Mieux connaître les virus présents sur Terre grâce aux métagénomes

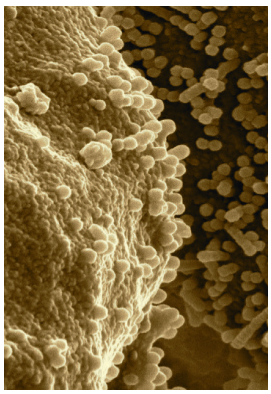
Éric Olo Ndela, Louis-Marie Cobigo, Simon Roux, François Enault

médecine/sciences 2022 ; 38 : 999-1007

Mieux connaître les virus présents sur Terre grâce aux métagénomes

Éric Olo Ndela[1], Louis-Marie Cobigo[1], Simon Roux[2], François Enault[1]

> En dépit de leur très grand nombre, les virus qui peuplent l'environnement restent largement méconnus. Les approches de métagénomique ont permis depuis vingt ans de mieux connaître la composition des communautés virales naturelles, notamment les groupes viraux les plus fréquemment trouvés, et de lever peu à peu le voile sur l'étendue de leur diversité, révélant le grand nombre d'espèces, de genres et même de familles virales, pour la plupart identifiés pour la première fois. Au sein de ces groupes, le contenu en gènes, les hôtes infectés et les écosystèmes habités sont souvent cohérents avec l'histoire évolutive, reflet de l'origine très ancienne des virus et de leur très longue coévolution avec leurs hôtes, plus que de leur capacité à muter rapidement. <



[1] Université Clermont Auvergne, CNRS, LMGE, F-63000 Clermont-Ferrand, France
[2] *Department of Energy Joint Genome Institute, Lawrence Berkeley National Laboratory*, 1 Cyclotron Road, Berkeley, CA 94720, États-Unis.
francois.enault@uca.fr

Connaître le monde viral à travers les virus cultivés

Tous les organismes vivants connus peuvent être infectés par des virus. Aussi, partout où il y a de la vie, il y a des virus. Des particules virales libres (ou virions) peuvent donc être observées dans tous les écosystèmes, depuis les environnements associés à l'homme (microbiote humain, environnements impactés par l'activité humaine) aux milieux naturels (sol, milieux lacustres, océans), y compris les environnements les plus extrêmes (milieux hypersalins, fortement acides, à température élevée, etc.). Les virus sont considérés comme les entités biologiques les plus abondantes de la biosphère, avec un nombre estimé à 10^{31} [1]. Même si différentes morphologies existent (par exemple en forme d'ampoule, de goutte ou encore de bacille), les virus les plus fréquemment observés dans la plupart des écosystèmes par microscopie électronique à transmission sont des virions icosaédriques (polyèdres réguliers à 20 faces), avec ou sans queue. Bien qu'une majorité de microbes, et donc de virus qui les infectent, restent encore incultivables, les approches d'isolement et de culture de souches virales ont permis, au cours des cinquante dernières années, de caractériser des milliers de virus et de séquencer leur génome. Ces génomes viraux sont très variables en termes de support, de taille, ainsi que de contenu en gènes. Pas moins de sept supports génétiques différents ont en effet été décrits pour les génomes viraux (ADN ou ARN double brin [db], simple brin [sb], sens et antisens pour l'ARN, et rétro-transcrit), et ces génomes, circulaires ou linéaires, segmentés ou non, ont une taille très variable, allant de 1 680 à 2,5 millions de nucléotides. Les critères morphologiques ne permettant pas de les classer en groupes monophylétiques, les virus sont depuis peu classés principalement selon leur génome et les gènes qu'ils codent. L'analyse de ces génomes a permis d'établir une taxonomie complète de ces virus, et une différence majeure par rapport aux organismes cellulaires (bactéries, archées et eucaryotes) est l'existence de nombreux groupes de virus n'ayant aucun gène en commun [2]. Ces groupes, apparus de manière indépendante, ne contiennent pour la plupart que des virus infectant des hôtes appartenant à un seul des trois domaines du vivant (archées, bactéries, eucaryotes), reflet de leur spécificité d'hôte. Ces différents virus sont *a priori* anciens et leur histoire évolutive commune avec leurs hôtes longue. Les nombreux événements de spéciation ou d'adaptation de leurs hôtes à différentes niches écologiques ont ainsi créé des barrières entre organismes cellulaires, et donc entre leurs virus, favorisant l'évolution parallèle de très nombreux sous-groupes pour chaque grande lignée virale. Cela se traduit par une diversité génomique et génétique des virus très grande [3]. Ainsi les milliers de virus cultivés sont très largement insuffisants pour que nous ayons une vision représentative du monde viral présent sur la Terre.

Vignette (© Philippe Roingeard).

Étudier les virus présents dans l'environnement

Pour mieux connaître les virus présents sur la Terre, il a fallu développer des méthodes contournant l'étape initiale de mise en culture. Historiquement, les premières analyses des communautés virales présentes dans un environnement ont été réalisées par des observations en microscopie électronique associées à des comptages. Cette méthode étant limitée à la caractérisation de la morphologie des virus, des approches d'écologie moléculaire ont été par la suite utilisées afin d'accéder à la diversité génétique présente dans une communauté virale, notamment par la création de profils génétiques (par électrophorèse sur gel en gradient dénaturant ou DGGE, pour *denaturing gradient gel electrophoresis,* ou par électrophorèse en champs pulsés ou PFGE, pour *pulsed-field gel electrophoresis*). Cependant, ces approches globales ne donnant pas, non plus, accès à l'information génétique contenue dans les génomes viraux, des gènes ont été ciblés, amplifiés par PCR (*polymerase chain reaction*) puis séquencés pour mieux appréhender leur diversité et celle des organismes les abritant. Ces méthodes restent néanmoins limitées à l'étude de certaines familles virales puisqu'aucun gène n'est ubiquitaire et conservé dans l'ensemble des génomes viraux, comme c'est le cas pour le gène codant l'ARN ribosomique 16S des bactéries, utilisé pour les études de populations bactériennes. Ce type d'analyse ne peut donc pas être utilisé dans le cadre de recherches exploratoires, puisqu'il est nécessaire pour l'appliquer de disposer préalablement de séquences de référence d'un gène marqueur qui sera ciblé afin de pouvoir réaliser les amorces utilisées pour l'amplification par PCR.

