Déchiffrer le fonctionnement des communautés microbiennes de la méthanisation, et leurs virus, en combinant écologie moléculaire et métagénomique

Ariane Bize

HAL Id: tel-04317554
https://hal.inrae.fr/tel-04317554

Submitted on 1 Dec 2023

université
PARIS-SACLAY

Déchiffrer le fonctionnement des communautés microbiennes de la méthanisation, et leurs virus, en combinant écologie moléculaire et métagénomique

**Habilitation à diriger des recherches
de l'Université Paris-Saclay**

présentée et soutenue à Antony, le 16 mai 2023, par

Ariane BIZE

Composition du jury

Didier DEBROAS Professeur, Université Clermont Auvergne	Rapporteur
Nathalie DESMASURES Professeure, Université de Caen Normandie	Rapportrice
Jérôme HAMELIN Directeur de recherche, INRAE	Rapporteur
Ludwig JARDILLIER Professeur, Université Paris-Saclay	Examinateur
Emilie MULLER Chargée de recherche, Université de Strasbourg	Examinatrice

Remerciements

Je remercie vivement Didier Debroas, Nathalie Desmasures et Jérôme Hamelin d'avoir accepté de rapporter mes travaux, ainsi que Ludwig Jardillier et Emilie Muller qui ont bien voulu les examiner. Je suis très honorée de leur participation à mon jury d'habilitation à diriger des recherches.

Je remercie particulièrement Théodore Bouchez, Jean-Philippe Torterotot et Gérard Sachon, qui ont permis mon arrivée au Cemagref en 2008. Je suis très reconnaissante envers Olivier Chapleur, Laurent Mazéas et à nouveau Théodore Bouchez, pour leur accueil lors de mon arrivée dans l'unité. Ce trio qui décryptait alors le fonctionnement des communautés microbiennes de la méthanisation m'a grandement aidée à m'intégrer et à développer ma propre thématique de recherche. J'ai bénéficié de leurs conseils judicieux, du partage de leurs connaissances et de leurs méthodes, des discussions scientifiques, et de leur soutien continu. Tout ceci perdure encore aujourd'hui et constitue pour moi un environnement très structurant.

Je voudrais plus largement remercier l'ensemble des personnes qui ont rendu ces travaux possibles, notamment les membres du pôle analytique de l'unité PROSE : Cédric Midoux en bioinformatique, Véronique Jamilloux en informatique, Chrystelle Bureau en biologie moléculaire, Angeline Guenne et Nadine Derlet en chimie analytique, et par le passé Anne Goubet et Céline Madigou. Leur implication et leur disponibilité sont très précieuses. Je remercie de plus les étudiants que j'ai eu le plaisir d'encadrer ou co-encadrer, en particulier les anciens et actuels doctorants, Nelly Badalato, Andreia Salvador, Hoang Ngo, Franciele Camargo, Lays Leonel et Marion Covès, et les anciens post-doctorants Lü Fan, Sabine Podmirseg, Hao Liping et Tiago Delforno. J'exprime plus généralement ma reconnaissance envers l'ensemble des personnels de l'unité PROSE, où règne une atmosphère conviviale et d'entraide.

Je pense également aux collaborateurs scientifiques, en particulier Mahendra Mariadassou, que je remercie pour ses conseils avisés en statistiques, ainsi que Marie-Agnès Petit, François Enault, et Mart Krupovic, avec qui les échanges sur les virus et l'analyse de leurs génomes sont très enrichissants. Je tiens également à remercier Céline Roose-Amsaleg, pour notre collaboration très agréable initiée il y a plusieurs années et pour ses conseils. Son attitude toujours constructive est très motivante. Je remercie Ludwig Jardillier et Maria Ciobanu, pour les discussions scientifiques stimulantes en écologie microbienne environnementale. Un grand merci enfin à Violette Da Cunha et Patrick Forterre, grâce à qui je reste connectée au monde des archées.

Pour terminer, je souhaiterais remercier ma famille, pour son soutien inconditionnel et pour le rôle déterminant qu'elle a joué dans mon orientation scientifique. Mes derniers remerciements, et non les moindres, vont à Pol, pour sa présence quotidienne, ses conseils en informatique et en Anglais, ainsi que sa patience sans cesse renouvelée envers les contraintes du métier de la recherche, où le temps est dérythmé par de joyeuses *deadlines*.

Table des matières

Liste des figures

Liste des tableaux

A. Introduction – la méthanisation

Dans ce manuscrit, je présente les principaux travaux de recherche auxquels j'ai contribué depuis le début de mon doctorat en 2005, jusqu'à aujourd'hui. J'expose tout d'abord des recherches en écologie microbienne de la méthanisation, en particulier sur la bioconversion de substrats cellulosiques. Il s'agit de travaux que j'ai menés pendant les années qui ont suivi mon arrivée en poste au Cemagref (devenu Irstea, puis INRAE). Dans un second temps, je décris brièvement un projet de développement d'un système d'information, que je coordonne, et qui a vocation à permettre une meilleure valorisation des données méta-omiques de procédés de biotechnologies environnementales. Par la suite, je synthétise mes recherches en lien avec la virologie microbienne : d'une part, mes travaux de thèse sur les virus d'archées acidothermophiles, et d'autre part, les recherches que j'ai développées plus récemment, en écologie virale de la méthanisation. Enfin, je présente des perspectives de recherche. Le manuscrit s'achève avec mon CV, et une sélection de publications en annexe. Certains des travaux décrits dans ce manuscrit ne sont pas encore publiés : j'ai choisi de les inclure en raison de la place importante qu'ils ont occupée dans mon activité, et également pour la cohérence qu'ils apportaient au manuscrit. La majorité de mes travaux de recherche porte sur la caractérisation des communautés microbiennes de la méthanisation. Dans cette section d'introduction, j'apporte donc quelques informations clés sur le processus et procédé de méthanisation.

La méthanisation, encore appelée digestion anaérobie, est un processus naturel de conversion de la matière organique qui se produit en conditions anaérobies et réductrices. Une synthèse des connaissances sur la méthanisation est disponible dans le livre « La méthanisation » (Moletta, 2015). Cette conversion est catalysée par des communautés microbiennes complexes et aboutit à la production de biogaz, riche en dioxyde de carbone (CO_2) et en méthane (CH_4). Le biogaz contient d'autres composés en faibles quantités, tels que du dihydrogène (H_2) ou du sulfure d'hydrogène (H_2S). La méthanisation est décomposée en 4 étapes successives : l'hydrolyse, la fermentation (ou acidogenèse), l'acétogenèse et la méthanogenèse (Figure 1). La méthanogenèse est le fait d'archées uniquement, et elle est considérée comme l'un des plus anciens métabolismes impliqués dans le cycle du carbone (Sauterey *et al.*, 2020). La méthanisation est active dans une grande variété d'environnements compatibles avec la vie, telles que rizières, marais, tourbières, sédiments marins et lacustres, ou encore tractus digestif de divers animaux (ruminants, termites, humain, …).

De nombreux facteurs abiotiques sont susceptibles d'affecter la méthanisation, comme le pH, la température et le potentiel d'oxydoréduction. Un pH proche de la neutralité est optimal. Une large gamme de température est compatible avec la méthanisation, sous l'action de communautés microbiennes psychrophiles, mésophiles ou thermophiles. Cependant, les dynamiques peuvent différer selon la température, avec généralement des cinétiques plus lentes à basse température. Concernant le potentiel d'oxydoréduction, on notera qu'en présence de certains accepteurs d'électrons, un phénomène de compétition peut se produire pour la consommation d'acétate ou d'H_2/CO_2. Ceci est illustré sur la Figure 1, dans le cas des ions sulfate (SO_4^{2-}) et de consommation d'H_2/CO_2 par les bactéries sulfato-réductrices. Ce phénomène est tout simplement lié à la position des différents accepteurs d'électron sur l'échelle des potentiel redox, le CO_2 étant l'un des plus « piètres » oxydants.
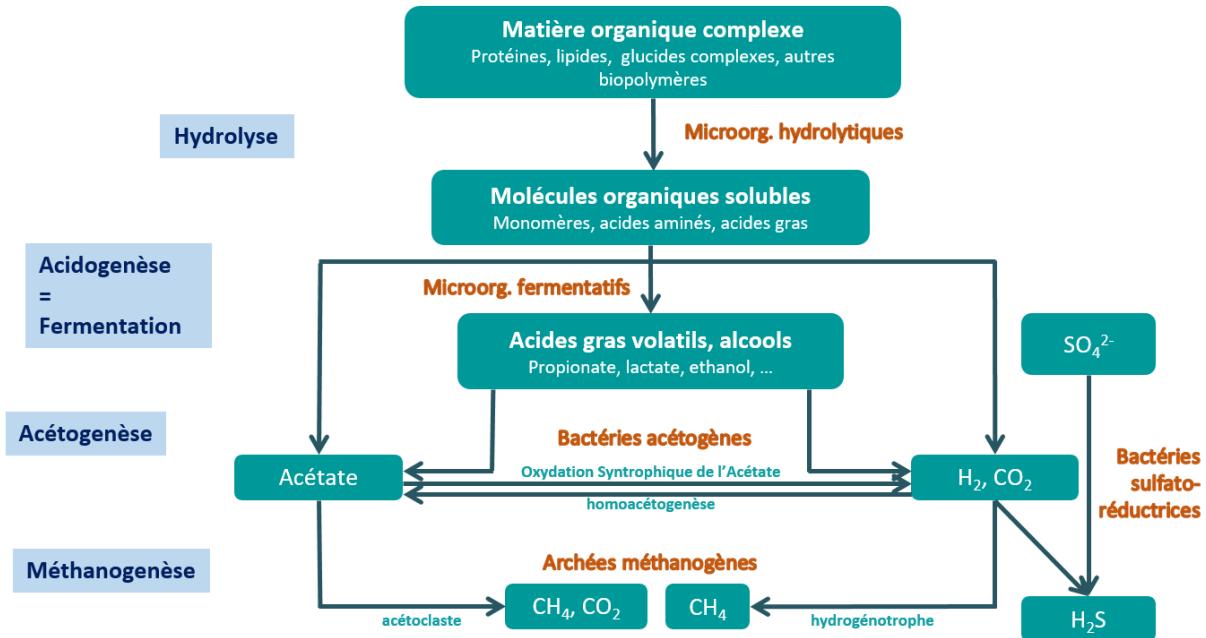
Figure 1. Les étapes du processus de méthanisation.

Le processus de méthanisation est exploité par l'homme dans des procédés de traitement et valorisation de déchets et effluents organiques (Figure 2). En effet, le biogaz obtenu est une énergie renouvelable, dont le pouvoir combustible dépend de sa teneur en méthane, qui est typiquement de l'ordre de 55%. Le pouvoir calorifique inférieur (PCI) du méthane est d'environ 9,94 kWh/normo m^3 dans les conditions normales de température et de pression (0°C, 1 atm). Différents modes de valorisation du biogaz sont possibles après son épuration. Celui-ci peut être injecté dans les réseaux de gaz de ville, il peut alternativement permettre la production de chaleur et d'électricité par cogénération, ou encore être utilisé comme carburant pour véhicule. Par ailleurs, les boues de digestion, si elles respectent certaines normes, peuvent être valorisées par épandage.
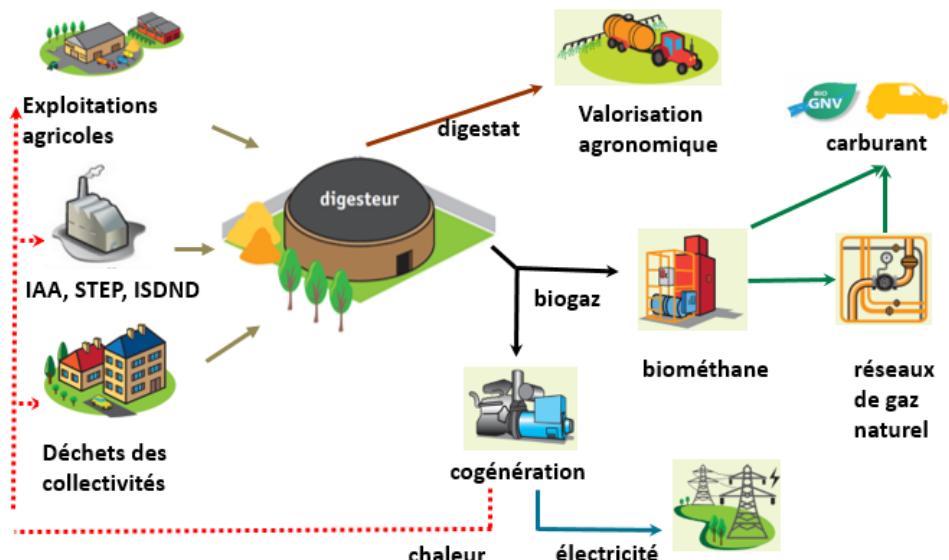
Figure 2. La méthanisation : sources d'intrants et valorisation du biogaz et du digestat produit.

