coding genes [1–12]. Among these genes, some code for unexpected functions for a virus, the most surprising being homologues to cellular histones [1,2]. Viruses from this family belong to the NCLDVs (for nucleocytoplasmic large DNA viruses), i.e., the *Nucleocytoviricota* phylum, according to the latest International Committee on Taxonomy of Viruses (ICTV) classification [13,14]. Marseilleviruses' replication cycles start with their phagocytosis by the *Acanthamoeba* host. Once in the cytoplasm, they form the so-called "viral factory" in the vicinity of the nucleus where virion assembly and DNA packaging occur simultaneously [1]. Mature particles are then released through cell lysis roughly 8 h post-infection (pi) [1]. However, the duration of the replication cycle is variable among *Marseilleviridae* with strains for which virions are released at 13–16 h up to 24 h pi. Since marseilleviruses encode a complete transcription apparatus and the host nucleus appears to remain intact during the entire cycle, it was initially assumed that marseilleviruses were *bona fide* cytoplasmic viruses, without a nuclear phase. However, it was subsequently shown that virally encoded RNA polymerase subunit proteins are not packaged within the virions, thus precluding the transcription of viral genes to start [10]. As a workaround, nuclear proteins are actively, albeit transiently, recruited by the viral factory to initiate the transcription of viral genes,

thus placing marseilleviruses between viruses strictly replicating within the cytoplasm and those involving an intranuclear phase [10].

Marseillevirus T19 was the first *Marseilleviridae* to be isolated by co-culturing with *Acanthamoeba castellanii* [1]. Since then, several strains were isolated using the same approach, mainly from aquatic samples of different continents (Asia [9,12,15], Africa [4,5], South America [6,11], Europe [1–3,8] and Australia [7,10]). In addition, marseillevirus-like genomic sequences were identified in environmental metagenomics assembled data [16]. Among the isolated strains, thirteen were fully sequenced (Table S1), and their phylogeny shows that they belong to five distinct clades [10,15] (Figure S1). From the analysis of the genes encoded in these genomes, it was estimated that roughly 25% of them are of potential cellular origin, making horizontal gene transfers (HGT) a contributing factor shaping the *Marseilleviridae* genomes [1]. Surprisingly, only 23% of these exchanges involve the *Amebozoa* host, as opposed to 45% for bacteria and bacteriophages [1]. Even more remarkably, this large fraction of bacteria-related genes are subjected to strong purifying selection, and thus probably contribute to viral fitness [7]. One striking example is the *Marseilleviridae*-encoded restriction–modification (RM) system that involves restriction endonucleases and DNA methyltransferases of bacterial origin [17]. It is suspected that it serves as a weapon against amoeba intracellular parasites, thus giving to the virus a selective advantage.

Besides evolutionary questions, marseilleviruses' physiology has been examined through several genome-wide surveys using various omics data. First, proteomic data of the viral particles of three *Marseilleviridae* members were produced, namely marseillevirus [1], noumeavirus and melbournevirus [10]. This not only revealed the proteins that build the structure of the *Marseilleviridae* virions, but also those packaged within it that could be essential for initiating the viral replication. In addition, the marseillevirus' transcriptional activity during an infection cycle in *A. castellanii* was recently surveyed by RNA sequencing (RNA-seq), showing that the host translation apparatus is downregulated during the infection [18]. This now provides us with a sufficient body of data to conduct an in-depth comparative genomics study of the *Marseilleviridae* family.

In this study, we first experimentally confirm the circular structure of the marseilleviruses genomes. Using available genomic, proteomic and transcriptomic data, we then reveal a strong bias in the distribution of the marseilleviruses' genes. We examine the genomic rearrangements as well as the genomic distribution of several gene categories along the genomes. More specifically, we unveil the uneven distribution of the core genes (i.e., genes conserved in all *Marseilleviridae*), the virion-associated genes and the paralogous genes (i.e., genes that were duplicated during the *Marseilleviridae* evolution). This work helps us to better understand the global organization of the *Marseilleviridae* genomes, as well as the evolution and physiology of this viral family.

2. Materials and Methods

2.1. Pulse-Field Gel Electrophoresis

A viral suspension of noumeavirus was prepared according to [10]. The viral suspension was calibrated at an OD_{600} of 0.24. Drops of 45 µL of the viral suspension were embedded in 1% low melting agarose, and the plugs were incubated in lysis buffer (50 mM Tris-HCl pH 8.0, 50 mM EDTA, 1% (*v*/*v*) laurylsarcosine, and 1 mg/mL proteinase K) for 24 h at 50 °C with light shaking (500 rpm). The lysis buffer was renewed every 8 h and 1 mM DTT was added 30 min before the second buffer change. After lysis, the plugs were washed once in sterile water and twice in TE buffer (10 mM Tris HCl pH 8.0 and 1 mM EDTA) with 1 mM PMSF, for 15 min at 50 °C. The plugs were then equilibrated in the appropriate restriction buffer and digested with 20 units of ApaI at 25 °C over night (o/n) and for 3 more hours in fresh reagent. Double digested plugs were then equilibrated in the appropriate restriction buffer and digested with 20 units of SwaI at 25 °C o/n and for 3 more hours in fresh reagent. All digested plugs were washed once in sterile water for 15 min, once in lysis buffer for 2 h at 50 °C and three times in TE buffer. Electrophoresis was carried out in 0.5× TBE using a 1% agarose gel for 20 h

18 min at 6 V/cm, 120° included angle and 14 °C constant temperature in a CHEF-MAPPER system (Bio-Rad) with pulsed times ramped from 0.47 s to 54.17 s.

2.2. Genome Analysis

We gathered the 13 publicly available *Marseilleviridae* complete genomes from the GenBank database (Table S1). Suspecting the tokyovirus assembly to be contaminated with *A. castellanii* mitochondrion sequences, we reassembled the original Illumina sequences [9] using the Spades assembler [19] version 3.13.0 with the "meta" parameter. This resulted in two highly covered assembled scaffolds—a 362,593 nt one, corresponding to the tokyovirus genome (Dataset S1), and a second of 41,646 nt, corresponding to the *A. castellanii* mitochondrion.

We performed a protein-coding gene re-annotation of all the sequences using the same gene-finding algorithm—GeneMarkS [20] version 4.32—with the "virus" parameter and kept the open reading frames (ORF) coding for proteins of at least 50 amino acids.

Global analysis of nucleotide sequence conservation along the genomes was performed using the mVista online tool [21] with the "Shuffle-LAGAN" alignment program and the "translated anchoring" option.

Genomic rearrangements were visualized using the ACT genome viewer [22]. We first generated pairwise genome-wide protein alignments using Promer from the Mummer package [23], converted the alignments to the "crunch" file format, and visualized the genome-wide alignments in ACT.

2.3. Homologous Proteins Clusering and Pangenome Analysis

Protein clustering was performed using OrthoFinder [24] version 2.4.0 with the following options: "-M msa -S blast". For each cluster (referred to as "Orthogroup"), protein sequences were aligned using Clustal Omega [25], and phylogenetic trees were computed using IQtree [26]. The pangenome analysis was performed using the PanGP tool [27] and the Micropan R package [28]. The core genes were extracted from the orthogroups where at least one gene from each virus was present. The strain-specific genes correspond to orthogroups where genes belonged to a single virus (i.e., singletons). To avoid false positive singletons we only kept the genes that had no blastP match (E-value < 10^{-5}) in the other viruses.

2.4. Nucleotide Biais Composition

Cumulated AT-skews and GC-skews were computed using an in-house script provided in supplements. Breakpoints of GC-skews and AT-skews in artificially rearranged chromosomes were performed using the "rearranged.oriloc" function (see [29] for a detailed explanation) from the SeqinR R package.

2.5. Transcriptomic Data Analyses

Raw RNA-seq data from the PRJEB34467 sequencing project [18] of the *A. castellanii* infection by marseillevirus T19 were obtained from the SRA database. The dataset covers the marseillevirus infection cycle through 9 time points: 0 h pi (ERR3528397), 1 h pi (ERR3528398), 2 h pi (ERR3528399), 4 h pi (ERR3528400), 5 h pi (ERR3528401), 6 h pi (ERR3528402), 8 h pi (ERR3528403), 10 h pi (ERR3528404) and 12 h pi (ERR3528405). Paired-end reads were mapped to the genomes of marseillevirus (GU071086), *A. castellanii* (GCA_000193105 assembly) and *A. castellanii* mitochondrion (U12386) using Hisat2 version 2.1.0 with the following options: "—rna-strandness FR—no-discordant –max-intronlen 1500". This resulted in 95.3% of the reads being correctly aligned. Read counts and normalization in TPM values were performed using TPMCalculator [30]. Heatmap and gene-expression clustering was done on scaled log(TPM) values (centered by gene average expression and normalized by its standard deviation) using the "ComplexHeatmap" R package [31] with the following parameters: "clustering_distance_row = 'spearman'" and "km = 3".