En s'affranchissant des limites de la mise en culture ainsi que de la nécessité de connaissances préalables, la métagénomique s'est imposée depuis une vingtaine d'années comme la méthode la plus efficace pour explorer la diversité virale de l'environnement. Cette approche consiste à séquencer des fragments aléatoires de génomes issus d'un échantillon et permet de mieux connaître les virus présents dans un écosystème donné.

Obtenir des génomes viraux par métagénomique

Deux voies sont possibles pour étudier les virus par métagénomique (*Figure 1*). Le séquençage de matériel génétique peut en effet être réalisé soit en ciblant spécifiquement les virus, grâce à une filtration préalable, on parle alors de métagénome viral ou virome, soit sur l'ensemble des microbes (virus compris) : on parlera alors de métagénome microbien ou microbiome[1]. Il sera ensuite nécessaire de trier *in silico* et *a posteriori*, les génomes viraux des autres génomes de microbes.

Pour obtenir un virome, un échantillon est prélevé, par exemple quelques litres d'eau ou grammes de sols ou de fèces. Une filtration est ensuite effectuée afin d'éliminer les cellules « contaminantes » et de conserver la plupart des virus contenus dans l'échantillon. Généralement, la fraction utilisée pour la préparation d'un virome est la fraction dont le diamètre est inférieur à celui des pores du filtre utilisé, 0,45 µm, voire 0,2 µm, ce qui exclut *de facto* les virus géants qui présentent un diamètre nettement plus grand. Différentes techniques sont ensuite appliquées afin de concentrer les particules virales, notamment la filtration en flux tangentiel, la précipitation chimiques des capsides virales, par le polyéthylène glycol (PEG) ou par floculation au fer, ou encore la centrifugation différentielle. Des traitements, comme les gradients de densité de chlorure de césium ou de sucrose, permettent ensuite de purifier les capsides virales et de les séparer des matériaux et éléments « contaminants » qui ont été co-concentrés. Enfin, différents traitements enzymatiques (DNAse et RNAse) permettront d'éliminer le matériel génétique resté libre, hors des capsides. Les particules virales ainsi obtenues peuvent alors être ouvertes par choc thermique afin de libérer le matériel génétique qu'elles contiennent. Fortement enrichi, voire exclusivement constitué du génome des virus qui est par nature court, l'ADN récupéré sera amplifié afin d'en obtenir une quantité suffisante. Dans la majorité des cas, seules les molécules d'ADN, simple-brin et double-brin, sont ciblées lors de cette étape d'amplification. En effet, les virus présents dans l'environnement sont principalement des virus de microbes, les procaryotes étant infectés très majoritairement par des virus à ADN.

Dans la grande majorité des cas, l'amplification de l'ADN viral est désormais réalisée par une technique dite de déplacements multiples (ou en anglais, MDA pour *Multiple Displacement Amplification*). Cette méthode, fondée sur les propriétés particulières d'une ADN polymérase issue du bactériophage Φ29 (activité de polymérisation très rapide, grande fidélité de recopiage, activité de déplacement de brin) permet d'amplifier rapidement des quantités très faibles d'ADN, amplifiant aléatoirement la totalité de l'ADN présent dans l'échantillon, ce qui n'est pas le cas des techniques classiques d'amplification par PCR qui nécessitent des amorces spécifiques et ciblent donc une séquence particulière.

Au cours de ces dernières années, différentes techniques permettant de séquencer l'ADN viral amplifié se sont succédées. Actuellement, la méthode la plus utilisée permet d'obtenir, pour chaque échantillon, des dizaines de millions de copies de séquences courtes d'ADN (ou lectures). S'ensuit alors un travail d'analyse de ces séquences par bioinformatique. Les pre-

[1] En français, on parle de microbiote pour l'ensemble des microbes sans définir les gènes qu'ils présentent. Le microbiome est quant à lui défini par les gènes exprimés par ces microbes.

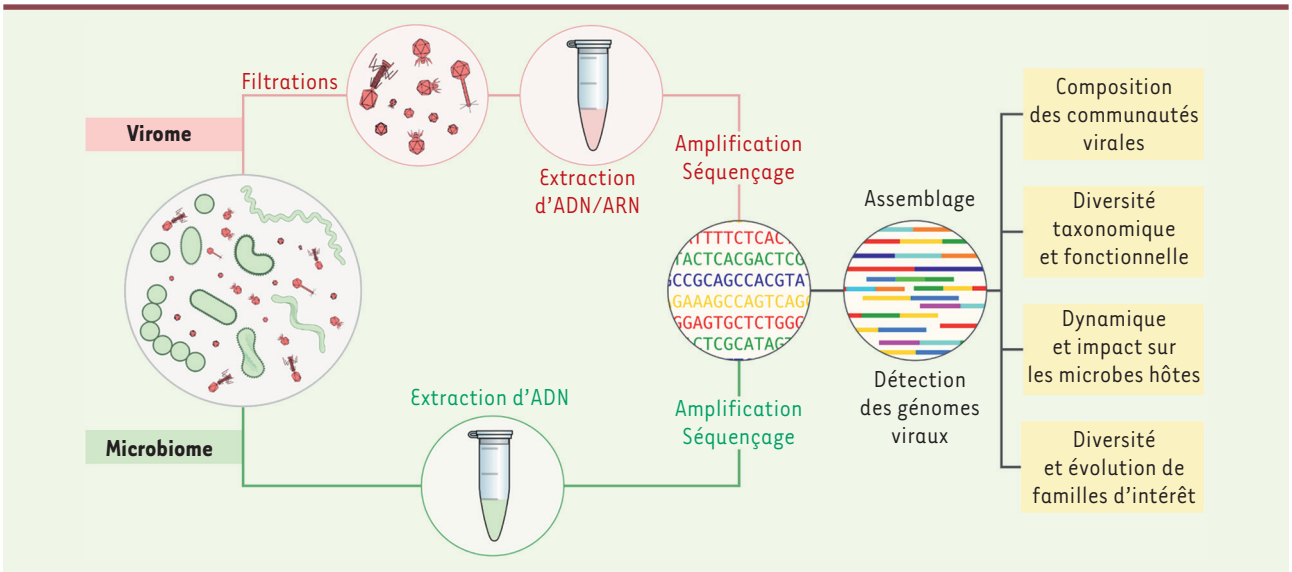


Figure 1. ***Schéma simplifié de l'étude des virus présents dans un écosystème par métagénomique.*** À partir d'un échantillon environnemental, un virome peut être obtenu en enrichissant la fraction virale par des étapes de filtration, le matériel génétique présent dans les particules virales ainsi récupérées étant ensuite extrait, amplifié puis séquencé. Il est également possible de séquencer tout le microbiome (virus inclus). Dans les deux cas, les séquences génétiques générées sont ensuite assemblées et les (fragments de) génomes viraux sont identifiés et séparés du reste des génomes microbiens. Ces génomes viraux permettent d'étudier la composition et la diversité de la communauté virale, d'éventuellement suivre sa dynamique, d'évaluer son impact sur les communautés microbiennes, ou encore d'étudier de manière approfondie des familles virales d'intérêt.