IAA : industrie agroalimentaire – STEP : station d'épuration – ISDND : installation de stockage de déchets non-dangereux.

Source : ATEE (Association Technique Energie Environnement)

Le déploiement de la méthanisation à l'échelle industrielle a débuté à la fin du 19ème siècle, avec la construction d'une station de méthanisation à Bombay, en Inde, en 1859. Cela survient moins d'un siècle après la découverte, par Alessandro Volta, de gaz inflammable émis par des fonds boueux du lac Majeur en Italie (1776). Ainsi, les procédés de méthanisation ont été développés et améliorés progressivement, sans connaissance fine sur les communautés microbiennes impliquées, qui ont constitué pendant longtemps une boîte noire. En France, c'est d'abord la méthanisation en voie liquide, des boues de station d'épuration urbaine, qui s'est établie. Depuis les années 2000, la méthanisation connaît un développement rapide, soutenu par des politiques incitatives. En 2020, on comptait en France environ un millier d'installations de production de méthane, dont la majorité étaient des méthaniseurs à la ferme. La production d'énergie de l'ensemble de ces installations s'élevait à environ 12 TWh, soit 4% de la production primaire d'énergie renouvelable en France.

Améliorer l'opération des méthaniseurs demeure un enjeu important, notamment dans les cas d'hétérogénéité spatiale et temporelle des substrats d'alimentation, comme cela peut fréquemment être le cas à la ferme ou en méthanisation territoriale. En particulier, des inhibiteurs, présents dans le substrat ou se formant au sein des méthaniseurs, sont susceptibles de perturber l'activité des communautés microbiennes et par conséquent d'affecter les performances du procédé. La recherche en écologie microbienne de la méthanisation a un rôle à jouer, car le développement de ce type de connaissances peut permettre d'augmenter le niveau de compréhension du processus et ainsi apporter des bases scientifiques pour améliorer les performances et la robustesse de ces procédés.

Le développement des approches isotopiques et des techniques d'écologie moléculaire reposant sur l'ARNr16S, dans les années 90, a permis de déchiffrer le fonctionnement des communautés microbiennes de la méthanisation. Ces recherches se sont renforcées à partir des années 2000, avec l'arrivée du séquençage nouvelle génération et d'autres méthodes analytiques à haut-débit. Actuellement, les connaissances descriptives sur l'écologie microbienne des méthaniseurs sont bien développées, et leur acquisition se poursuit. De façon surprenante, les recherches en écologie virale des méthaniseurs ont émergé seulement très récemment (Calusinska *et al.*, 2016), alors que le rôle important des virus dans d'autres environnements est déjà bien établi : dans les océans, il est estimé que la lyse cellulaire liée aux infections virales aboutirait à un *turn-over* d'au moins 20% de la biomasse photosynthétique (Suttle, 2005). Concernant la méthanisation, les recherches en écologie virale mériteraient donc d'être encore développées. Un autre enjeu est de mobiliser les connaissances sur les communautés microbiennes pour développer des outils opérationnels, tels que des biomarqueurs, afin de mettre en place un *management* microbien des méthaniseurs (Carballa *et al.*, 2015) et ainsi améliorer leurs performances globales. Cela rejoint plus généralement des enjeux clés bien identifiés en écologie microbienne, en particulier de parvenir à développer une écologie plus prédictive (Widder *et al.*, 2016).

B. Synthèse des travaux de recherche

I. Déchiffrer les communautés microbiennes impliquées dans la méthanisation de la cellulose par des méthodes méta-omiques : de l'exploration aux approches comparatives

La lignocellulose, composant de la paroi végétale, est le biopolymère le plus abondant sur Terre. Elle représente une source majeure de matériaux et d'énergie biochimique renouvelable. Sa bioconversion est donc étudiée depuis plusieurs décennies (Lynd Lee *et al.*, 2002). Lorsque j'ai rejoint le Cemagref en

2008, la vitesse d'hydrolyse de la cellulose était déjà identifiée comme limitante dans les procédés de méthanisation (ex : (Noike *et al.*, 1985)). Les liens entre dynamique de colonisation microbienne de la cellulose, vitesse d'hydrolyse, et identité des microorganismes cellulolytiques avaient notamment été caractérisés sur un substrat modèle, la cellulose cristalline (O'Sullivan *et al.*, 2005, Song *et al.*, 2005, Jensen *et al.*, 2008), en combinant par exemple de l'hybridation *in situ* (FISH, *Fluorescent In Situ Hybridization*) et des mesures de vitesse d'hydrolyse de la cellulose. Au sein de notre unité, Tianlun Li et Olivier Chapleur, avaient décrypté, lors de leur doctorat, le fonctionnement de communautés microbiennes de la méthanisation de cellulose modèle, par des techniques isotopiques (Li *et al.*, 2009, Chapleur *et al.*, 2014).

Comprendre le déterminisme des performances de la méthanisation des déchets lignocellulosiques, en lien avec l'activité des communautés microbiennes, demeurait un sujet de fort intérêt. Les approches méta-omiques commençaient à percer dans nos domaines appliqués (Schlüter *et al.*, 2008, Hanreich *et al.*, 2012) et apportaient la possibilité d'identifier sans *a priori* les fonctions biologiques des communautés microbiennes de la méthanisation. Aussi, ai-je participé pendant plusieurs années au développement et à la mise en œuvre d'approches méta-omiques au sein de l'unité. Pour leur application, je me suis intéressée plus particulièrement à la méthanisation de matériaux cellulosiques manufacturés (de type papier, mouchoirs en papier) qui était moins étudiée que celle des fibres végétales natives ou des substrats modèles tels que la cellulose cristalline. Ces substrats manufacturés étaient en outre en bonne adéquation avec une thématique majeure de l'unité à cette période, la méthanisation des déchets ménagers. Ces derniers contiennent en effet une fraction importante de papiers, cartons, textiles et textiles sanitaires, tous riches en cellulose (campagnes MODECOM de caractérisation des déchets ménagers menées par l'Ademe en 1993, 2007 et 2017). Dans un premier temps, j'ai participé à une analyse métaprotéomique exploratoire de la méthanisation de papier en conditions thermophiles, en co-encadrant un post-doctorat. Dans un second temps, j'ai co-encadré un doctorat visant à étudier les liens entre structure de la lignocellulose, colonisation microbienne, et performances de méthanisation. Ces recherches sont présentées ci-dessous.

I.I   Analyse métaprotéomique des communautés microbiennes lors de la digestion anaérobie thermophile de déchets lignocellulosique

La méthanisation de papier blanc a été étudiée en conditions thermophiles, dans le cadre du post-doctorat de Lü Fan (Université de Tongji, Chine), que j'ai co-encadré avec Théodore Bouchez. Les analyses métaprotéomiques ont été réalisées en collaboration avec la plate-forme INRA PAPPSO. Ces travaux ont fait l'objet d'une publication (Lü *et al.*, 2014).

Le papier de bureau a été sélectionné comme substrat cellulosique, dans la mesure où il est abondant dans les déchets ménagers et a une composition relativement stable, de 70% d'hemicellulose et 30% de cellulose. Les incubations ont été menées à 55°C, en microcosmes de méthanisation discontinus (*batch*) de 1 litre, en 5 réplicas. Dans ces 5 microcosmes, des dynamiques similaires et classiques ont été obtenues en termes de production de méthane et de concentrations en acides gras volatils, composés intermédiaires formés lors de la méthanisation (Figure 1). L'un des réplicas a été sacrifié après 60 jours d'incubation, en phase production de méthane, afin de réaliser l'analyse métaprotéomique : 62% du carbone initialement introduit avait alors été dégradé. Les quatre autres réplicas ont été incubés pendant un total de 120 jours, afin que la production cumulée de méthane atteigne un plateau et que l'ensemble de la dynamique puisse ainsi être observé.

Une analyse métaprotéomique de type *bottom-up* a été menée. Elle consiste à digérer les protéines en fragments polypeptidiques, à identifier ces fragments par spectrométrie de masse MS/MS, après

séparation par chromatographie liquide, par comparaison à des bases de données de protéines de références. Enfin, les protéines sont identifiées *in silico* à partir de leurs composants peptidiques. Le développement du protocole a constitué une partie importante du projet et a été réalisé en collaboration avec INRA PAPPSO. L'enjeu était de parvenir à exploiter ces échantillons complexes malgré la présence probable de composés susceptibles de perturber l'analyse. Les cellules microbiennes présentes dans les échantillons ont été lysées mécaniquement et chimiquement, et les protéines du surnageant ont été purifiées par précipitation à l'acide trichloroacétique. Nous avons testé trois stratégies différentes de séparation des protéines. La première consistait à séparer classiquement les protéines selon leur poids moléculaire, par électrophorèse en gel précoulé à gradient d'acrylamide. Dans le second cas, les protéines étaient séparées selon le point isoélectrique avec un système hors-gel (Agilent OFFGEL Fractionator), et chacune des fractions collectées était migrée très brièvement dans un gel précoulé à gradient d'acrylamide, dans le but d'éliminer les impuretés, sans toutefois viser à séparer davantage les protéines. Enfin, la troisième stratégie a consisté à ne pas séparer les protéines, en se contentant de l'étape de migration très courte dans un gel. La deuxième et la troisième stratégie ont été réalisées en 3 réplicas techniques, si bien qu'un total de 65 fractions de protéines a été produit : 26 (stratégie 1) + 3x12 (stratégie 2) + 3 (stratégie 3).

Pour ces échantillons complexes, la meilleure sensibilité a été obtenue avec la deuxième stratégie, c'est-à-dire la séparation OFFGEL, suivie de la migration courte sur gel. Pour la présente étude, nous avons combiné les résultats obtenus avec l'ensemble des 65 fractions. Leur analyse par spectrométrie de masse a été conduite par nanoLC-MS/MS (LTQ-Orbitrap, Thermo Fisher, Waltham, MA, USA, INRA PAPPSO, Jouy-en-Josas), après digestion trypsique des protéines. Pour l'identification des protéines, nous avons fait le pari d'utiliser uniquement les bases de données de protéines publiques, en l'occurrence UniprotKB. Un total de 717 065 spectres a été obtenu, dont 40 818 (~6%) ont pu être assignés à des protéines. *In fine*, ce sont 13 090 peptides correspondant à 2 541 protéines, soit 514 groupes de protéines non redondantes, qui ont été retenus, en prenant en compte le taux d'erreur (*False Discovery Rate, FDR*) pour filtrer les résultats d'identification.

Parmi les microorganismes auxquels ces 514 groupes non-redondants ont pu être attribués, 4 taxons ou groupes fonctionnels dominants sont ressortis (Figure 3), parmi lesquels deux microorganismes cellulolytiques, *Acetivibrio thermocellus* (alors nommé *Clostridium thermocellum*) et *Caldicellulosiruptor* spp. Des archées méthanogènes hydrogénotrophes étaient également actives, dominées par *Methanothermobacter*. Enfin, de façon surprenante, l'un des groupes les plus actifs correspondait à des microorganismes protéolytiques de l'espèce *Coprothermobacter proteolyticus*.