2.6. Phylogeny and Selection Pressure Analysis

The *Marseilleviridae* phylogeny was computed using the concatenated multiple alignments of single-copy orthologous core genes and the IQtree software [26]. Bootstrap values were calculated using the ultrafast bootstrap approximation with 1000 replicates.

Selection pressure was measured based on the dN/dS (ω) ratios of marseilleviruses single copy orthologous genes using the Codeml algorithm [32] through the ete3 package [33]. For each orthogroup we computed a codon alignment based on nucleotide sequences and protein alignments. We then calculated ω values using two models: the M0 model (single ω for the whole tree) and the b_free model (distinct ω values for the gene of interest and for the rest of the tree). Each ω value was selected according to the LRT *p*-value between the models. To avoid saturation, ω values were only considered if $dS \leq 2$, $0.01 \leq dN \leq 2$ and $\omega \leq 10$.

3. Results

3.1. Marseilleviridae Genomes Are Circular

In its initial description, it was proposed that marseillevirus had a circular genome, albeit without experimental data supporting this predicted architecture [1]. The replication of a dsDNA genome involves different mechanisms depending on whether it is linear or circular. The analysis of genomic rearrangements also differs depending on the topology of the chromosome. Therefore, we first sought to experimentally confirm the circular structure of the *Marseilleviridae* genomes. We thus performed a Pulse-Field Gel Electrophoresis (PFGE) experiment combined with the restriction digestion of noumeavirus DNA, a *Marseilleviridae* belonging to the B clade (Figure S1). We used two restriction enzymes to cleave noumeavirus DNA: ApaI and SwaI. The first enzyme (ApaI) is predicted to cleave noumeavirus DNA at a single position. If the genome is linear, the digestion should result in two fragments, whereas a circular genome is expected to produce a single fragment of the size of the genome (376,207 nt). Figure 1C clearly shows a single band at the expected size of approximately 380 kb. The second enzyme (SwaI) is also predicted to cleave the DNA only once. The double digestion with both enzymes should thus produce two fragments in the case of a circular genome—one of 143 kb and a second one of 233 kb. Again, as expected, the migration of noumeavirus DNA subjected to double digestion resulted in two bands of the proper size (Figure 1D). It is well known that closed-circular supercoiled DNA moves very slowly in pulse-field gels [34]. Accordingly, the undigested noumeavirus DNA migrates slower than the single-cut one (Figure 1B,C). Altogether, these data confirm that noumeavirus DNA is circular, as are most likely all the *Marseilleviridae*.

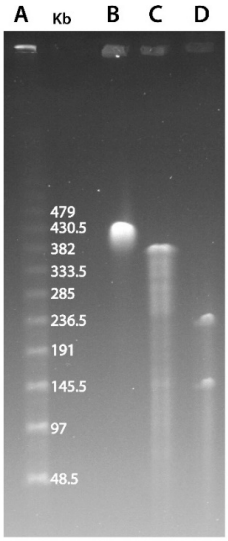


Figure 1. PFGE resolution of noumeavirus genomic DNA. (**A**) λ DNA ladder. (**B**) Undigested

noumeavirus DNA. (**C**) Digested DNA using the ApaI restriction enzyme. (**D**) Double digestion using ApaI and SwaI restriction enzymes.

3.2. Asymetry in Sequence Conservation along the Genomes

Among the isolated marseilleviruses, thirteen have a complete genome sequence (Table S1). They belong to the five currently established *Marseilleviridae* clades (Figure S1). Since the gene annotation tools and procedures used to annotate the available marseillevirus genomes are not standardized, we performed a re-annotation of the genomic sequences using the same protocol (see Materials and Methods). As shown in Table S1, gene density was consistent between strains, except for the insectomine virus that contained much more ORF, hinting at potential sequencing errors. Accordingly, the average predicted protein length was significantly smaller (Mann–Whitney p-value $= 5.6 \times 10^{-79}$) in this genome (89 aa) than in the other genomes (159 aa). We thus safely excluded it for the rest of the study, and only kept the twelve complete marseillevirus genomes that could be reliably compared.

We first sought to explore the large genomic rearrangements and insertions/deletions that occurred within these strains. Since *Marseilleviridae* genomes are circular, there is no reason for the assembler algorithms to start the assembly at the same position. Therefore, the genomes have to be aligned to a common starting point to be compared. We chose the strictly conserved Major Capsid Protein (MCP) gene which is encoded in a single copy in all marseilleviruses to define the starting position of the linearized genomes.

As shown in Figure 2, the pairwise comparison of *Marseilleviridae*, ordered according to their phylogeny, depicts a disparate frequency in genomic rearrangements. Unsurprisingly, rearrangement events were much more frequent between strains belonging to diverging clades than strains from the same clades. One exception, though, is tokyovirus. It contains a large inversion compared to the other marseilleviruses of the clade A. The *Marseilleviridae* phylogeny also shows that tokyovirus is the most divergent when viruses from the same clade are compared (Figure S1). All the other viruses belonging to the same clade (either clade A or B) exhibit almost perfectly collinear genomes. Oppositely, most of the inter-clade comparisons display a large amount of inversions and intrachromosomal translocations. Surprisingly, most of these rearrangements as well as insertions/deletions are not uniformly distributed along the genomes. They mostly occur in the leftmost two thirds of the genomic sequences (Figure 2). Conversely, the rightmost region of the *Marseilleviridae* genomes is virtually devoid of rearrangements. This region thus seems to be in a distinct evolutionary regime compared to the rest of the genome.

We next explored the sequence conservation of the marseilleviruses genomes at the nucleotide level. As expected, the pairwise comparison of the average nucleotide identity (ANI) computed using the OrthoANI tool [35] follows the *Marseilleviridae* phylogeny. The matrix in Figure S1 shows that pairwise ANI values range from 65.7% to 99.2%. Not surprisingly, the average pairwise ANI was significantly higher (Mann–Whitney p-value $= 2.7 \times 10^{-9}$) in the intra-clade than inter-clades comparisons (on average 88.75% and 69.56%, respectively).

The ANI matrix (Figure S1) only gives an average estimate of sequence identity between pairs of genomes. To further explore marseilleviruses' nucleotide sequence conservation, we analyzed its variations along the genomes. However, as mentioned earlier, several chromosomal rearrangements occurred during the *Marseilleviridae* evolution (Figure 2). In this context, a global genome alignment would not allow us to measure sequence conservation in a meaningful way. Instead, we used the Shuffle-LAGAN method [36] from the mVista tool [21]. This algorithm performs "glocal" genome alignment, which is a hybrid between local and global alignments. It first models the rearrangements between a pair of sequences and then aligns them. We compared a representative genome from each clade (tokyovirus, lausannevirus, tunisvirus, brazilian marseillevirus and golden marseillevirus) to the marseillevirus reference. Given that the viruses from clade A are highly conserved (Figure S1), we chose the most divergent one, namely tokyovirus, to compare to the marseillevirus reference and highlight potential divergent regions. It is clear from Figure 3 that sequence conservation is not uniform

along the genome. Even when rearrangements are taken into account, the rightmost parts of the marseilleviruses genomes are more conserved at the nucleotide level. This mirrors our observations of the localized lower density of rearrangements in this region.