mières étapes de ce travail sont la suppression des lectures de mauvaise qualité, puis la réalisation d'un assemblage, ce qui consiste à mettre « bout-à-bout » les lectures courtes se chevauchant et ainsi reconstituer les fragments de génomes les plus longs possibles, parfois plusieurs centaines de milliers de nucléotides.

Les principaux avantages des métagénomes produits après enrichissement de virus sont l'accès aux virus abondants mais aussi aux virus rares. Ces métagénomes permettent également de produire une cartographie avec une confiance accrue quant à l'origine virale de la séquence examinée [4]. Ces données présentent cependant quelques limites puisqu'elles peuvent être à l'origine d'une surreprésentation des virus virulents (qui se répliquent rapidement dans le milieu, contrairement aux virus peu virulents), et surtout, une sous-représentation des grands virus qui ont de grandes capsides et ont été éliminés lors de l'étape de filtration.

De manière complémentaire, des séquences virales peuvent être assemblées dans des métagénomes microbiens, plus faciles à produire. Ces métagénomes sont constitués des génomes des virus en cours de réplication dans les cellules qu'ils ont infectées, mais aussi des virus intégrés dans les génomes de l'hôte sous forme de provirus, ou des virus présents en tant qu'épisomes dans leur cellule hôte, ou même des particules virales libres présentes dans les échantillons. L'analyse de ces microbiomes permet de détecter les infections lytiques, tempérées et persistantes. Elle permet également d'éviter les biais découlant de la sélection des grandes particules virales en fonction de leur taille lors de l'étape de filtration. Cependant, les virus qui seront détectés seront ceux qui ont infecté les microbes dominants dans l'échantillon. Les virus rares ou ceux qui infectent des hôtes rares resteront donc sous-représentés ou absents. Les comparaisons entre virome et microbiome suggèrent que ces différentes approches sont donc complémentaires et permettent une exploration exhaustive de l'espace des génomes des virus [4].

Identifier les séquences virales *in silico*

Qu'ils proviennent de microbiomes ou de viromes, l'origine virale des (fragments de) génomes doit être vérifiée car même les échantillons enrichis en particules virales contiennent souvent une quantité non négligeable d'ADN cellulaire « contaminant » [5]. Cette contamination des viromes par l'ADN cellulaire peut provenir (1) de la difficulté à séparer les particules virales des fractions cellulaires, par exemple à cause de la présence de microbes nanométriques (de diamètre inférieur à 0,4 nm) ; (2) de la difficulté à traiter les échantillons riches en particules minérales (sol et sédiment), l'ADN extracellulaire étant fixé et protégé par ces particules minérales ; (3) de la présence de vésicules membranaires produites par des microbes ; ou du fait que des particules virales contiennent de l'ADN cellulaire, les microbes échangeant en effet des fragments de leur patrimoine génétique par le biais de

virus actifs (transduction), ou de virus domestiqués (les agents de transfert de gènes). Bien évidemment, les métagénomes microbiens sont eux majoritairement constitués d'ADN cellulaire.

Plusieurs outils et protocoles bioinformatiques ont été mis au point pour identifier les séquences de virus [6, 7]. Pour pouvoir dire qu'une séquence d'ADN est bien un fragment de génome viral et non microbien, ces logiciels se fondent : 1) sur la présence de gènes caractéristiques de virus ; 2) sur un enrichissement en gènes similaires à des gènes de virus connus ; 3) sur des profils en oligonucléotides similaires à ceux de virus connus ; 4) sur une sous-représentation des gènes similaires à des génomes cellulaires connus ; ou 5) sur des caractéristiques structurelles typiques des génomes viraux, comme le peu de changements de brins pour les zones codantes, ou la présence de nombreux petits gènes. L'identification des génomes viraux intégrés au génome de leur hôte (ou provirus) et des bornes précises de ces génomes intégrés, est encore plus difficile. L'identification de ces provirus ne permet pas de distinguer ceux qui sont actifs (encore capables de se répliquer et de produire des virions) de ceux qui sont inactifs (trace d'une infection passée, en cours d'effacement du génome de l'hôte).

Des virus ubiquitaires dans les milieux naturels

Pour mieux définir la composition de la communauté virale d'un échantillon, les génomes (ou fragments de génomes) qui ont été obtenus seront comparés aux génomes de virus connus et référencés dans les bases de données comme la base RefSeq du *National Center for Biotechnology Information* (NCBI). Certains groupes de virus sont souvent dominants quel que soit le type d'écosystème examiné (*Figure 2*) [8-15]. En effet, la majorité des séquences générées sont similaires à des génomes de virus qui infectent des bactéries comme les *Caudoviricetes* (virus à queue à ADNdb) et les *Microviridae* (petits virus icosaédriques à ADNsb). Viennent ensuite des virus qui infectent les cellules eucaryotes, comme les *Cressdnaviricota* (petits virus icosaédriques à ADNsb contenant un gène conservé initiant leur réplication, comme les *Circo-*, *Nano-*, ou *Geminiviridae*), et les *Phycodnaviridae* (grands virus nucléocytoplasmiques).