L'examen détaillé des fonctions biologiques putatives de ces protéines a montré la complémentarité des deux microorganismes cellulolytiques pour la déconstruction du papier. *A. thermocellus*, qui synthétise des cellulosomes, semblait plutôt impliqué dans l'hydrolyse de la cellulose, en particulier de de ses parties cristallines. Les membres du genre *Caldicellulosiruptor* sont des microorganismes cellulolytiques non producteurs de cellulosomes, qui secrètent des enzymes multifonctionnelles et utilisent une large gamme de composants végétaux, incluant cellulose, cellulose cristalline, hémicellulose, amidon et pectine. Dans le cas présent, les protéines détectées pour ce genre suggéraient un rôle actif dans la dégradation de l'hémicellulose. L'abondance et l'activité de ces deux groupes microbiens a été confirmée par FISH. De façon cohérente, des proportions significatives du genre *Caldicellulosiruptor* et de la famille Ruminococcaceae ont été observées par séquençage métabarcoding ADNr 16S réalisé sur les mêmes réacteurs, au même point de temps. Notons que la famille Ruminococcaceae incluait à l'époque *A. thermocellus*, qui est désormais classé dans la famille Oscillospiraceae.
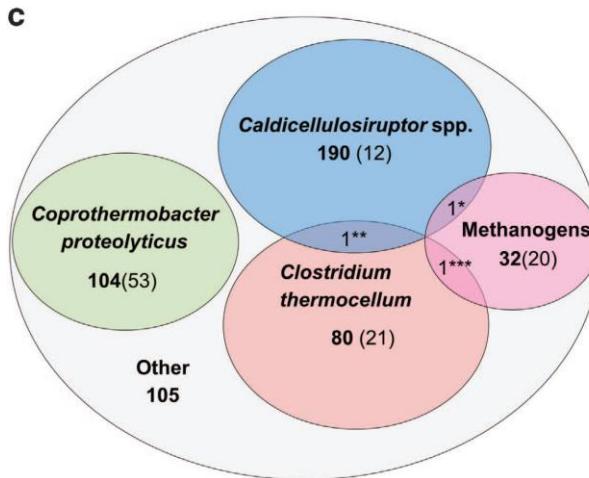
Figure 3 Distribution taxonomique des groupes non-redondants de protéines identifiées.
Les nombres en caractères gras représentent les nombres de groupes non-redondants de protéines. Les nombres entre parenthèses indiquent le nombre de clusters UniRef50 spécifiques au groupe taxonomique ou fonctionnel considéré. *Clostridium thermocellum* correspond désormais à *Acetivibrio thermocellus*.

En ce qui concerne la méthanogenèse, la nature des protéines identifiées était cohérente avec le caractère hydrogénotrophe, déjà connu, des espèces microbiennes les produisant. En particulier, des protéines spécifiques de la voie hydrogénotrophe ont été identifiées, codées par *hdrA*, *hdrC* et *mvhD*, catalysant la réduction de CoM–S–S–CoB par H_2. L'activation exclusive de la voie hydrogénotrophe a été confirmée par des analyses de composition isotopique du biogaz, en mesurant le facteur de fractionnement apparent du méthane (Conrad, 2005). Les résultats de métabarcoding ADNr16S ont confirmé que *Methanobacter* était extrêmement dominant parmi les archées méthanogènes détectées.

Le fait d'une part que la méthanogenèse soit exclusivement hydrogénotrophe, et d'autre part que de l'acétate soit présent dans le milieu (concentrations maximales de l'ordre de 300-400 mg-C/L) indiquait que l'oxydation syntrophique de l'acétate devait se produire (Figure 1). En accord avec cette hypothèse, des protéines bactériennes de la voie de l'Acétyl-CoA ont été identifiées. Cependant, les assignations taxonomiques de ces protéines pointaient vers des microorganismes variés, et il n'a pas été possible d'identifier avec certitude l'espèce majoritaire catalysant l'oxydation syntrophique de l'acétate. D'après les analyses de métabarcoding ADNr16S, le genre *Gelria* faisait partie des groupes dominants dans l'échantillon considéré, et il pourrait avoir joué ce rôle. Peu de protéines issues de ce genre ont été détectées, vraisemblablement en raison de l'absence de représentant suffisamment proche dans les bases de données publiques de protéines. En effet, l'identification des peptides est très sensible aux différences de séquences, puisqu'on recherche une égalité parfaite des masses des peptides correspondants.

Enfin, le rôle protéolytique de *C. proteolyticus* a été confirmé par l'identification d'une protéase putative, extracellulaire et attachée à la paroi cellulaire (13 peptides identifiés) et de 3 groupes de protéines issus de transporteurs *ABC* liés au transport de peptides. L'abondance et l'activité de membres de *Coprothermobacter* ont été confirmées par FISH. La forte activité protéolytique peut paraître surprenante sachant que le seul substrat introduit était carboné. D'après les analyses de métabarcoding ADNr16S, la composition des communautés microbiennes avait grandement évolué entre le jour initial et le jour 60, reflétant l'adaptation de l'inoculum aux conditions précises du microcosme. On peut penser que cela est associé à une certaine mortalité des microorganismes non-sélectionnés, dont les constituants protéiques alors libérés permettent la croissance de microorganismes protéolytiques. Cette évolution des communautés microbiennes entre le début de

l'incubation et les jours suivants est un phénomène que nous observons très fréquemment en microcosmes *batch*. De façon très intéressante, trois des peptidases putatives identifiées pendant l'étude, provenant de *C. proteolyticus* pourraient être impliquées dans la synthèse de microcines. Il s'agit de toxines bactériennes composées de quelques peptides. Cela suggère que *C. proteolyticus* pourrait avoir été un prédateur actif, et ne pas seulement utiliser pour sa croissance du matériel protéique extracellulaire disponible. Cette hypothèse resterait à prouver.
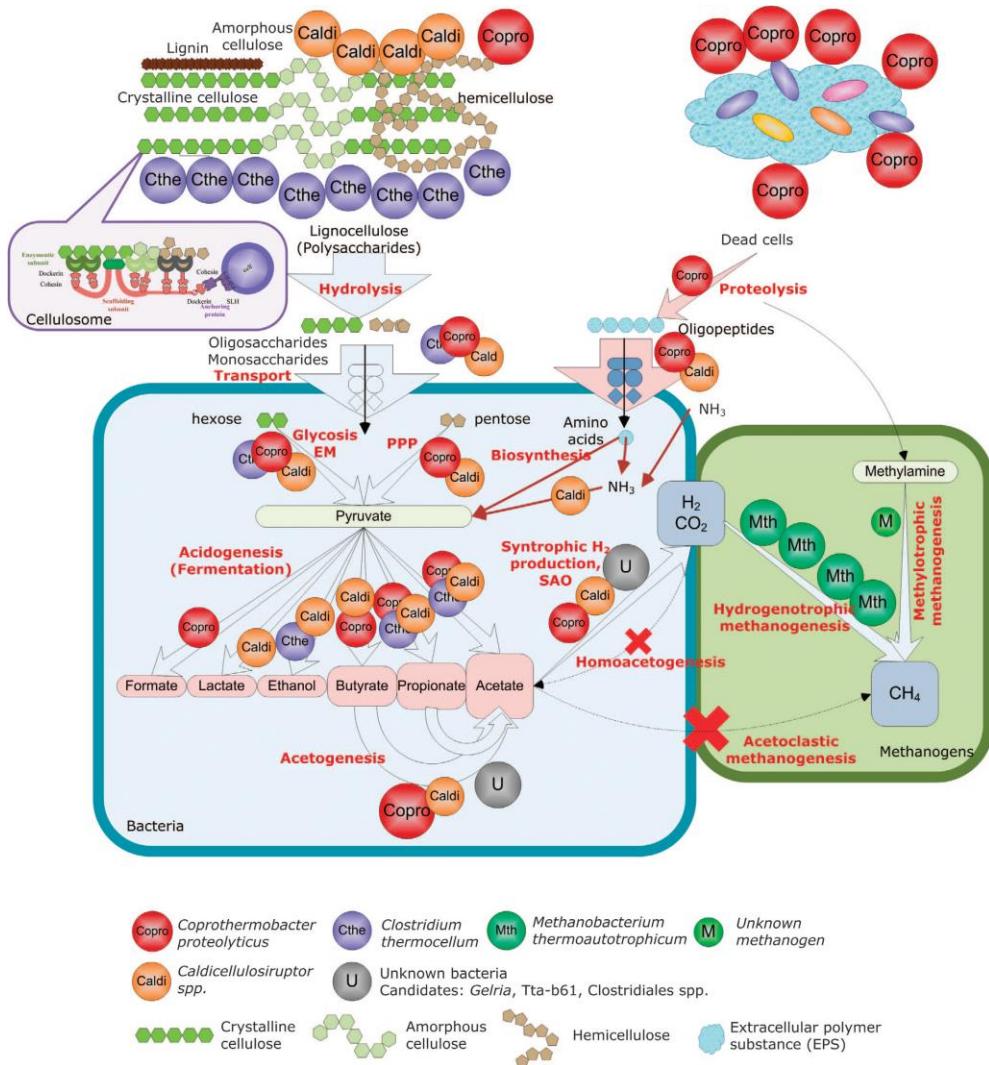
Figure 4. Modèle fonctionnel de la digestion anaérobie de lignocellulose par des communautés microbiennes thermophiles.

La Figure 4 présente le modèle fonctionnel établi à partir des résultats de l'étude. Ainsi, par une approche polyphasique et en s'appuyant uniquement sur les bases de données publiques de protéines, nous avons pu proposer un modèle fonctionnel qui n'était pas complet mais constituait une avancée très significative, suggérant en particulier la nécessité de reconsidérer certains schémas fonctionnels.

I.II   Effet du substrat cellulosique sur la dynamique d'hydrolyse et de fermentation par une souche bactérienne ou par une communauté microbienne anaérobie

Forte de cette première expérience, j'ai souhaité employer la métaprotéomique dans un but comparatif, dans le cadre du co-encadrement de la thèse de Nelly Badalato, qui visait à élucider les liens entre colonisation de la lignocellulose, activité des communautés microbiennes et performances de méthanisation. Trois substrats manufacturés distincts ont été choisis, contenant majoritairement de la cellulose : du papier filtre Whatman, des mouchoirs en papier, et des disques de coton. Des conditions mésophiles ont été retenues (35°C), car elles représentent la majorité des installations industrielles de méthanisation. Pour l'analyse métaprotéomique, nous avons globalement conservé le même protocole que précédemment, toujours en collaboration avec INRA PAPPSO, avec quelques évolutions. Par simplicité, nous avons opté pour une séparation des protéines sur gel SDS-PAGE, selon le poids moléculaire : dans la précédente étude, cette méthode avait abouti à une sensibilité, qui, sans être la meilleure, était satisfaisante. Pour l'identification des protéines, nous avons combiné les séquences protéiques issues de bases de données publiques, et de quelques jeux de données de métagénomique *shotgun* acquis spécifiquement à partir de nos échantillons, afin de couvrir de façon plus homogène les différents groupes fonctionnels microbiens actifs. Enfin, nous avons décidé de travailler à plusieurs niveaux de diversité : outre les microcosmes de méthanisation, nous avons également réalisé des incubations anaérobies de fermentation avec une souche cellulolytique bactérienne pure, en utilisant les mêmes substrats cellulosiques. Ce système de moindre complexité pouvait faciliter la compréhension mécanistique. A nouveau, une approche polyphasique a été adoptée, combinant notamment des observations de la colonisation de la cellulose par microscopie, du métabarcoding ADNr16S et de la PCR quantitative, en plus de l'analyse des protéines. Les résultats obtenus en culture de souche pure ont été publiés (Badalato *et al.*, 2017), ce qui n'est pas encore le cas de l'étude des communautés complexes. Les résultats sont présentés ici de façon croisée, en se focalisant sur quelques points clés.