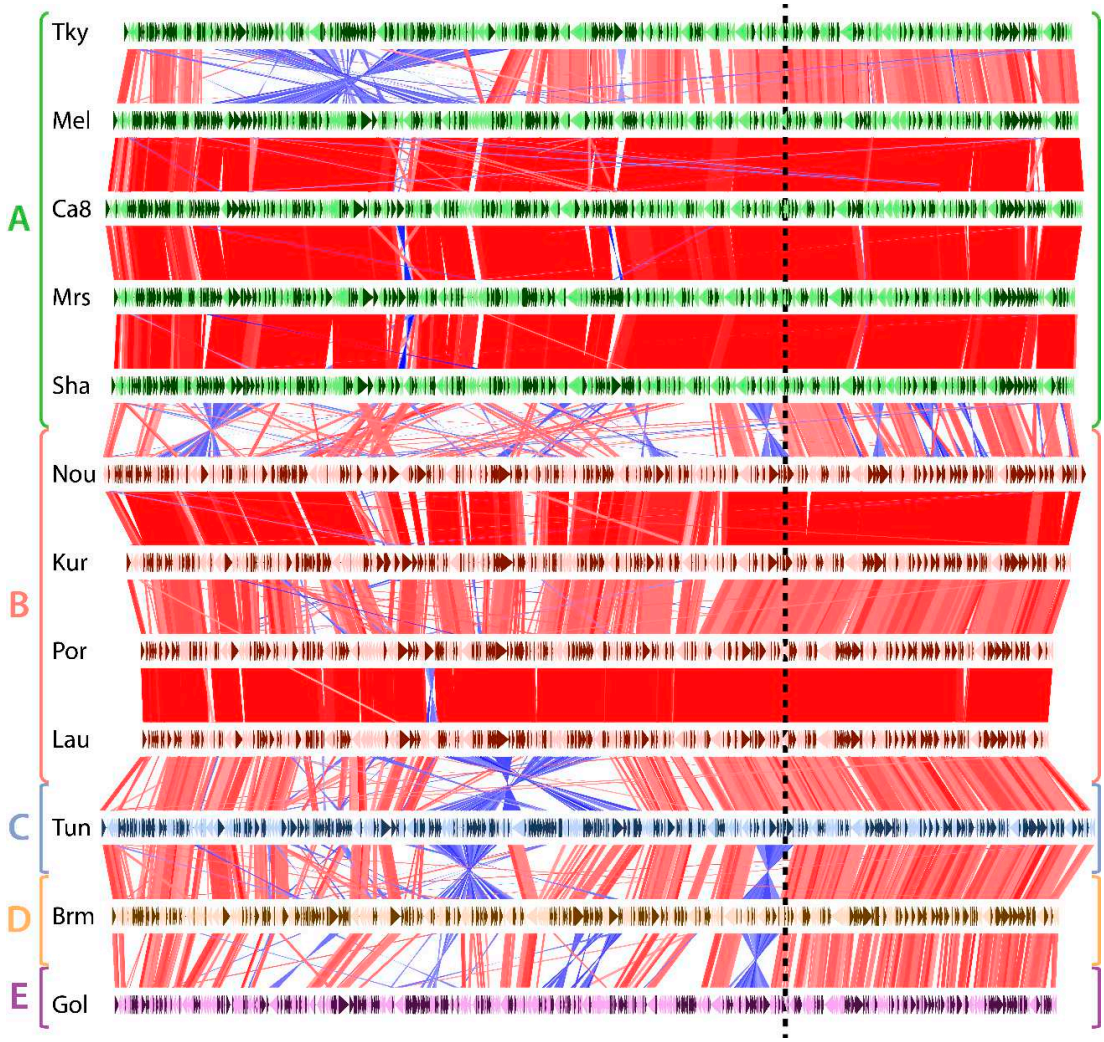


Figure 2. Genomic rearrangements in marseilleviruses. Each horizontal line represents a viral genome, namely: tokyovirus (Tky), melbournevirus (Mel), cannes 8 virus (Ca8), marseillevirus (Mrs), marseillevirus shanghai (Sha), noumeavirus (Nou), kurlavirus (Kur), port-miou virus (Por), lausannevirus (Lau), tunisvirus (Tun), brazilian marseillevirus (Brm) and golden marseillevirus (Gol), grouped and color-coded according to the clade it belongs to (shown on le left). Genes encoded on the forward strain are shown in dark colors and genes on the reverse strand in light colors. Vertical red and blue lines represent homologous genes between a pair of genomes. Red lines correspond to genes that are in the same direction, while blue lines represent inverted genes. The dashed vertical line separates the region prone to genomic rearrangements (on the left) from the one relatively depleted in rearrangements (on the right).

Owing to the apparent dichotomous distribution of sequence conservation within marseilleviruses genomes, we next examined potential variations in nucleotide composition. However, the overall GC-content was not found to be different between the most conserved regions (rightmost third of the genomes) and the rest of the genomes, with 44% and 43.4%, respectively. Beyond global nucleotide composition, though, asymmetries can occur over strands, with an excess of G over C (or A over T) and vice versa. Such asymmetries can be unveiled by computing the so-called cumulated GC-skew $((G - C)/(G + C))$ and AT-skew $((A - T)/(A + T))$ along the genomes. The Figure S2 shows the cumulated

AT- and GC-skews in *Marseilleviridae* normalized by the length of each genome in order to compare all the viruses on the same scale. Although the curves are noisy, one can see a general trend for the AT-skews of all viruses with roughly constant values from the leftmost extremity of the linearized genomes to the middle, followed by a drop, a plateau and a subtle increase by the end of the genome. The amplitude of the variations is variable between the strains, but the minimal values are all roughly located from 70% to 90% of the genome lengths. A similar although more blurry trend is depicted by the cumulative GC-skew.

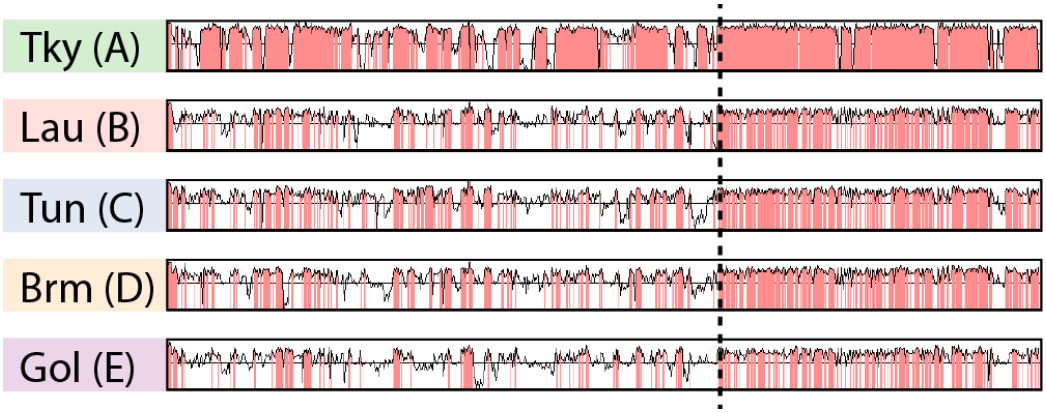


Figure 3. Nucleotide sequence conservation in marseilleviruses. Each row represents the nucleotide sequence identity from the glocal alignment of a *Marseilleviridae* against the marseillevirus reference. Regions with a sequence identity above 75% are highlighted in red. The following *Marseilleviridae* were used: tokyovirus (Tky), lausannevirus (Lau), Tunisvirus (Tun), Brazilian marseillevirus (Brm) and golden marseillevirus (Gol). The letters in parenthesis represent the clades the *Marseilleviridae* belong to. The dashed line separates the most conserved region from the most divergent one.

Nucleotide composition asymmetry is associated with several factors. The first one is the protein-coding gene orientation bias, which, due to the asymmetry of the transcription process, can lead to compositional asymmetries. Likewise, codon usage bias may cause nucleotide skews related to the asymmetry in encoded gene strands. Another main explanation is the mutation bias associated with DNA replication. In prokaryotes, the shifts in GC- and AT-skews are often correlated with the replication origin and termination sites. Analysis of nucleotide skews is thus frequently used to predict replication origins, but due to the multiple factors involved in composition asymmetry it is often a poor predictor. A workaround to uncouple the confounding factors is to artificially rearrange the genes to follow a perfect strand orientation, and analyze the GC- and AT-skews in this rearranged chromosome [29,37]. Deviations from the correlation between gene orientation skew and AT- or GC-cumulated skews are signs of replication-related asymmetries. We used this method on the *Marseilleviridae* genomes to identify breakpoints in skews and thus potential replication origin sites. As shown in Figure S2, there is a hot-spot of AT-skew breakpoints toward 80% of the marseilleviruses genomes, but there are also many AT and GC breakpoints outside this location that are distributed along the genome. Moreover, the breakpoints found using the forward and reverse strands should theoretically be co-localized, which is not the case here. Our interpretation is that a replication origin is probably present in the marseilleviruses conserved region, at roughly 20% of the rightmost extremity, but also that there are potentially multiple replication origins.

In agreement with this, we found that *Marseilleviridae* encode several copies of the predicted origin of replication binding proteins, containing the PFAM02399 protein domain. It is actually one of the *Marseilleviridae* protein families that contains the largest number of paralogs. For instance, noumeavirus encodes for as much as five different full-length copies of this protein with a recognizable protein domain. The PFAM02399 domains containing genes are evenly distributed along the genomes with no specific trend in their genomic distribution. In addition to this, there are several truncated

proteins within this protein family that potentially correspond to pseudogenes, with up to six in golden marseillevirus. One can hypothesize that if the different encoded copies are functional, they may recognize different regions of the genome, in line with our suggestion of the potential multiple replication origins in *Marseilleviridae*.