Le fait de retrouver principalement des bactériophages (ou phages) dans les échantillons de milieux naturels n'est pas si surprenant : les bactéries peuplent en effet tous les écosystèmes et leur nombre est estimé à 10^{30} sur la Terre. Parmi ces phages, il est aussi logique de retrouver principalement des phages à queue, puisque ces derniers représentent la majorité des phages cultivés et qu'ils sont les plus fréquemment observés en microscopie. La présence ubiquitaire des phages de la famille des *Microviridae* et des *Cressdnaviricota* est moins attendue. Leur abondance, très importante dans de nombreux viromes examinés, reste ainsi un sujet de débat. En effet, l'ADN viral est amplifié avant séquençage, souvent réalisé par MDA, une méthode qui repose sur l'utilisation d'une enzyme qui amplifie préférentiellement les petits génomes circulaires à ADN simple brin, comme ceux des *Microviridae* ou des *Cressdnaviricota*. Les virus d'eucaryotes (*Cressdnaviricota* et *Phycodnaviridae*) qui sont observés, sont capables d'infecter une grande variété de cellules eucaryotes, mais ceux qui sont trouvés de manière abondante sous forme de virions dans ces écosystèmes infectent probablement des microbes, c'est-à-dire de petits eucaryotes unicellulaires, comme les diatomées, principaux membres du phytoplancton des milieux aquatiques. Comme nous l'avons vu, les grands virus à ADN nucléocytoplasmiques (NCLDV ou *Nucleocytoviricota*) sont trop grands pour être trouvés dans les viromes, mais une fouille de milliers de microbiomes a révélé leur présence récurrente [16].

Les virus d'archées, un exemple de génomes difficiles à identifier

Les virus d'archées sont retrouvés dans les milieux extrêmes, dans lesquels les archées sont souvent abondantes. Par exemple, des génomes similaires à ceux des virus en forme de citron, infectant des archées halophiles, ont été trouvés dans les viromes de milieux hypersalins [17]. Pourtant, l'étude des quelques (moins d'une centaine) virus d'archées cultivables montre que ces derniers sont plus diversifiés en termes de morphologie et de génome que les phages, dont la diversité est dominée par les phages à queue qui partagent tous plusieurs gènes. Ce manque de connaissance et l'hétérogénéité des génomes connus suggèrent qu'il existe de nouvelles familles de virus d'archées qui restent à découvrir. Mais comment identifier leurs génomes, puisqu'ils ne partagent sans doute que peu de gènes (voire aucun) avec les virus connus, que les rares gènes conservés sont très distants et difficiles à identifier, et que la morphologie de ces nouveaux virus est impossible à déterminer uniquement à partir de leur génome ? Ce problème n'est pas l'apanage des virus d'archées : des groupes de virus de bactéries ou d'eucaryotes restent également à découvrir [18]. Ces nouveaux virus n'ayant que des traces ténues de ressemblance (voire aucune ressemblance détectable) avec ceux déjà connus, ils sont difficiles à extraire des métagénomes et seront même, pour la plupart, exclus lors de l'étape de décontamination (élimination des séquences supposées contaminantes) qui repose sur les ressemblances avec des virus connus.

Quid des virus à ARN ?

Les virus à ARN ont jusqu'à présent été ciblés par peu d'études de métagénomique virale, principalement effectuées afin d'identifier des agents pathogènes

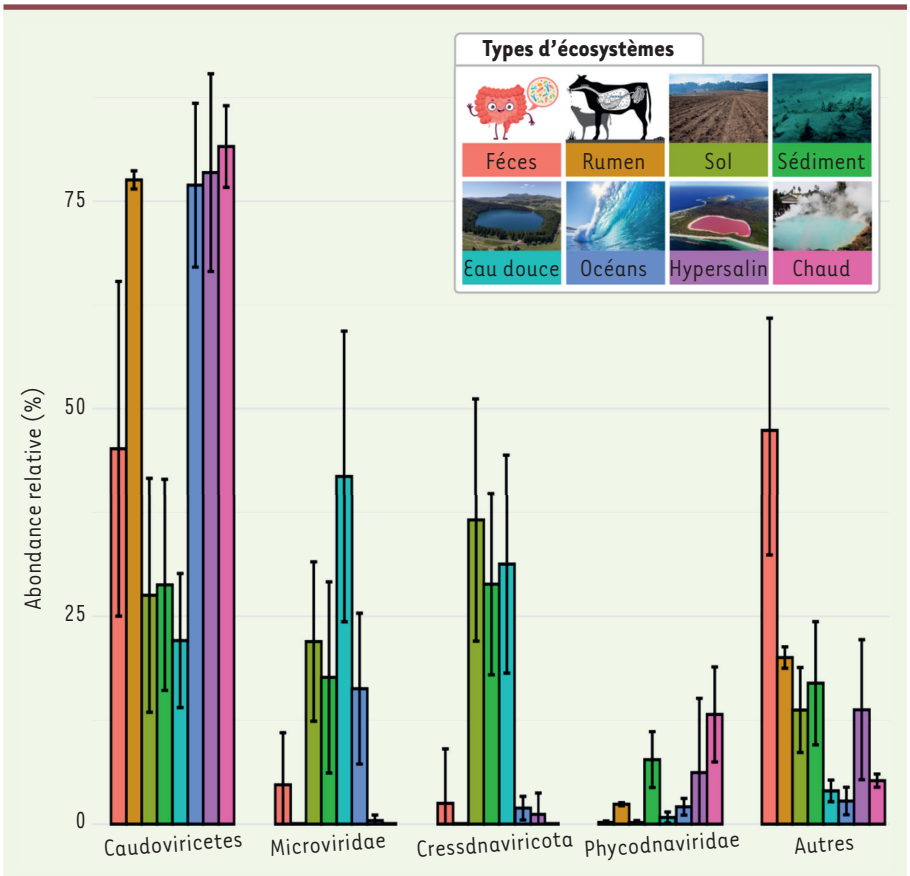


Figure 2. ***Abondance des groupes viraux fréquemment retrouvés dans différents écosystèmes.*** Abondance relative des quatre groupes viraux les plus fréquemment retrouvés dans les communautés virales présentes dans différents types d'écosystèmes : les fèces (représentatifs du tube digestif) de mammifères, comme le panda [8] ou le chien [9] ; le rumen de vache [10] ; les sols [11] ; l'eau douce de lacs [12] ou de rivières [13] ; les océans [14] ; ou des environnements considérés comme extrêmes avec, par exemple, les sédiments, les milieux hypersalins ou encore les sources hydrothermales [15].

chez l'homme, chez les plantes et les animaux importants pour l'agriculture. Aucun virus d'archées et très peu de phages cultivés sont constitués d'un génome à ARN et la diversité des virus à ARN dans l'environnement était jusqu'alors considérée comme relativement faible. Plusieurs études récentes ont cependant permis d'avoir une vision plus précise sur ces virus [19]. En utilisant le gène codant une polymérase à ARN dépendante de l'ARN (RdRp), un gène indispensable pour les virus à ARN et absent chez les autres virus et chez les cellules, plusieurs milliers d'espèces de virus à ARN ont pu être identifiées au sein de nombreux écosystèmes différents. Mieux comprendre la diversité réelle de ces virus à ARN devrait donc permettre d'identifier de nouveaux groupes à différents niveaux taxonomiques, et de retracer leur histoire évolutive ancienne, potentiellement jusqu'aux premières heures de l'existence des virus et des organismes cellulaires sur la Terre.