Concernant les incubations en présence d'une souche bactérienne pure, c'est *Ruminiclostridium cellulolyticum* (famille Oscillospiraceae) qui a été retenue. Elle était à l'époque nommée *Clostridium cellulolyticum* et classée dans la famille Ruminococcaceae. Il s'agit d'une bactérie cellulolytique mésophile modèle, productrice de cellulosomes, qui est notamment très étudiée par une équipe du Laboratoire de Biochimie Bactérienne, dirigée par Chantal Tardif, à Marseille. Cette équipe a en particulier caractérisé finement la composition du cellulosome de *R. cellulolyticum* et sa modulation (ex : (Gal *et al.*, 1997, Parsiegla *et al.*, 1998)). Une série d'études sur le métabolisme *R. cellulolyticum* avait par ailleurs été menée par Michael Desvaux et d'autres chercheurs du Laboratoire de Biochimie des Bactéries Gram + près de Nancy, en conditions continues (ex : (Desvaux *et al.*, 2001)) ou discontinues (*batch)* (ex : (Desvaux *et al.*, 2000)). Nous avons retenu des conditions d'incubation *batch*.

Le premier point marquant était la similarité des cinétiques de dégradation de chaque substrat, d'un point de vue qualitatif, lorsque l'on comparait la fermentation par la souche pure à la méthanisation par des communautés microbiennes complexes. L'hydrolyse du mouchoir en papier était la plus rapide, puis celle du papier filtre Whatman, suivie de près pas celle du coton. En présence de *R. cellulolyticum*, la dynamique est illustrée ci-dessus par l'évolution des concentrations en carbone organique dissous et des produits solubles majoritaires de la fermentation, dont on peut estimer qu'ils s'accumulent d'autant plus rapidement que la cellulose a une vitesse d'hydrolyse élevée (Figure 5). Pour les communautés complexes, nous nous sommes basés sur le même type d'analyses, ainsi que sur le suivi de la production de biogaz, permettant d'établir le bilan carbone à chaque point de temps (Figure 6). Ainsi, la diversité microbienne élevée, dans le cas de l'inoculum complexe, n'a pas permis de surmonter le caractère relativement récalcitrant des disques de coton et du papier filtre Whatman, par rapport

au mouchoirs en papier. Les caractéristiques propres des substrats cellulosiques semblent donc constituer un déterminant très fort de la cinétique d'hydrolyse. On peut supposer que la densité et l'accessibilité des sites d'ancrages pour les enzymes cellulolytiques jouent un rôle prépondérant.
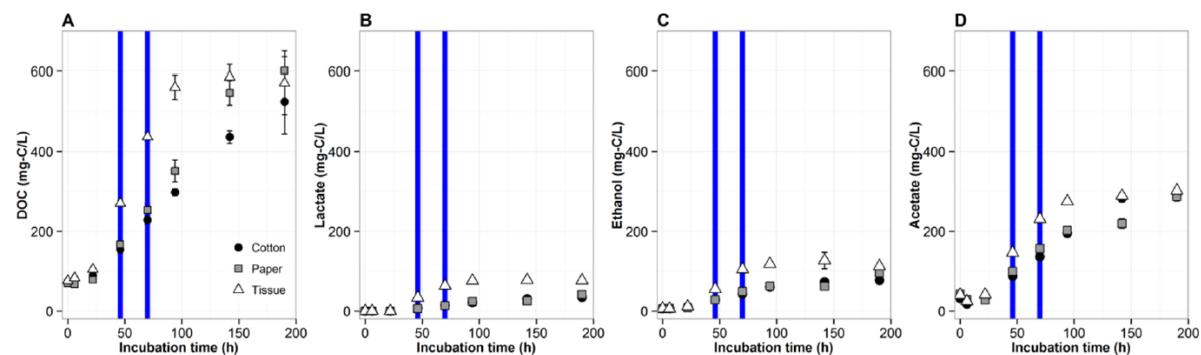
Figure 5. Dynamique des concentrations de composés organiques carbonés solubles dans les réacteurs contenant R. cellulolyticum.

DOC, *Dissolved Organic Carbon* (Carbone Organique Dissous) – La moyenne et l'écart-type obtenus pour les 3 réplicas de chaque condition sont montrés. Les lignes verticales indiquent les points de temps retenus pour les analyses protéomiques.
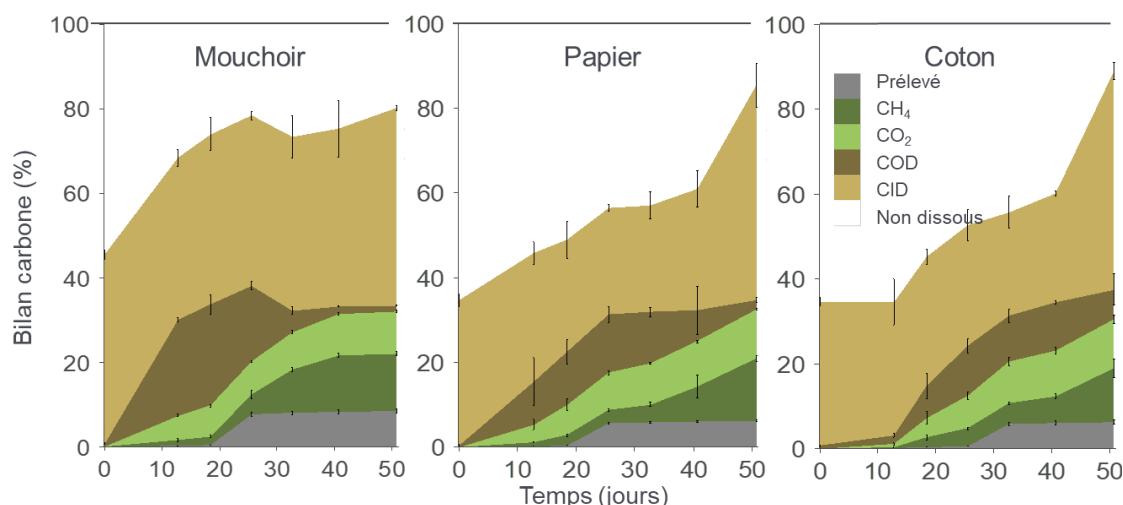
Figure 6. Bilan carbone pour les différents microcosmes de méthanisation, au cours de l'incubation.

COD: carbone organique dissous, CID : carbone inorganique dissous. La moyenne et l'écart-type obtenus pour les 3 réplicas de chaque condition sont montrés.

Afin de mieux comprendre l'origine des différences de vitesse d'hydrolyse, nous avons caractérisé finement les 3 substrats cellulosiques (Tableau 1, Figure 7), en collaboration scientifique avec Gérard Mortha (INP-Pagora, Laboratoire de Génie des Procédés pour la Bioraffinerie, les Matériaux Bio-sourcés et l'Impression Fonctionnelle, LGP2, Grenoble) et Alain Buléon (INRA Biopolymères Interactions Assemblages, BIA, Nantes).

Les principales différences entre les substrats concernaient l'indice de cristallinité, la distribution des masses molaires et le contenu en hémicellulose. L'indice de cristallinité est communément mesuré pour estimer les quantités de régions cristallines dans la cellulose, moins aisément dégradables que les régions amorphes. Le mouchoir en papier présentait à la fois l'indice de cristallinité et le degré moyen de polymérisation les plus faibles, ce qui pouvait expliquer sa biodégradation plus rapide. De plus, il était le seul substrat à contenir des proportions significatives d'hémicelluloses (contenu en

pentoses et xylose, Tableau 1), ce qui pouvait contribuer à améliorer l'accessibilité aux enzymes et/ou l'hydrophilie au niveau supramoléculaire. En effet, les réseaux de cellulose et hémicellulose sont moins ordonnés et cristallins que les réseaux de cellulose pure, ce qui pourrait donc conduire à une biodégradation plus rapide. De manière cohérente, les disques de cotons avaient le plus haut degré de polymérisation, et un indice de cristallinité élevé (pas le plus élevé toutefois).

Tableau 1. Caractéristiques détaillées du mouchoir papier, du papier filtre Whatman, et des disques de coton.

	Mouchoir papier	Filtre Whatman	Disques de coton
Matières sèches / matières volatiles (%/%)	94,9/94,2	96,0/95,8	96,5/96,4
Carbone/Azote (%/%)	41,9/0,07	43,1/0,03	41,1/0,05
Demande Chimique en Oxygène (g/g)	1,06	1,14	1,11
Index de cristallinité (%)	50	94	74
Degré de polymérisation*	970	1300	2730
Contenu total en sucres (% matières sèches)	97,23	99,22	99,47
Hexoses (% matières sèches)	83,35	98,67	98,94
dont glucose (% matières sèches)	81,54	98,58	98,69
Pentoses (% matières sèches)	13,87	0,55	0,54
dont xylose (% matières sèches)	13,72	0,51	0,45
Fractionnement Van Soest			
fraction soluble dans un détergent neutre (%)	0,05	0,01	5,3
fraction soluble dans un détergent acide (%)	14,89	4,01	43,6
fraction soluble dans l'acide sulfurique (%)	85,06	84,84	49,8
matières volatiles insolubles (%)	0	11,13	1,2

* Les degrés de polymérisation moyens sont calculés à partir des distributions des masses molaires montrées en Figure 7.
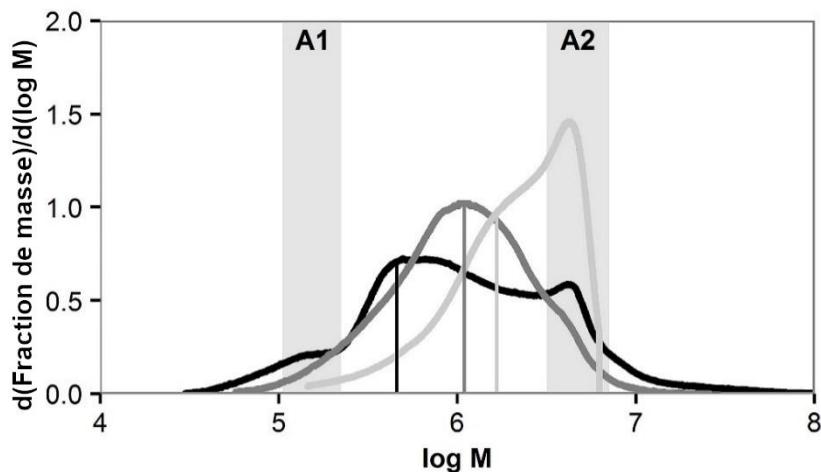
Figure 7. Distribution des masses molaires des chaînes de cellulose et d'hémicellolose dans les trois substrats employés.
Lignes noires : mouchoir en papier – Lignes grises : papier filtre Whatman – Lignes gris clair : disques de coton. M : masse molaire. Les lignes verticales indiquent les pics de masse molaire pour les chaines de cellulose individuelles. Aire grise A1 : pic observé pour l'hémicellulose dans le mouchoir en papier (fabriqué à partir de pulpe de bois blanchie). Aire grise A2 : pic correspondant à des polymères de très haut poids moléculaire, vraisemblablement des agrégats de chaînes de celluloses.

Les observations en microscopie en état frais effectuées pour les incubations avec *R. cellulolyticum* ont mis en évidence, pour les 3 substrats, un développement progressif de la colonisation des fibres au cours du temps, et les niveaux de colonisation atteints étaient bien plus élevés dans le cas du mouchoir en papier, suivi par le papier filtre Whatman et enfin, par les disques de coton. Ces derniers présentaient une colonisation bien plus éparse. Les images peuvent être visualisées dans l'article correspondant en annexe. Ces résultats étaient très cohérents avec l'hypothèse de forte limitation de l'hydrolyse par les propriétés d'accessibilité du substrat.