3.3. Biaised Distribution of Core Genes

The analysis of DNA sequence conservation within *Marseilleviridae* highlighted large regions of sequence divergence (Figure 3). This prompted us to explore the pangenome of this viral family. To this end, we clustered the protein-coding genes into homologous gene families (orthogroups) using OrthoFinder [24]. Such a clustering delineates different categories of genes that are traditionally coined as "core" when they are present in all the studied strains, and "accessory" for genes not strictly conserved within strains. Among this last category, genes found in a single genome are referred as "strain-specific" genes. As shown in Table 1, the proportion of core genes is fairly constant among the marseilleviruses, with an average of 54%. Conversely, the proportion of strain-specific genes is much more variable, ranging from as low as 1% up to 14% for golden marseillevirus, with an average of 3%. These strain-specific genes can either correspond to genes only found in *Marseilleviridae*, the so-called "ORFans", or to genes with homologs outside of the family. Here we find that 98% of the strain-specific genes are genuine ORFans, the others being HGT candidates. The very high proportion of strain-specific genes in golden marseillevirus points to an unexplored diversity of *Marseilleviridae*. To confirm this, we performed an analysis of the marseilleviruses pangenome and coregenome. Figure S3 shows the number of shared (i.e., core) genes as a function of incrementally incorporated genomes. The curve is clearly asymptotic, meaning that the pool of core genes identified from the strains under study (on average 271 gene per strain) will not evolve as new marseilleviruses are discovered. By contrast, a similar analysis of the pangenome (i.e., the total of marseilleviruses genes) displays an unsaturated curve. With an α Heap's law parameter of 0.86 when fitting this data, the *Marseilleviridae* pangenome is considered open ($\alpha < 1$) [38]. This confirms that the *Marseilleviridae*'s diversity is not fully explored yet.

Table 1. Counts and frequencies of core, strain-specific, single-copy and paralogous genes in marseilleviruses.

	Core	Strain-Specific	Single Copy	Paralogs
Tokyovirus	269 (55%)	35 (7%)	396 (81%)	95 (19%)
Melbournevirus	265 (52%)	6 (1%)	409 (81%)	96 (19%)
Cannes 8 virus	268 (53%)	3 (1%)	407 (80%)	103 (20%)
Marseillevirus	264 (52%)	7 (1%)	402 (79%)	107 (21%)
Marseillevirus shanghai	266 (53%)	3 (1%)	404 (80%)	101 (20%)
Noumeavirus	277 (55%)	16 (3%)	394 (78%)	113 (22%)
Kurlavirus	272 (55%)	12 (2%)	381 (77%)	114 (23%)
Port-miou virus	269 (57%)	8 (2%)	383 (82%)	95 (18%)
Lausannevirus	266 (58%)	3 (1%)	375 (81%)	86 (19%)
Tunisvirus	281 (52%)	31 (6%)	385 (71%)	155 (29%)
Brazilian marseillevirus	272 (56%)	13 (3%)	373 (77%)	114 (23%)
Golden marseillevirus	282 (52%)	76 (14%)	373 (69%)	170 (31%)

Since core genes compose a large part of the marseilleviruses' gene repertoires, we next wanted to study their distribution along the genomes, seeking for potential hot-spots. To this end, we first normalized each core gene genomic position by the length of its cognate genome. For each genome we next measured the density of core genes in a sliding window. The resulting smoothed density was next centered and scaled to a z-score according to the median value of all the windows to highlight variations in core gene densities. The heatmap presented in Figure 4 clearly reveals a strong asymmetry in marseilleviruses core gene densities. Again, the rightmost part of the genomes is strongly enriched

in core genes compared to the rest of the genomes. This pattern is shared by all viruses, regardless of the clades they belong to. This region roughly corresponds to a third of the genomes.

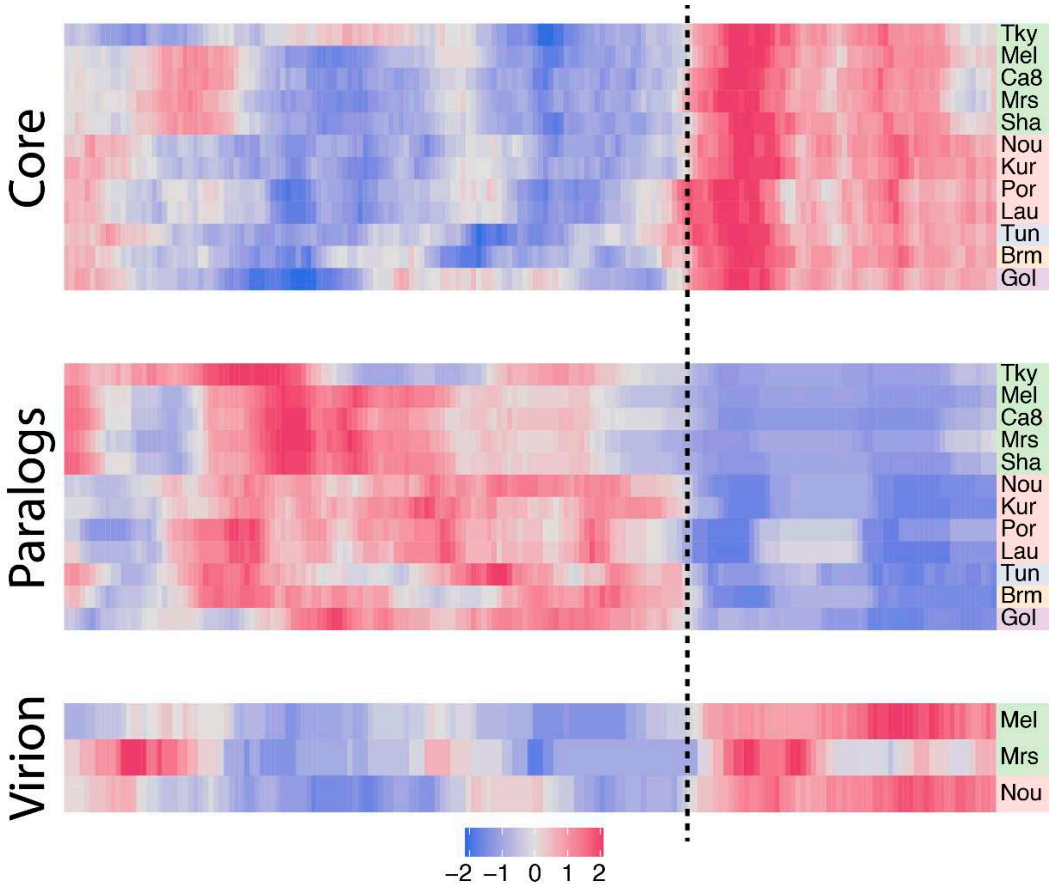


Figure 4. Density of core genes, paralogs and virion-associated protein-coding genes in marseilleviruses. Each row corresponds to a marseillevirus, namely: tokyovirus (Tky), melbournevirus (Mel), cannes 8 virus (Ca8), marseillevirus (Mrs), marseillevirus shanghai (Sha), noumeavirus (Nou), kurlavirus (Kur), port-miou virus (Por), lausannevirus (Lau), tunisvirus (Tun), brazilian marseillevirus (Brm) and golden marseillevirus (Gol). Strains are color-coded according to the clade they belong to, with A in green, B in red, C in blue, D in orange an E in purple. The z-score normalized density is color-coded from blue (low density) to pink (high density). The dash line separates the region of higher density in core genes, paralogs and virion-associated protein-coding genes.

Since the rightmost region of the *Marseilleviridae* genomes, now referred to as the "core region", is enriched in core genes (Figure 4) and is more conserved (Figure 3), we reasoned that it could be subjected to a different selection pressure. To test this hypothesis, we computed the ratios (ω) of non-synonymous mutation rates (dN) over synonymous mutation rates (dS) of orthologous genes using the Codeml program [32] (see Materials and Methods). The distribution of ω values along normalized genomic positions is shown in Figure S4. Globally, all the genomic positions are subjected to strong negative selection ($\omega << 1$) whatever the region, confirming the selection pressure previously measured on melbournevirus genes [7]. However, genes from the core region have more homogeneous ω values and seem to be under marginally, although statistically significant, stronger purifying selection compared to the rest of the genome, with an average ω of 0.097 and 0.156, respectively (Mann–Whitney p-value = 2.7×10^{-94}).