Un monde viral fait de millions d'espèces encore mal caractérisées

La répartition des génomes obtenus dans les études, dans des groupes qui soient cohérents et qui partagent des caractéristiques distinctives, est une étape essentielle pour mieux les décrire et les comprendre, et cela à différents niveaux taxonomiques.
Commençons par l'échelle la plus fine : l'espèce.

Même si la notion d'espèce est controversée pour les eucaryotes, pour les procaryotes et encore plus pour les virus, il a été montré que les flux de gènes et la sélection définissent des frontières claires entre des groupes de virus connus [20]. Des études fondées sur l'analyse de métagénomes confirment l'existence de ces groupes dans lesquels une certaine homogénéité existe et qui restent isolés des autres groupes [21]. Ces *espèces*, aussi appelées populations ou unités taxonomiques opérationnelles virales (*viral Operational Taxonomic Units* ou vOTU), sont déterminées en regroupant les génomes pour lesquels 95 % des nucléotides sont identiques, et cela sur la majorité de leur longueur (plus de 85 % de couverture réciproque) [4]. De la même manière, une identité nucléotidique moyenne supérieure à 70 % est utilisée pour définir les genres viraux. À partir de ces seuils, les études récentes ont mis en évidence l'étendue vertigineuse de la diversité virale. Ainsi, en 2016, la fouille de milliers de microbiomes issus d'écosystèmes variés a permis l'identification de 125 000 génomes viraux partiels, regroupés en plus de 80 000 espèces, dont seuls 800 contiennent au moins un virus isolé [22]. Encore plus récemment, 200 000 espèces virales ont été identifiées dans les océans [21], 54 000 espèces et 6 000 genres dans le tube digestif humain [23], et 4 000 espèces dans des tourbières [24]. Ces chiffres sont bien sûr dépendants du nombre de génomes obtenus, donc de la quantité d'échantillons et de la profondeur des séquençages. Pour les écosystèmes à partir desquels

peu ou aucun virus n'a été cultivé, peu de nouveaux génomes se sont révélés proches de virus cultivés : par exemple, seules 61 des plus de 4 000 espèces virales issues de tourbières contiennent un virus connu, illustrant le manque de connaissance que l'on a des virus peuplant ce type d'écosystème [24].

Les niveaux taxonomiques supérieurs sont eux aussi définis uniquement à partir de l'information génomique (contenu en gènes, phylogénies de gènes marqueurs, etc.), comme le montre la décision récente de ne plus utiliser les familles *sipho-*, *myo* – et *podo-viridae* au sein des *Caudoviricetes*, puisque ces différentes familles avaient été établies sur des critères morphologiques (respectivement, longue queue non contractile, queue contractile et queue courte) et n'étaient pas monophylétiques [25]. Définir des nouvelles familles, des nouvelles classes ou même de nouveaux ordres et phylums à partir de données métagénomiques ne peut se faire qu'à partir de génomes (quasi) complets, et aucune méthode n'est actuellement utilisée de manière systématique. Toutefois, si l'on reprend l'exemple des virus de tourbières, seules 2 % des 4 000 espèces sont rattachées à des virus cultivés au sein de groupes équivalents à des familles, 25 % des espèces sont rattachées à au moins un des centaines de milliers de génomes issus de métagénomes d'autres types d'écosystèmes, et 73 % de ces 4 000 espèces restent très distantes de tout génome cultivé ou métagénomique. Ce résultat est assez classique, et même au sein de l'écosystème le plus étudié, le tube digestif humain, et dans le groupe le plus connus, les *Caudoviricetes*, une majorité des 1 400 familles établies à partir des 54 000 espèces représentent des familles qui n'avaient pas encore été détectées [23]. Plusieurs millions d'espèces virales sont ainsi probablement présentes sur la Terre et une majorité d'entre elles restent à découvrir, ce qui est également le cas pour les familles de virus. Et il est bien sûr probable que de nombreux groupes d'un niveau taxonomique supérieur (classe, ordre, et même phylum) restent également inconnus.

Les gènes viraux sont diversifiés mais le contenu en gènes des virus est stable

Les génomes viraux sont souvent décrits comme le plus grand réservoir de matériel génétique inexploré sur la Terre. En effet, pour les 125 000 génomes viraux issus d'écosystèmes variés, 75 % des 2,8 millions de protéines qu'ils codent n'ont aucune similarité avec des protéines de virus cultivés [22] et sont diverses, puisque regroupées en plus de 1,2 million de groupes d'orthologues. Les méthodes bioinformatiques classiques ne sont pas suffisamment sensibles pour détecter les similarités souvent faibles entre protéines virales [3]. De nombreuses familles de protéines virales semblent en effet plus hétérogènes que les familles de protéines microbiennes, peut-être à cause des taux de mutations élevés des virus, de leur origine très ancienne, ou plus probablement à cause de contraintes plus faibles sur leurs structures tridimensionnelles. L'univers des familles de protéines virales, estimé à quatre millions [26], sera probablement revu à la baisse en utilisant des méthodes fondées sur les modèles de Markov cachés[2] (ou HMM), plus sensibles pour identifier les ressemblances entre protéines. Pour les virus complets, ces méthodes ont permis de montrer qu'en fait, moins de 40 000 familles de protéines existaient au sein des virus infectant les procaryotes, et non 60 000, comme présumé antérieurement [3]. Néanmoins, l'ensemble des gènes présents dans les génomes viraux est probablement extrêmement vaste puisque de nombreux gènes cellulaires sont intégrés, de manière pérenne, dans des génomes viraux, et que de nombreuses familles virales existent, dont beaucoup restent à découvrir, celles-ci étant constituées de gènes différents. Au sein de chaque famille de virus, les gènes sont en revanche souvent conservés, même si leur lien de parenté est parfois difficile à établir. Ce dernier point est d'ailleurs révélé par les résultats très différents obtenus en fonction de l'écosystème étudié. En effet, les génomes identifiés dans des écosystèmes peu étudiés, et pour lesquels peu de virus ont pu être cultivés (tels que le sol, les tourbières, etc.), contiennent de nombreux gènes qui ne sont similaires à aucun gène connu, contrairement à ceux issus d'échantillons océaniques ou de fèces humaines, ce qui témoignent du manque de connaissance actuelle sur les communautés virales de nombreux environnements et sur la diversité génétique des familles de protéines virales.