L'effet des différents substrats sur l'expression des fonctions biologiques microbiennes a par la suite été examiné dans les deux cas considérés : lors de la fermentation anaérobie par *R. cellulolyticum*, et lors de la méthanisation par des communautés microbiennes complexes. Les protéomes de *R. cellulolyticum* ont été comparés en présence de mouchoir en papier et de papier filtre Whatman, par une approche quantitative sans marquage (*label-free*, pas de marquage des protéines) et ascendante (*bottom-up*), c'est-à-dire qu'une digestion trypsique était appliquée préalablement à l'analyse de spectrométrie de masse MS/MS. La quantification relative a été basée sur le calcul des courants ioniques extraits (*eXtracted Ion Chromatogram*, XIC). Le papier filtre Whatman a été inclus en tant que substrat de référence, dans la mesure où il avait déjà été employé auparavant dans différentes études (ex : (Gehin *et al.*, 1996)). Le mouchoir papier a été quant à lui sélectionné car il était associé à la bioconversion la plus rapide. Un total de 151 protéines présentant des niveaux significativement différents a été identifié. Ces protéines incluaient 20 des 65 sous-unités du cellulosome de *R. cellulolyticum*. Huit CAZYmes (*carbohydrate active enzymes*, http://www.cazy.org/, (Drula *et al.*, 2022)) n'appartenant pas au cellulosome, ainsi que 44 groupes protéiques extracytoplasmiques distincts, étaient également différentiellement abondants.

La proportion élevée de composants du cellulosome différentiellement abondants démontre la large modulation de sa composition. Parmi les CAZYmes et les composants du cellulosomes, 10 endoglucanases avaient des niveaux plus faibles en présence de mouchoir en papier. La plupart d'entre elles appartenaient à la famille GH9. Cette dernière comprend principalement des cellulases (EC 3.2.1.4), avec majoritairement des endoglucanases (clivant les liaisons glycosidiques internes aux polymères de glucose) et quelques exoglucanases processives (agissant aux extrémités des polymères de glucose). Le profil quantitatif observé pour les composants du cellulosome et les CAZYmes semblait donc bien cohérent avec les plus faibles degrés de cristallinité du mouchoir en papier. Notons qu'en revanche, malgré le fort taux d'hémicellulose du mouchoir en papier, nous n'avons pas détecté de niveaux plus élevés d'enzymes xylanases.

Outre la modulation de composition du cellulosome et de CAZymes, le second résultat marquant était une hausse des niveaux d'enzymes liées au catabolisme du carbone en présence de mouchoir en papier. Cette hausse s'explique certainement par un débit entrant de glucides plus élevé. Plus en détail, 18 enzymes des voies cataboliques du xylose et du glucose de *R. cellulolyticum* ont pu être quantifiées et évaluées par les modèles statistiques. Parmi elles, 8 présentaient des niveaux ou évolutions significativement différents, dont 7 à la hausse en présence de mouchoir papier.

Si l'on considère ensuite le cas des communautés microbiennes complexes, une première observation globale était l'adaptation de la composition des communautés aux substrats, de manière différenciée (Figure 8). Nous avons tout d'abord observé une évolution importante entre le jour initial et les jours suivants, avec par exemple la diminution de la part du phylum Cloacimonetes. Les 4 phylums majoritaires étaient par la suite Bacteroidetes, Firmicutes, Spirochaetae et Synergystetes. L'abondance de Firmicutes et Spirochaetae étant bien plus importante en présence des disques de coton, et dans une certaine mesure, des papiers filtres Whatman, on peut supposer le rôle important de certains de leurs membres dans l'hydrolyse de régions récalcitrantes de la cellulose.

Des analyses multivariées (Figure 9) ont confirmé ces différences de composition, avec un premier axe séparant les échantillons prélevés au point de temps initial et aux points de temps suivants, et un second axe séparant les échantillons associés au mouchoir en papier d'une part, et ceux associés aux disques de coton ou au papier filtre Whatman, d'autre part. Les échantillons associés au coton et aux filtres Whatman formaient des nuages de points très proches mais néanmoins distincts, suggérant des différences fines de composition entre les communautés sélectionnées par ces deux substrats. De manière plus quantitative, des analyses de type PERMANOVA (analyse de variance multivariée par

permutation) ont montré que la nature du substrat expliquait, de manière statistiquement significative, 37% de la variance des données de composition microbienne, et la durée d'incubation 24%, des proportions élevées.
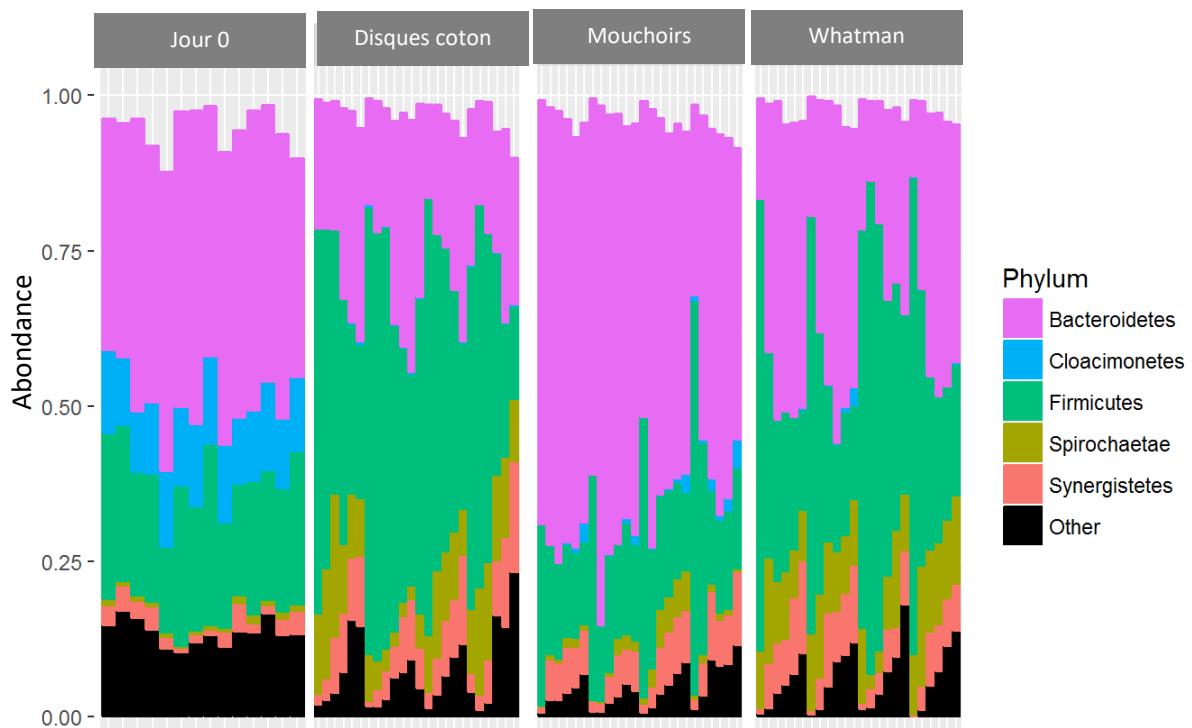
Figure 8. Composition des communautés microbiennes dans les microcosmes de méthanisation.
Il s'agit ici de la composition en bactéries, sur la base des analyses de métabarcoding ADNr16S réalisées avec des amorces ciblant archées + bactéries. Les échantillons du jour 0 sont montrés séparément, pour mettre en évidence l'adaptation des communautés entre le jour initial et les suivants. Pour chaque substrat, les échantillons sont classés de gauche à droite par par réplicas puis par point de temps. Les derniers prélèvements ont été effectués au jour 50.
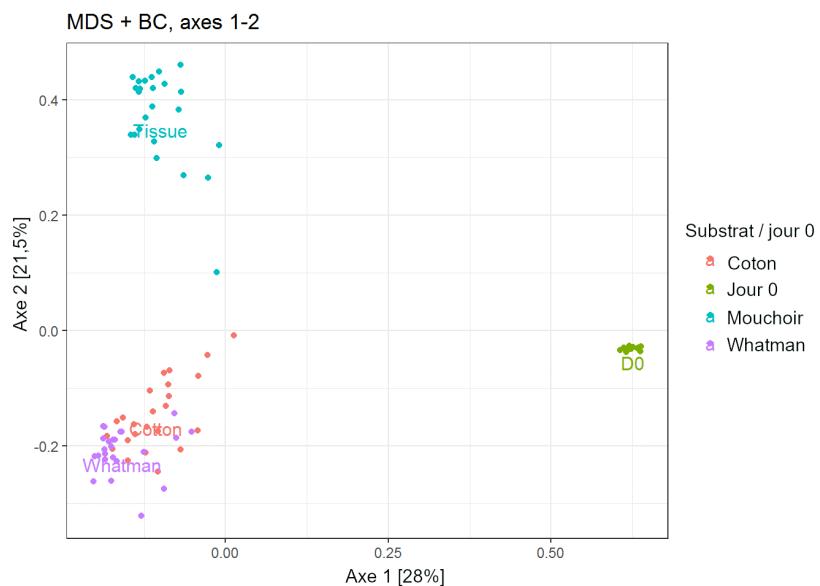
Figure 9. Analyse en coordonnées principales de la composition des communautés microbiennes, pour les échantillons issus des 3 conditions différentes.
La distance de Bray-Curtis a été utilisée. Mouchoir : mouchoirs en papier – Coton : disques de coton – Whatman : papier filtre Whatman – Jour 0 : jour 0 de l'incubation.