Following the analysis of core genes, we next explored gene duplication events. More specifically, we analyzed the orthogroups previously defined (see Materials and Methods) and categorized genes into two bins: single copy genes and duplicated genes (paralogs). The vast majority of marseilleviruses

genes are single copy genes with an average of 78% per virus (Table 1). Symmetrically, the proportion of duplicated genes is low (22%), and is even significantly lower in strain-specific genes, with only 2.8% in this category (Fisher's exact test p-value = 6.5×10^{-9}). We then again investigated the densities of duplicated genes along the genomes using the method previously described (Figure 4). Contrary to core genes, the paralogs are mostly present in the leftmost part of the genomes.

3.4. Biaised Distribution of Virion-Associated Proteins and Late-Expressed Transcripts

In viruses, genes coding for the proteins present in the virions, in particular the structural proteins that build the particles, are thought to be among the most conserved ones. There are now three *Marseilleviridae* members for which the viral particles' proteome compositions have been analyzed by mass spectroscopy [1,10]. Two of them, marseillevirus and melbournevirus, belong to clade A, and the third one, noumeavirus, to clade B. To verify this assumption in *Marseilleviridae*, we analyzed the overlap between the core genes and genes coding for virion-associated proteins. The proportion of core genes in virion-associated proteins is remarkably constant among strains, with 83.5% in melbournevirus, 83.7% in marseillevirus and 83% in noumeavirus. These values are also significantly higher than the proportion of core genes in proteins not identified in virion proteomes, corresponding to 43.9%, 48.9% and 44.4%, respectively, with Fisher's exact test p-values of 4.2×10^{-14}, 9×10^{-6} and 1.9×10^{-15}. Furthermore, we know from previous work that *Marseilleviridae* particle proteomes are well conserved, with a high correlation in their respective protein contents [10]. Altogether this means that *Marseilleviridae* particles are not strictly composed of core genes, they mostly contain these types of proteins, and the same core genes are used in different viruses' particles. Knowing this, we expected the virion-associated protein-coding genes to be asymmetrically distributed along the marseilleviruses genomes, and indeed, the density of virion-associated genes is clearly biased to the core region previously identified (Figure 4).

The global transcriptome of *A. castellanii* infection by marseillevirus has been studied by RNA-seq through a replicative cycle [18]. We used this dataset to test whether the marseillevirus' transcriptional activity was also regionalized along the genome. Starting from the raw sequence data, we mapped the reads to the *A. castellanii* and marseillevirus genomes, and computed a normalized expression value for each gene (see Materials and Methods). As previously observed [18], we found the marseillevirus' genes expressions to be clustered into three main classes: early, intermediate and late (Figure S5). Early genes are expressed from the beginning of the cycle, with a peak of transcriptional activity between 1 h pi and 2 h pi, intermediate genes are mostly expressed between 1 h pi and 4 h pi, and late genes from 4 h pi until the end of the cycle. We next focused on the marseillevirus virion-associated protein-coding genes to check whether they were expressed in a time-dependent manner. Confirming the results previously obtained [18], we found an enrichment in virion-associated genes in the late expression class. In addition, we also found that marseillevirus genes orthologous to melbournevirus and noumeavirus virion-associated genes were enriched in that category (Figure S5). Finally, although the bias was less pronounced, we found that core genes were statistically enriched in the late expression class (61%) compared to the intermediate (41%) and early (43%) expression classes. Conversely, strain- and clade-specific genes are not specifically enriched in one of those classes (Chi-square p-value = 0.1).

We next wondered whether the genes encoded in the core region had a higher transcriptional activity than those in the rest of the genome. To this end we measured the global expression level of each marseillevirus gene by summing the expression value of all time points. As shown in Table 2, when comparing the genes from the core region from the rest, we found no statistical difference in global expressions (Student's t-test p-value = 0.29). However, taking the summed time-point expressions as a proxy of global gene expression might introduce a bias, since late genes are only expressed at the end of the cycle, thus they might contribute to a lesser extent. To overcome this potential bias, we also analyzed the maximal time-point expression of each gene. Again, we did not find a higher expression for core region-encoded genes with this metric (Student's t-test p-value = 0.43). Altogether, these data suggest that core region-containing genes had no particular behavior in terms of expression strength.

Table 2. RNA-seq gene expression in marseillevirus.

	Early Expressed Genes Counts (%)	Intermediate Expressed Genes Counts (%)	Late Expressed Genes Counts (%)	* Maximal Expression (Mean ± SD)	* Total Expression (Mean ± SD)
Core-region	11 (7%)	39 (23%)	116 (70%)	7.4 ± 1.3	52 ± 9.8
Other region	50 (15%)	143 (42%)	150 (44%)	7.3 ± 1.3	51 ± 9.8

* RNA-seq expression is measured in log(TPM).

Finally, we analyzed the transcriptional activity of the core region regarding the expression timing. As shown in Table 2, the frequency of late genes (70%) was significantly higher (Chi-square p-value = 1.96×10^{-7}) in the core region compared to the rest of the genome (44%). Thus the core region seems to be mostly expressed at the end of the replication cycle.

4. Discussion

Double-stranded DNA viruses have structurally diverse genomes that are either linear or circular. The first *Marseilleviridae* to be isolated (marseillevirus) was predicted to have a circular genome [1]. However, no formal proof was given in this initial work to validate this assumption, nor in the following studies describing new isolates from this family [1–12]. Assuming a genome structure without experimental evidence can impair our understanding of the physiology of the viral family, as the mechanism by which genetic material is replicated depends on that topology. A recent study of the faustoviruses, for instance, showed that the initially assumed circular viral chromosomes were actually linear [39]. It is thus essential to experimentally validate predicted genome structures. In this work, we confirmed and demonstrated that *Marseilleviridae* have circular genomes.

Surprisingly, circular genome topology is rather unusual among the numerous large and giant viruses infecting amoeba. From the eight viral families described so far, only two exhibit a circular genome: the Pithoviruses (with pithoviruses [40], cedratviruses [17,41] and orpheovirus [42]) and the *Marseilleviridae*. The six remaining families, namely the *Mimiviridae* [43], the pandoraviruses [44], the molliviruses [45], the faustoviruses [39], the pacmanviruses [46] and medusavirus [47], are all predicted, based on sequencing read mapping and genome assembly, to exhibit linear genomes. Then what would be the advantage, if any, of a virus encoding its genes in a circular chromosome? One possibility would be to escape exonuclease enzymatic activity. An example can be found in the *Escherichia coli* bacteria, which use the RecBCD exonuclease as a weapon against invading bacteriophages that contain free-ends DNA [48]. With the bacterial genome being circular, it is not subjected to exonuclease activity. Some bacteriophages counteract this attack by encoding inhibitors of RecBCD, such as the Gam protein encoded by the phage lambda [48,49]. An analogous escape mechanism could be at play here, whereby the circular structure of the *Marseilleviridae* genomes could lead them to escape exonucleases either encoded by the host, intracellular bacteria infecting the amoeba, or even other viruses in the case of coinfection. These kinds of virus–host and virus–pathogens interactions somehow relates to the ones driven by the *Marseilleviridae* encoded restriction–modification systems [17]. In that case, the viruses use endonucleases to digest competing pathogens' DNA inside the amoeba while protecting themselves against degradation by methylating their own DNA. Here, the genome topology by itself would be sufficient to escape exonuclease activity.

Marseilleviridae are thought to be prone to frequent HGT, with roughly a quarter of their genes suspected to be acquired through this route [1]. Surprisingly, supposed gene exchanges with bacteria are even more frequent than the ones involving their amoebic host [1]. Indeed, with as much as 45% of all potential cell–virus gene exchanges, this represents an unexpected proportion. This might relate to the fact that *Acanthamoeba* are infested by a large variety of bacterial parasites or symbionts [50]. However, in other viruses, such as the giant pandoraviruses, which have been scrutinized for HGTs, it was shown that bacteria only account for 20% of the exchanges related to cellular organisms [51]. This proportion even drops to 13% when considering cell-to-virus transfers specifically. Yet these two viral families

infect the exact same host, and thus face the same environment. So, there might be other determinants explaining the higher proportion of bacteria-related exchanges in *Marseilleviridae*. The genome structure could be one of them. Considering that *Marseilleviridae* adopt a bacterial-like circular genome, one could hypothesize that this topology somehow favors genetic exchanges, leading in certain cases to a selective advantage, as exemplified by the negative selection pressure acting on bacterial-like genes [7]. The *Marseilleviridae* RM systems are a striking example of such transfers [17]. It is noteworthy that the only giant virus family exhibiting circular genomes, the pithoviruses, also contains a high proportion (38%) of cell–virus potential gene exchanges related to bacteria, although it only accounts for 8% of the total gene set [40]. This again supports the hypothesis that a circular genome topology might facilitate gene transfers with this domain.