Étude de familles d'intérêt grâce aux génomes complets

Si l'étude des communautés virales et de l'ensemble des gènes qu'elles portent est intéressante, mieux comprendre la diversité et l'histoire évolutive de familles particulières de virus à partir de données métagénomiques reste crucial. Cela peut être réalisé par une analyse phylogénétique reposant sur des gènes marqueurs, c'est-à-dire des gènes que tous les virus d'un groupe possèdent. Par exemple, tous les *Caudoviricetes* bactériens et archéens contiennent les gènes qui codent la grande sous-unité de la terminase (TerL), la protéine majeure de capside de type HK97, et la protéine portail (portal). Les *Microviridae* possèdent tous, quant à eux, une protéine de capside et une protéine initiant leur réplication, cette dernière étant également conservée chez tous les *Cressdnaviricota*. Les relations entre virus ainsi établies grâce à cette phylogénie, voire la

[2] Les modèles de Markov cachés sont des outils mathématiques très utilisés en bioinformatique, permettant ici de modéliser un alignement multiple de séquences homologues puis de vérifier si une nouvelle séquence correspond à ce modèle ou non et fait donc également partie de cette famille protéique.

définition de clades[3], pourront être ensuite utilisées pour étudier la conservation du contenu en gènes au sein de ces clades.

La fouille récente de métagénomes issus de divers écosystèmes a, par exemple, permis la découverte de centaines de génomes de phage « géants », d'une longueur supérieure à 200 kilobases (kb), en particulier la découverte du plus grand génome de phage observé jusqu'à présent (735 kb) [27]. Les génomes de ces *Caudoviricetes* codent des systèmes CRISPR-Cas, typiquement utilisés par les bactéries et archées pour se défendre contre les virus. Ces systèmes CRISPR-Cas codés par des virus sans doute utilisés pour reconfigurer la machinerie cellulaire, favoriseraient la production de protéines de phages, ou encore éliminerait des phages concurrents. La plupart des clades définis par la phylogénie, réalisée sur la base de la terminase TerL et de la protéine de capside, intègrent des phages « géants » issus de divers écosystèmes et infectant différents hôtes bactériens. Le fait que les phages de ces clades particuliers ont une taille de génome comparable, suggère que ces phages sont anciens et que leur grand génome est une caractéristique qui a été conservée au cours de l'évolution.

L'assemblage de génomes (quasi) complets est encore plus aisé pour les virus constitués d'un génome de petite taille, en particulier les *Microviridae*, fréquemment identifiés dans les métagénomes [28], et dont le génome est formé d'une molécule d'ADN simple-brin circulaire d'environ 5 kb. Plusieurs virus de cette famille ont été décrits à partir de culture d'*Escherichia coli*, de bactéries pathogènes de type *Chlamydia*, ou encore de bactéries marines du type *Citromicrobium*. D'autres ont été découverts intégrés sous forme de prophages dans des génomes de *Bacteroidetes*, des bactéries commensales de la flore intestinale humaine. De nombreux génomes complets de *Microviridae* sont souvent assemblés à partir de métagénomes, comme par exemple les 598 nouveaux génomes issus de 78 viromes isolés de fèces et de tissus échantillonnés chez divers mammifères, oiseaux, poissons et crustacés [29]. La diversité de ces génomes est grande, chacun des 598 nouveaux génomes formant une nouvelle espèce, l'ensemble se répartissant en 566 genres différents. Une phylogénie fondée sur la protéine de capside et incluant 98 de ces nouveaux génomes pris au hasard montre que certains des nouveaux génomes sont proches de génomes de référence, mais que de nouveaux groupes monophylétiques encore inconnus se dessinent (*Figure 3*). Certains des nouveaux génomes assemblés sont représentés par de longues branches séparées des autres génomes, indiquant, là encore, une diversité très vaste et encore mal connue. La plupart des caractéristiques de ces virus sont cohérentes au sein des clades définis par phylogénie. En effet, les *Microviridae* qui sont proches ont des taux de GC similaires ; ils proviennent d'écosystèmes similaires ; ils codent les mêmes gènes, l'ordre de ceux-ci étant de plus conservé le long du génome ; et ils infectent des hôtes bactériens qui sont proches.

[3] Un clade, aussi appelé groupe monophylétique, est un groupe d'organismes, vivants ou ayant vécu, comprenant un organisme particulier et la totalité de ses descendants.

Alors que les trois premiers caractères spécifiques de ces virus peuvent être facilement étudiés sur des séquences obtenues à partir d'échantillons de milieu environnemental, l'étude portant sur les hôtes qu'ils infectent représente un obstacle majeur. La solution *in silico* la plus simple et la plus fiable est la recherche de prophages, dont la séquence est très similaire à celle des particules présentes dans l'environnement mais intégrée dans le génome d'une bactérie connue. D'autres méthodes permettent de prédire l'hôte selon chaque séquence métagénomique, en cherchant des *spacers* CRISPR similaires, trace d'une infection antérieure dans le génome bactérien, ou en cherchant un génome microbien ayant une signature génomique similaire, témoin d'une adaptation du génome viral à son hôte [30].

En ce qui concerne les 598 *Microviridae* auxquels nous avons fait référence, aucun hôte n'a pu être prédit, sans doute du fait que les bases de données de génomes microbiens indiquent actuellement que ces virus infectent des souches de microbes non cultivables. Au vu des groupes monophylétiques (*Figure 3*), les virus semblent se répartir par type d'écosystèmes, indépendamment des localisations géographiques des échantillons prélevés, sans doute le reflet de l'ancestralité de ces virus, de leur spécificité à leur hôte, et des échanges entre ces communautés microbiennes au niveau mondial.