Afin d'identifier plus systématiquement, et à un niveau taxonomique fin, les microorganismes dont les abondances étaient affectées par le substrat, nous avons réalisé une analyse statistique avec DESeq2 (*package* R). Le modèle prenait en compte deux variables qualitatives : la nature du substrat (3 valeurs possibles) et le point de temps (5 valeurs possibles, le temps 0 ayant été retiré pour l'analyse). Un total de 595 clusters d'archées et de bactéries avaient été obtenus avec *swarm* (Mahé *et al.*, 2014), au sein d'un pipeline FROGS (Escudié *et al.*, 2018)). Parmi eux, 134 présentaient des niveaux significativement différents selon le substrat ou le point de temps. Certains d'entre eux étaient abondants et présentaient des niveaux très contrastés selon le substrat (Figure 10). On peut citer un cluster de la famille Marinilabiaceae (phylum Bacteroidetes, Cluster 1) qui représentait de l'ordre de 40% des bactéries et archées en présence de mouchoir en papier, et moins de 10% en présence des autres substrats. Un autre cluster, de l'ordre Clostridiales (phylum Firmicutes à l'époque, Cluster 3), était abondant uniquement en présence de disques de coton, représentant de l'ordre de 30% de la communauté microbienne à des points de temps précoces, puis déclinant au cours du temps : cette dynamique suggère un rôle clé dans l'hydrolyse de la cellulose ou son initiation. Un troisième cluster, de la famille Rikenellaceae (phylum Bacteroidetes, Cluster 7), présentait une abondance décroissante au cours de l'incubation, et était plus abondant en présence de mouchoirs en papier, puis de papier filtre Whatman, et enfin de disques de coton. Enfin, un cluster du genre *Treponema* (phylum Spirochaetae, Cluster 8), atteignait des niveaux de l'ordre de 10% dans les incubations de disques de coton ou de papier filtre Whatman, et beaucoup plus faibles avec le mouchoir en papier, d'environ 1%.
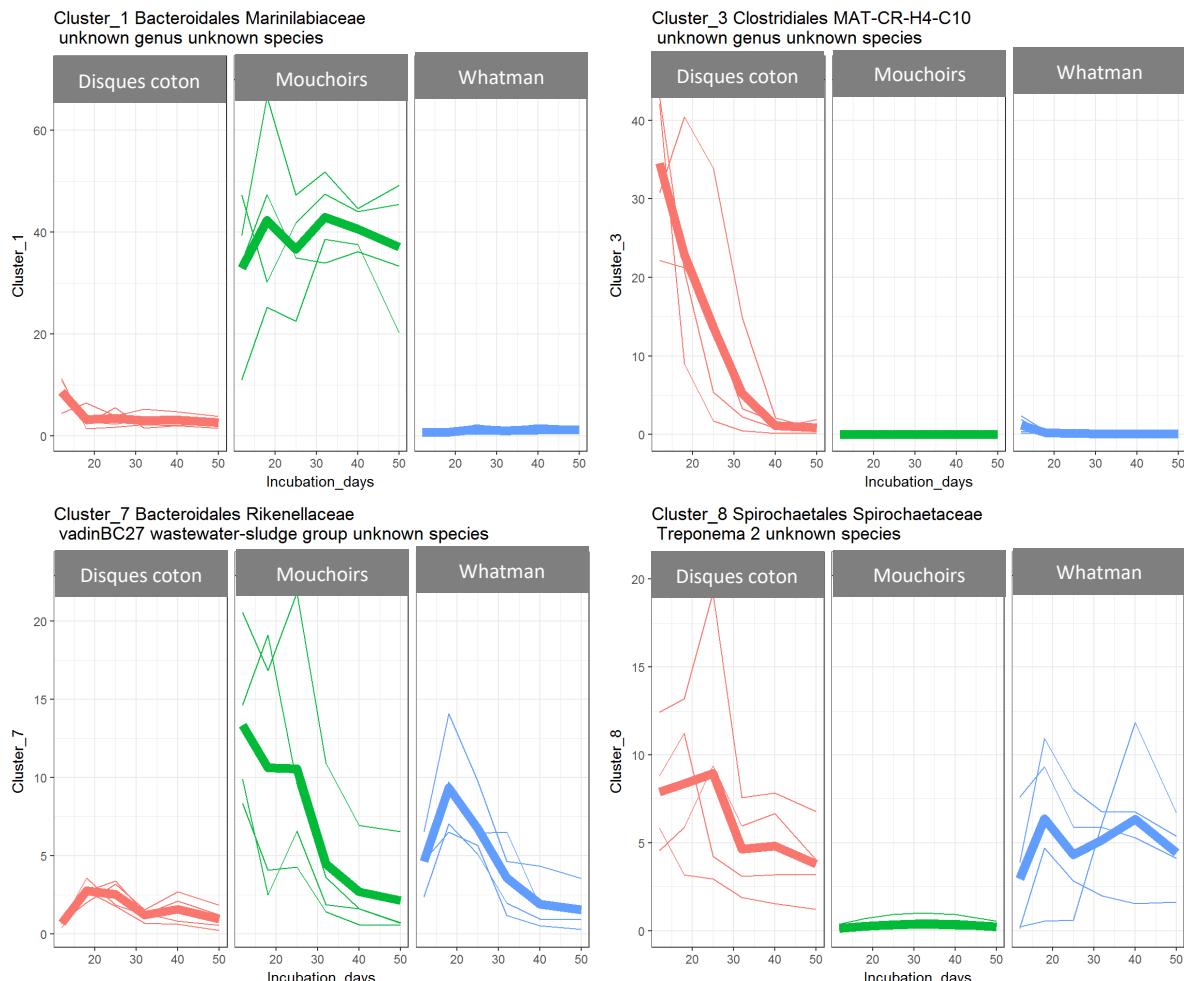
Figure 10. Dynamique de quelques clusters abondants et présentant des différences significatives de niveaux selon le substrat.

Les traits fins représentent chacun de 4 réplicas. Les traits épais représentent la moyenne obtenue pour les 4 réplicas.

Ainsi, des variations quantitatives de composition des communautés microbiennes ont été observées selon les substrats. Notons que ces variations peuvent provenir aussi bien d'effets directs, typiquement liés aux propriétés des substrats, aboutissant la sélection de microorganismes cellulolytiques spécifiques, que d'effets indirects : l'hydrolyse plus rapide du mouchoir en papier a conduit, de façon transitoire, à de plus fortes concentrations en acides gras volatils, et ces derniers peuvent certainement constituer un facteur important de sélection de microorganismes. On peut également imaginer une sélection de microorganismes fermentatifs selon leur compatibilité avec les microorganismes cellulolytiques premièrement sélectionnés.

La sélection observée au niveau de la composition des communautés microbiennes s'accompagnait-elle de différences au niveau fonctionnel ? Nous avons réalisé des analyses de PCR quantitatives ciblant la famille de CAZYmes GH48. Celle-ci comprend des cellulases qui sont généralement les sous-unités les plus abondantes des cellulosomes bactériens. Elles dégradent préférentiellement la cellulose cristalline et amorphe et jouent un rôle clé pour la synergie du système de cellulases. Les résultats ont mis en évidence des niveaux de CAZYmes GH48 plus élevés lors de la méthanisation de disques de coton (jusqu'à 10^8 copies/mL de culture), puis de papier filtre Whatman (~10^7 copies/mL), les niveaux les plus bas étant observés en présence de mouchoir en papier (~10^6 copies/mL). Ainsi, l'utilisation de disques de coton aboutissait à la sélection la plus forte de microorganismes possédant un gène clé dans l'hydrolyse de cellulose cristalline, ce qui était cohérent avec sa plus forte récalcitrance.

Comme évoqué en introduction de cette section, nous avons également mené des analyses métaprotéogénomiques, qui ont mis en évidence des tendances intéressantes, mais dont les résultats restent à examiner de manière plus approfondie. Elles ont tout d'abord montré une grande cohérence avec les analyses de métabarcoding ADNr16S. En effet, les 3 classes les plus abondantes, d'après l'assignation taxonomique des protéines identifiées, étaient Bacteroidia, Clostridia et Spirochaetes (Figure 11), avec les mêmes tendances que lors de l'analyse de métabarcoding : Bacteroidia plus représenté lors de la méthanisation de mouchoir en papier ; Clostridia et Spirochaetes *a contrario* moins présents dans les microcosmes contenant du mouchoir en papier. Ainsi, les conclusions concernant l'abondance de ces grands groupes dominants étaient *a priori* également valables en termes d'activité biologique.
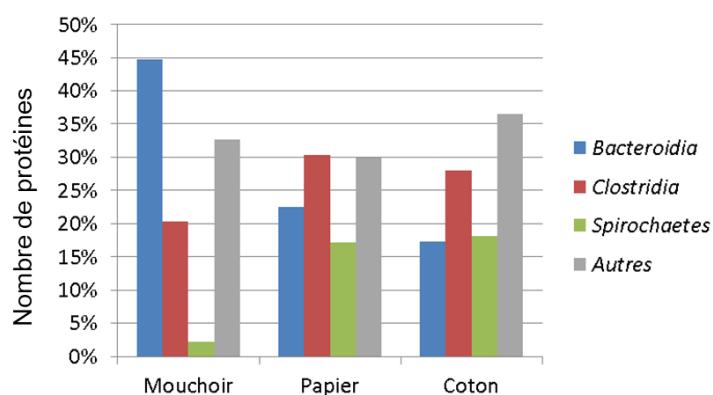
Figure 11. Assignation taxonomique des protéines identifiées par métaprotéomique shotgun

Un second résultat d'intérêt est issu de l'analyse de gènes de cellulosome dans les séquences métagénomiques. Pour en faciliter la compréhension, nous présentons d'abord un schéma de la structure générale des cellulosomes (Figure 12). Sur cette figure, on remarque en particulier les domaines de liaison au substrat cellulosique (CBM, *Carbohydrate Binding Modules*), qui permettent au

cellulosome de s'ancrer au substrat. Les cohésines sont des unités participant à la structure basale du cellulosome. Les dockérines sont des protéines qui assurent le lien structural entre les cohésines d'une part et les sous-unité enzymatiques. Ces sous-unités enzymatiques incluent des *glycosides hydrolases* (GH). Enfin, le domaine d'homologie à la couche S (SLH, *S-layer homology*) participe à l'ancrage du cellulosome dans la paroi bactérienne.
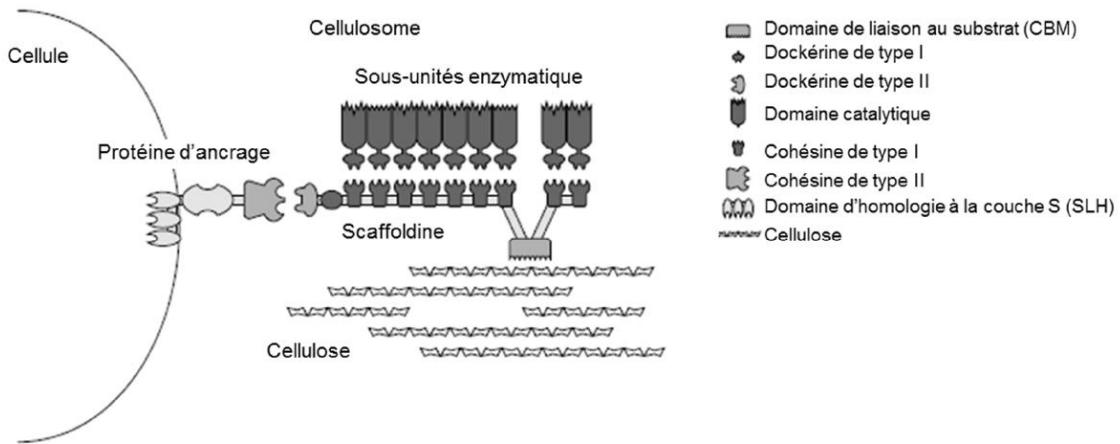
Figure 12. Structure générale des cellulosomes.
Adapté de (Shoham *et al.*, 1999).

Ces différents gènes de composants cellulosomiques ont été recherchés dans les métagénomes qui avaient été séquencés dans le but d'enrichir l'analyse métaprotéomique. Les résultats sont présentés dans le Tableau 2, et de manière plus visuelle dans la Figure 13. Bien que préliminaires, ils suggèrent des différences de profil en présence du mouchoir en papier et du papier filtre Whatman.

Tableau 2. Nombre de gènes de CAZymes détectés dans les métagénomes issus des microcosmes de méthanisation.

Module	Famille	Total	Mouchoirs	Whatman
	3	7	0	7
	30	2	0	2
	4	2	0	2
CBM	50	6	0	6
	54	4	2	2
	67	7	3	4
	9	1	0	1
cohésine		20	2	18
dockérine		3	2	1
	109	2	1	1
GH	3	4	4	0
	43	7	7	0
SLH		12	1	11
Total		77	22	55

Les résultats demeurent difficiles à interpréter à ce stade, car certaines familles clés, comme GH9 ou GH48, n'ont pas été détectées ici, alors que GH48 avait été détectée par PCR quantitative. Il peut être supposé que ces données de métagénomiques, générées avec un séquenceur « de paillasse » à débit moyen (Ion torrent PGM), ont une profondeur limitée, qui ne permet pas d'atteindre une vision exhaustive des gènes de cellulosome présents dans les échantillons.
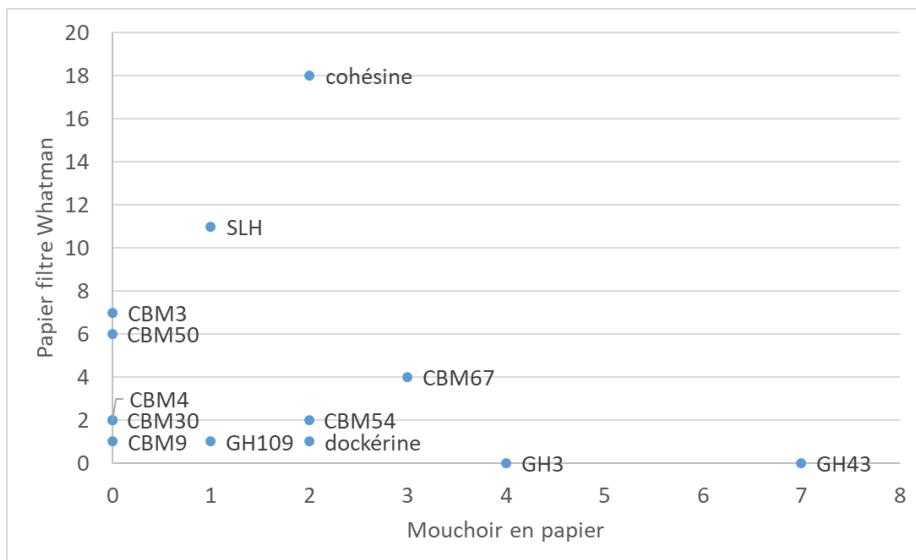
Figure 13. Nuage de point représentant le nombre de gènes de cellulosome détectés dans les séquences métagénomiques.
Les données correspondent à celles du Tableau 2.