Following the discovery of the second *Marseilleviridae* strain (lausannevirus) [2], the authors noticed an asymmetry in the distribution of its genes along the genome. They unveiled an enrichment in annotated genes (i.e., with a predictable function based on sequence homology) on one side of the lausannevirus genome, and an opposite enrichment in "hypothetical protein" genes on the other side. They also noticed localized hot-spots of sequence rearrangements between lausannevirus and marseillevirus [2]. In this study, we expanded the comparative analysis to the twelve complete *Marseilleviridae* genomes that could be reliably compared. Our data clearly show a strong asymmetry in the *Marseilleviridae* genomes, with one region, namely the core region, corresponding to roughly a third of the genome that exhibits several peculiar properties. We first revealed that this region is virtually devoid of genomic rearrangements, while these frequently occurred in the course of the *Marseilleviridae* evolution. Accordingly, this region is also more conserved at the nucleotide level. Strikingly, the density of core genes is also much higher in this region. Such a regionalized distribution of family core genes has already been observed in other amoeba-infecting giant viruses, all of which having linear genomes. For instance, the viruses with the largest known genomes so far, the pandoraviruses, exhibit a regionalized enrichment of core genes in the first half of their genomes [51]. A similar dichotomous distribution was revealed in the distantly related molliviruses, where genes shared between molliviruses and pandoraviruses are also co-localized in half of their genomes [52]. Likewise, in faustoviruses, sequence conservation is not uniformly distributed, although it displays a different pattern with greater sequence divergence in the middle of the genome and at the extremities [39]. Somehow this relates to the strongly biased distribution of conserved gene order observed in the central part of the *Mimiviridae* genomes, as compared to the shuffled extremities [53]. Thus there are clearly different patterns of sequence conservation asymmetry in giant viruses infecting amoeba. Yet regardless of the genome structure, be it circular or linear, this asymmetry seems to be a common trait. Beyond amoeba-infecting giant viruses, the *Poxviridae*, also members of the *Nucleocytoviricota* phylum, retained most of the conserved genes in the central part of their genomes [54].

Besides the regionalized enrichment of core genes at a specific genomic location, we showed that *Marseilleviridae* genes coding for proteins detected in viral particles are also clustered together. Thus, whether they build the particles or are involved in the early phase of the infection, "important" viral genes are clustered in the core region. Then how could we explain such a regionalization? The globally late expression of the core region encoded genes might be a key to understanding this pattern. In mimivirus, the transcriptional time-dependent activity is clearly governed by the strict conservation of sequence motifs in gene promoters [55,56]. Thus genes do not need to be located in a specific genomic region to activate their expression in a time-dependent manner. On the contrary, the analysis of the marseillevirus' transcriptome failed to unveil sequence motifs explaining gene expression patterns [18]. In that case, clustering the genes in a confined region of the genome might be a good strategy to activate gene expression at the right time. The transcriptional switch could then be done thanks to a particular topology of the DNA in that region. In that context, one has to keep in mind that *Marseilleviridae* have the astonishing ability to encode histones [1,2]. These could play a role in the transcriptional regulation of this specific genomic region. In other words, one can hypothesize that the core region's transcriptional dynamics are controlled by DNA-dependent topological properties. It is

also worth mentioning that *Marseilleviridae* probably use the host-encoded transcription apparatus in the early phases of the infection, and then switch to the viral encoded apparatus as soon as the viral RNA polymerase is available [10]. Thus genes from the core region might by controlled by the latter transcriptional system.

In circular bacterial genomes, genes tend to be less conserved with the increasing distance from the origin of replication [57,58]. Essential and highly expressed genes are usually located near the replication origin, and this is especially true for transcription- and translation-related genes [58]. One explanation for this correlation is the replication-associated gene dosage. As replication starts, genes located near the site of replication origin are in two copies, thus are more expressed than the ones located near the replication terminus, which remain in one copy [59]. Owing to the fact that *Marseilleviridae* genomes are (i) circular and (ii) highly asymmetrical, with core genes being clustered together, we explored the possibility of a relation with the distance to their origin of replication, akin to what was observed in bacterial chromosomes. Our work suggests that *Marseilleviridae* might share a replication origin located in the core region, thus resembling what is observed in the bacterial world. Next, we analyzed the proteins potentially involved in replication origin recognition. Our work revealed that *Marseilleviridae* surprisingly encode many copies of these types of proteins. If the different copies are functional, one can hypothesize that they recognize different sequence-specific sites. Our detection of several compositional strand biases argues for the presence of multiple dispersed replication origins, as seen in the circular archaea genomes [60]. However, sequence analysis is clearly limited to uncover such subtle genomic signals. A solution would be to use deep sequencing methods to uncover replication origins. The application of such a method to another member of the *Nucleocytoviricota*, the vaccinia *Poxviridae*, allowed the replication origins to be mapped at a single base pair resolution [61]. They happen to be located near the ends of this covalently closed linear genome, at the concatemer junctions. This mechanism could be shared with giant viruses with the same DNA topology. However, the question remains open for viruses with circular genomes, such as the *Marseilleviridae*, and deserves to be approached experimentally.

Our analysis of the *Marseilleviridae*'s gene content highlighted an open pangenome, meaning that the *Marseilleviridae*'s diversity has not been fully uncovered yet. This is mainly exemplified by the relatively large fraction of strain-specific genes in golden marseillevirus. Paradoxically, the recent works in the metagenomic data analysis of giant viruses through the assembly of huge datasets revealed very few metagenome-assembled genomes (MAG) related to the *Marseilleviridae* [62]. They seem to be nearly absent from the environmental microbial data. This might highlight the limits of such methods in revealing the true diversity of giant viruses in the wild, or indicate that environments containing *Marseilleviridae* have not been correctly sampled yet. We believe that future studies will be needed to isolate and characterize new *Marseilleviridae* members so as to fully comprehend this viral family.

Supplementary Materials: The following are available online at http://www.mdpi.com/1999-4915/12/11/1270/s1, Figure S1: Nucleotide-level genomic similarity between *Marseilleviridae*. Figure S2: Identification of potential origins of replication using cumulative AT-skew and GC-skew, Figure S3: Core-genome and pan-genome of the *Marseilleviridae*, Figure S4: Selection pressure along the *Marseilleviridae* genomes, Figure S5: RNA-seq marseillevirus gene expression, Table S1: Complete marseilleviruses sequenced genomes. Computer code: In-house python script used to compute cumulative AT-skew and GC-skew. Dataset S1: Reassembled genomic sequence of tokyovirus.

Author Contributions: conceptualization, M.L.; experimental analyses, L.B. and E.C.-F.; computational analyses, L.B., S.R. and M.L.; manuscript writing M.L. All authors have read and agreed to the published version of the manuscript.

Funding: L.B. received an internship compensation through the Fondation Bettencourt Schueller (OTP51251). E.C. is the recipient of a DGA-MRIS scholarship (scholarship 201760003) and S.R. is supported by a doctoral fellowship obtained from Aix-Marseille University.

Acknowledgments: We would like to acknowledge Masaharu Takemura for kindly providing the raw tokyovirus sequence data. We also would like to thank Chantal Abergel and Sandra Jeudy for their advice on the manuscript, as well as Jean-Michel Claverie for the initial discussions in this project.

Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results.