Conclusion

À côté des études de macrodiversité, les données de séquençage devraient permettre, par leur volume et la taille des séquences générées, d'étudier la microdiversité présente au sein de chaque espèce virale. Elles devraient également permettre de mieux prédire les hôtes que ces virus infectent à partir de leur génome. Relier les virus identifiés par métagénomique, et leur(s) hôte(s), sur la seule base de leur séquence génomique, reste, à l'heure actuelle, encore limité par le manque de techniques fiables et les biais de représentation pour les virus et les microbes dans les bases de données génomiques [4]. L'intégration entre approches expérimentales, complémentaires des analyses métagénomiques, telles que la génomique sur « cellule-unique » et la fluorescence *in situ*, sera très certainement nécessaire pour reconstruire de manière plus complète les réseaux entre virus et hôtes, dans différents environnements [31]. Cela permettra de pouvoir suivre et mesurer l'impact des communautés virales sur les communautés microbiennes et la dynamique conjointe de ces communautés, objectif central de l'écologie virale. ◊

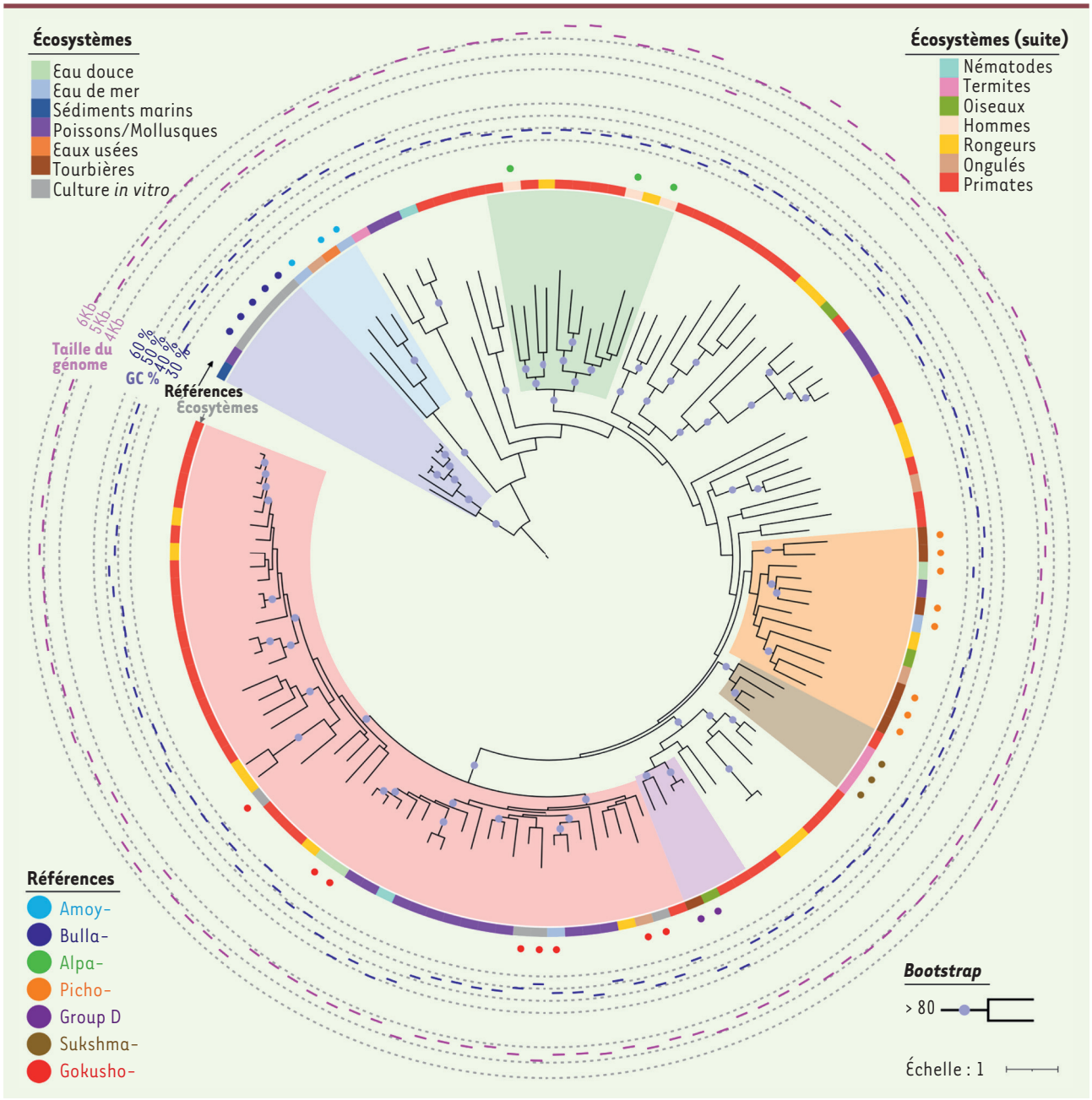


Figure 3. ***Phylogénie de 130 génomes de microvirus.*** Cette phylogénie a été réalisée à partir de la protéine de capside de 32 microvirus de références, classés précédemment dans une sous-famille connue, et 98 génomes complets assemblés à partir de 78 viromes associés à divers hôtes animaux [29]. Une valeur de « *bootstrap* » (pourcentage de 0 à 100 qui évalue la robustesse) est associée à chaque branche de l'arbre indiquant le nombre de fois où cette branche a été retrouvée lorsque l'on utilise différentes positions de l'alignement des protéines de capside. Les branches ayant une valeur de « *bootstrap* » supérieure à 80 sont indiquées avec un cercle.

À l'extérieur de la phylogénie sont présentées quatre pistes : les écosystèmes dans lesquels chaque génome a été échantillonné ; les génomes de référence, indiqués par des cercles colorés selon la sous-famille ; le pourcentage en nucléotides G et C des génomes ; et la taille de chacun des génomes. Les zones colorées indiquent les groupes monophylétiques contenant un ou plusieurs génomes de référence et éventuellement des séquences environnementales.

SUMMARY

A better understanding of Earth's viruses thanks to metagenomes

Despite their large number, viruses present in the environment remain largely unknown. Metagenomic approaches, targeting viruses specifically or not, have allowed us a better understanding of the composition of natural viral communities, with *Caudoviricetes*, *Microviridae*, *Cressdnaviricota* or *Phycodnaviridae* being the most frequently found viral groups. Metagenomes are gradually revealing the extent of the diversity of these groups and their structure, highlighting the large number of species, genera and even viral families, most of which being seen for the first time. Within these groups, the gene content, infected hosts and inhabited ecosystems are often consistent with the evolutionary history traced with marker genes. Thus, the diversity of viruses and their genes is more a reflection of their ancient origin and long coevolution with their hosts than of their ability to mutate rapidly. ◊

LIENS D'INTÉRÊT

Les auteurs déclarent n'avoir aucun lien d'intérêt concernant les données publiées dans cet article.