En conclusion de cette étude comparative de l'effet du substrat cellulosique sur les dynamiques physico-chimiques et biologiques de la fermentation et de la méthanisation, on pourra retenir que des substrats en apparence relativement similaires (produits manufacturés, absence de lignine, cellulose majoritaire), aboutissent à des différences significatives de dynamique de bioconversion. Ils sont également associés à des différences significatives de composition du cellulosome, dans le cas de la fermentation par une souche bactérienne pure, et de composition des communautés microbiennes dans le cas de la méthanisation par des communautés microbiennes complexes. Les aspects fonctionnels restent à approfondir dans ce dernier cas. Cela montre l'adaptation fine des microorganismes aux propriétés des substrats cellulosiques. Cette adaptation semble plus précisément liée aux différences d'accessibilité des fibres de cellulose aux enzymes microbiennes. On notera également que la diversité microbienne élevée, dans le cas de la méthanisation, ne suffit pas à gommer les différences de cinétique de bioconversion observées entre les 3 substrats. On aurait pu imaginer que la sélection de microorganismes plus performants ou plus adaptés, en présence des substrats les plus récalcitrants, masque cet effet substrat. Ainsi, les caractéristiques d'accessibilité des substrats semblent être des déterminants très forts de la cinétique d'hydrolyse, ce qui est en accord avec la littérature (O'Sullivan *et al.*, 2006). Cette étude constitue un exemple spécifique de couple inoculum – substrats, et ne peut être généralisée directement. Néanmoins, ils inciteraient, pour améliorer les performances de méthanisation de déchets cellulosiques, à jouer en priorité sur les propriétés des substrats cellulosiques, par exemple par des prétraitements.

I.III   Analyses de séquençage haut-débit de fermentation de la celllulose dans un contexte finalisé

En parallèle de ces études situées à un niveau très amont par rapport à l'application, j'ai à cœur de participer à des projets plus directement tournés vers des questions opérationnelles. Ces activités me semblent très complémentaires, et essentielles dans une unité de recherche finalisée, afin de conserver une vision relativement réaliste des enjeux opérationnels et d'être conscient du positionnement de nos recherches par rapport à ce contexte. En lien avec la valorisation de déchets cellulosiques, j'ai contribué à une analyse métagénomique des communautés microbiennes dans le

cadre d'une étude visant à optimiser les conditions de production d'hydrogène à partir de déchets d'écorces d'agrumes, qui sont riches en cellulose. J'ai pour cela encadré Franciele Camargo, doctorante de l'université brésilienne de São Paulo, pour l'analyse de ses données métagénomiques, lors de son séjour de quelques mois dans notre unité (Camargo *et al.*, 2021). Concernant les communautés microbiennes, ces travaux ont notamment mis en évidence, en conditions optimales, la sélection reproductible de quelques genres bactériens dominants, tels que *Clostridium*, *Paraclostridium*, *Coprothermobacter* et *Defluvitoga.* Cette étude constitue ainsi un bel exemple d'ingénierie écologique.

Depuis plusieurs années, je travaille par ailleurs en collaboration avec ECOBIO (UMR 6553 CNRS – Université de Rennes) et le SIAAP (Syndicat Interdépartemental pour l'Assainissement de l'Agglomération Parisienne), dans le cadre d'un projet d'optimisation de la co-digestion, en voie sèche, de la fraction organique d'ordures ménagères et de fumier équin très riche en paille. Ce projet s'inscrit dans un contexte de développement de la méthanisation territoriale. L'effet de différents paramètres a été évalué en pilotes de 60 litres opérés au SIAAP, en particulier la proportion des différents co-substrats. Une inhibition acide transitoire a été observée aux fortes proportions d'ordures ménagères, liée à une accumulation d'acides gras volatils qui s'explique certainement par une vitesse de fermentation des ordures ménagères plus élevée que celle du fumier, très riche en lignocellulose. Par des analyses de métabarcoding ADNr 16S, nous avons montré que la composition des communautés microbiennes s'adaptait, dans une large mesure, aux conditions acides, permettant alors la résorption du pic d'acides gras volatils, et ainsi, la levée de l'inhibition. Outre des perspectives en terme de bioindication, ces recherches suggèrent également la possibilité d'adapter progressivement un inoculum à des proportions élevées d'ordures ménagères lors de la co-digestion. Un manuscrit d'article présentant ces résultats a été soumis récemment.

J'ai également contribué à plusieurs autres travaux impliquant la caractérisation des communautés microbiennes de la méthanisation par des approches méta-omiques, sans nécessairement me limiter aux substrats lignocellulosiques (Bize *et al.*, 2015, Delforno *et al.*, 2019). Dans un autre registre, j'ai apporté mon expertise au SIAAP, pendant la pandémie de Covid-19, pour étudier le devenir de SARS-CoV-2 dans les boues de digestion au cours de leur stockage ou de leur méthanisation thermophile (Guérin-Rechdaoui *et al.*, 2022).

Ces différentes activités représentent une partie significative, mais non majoritaire, de mon temps de travail et m'aident à développer une vision plus transversale de l'écologie microbienne appliquée aux procédés de biotechnologies environnementales.

II. Développement d'un système d'information pour capitaliser sur les données méta-omiques de procédés de biotechnologies environnementales

Observant que les données méta-omiques issues de bioprocédés environnementaux s'accumulaient au sein de notre unité, et plus généralement au sein de la communauté, sans réelle possibilité de les exploiter au-delà des projets individuels, j'ai souhaité coordonner le développement d'un outil qui permettrait de capitaliser sur ces données et en favoriserait notamment le partage et la ré-utilisation. Bien que ce projet soit présenté ici de manière très synthétique, il constitue une partie importante de mon activité. Il peut avoir des retombées intéressantes d'un point de vue scientifique, en particulier dans l'optique de développer des biomarqueurs microbiens pour la conduite des méthaniseurs.

Le système d'information est nommé DeepOmics (*Digital Environmental Engineering Platform for Omics data*), et il permet à l'heure actuelle d'entreposer des données de séquençage d'amplicon (données brutes au format *fastq*, données traitées au format *biom*), et leurs métadonnées basées sur le standard MIxS (https://github.com/GenomicsStandardsConsortium/mixs) (Figure 14). Dans DeepOmics, ces données « omiques » peuvent être accompagnées de données « métier » très riches : *design* du procédé, conditions opératoires, et mesures physico-chimiques. Ce système d'information est de plus adapté aux données de laboratoire et de terrain.

Actuellement au stade de pré-production, une instance de DeepOmics est accessible sur demande à différents utilisateurs et est hébergée sur le *data center* INRAE de Toulouse (https://deepomics-test.solapp.inrae.fr/). Les utilisateurs, principalement internes à INRAE dans cette première phase, peuvent depuis plus d'un an entrer des données qui seront conservées lors du passage en production.

Figure 14. Aperçu de l'interface de DeepOmics.
Projet de démonstration pour les données de laboratoire, volet *meta-omics analysis*, *biosample results*.

Le développement de DeepOmics bénéficiait d'une organisation pérenne au sein d'Irstea, la DSI assurant le développement informatique et la maintenance. Depuis la création d'INRAE, une nouvelle organisation pérenne est à mettre en place, car la DSI, de plus grande échelle dans le nouvel institut, ne prend pas en charge ce type de développements. Ce projet relève désormais de l'échelle du département. Cependant, la DSI a accepté d'accompagner encore le projet pendant cette période de transition. De plus, DeepOmics a bénéficié récemment de différents financements sur projets, mais cela ne garantit pas encore sa stabilité. Il serait nécessaire pour cela de disposer de moyens humains pérennes et continus en informatique, afin d'assurer l'exploitation de l'instance de production. Cette expérience met en lumière la question (bien connue) des moyens nécessaires à assurer la pérennité des logiciels informatiques : nombre d'outils bioinformatiques ont été développés par le passé mais n'ont plus été maintenus faute de moyens. En ce sens, des outils génériques, développés par des grands acteurs du domaine, ont certainement plus de facilité à perdurer.

La question du degré de spécialisation des entrepôts de données mérite d'être abordée. Il est incontournable aujourd'hui de déposer les séquences dans les grands entrepôts généralistes que sont INSDC (*International Nucleotide Sequence Database Collaboration*), DDBJ (*DNA Data Bank of Japan*) ou ENA (*European Nucleotide Archive*). Néanmoins, des entrepôts complémentaires spécialisés

peuvent être nécessaires. Nous en avons fait l'expérience dans notre domaine. Sleheddine Kastalli, stagiaire de M2 à INRAE PROSE en 2021, que j'ai co-encadré, a réalisé une méta-analyse statistique sur les communautés microbiennes des systèmes bio-électrochimiques. Parmi 276 articles sur cette thématique, et mentionnant le séquençage, seuls 24 ont *in fine* pu être exploités. Le plus souvent, cela s'expliquait par l'absence de séquences déposées dans les bases de données. Mais dans plusieurs cas, il n'était pas possible de faire le lien entre les séquences déposées et les résultats de l'article, ce qui empêchait de constituer des méta-données, et donc d'exploiter les données. Lors de ce travail de M2, des pistes intéressantes ont été identifiées grâce à cette méta-analyse, comme l'influence de la géométrie de l'électrode sur la composition des communautés microbiennes et sur les performances. Mais la puissance statistique aurait été bien plus élevée, et les conclusions plus robustes, si la totalité des 276 études avait pu être intégrée à la méta-analyse.

Ainsi, il me semble que les entrepôts spécialisés ont un rôle à jouer pour fournir des données homogènes, bien renseignées du point de vue du domaine considéré, aisément identifiables et réutilisables, ce qui n'empêche pas de prévoir une interopérabilité avec les grandes bases de données plus généralistes.

III. Diversité des virus d'archées : des environnements acidothermophiles aux procédés de méthanisation

Après une première phase d'activité centrée principalement sur la méthanisation de substrats lignocellulosiques, j'ai souhaité faire évoluer ma ligne de recherche vers des nouvelles thématiques. S'il est vrai que les méthodes méta-omiques avaient été encore peu employées à cette époque dans le domaine des biotechnologies environnementales, et constituaient donc en elles-mêmes un point d'originalité, la méthanisation de la cellulose était une thématique déjà très largement étudiée au sein de la communauté scientifique (Noike *et al.*, 1985, Adney *et al.*, 1991, Lai *et al.*, 2001, O'Sullivan *et al.*, 2005, Song *et al.*, 2005, Jensen *et al.*, 2009). Dans notre unité, ce sujet était également abordé depuis plusieurs années (Li *et al.*, 2009, Qu *et al.*, 2009, Chapleur *et al.*, 2014).