References

1. Boyer, M.; Yutin, N.; Pagnier, I.; Barrassi, L.; Fournous, G.; Espinosa, L.; Robert, C.; Azza, S.; Sun, S.; Rossmann, M.G.; et al. Giant Marseillevirus highlights the role of amoebae as a melting pot in emergence of chimeric microorganisms. *Proc. Natl. Acad. Sci. USA* **2009**, *106*, 21848–21853. [CrossRef]
2. Thomas, V.; Bertelli, C.; Collyn, F.; Casson, N.; Telenti, A.; Goesmann, A.; Croxatto, A.; Greub, G. Lausannevirus, a giant amoebal virus encoding histone doublets. *Environ. Microbiol.* **2011**, *13*, 1454–1466. [CrossRef]
3. Aherfi, S.; Pagnier, I.; Fournous, G.; Raoult, D.; La Scola, B.; Colson, P. Complete genome sequence of Cannes 8 virus, a new member of the proposed family "Marseilleviridae". *Virus Genes* **2013**, *47*, 550–555. [CrossRef] [PubMed]
4. Boughalmi, M.; Pagnier, I.; Aherfi, S.; Colson, P.; Raoult, D.; La Scola, B. First isolation of a Marseillevirus in the Diptera Syrphidae Eristalis tenax. *Intervirology* **2013**, *56*, 386–394. [CrossRef]
5. Aherfi, S.; Boughalmi, M.; Pagnier, I.; Fournous, G.; La Scola, B.; Raoult, D.; Colson, P. Complete genome sequence of Tunisvirus, a new member of the proposed family Marseilleviridae. *Arch. Virol.* **2014**, *159*, 2349–2358. [CrossRef]
6. Dornas, F.P.; Assis, F.L.; Aherfi, S.; Arantes, T.; Abrahão, J.S.; Colson, P.; La Scola, B. A Brazilian Marseillevirus Is the Founding Member of a Lineage in Family Marseilleviridae. *Viruses* **2016**, *8*, 76. [CrossRef]
7. Doutre, G.; Philippe, N.; Abergel, C.; Claverie, J.-M. Genome analysis of the first Marseilleviridae representative from Australia indicates that most of its genes contribute to virus fitness. *J. Virol.* **2014**, *88*, 14340–14349. [CrossRef]
8. Doutre, G.; Arfib, B.; Rochette, P.; Claverie, J.-M.; Bonin, P.; Abergel, C. Complete Genome Sequence of a New Member of the Marseilleviridae Recovered from the Brackish Submarine Spring in the Cassis Port-Miou Calanque, France. *Genome Announc.* **2015**, *3*. [CrossRef] [PubMed]
9. Takemura, M. Draft Genome Sequence of Tokyovirus, a Member of the Family Marseilleviridae Isolated from the Arakawa River of Tokyo, Japan. *Genome Announc.* **2016**, *4*. [CrossRef]
10. Fabre, E.; Jeudy, S.; Santini, S.; Legendre, M.; Trauchessec, M.; Couté, Y.; Claverie, J.-M.; Abergel, C. Noumeavirus replication relies on a transient remote control of the host nucleus. *Nat. Commun.* **2017**, *8*, 15087. [CrossRef]
11. Dos Santos, R.N.; Campos, F.S.; Medeiros de Albuquerque, N.R.; Finoketti, F.; Côrrea, R.A.; Cano-Ortiz, L.; Assis, F.L.; Arantes, T.S.; Roehe, P.M.; Franco, A.C. A new marseillevirus isolated in Southern Brazil from Limnoperna fortunei. *Sci. Rep.* **2016**, *6*, 35237. [CrossRef]
12. Chatterjee, A.; Kondabagil, K. Complete genome sequence of Kurlavirus, a novel member of the family Marseilleviridae isolated in Mumbai, India. *Arch. Virol.* **2017**, *162*, 3243–3245. [CrossRef]
13. International Committee on Taxonomy of Viruses (ICTV). Available online: https://talk.ictvonline.org/taxonomy/ (accessed on 16 September 2020).
14. International Committee on Taxonomy of Viruses Executive Committee. The new scope of virus taxonomy: Partitioning the virosphere into 15 hierarchical ranks. *Nat. Microbiol.* **2020**, *5*, 668–674. [CrossRef]
15. Aoki, K.; Hagiwara, R.; Akashi, M.; Sasaki, K.; Murata, K.; Ogata, H.; Takemura, M. Fifteen Marseilleviruses Newly Isolated From Three Water Samples in Japan Reveal Local Diversity of Marseilleviridae. *Front. Microbiol.* **2019**, *10*, 1152. [CrossRef]
16. Bäckström, D.; Yutin, N.; Jørgensen, S.L.; Dharamshi, J.; Homa, F.; Zaremba-Niedwiedzka, K.; Spang, A.; Wolf, Y.I.; Koonin, E.V.; Ettema, T.J.G. Virus Genomes from Deep Sea Sediments Expand the Ocean Megavirome and Support Independent Origins of Viral Gigantism. *mBio* **2019**, *10*, e02497-18. [CrossRef]
17. Jeudy, S.; Rigou, S.; Alempic, J.-M.; Claverie, J.-M.; Abergel, C.; Legendre, M. The DNA methylation landscape of giant viruses. *Nat. Commun.* **2020**, *11*, 2657. [CrossRef]
18. Rodrigues, R.A.L.; Louazani, A.C.; Picorelli, A.; Oliveira, G.P.; Lobo, F.P.; Colson, P.; La Scola, B.; Abrahão, J.S. Analysis of a Marseillevirus Transcriptome Reveals Temporal Gene Expression Profile and Host Transcriptional Shift. *Front. Microbiol.* **2020**, *11*, 651. [CrossRef]

19. Nurk, S.; Meleshko, D.; Korobeynikov, A.; Pevzner, P.A. metaSPAdes: A new versatile metagenomic assembler. *Genome Res.* **2017**, *27*, 824–834. [CrossRef]
20. Besemer, J.; Lomsadze, A.; Borodovsky, M. GeneMarkS: A self-training method for prediction of gene starts in microbial genomes. Implications for finding sequence motifs in regulatory regions. *Nucleic Acids Res.* **2001**, *29*, 2607–2618. [CrossRef]
21. Frazer, K.A.; Pachter, L.; Poliakov, A.; Rubin, E.M.; Dubchak, I. VISTA: Computational tools for comparative genomics. *Nucleic Acids Res.* **2004**, *32*, W273–W279. [CrossRef]
22. Carver, T.J.; Rutherford, K.M.; Berriman, M.; Rajandream, M.-A.; Barrell, B.G.; Parkhill, J. ACT: The Artemis Comparison Tool. *Bioinformatics* **2005**, *21*, 3422–3423. [CrossRef]
23. Kurtz, S.; Phillippy, A.; Delcher, A.L.; Smoot, M.; Shumway, M.; Antonescu, C.; Salzberg, S.L. Versatile and open software for comparing large genomes. *Genome Biol.* **2004**, *5*, R12. [CrossRef]
24. Emms, D.M.; Kelly, S. OrthoFinder: Phylogenetic orthology inference for comparative genomics. *Genome Biol.* **2019**, *20*, 238. [CrossRef]
25. Sievers, F.; Higgins, D.G. Clustal Omega, accurate alignment of very large numbers of sequences. *Methods Mol. Biol.* **2014**, *1079*, 105–116. [CrossRef]
26. Nguyen, L.-T.; Schmidt, H.A.; von Haeseler, A.; Minh, B.Q. IQ-TREE: A Fast and Effective Stochastic Algorithm for Estimating Maximum-Likelihood Phylogenies. *Mol. Biol. Evol.* **2015**, *32*, 268–274. [CrossRef]
27. Zhao, Y.; Jia, X.; Yang, J.; Ling, Y.; Zhang, Z.; Yu, J.; Wu, J.; Xiao, J. PanGP: A tool for quickly analyzing bacterial pan-genome profile. *Bioinformatics* **2014**, *30*, 1297–1299. [CrossRef]
28. Snipen, L.; Liland, K.H. Micropan: An R-package for microbial pan-genomics. *BMC Bioinform.* **2015**, *16*, 79. [CrossRef]
29. Necşulea, A.; Lobry, J.R. A new method for assessing the effect of replication on DNA base composition asymmetry. *Mol. Biol. Evol.* **2007**, *24*, 2169–2179. [CrossRef]
30. Vera Alvarez, R.; Pongor, L.S.; Mariño-Ramírez, L.; Landsman, D. TPMCalculator: One-step software to quantify mRNA abundance of genomic features. *Bioinformatics* **2019**, *35*, 1960–1962. [CrossRef]
31. Gu, Z.; Eils, R.; Schlesner, M. Complex heatmaps reveal patterns and correlations in multidimensional genomic data. *Bioinformatics* **2016**, *32*, 2847–2849. [CrossRef]
32. Yang, Z. PAML 4: Phylogenetic analysis by maximum likelihood. *Mol. Biol. Evol.* **2007**, *24*, 1586–1591. [CrossRef] [PubMed]
33. Huerta-Cepas, J.; Serra, F.; Bork, P. ETE 3: Reconstruction, Analysis, and Visualization of Phylogenomic Data. *Mol. Biol. Evol.* **2016**, *33*, 1635–1638. [CrossRef] [PubMed]
34. Barton, B.M.; Harding, G.P.; Zuccarelli, A.J. A general method for detecting and sizing large plasmids. *Anal. Biochem.* **1995**, *226*, 235–240. [CrossRef]
35. Lee, I.; Kim, Y.O.; Park, S.-C.; Chun, J. OrthoANI: An Improved Algorithm and Software for Calculating Average Nucleotide Identity. Available online: http://pubmed.ncbi.nlm.nih.gov/26585518/ (accessed on 21 September 2020).
36. Brudno, M.; Malde, S.; Poliakov, A.; Do, C.B.; Couronne, O.; Dubchak, I.; Batzoglou, S. Glocal alignment: Finding rearrangements during alignment. *Bioinformatics* **2003**, *19*, i54–i62. [CrossRef] [PubMed]
37. Nikolaou, C.; Almirantis, Y. A study on the correlation of nucleotide skews and the positioning of the origin of replication: Different modes of replication in bacterial species. *Nucleic Acids Res.* **2005**, *33*, 6816–6822. [CrossRef]
38. Tettelin, H.; Riley, D.; Cattuto, C.; Medini, D. Comparative genomics: The bacterial pan-genome. *Curr. Opin. Microbiol.* **2008**, *11*, 472–477. [CrossRef]
39. Geballa-Koukoulas, K.; Boudjemaa, H.; Andreani, J.; La Scola, B.; Blanc, G. Comparative Genomics Unveils Regionalized Evolution of the Faustovirus Genomes. *Viruses* **2020**, *12*, 577. [CrossRef]
40. Legendre, M.; Bartoli, J.; Shmakova, L.; Jeudy, S.; Labadie, K.; Adrait, A.; Lescot, M.; Poirot, O.; Bertaux, L.; Bruley, C.; et al. Thirty-thousand-year-old distant relative of giant icosahedral DNA viruses with a pandoravirus morphology. *Proc. Natl. Acad. Sci. USA* **2014**, *111*, 4274–4279. [CrossRef]
41. Andreani, J.; Aherfi, S.; Bou Khalil, J.Y.; Di Pinto, F.; Bitam, I.; Raoult, D.; Colson, P.; La Scola, B. Cedratvirus, a Double-Cork Structured Giant Virus, is a Distant Relative of Pithoviruses. *Viruses* **2016**, *8*, 300. [CrossRef]
42. Andreani, J.; Khalil, J.Y.B.; Baptiste, E.; Hasni, I.; Michelle, C.; Raoult, D.; Levasseur, A.; La Scola, B. Orpheovirus IHUMI-LCC2: A New Virus among the Giant Viruses. *Front. Microbiol.* **2017**, *8*, 2643. [CrossRef]