RÉFÉRENCES

1. Mushegian AR. Are there 1031 virus particles on earth, or more, or fewer? *J Bacteriol* 2020.
2. Koonin EV, Dolja VV, Krupovic M, *et al.* global organization and proposed megataxonomy of the virus world. *Microbiol Mol Biol Rev* 2020 ; 84 : e00061-19.
3. Terzian P, Olo Ndela E, Galiez C, *et al.* PHROG: families of prokaryotic virus proteins clustered using remote homology. *NAR Genom Bioinform* 2021 ; 3 : lqab067.
4. Roux S, Adriaenssens EM, Dutilh BE, *et al.* minimum information about an uncultivated virus genome (MIUViG). *Nat Biotechnol* 2019 ; 37 : 29-37.
5. Roux S, Krupovic M, Debroas D, *et al.* Assessment of viral community functional potential from viral metagenomes may be hampered by contamination with cellular sequences. *Open Biol* 2013 ; 3 : 130160.
6. Roux S, Enault F, Hurwitz BL, *et al.* VirSorter: mining viral signal from microbial genomic data. *PeerJ* 2015 ; 3: e985.
7. Guo J, Bolduc B, Zayed AA, *et al.* VirSorter2: a multi-classifier, expert-guided approach to detect diverse DNA and RNA viruses. *Microbiome* 2021 ; 9 : 37.
8. Ning S, Lu X, Zhao M, *et al.* Virome in fecal samples from wild giant pandas (Ailuropoda Melanoleuca). *Front Vet Sci* 2021 ; 8 : 767494.
9. Shi Y, Tao J, Li B, *et al.* the gut viral metagenome analysis of domestic dogs captures snapshot of viral diversity and potential risk of coronavirus. *Front Vet Sci* 2021 ; 8 : 695088.
10. Berg Miller ME, Yeoman CJ, Chia N, *et al.* Phage-bacteria relationships and CRISPR elements revealed by a metagenomic survey of the rumen microbiome. *Environ Microbiol* 2012 ; 14 : 207-27.
11. Han LL, Yu DT, Bi L, *et al.* Distribution of soil viruses across China and their potential role in phosphorous metabolism. *Environ Microbiome* 2022 ; 17 : 6.
12. Roux S, Enault F, Robin A, *et al.* Assessing the diversity and specificity of two freshwater viral communities through metagenomics. *PLoS One* 2012 ; 7 : e33641.
13. Colombo S, Arioli S, Neri E, *et al.* Viromes as genetic reservoir for the microbial communities in aquatic environments: A focus on antimicrobial-resistance genes. *Front Microbiol* 2017 ; 8 : 1095.
14. Kim Y, Aw TG, Rose JB. Transporting ocean viromes: invasion of the aquatic biosphere. *PLoS One* 2016 ; 11 : e0152671.
15. Dávila-Ramos S, Castelán-Sánchez HG, Martínez-Ávila L, *et al.* A review on viral metagenomics in extreme environments. *Front Microbiol* 2019 ; 10 : 2403.
16. Schulz F, Roux S, Paez-Espino D, *et al.* Giant virus diversity and host interactions through global metagenomics. *Nature* 2020 ; 578 : 432-6.
17. Roux S, Enault F, Ravet V, *et al.* Analysis of metagenomic data reveals common features of halophilic viral communities across continents. *Environ Microbiol* 2016 ; 18 : 889-903.
18. Aevarsson A, Kaczorowska AK, Adalsteinsson BT, *et al.* Going to extremes – a metagenomic journey into the dark matter of life. *FEMS Microbiol Lett* 2021 ; 368 : fnab067.
19. Zayed AA, Wainaina JM, Dominguez-Huerta G, *et al.* Cryptic and abundant marine viruses at the evolutionary origins of Earth's RNA virome. *Science* 2022 ; 376 : 156-62.
20. Bobay LM, Ochman H. Biological species in the viral world. *Proc Natl Acad Sci USA* 2018 ; 115(23) : 6040-5.
21. Gregory AC, Zayed AA, Conceição-Neto N, *et al.* Marine DNA viral macro – and microdiversity from pole to pole. *Cell* 2019 ; 177 : 1109-23.e14.
22. Paez-Espino D, Eloe-Fadrosh EA, Pavlopoulos GA, *et al.* Uncovering Earth's virome. *Nature* 2016 ; 536 : 425-30.
23. Nayfach S, Páez-Espino D, Call L, *et al.* Metagenomic compendium of 189,680 DNA viruses from the human gut microbiome. *Nat Microbiol* 2021 ; 6 : 960-9 70.
24. Ter Horst AM, Santos-Medellín C, Sorensen JW, *et al.* Minnesota peat viromes reveal terrestrial and aquatic niche partitioning for local and global viral populations. *Microbiome* 2021 ; 9 : 233.
25. Barylski J, Enault F, Dutilh BE, *et al.* Analysis of Spounaviruses as a case study for the overdue reclassification of tailed phages. *Syst Biol* 2020 ; 69 : 110-23.
26. Ignacio-Espinoza JC, Solonenko SA, Sullivan MB. The global virome: Not as big as we thought? *Curr Opin Virol* 2013 ; 3 : 566-71.
27. Al-Shayeb B, Sachdeva R, Chen LX, *et al.* Clades of huge phages from across Earth's ecosystems. *Nature* 2020 ; 578 : 425-31.
28. Roux S, Krupovic M, Poulet A, *et al.* Evolution and diversity of the Microviridae viral family through a collection of 81 new complete genomes assembled from virome reads. *PLoS One* 2012 ; 7 : e40418.
29. Tisza MJ, Pastrana DV, Welch NL, *et al.* Discovery of several thousand highly diverse circular DNA viruses. *Elife* 2020 ; 9 : e51971.
30. Galiez C, Siebert M, Enault F, *et al.* WIsH: who is the host? Predicting prokaryotic hosts from metagenomic phage contigs. *Bioinformatics* 2017 ; 33 : 3113-4.
31. Coclet C, Roux S. Global overview and major challenges of host prediction methods for uncultivated phages. *Curr Opin Virol* 2021 ; 49 : 117-26.

TIRÉS À PART

F. Enault