43. Raoult, D.; Audic, S.; Robert, C.; Abergel, C.; Renesto, P.; Ogata, H.; La Scola, B.; Suzan, M.; Claverie, J.-M. The 1.2-megabase genome sequence of Mimivirus. *Science* **2004**, *306*, 1344–1350. [CrossRef]
44. Philippe, N.; Legendre, M.; Doutre, G.; Couté, Y.; Poirot, O.; Lescot, M.; Arslan, D.; Seltzer, V.; Bertaux, L.; Bruley, C.; et al. Pandoraviruses: Amoeba viruses with genomes up to 2.5 Mb reaching that of parasitic eukaryotes. *Science* **2013**, *341*, 281–286. [CrossRef]
45. Legendre, M.; Lartigue, A.; Bertaux, L.; Jeudy, S.; Bartoli, J.; Lescot, M.; Alempic, J.-M.; Ramus, C.; Bruley, C.; Labadie, K.; et al. In-depth study of Mollivirus sibericum, a new 30,000-y-old giant virus infecting Acanthamoeba. *Proc. Natl. Acad. Sci. USA* **2015**, *112*, E5327–E5335. [CrossRef]
46. Andreani, J.; Khalil, J.Y.B.; Sevvana, M.; Benamar, S.; Di Pinto, F.; Bitam, I.; Colson, P.; Klose, T.; Rossmann, M.G.; Raoult, D.; et al. Pacmanvirus, a New Giant Icosahedral Virus at the Crossroads between Asfarviridae and Faustoviruses. *J. Virol.* **2017**, *91*. [CrossRef]
47. Yoshikawa, G.; Blanc-Mathieu, R.; Song, C.; Kayama, Y.; Mochizuki, T.; Murata, K.; Ogata, H.; Takemura, M. Medusavirus, a Novel Large DNA Virus Discovered from Hot Spring Water. *J. Virol.* **2019**, *93*, e02130-18. [CrossRef]
48. Dillingham, M.S.; Kowalczykowski, S.C. RecBCD enzyme and the repair of double-stranded DNA breaks. *Microbiol. Mol. Biol. Rev.* **2008**, *72*, 642–671. [CrossRef]
49. Murphy, K.C. Lambda Gam protein inhibits the helicase and chi-stimulated recombination activities of Escherichia coli RecBCD enzyme. *J. Bacteriol.* **1991**, *173*, 5808–5821. [CrossRef]
50. Schmitz-Esser, S.; Toenshoff, E.R.; Haider, S.; Heinz, E.; Hoenninger, V.M.; Wagner, M.; Horn, M. Diversity of bacterial endosymbionts of environmental acanthamoeba isolates. *Appl. Environ. Microbiol.* **2008**, *74*, 5822–5831. [CrossRef]
51. Legendre, M.; Fabre, E.; Poirot, O.; Jeudy, S.; Lartigue, A.; Alempic, J.-M.; Beucher, L.; Philippe, N.; Bertaux, L.; Christo-Foroux, E.; et al. Diversity and evolution of the emerging Pandoraviridae family. *Nat. Commun.* **2018**, *9*, 2285. [CrossRef]
52. Christo-Foroux, E.; Alempic, J.-M.; Lartigue, A.; Santini, S.; Labadie, K.; Legendre, M.; Abergel, C.; Claverie, J.-M. Characterization of Mollivirus kamchatka, the First Modern Representative of the Proposed Molliviridae Family of Giant Viruses. *J. Virol.* **2020**, *94*. [CrossRef]
53. 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* **2011**, *108*, 17486–17491. [CrossRef]
54. McLysaght, A.; Baldi, P.F.; Gaut, B.S. Extensive gene gain associated with adaptive evolution of poxviruses. *Proc. Natl. Acad. Sci. USA* **2003**, *100*, 15655–15660. [CrossRef]
55. Suhre, K.; Audic, S.; Claverie, J.-M. Mimivirus gene promoters exhibit an unprecedented conservation among all eukaryotes. *Proc. Natl. Acad. Sci. USA* **2005**, *102*, 14689–14693. [CrossRef]
56. Legendre, M.; Audic, S.; Poirot, O.; Hingamp, P.; Seltzer, V.; Byrne, D.; Lartigue, A.; Lescot, M.; Bernadac, A.; Poulain, J.; et al. mRNA deep sequencing reveals 75 new genes and a complex transcriptional landscape in Mimivirus. *Genome Res.* **2010**, *20*, 664–674. [CrossRef]
57. Lato, D.F.; Golding, G.B. Spatial Patterns of Gene Expression in Bacterial Genomes. *J. Mol. Evol.* **2020**, *88*, 510–520. [CrossRef]
58. Couturier, E.; Rocha, E.P.C. Replication-associated gene dosage effects shape the genomes of fast-growing bacteria but only for transcription and translation genes. *Mol. Microbiol.* **2006**, *59*, 1506–1518. [CrossRef] [PubMed]
59. Rocha, E.P.C. The replication-related organization of bacterial genomes. *Microbiology* **2004**, *150*, 1609–1627. [CrossRef]
60. Kelman, L.M.; Kelman, Z. Multiple origins of replication in archaea. *Trends Microbiol.* **2004**, *12*, 399–401. [CrossRef]
61. Senkevich, T.G.; Bruno, D.; Martens, C.; Porcella, S.F.; Wolf, Y.I.; Moss, B. Mapping vaccinia virus DNA replication origins at nucleotide level by deep sequencing. *Proc. Natl. Acad. Sci. USA* **2015**, *112*, 10908–10913. [CrossRef]

62. Schulz, F.; Roux, S.; Paez-Espino, D.; Jungbluth, S.; Walsh, D.A.; Denef, V.J.; McMahon, K.D.; Konstantinidis, K.T.; Eloe-Fadrosh, E.A.; Kyrpides, N.C.; et al. Giant virus diversity and host interactions through global metagenomics. *Nature* **2020**, *578*, 432–436. [CrossRef] [PubMed]