Ayant étudié des virus d'archées au cours de ma thèse, j'ai naturellement pensé à l'écologie virale des écosystèmes de méthanisation, et j'ai pu constater, au début des années 2010, que très peu d'études avaient été publiées sur ses aspects. Concernant les méthaniseurs, ce sont principalement l'abondance des particules virales et la diversité de leurs morphotypes qui avaient été rapportées, avec une dominance de caudovirus (Park *et al.*, 2007, Wu & Liu, 2009, Chien *et al.*, 2013). En revanche, le procédé de boues activées avait déjà été relativement bien étudié sous l'angle de l'écologie virale (Withey *et al.*, 2005), montrant que les virus de bactéries y infectaient des hôtes divers (Parsley Larissa *et al.*, 2010), qu'ils étaient abondants et/ou actifs (Hantula *et al.*, 1991, Wu & Liu, 2009), qu'ils présentaient dans une certaine mesure des variations temporelles (Lee *et al.*, 2007, Otawa *et al.*, 2007), et qu'ils étaient susceptibles d'affecter les dynamiques hôtes-virus (Lee *et al.*, 2007, Kunin *et al.*, 2008, Shapiro *et al.*, 2010). Ceci confirmait le potentiel scientifique des recherches en écologie virale des procédés de biotechnologie environnementale, et m'a encouragé à poursuivre dans cette direction. De plus, dans des environnements naturels, il était déjà bien établi que les virus étaient des entités biologiques très abondantes et qu'ils pouvaient avoir une grande influence sur les cycles biogéochimiques (Fuhrman, 1999, Weinbauer, 2004, Brussaard *et al.*, 2008, Kimura *et al.*, 2008), un aspect qui semblait tout-à-fait pertinent du point de vue du fonctionnement et de la performance de procédés. J'ai donc construit un premier projet de recherche (ANR JCJC VIRAME, 2017-2023), dont l'objectif était de caractériser *in situ* la diversité de virus d'archées méthanogènes, au sein de méthaniseurs.

Dans ce chapitre, je commence par donner des éléments d'information sur les virus d'archées, puis je présente une sélection de mes travaux de thèse qui concerne les interactions hôtes-virus chez des archées acidothermophiles. Enfin, je synthétise les résultats obtenus dans le cadre du projet VIRAME, sur la composition en k-mers des élément mobiles d'archées et sur la diversité des virus d'archées méthanogènes.

Les virus d'archées ont été étudiés en tant que tels à partir des années 1990, alors que le domaine *Archaea* défini par Carl Woese devenait bien établi (Woese *et al.*, 1978, Woese *et al.*, 1990). Wolfram Zillig a été pionnier dans ces recherches, isolant des plasmides et des virus d'archées acidothermophiles, provenant de sources géothermales terrestres, dans des zones de solfatares (Yellowstone, Pozzuoli, Kamchatka, Islande, …) (Martin *et al.*, 1984, Schleper *et al.*, 1992, Zillig *et al.*, 1993). L'idée première était de favoriser l'obtention d'outils de génétique pour les archées de l'ordre Sulfolobales. En effet, dans le cas des bactéries, les virus (bactériophages) ont joué un grand rôle dans le développement de la biologie moléculaire et de la génétique bactérienne, en lien avec les travaux de Max Delbrück et du groupe Phage (Stent, 1966). Les virus d'archées découverts par Wolfram Zillig ont intrigué par la diversité de leurs morphotypes et de leur contenu génétique, et sont devenus un sujet d'étude à part entière. Par la suite, David Prangishvili, après avoir travaillé avec Wolfram Zillig pendant plusieurs années, a tout particulièrement contribué à caractériser ces virus et à explorer leur diversité. Ses travaux ont abouti à la définition et la caractérisation de nombreuses familles de virus d'archées hyperthermophiles et acidothermophiles (Prangishvili *et al.*, 1999, Bettstetter *et al.*, 2003, Häring *et al.*, 2004, Prangishvili *et al.*, 2006). Depuis une dizaine d'années, Mart Krupovic a notamment poursuivi le développement de ce domaine de recherche, s'intéressant aux archées et à leurs virus (Krupovic *et al.*, 2014, Iranzo *et al.*, 2016, Krupovic *et al.*, 2018, Liu *et al.*, 2021, Medvedeva *et al.*, 2022), et plus largement à l'évolution des virus (Koonin *et al.*, 2015, Krupovic *et al.*, 2019). Les travaux de Mart Krupovic ont remarquablement contribué à faire évoluer les connaissances sur la diversité et l'évolution des virus, et à structurer ces connaissances (Koonin Eugene *et al.*, 2020). Mart Krupovic a par exemple mis en œuvre des approches de réseaux de gènes partagés pour représenter le caractère mosaïque et modulaire de virus et de plasmides, et ainsi mieux appréhender leurs liens évolutifs (Iranzo *et al.*, 2016, Iranzo *et al.*, 2016).

En 2020, 20 familles de virus d'archées étaient décrites (Figure 15), et les hôtes associés recouvraient 3 phylums. Des changements de taxonomie importants sont déjà survenus depuis lors, En particulier, les divisions *Siphoviridae*, *Myoviridae* et *Podoviridae,* fondées sur des caractères morphotypiques et qui ont été pendant de nombreuses années les 3 seules familles de l'ordre Caudovirales, n'existent plus. Les virus d'archées que l'on aurait auparavant affilié à ces trois familles sont désormais répartis en 14 familles distinctes, reflétant plus justement la diversité génétique de ces virus tête-queue. Par ailleurs, trois familles de virus d'archées Asgard ont été identifiées dans des MAGs (*metagenome-assembled genomes*) (Medvedeva *et al.*, 2022).

Parmi ces familles virales, plusieurs sont associées à des capsides icosaédriques, qui comportent soit des protéines avec un repliement de type HK97, soit avec un repliement *double jelly-roll*. Ces virus peuvent être qualifiés de cosmopolites, dans la mesure ou des structures de capsides similaires existent chez des virus d'eucaryotes ou de bactéries. En revanche, les autres familles virales sont spécifiques aux archées.
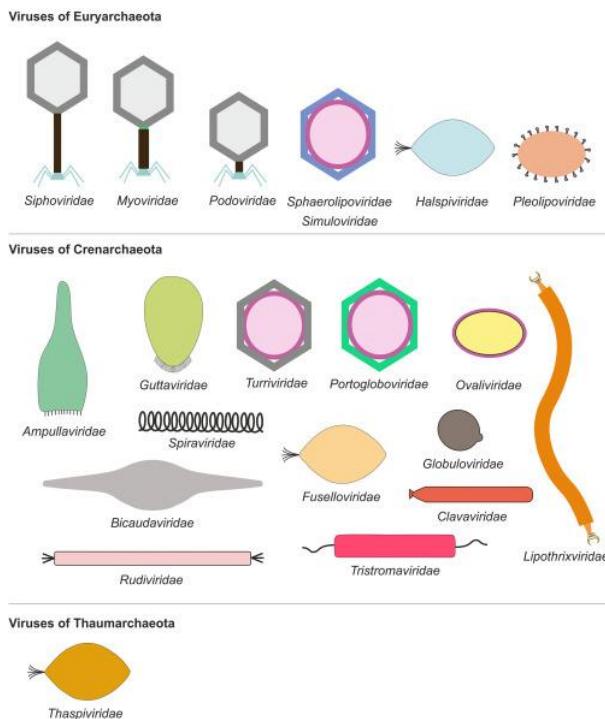
Figure 15: Représentation des morphotypes des virus d'archées infectant des membres de phylums Euryarchaeota, Crenarchaeota and Thaumarchaeota.
Les noms des familles virales sont indiqués sous les schémas des particules virales. D'après (Baquero et al., 2020).

IV.I   Découverte d'un nouveau mécanisme de sortie des virions, chez l'archée acidothermophile *Sulfolobus*

Mon doctorat s'est déroulé à l'Institut Pasteur, dans l'unité de Biologie moléculaire du gène chez les extrêmophiles (BMGE), sous la co-direction de David Prangishvili (BMGE) et Olivier Tenaillon (INSERM). Le cœur de mes travaux de thèse a consisté à déchiffrer les interactions hôte-virus dans le cas de *Sulfolobus islandicus* rod-shaped virus 2 (SIRV2), un virus en forme de baguette (famille *Rudiviridae)*, et son hôte acidothermophile *Sulfolobus islandicus* LAL14/1, isolé à partir d'échantillons de sources chaudes terrestres d'Islande (Zillig *et al.*, 1998). L'hôte croît de manière optimale en conditions aérobies, à un pH de 3.0 et à une température de l'ordre de 78°C. Comme pour la plupart des virus d'archées acidothermophiles alors caractérisés, SIRV2 était décrit comme un virus chronique, c'est-à-dire que des particules virales sont produites en continu au cours du cycle infectieux du virus, sans que cela ne provoque la mort de la cellule hôte. Appelé *carrier state* en Anglais, ce mode de vie des virus rappelle celui de M13 (famille *Inoviridae*) chez les bactéries.

Je devais étudier la co-évolution de SIRV2 et d'autres virus appartenant à différentes familles virales, avec un hôte commun. Lors des premières étapes du projet, plusieurs observations ont cependant soulevé des questions quant au caractère chronique de l'infection par SIRV2. Il est apparu que la nature du cycle infectieux de SIRV2 n'était pas encore clairement identifiée, et ma thèse s'est finalement centrée sur cette question. Ces travaux ont été publiés à l'issue de mon doctorat (Bize *et al.*, 2009).

Tout d'abord, je n'ai pas pu maintenir la co-évolution hôte-virus sur de longues périodes : des variants de l'hôte résistant au virus étaient rapidement sélectionnés. Cela suggérait que l'infection par le virus exerçait une très forte pression de sélection sur l'hôte, ce qui pouvait sembler contradictoire avec un cas d'infection chronique, que l'on pourrait supposer exercer une pression plus modérée sur l'hôte par rapport à un virus purement virulent. De plus, lors de l'étalement des virions de SIRV2 sur tapis

cellulaire, des plages étaient très clairement visibles, rappelant des plages de lyse (Figure 16 A). Pourtant, lorsque l'on mesurait la densité optique de cultures cellulaires en milieu liquide, après infection, cette dernière stagnait, mais ne chutait pas (Figure 16 B).
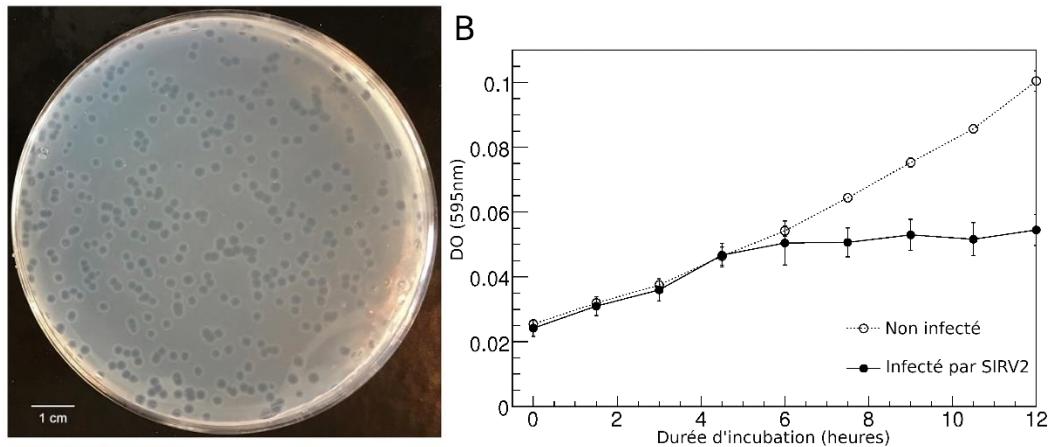
Figure 16: Observations de l'effet de l'infection par SIRV2 sur son hôte.
A. Plages de « lyse » de SIRV2 observées sur un tapis de cellules de *S. islandicus* LAL14/1 (photo issue de (Alfastsen *et al.*, 2021)). B. Effet de l'infection par SIRV2 sur la cinétique de croissance de cultures de S. islandicus. Les cultures, infectées avec une multiplicité d'infection de l'ordre de 7, ou non-infectées, ont été réalisées en triplicats. Les moyennes, +/- 1 écart-type, sont montrées. En cas d'infection, les virus ont été ajoutés après 4,5 heures d'incubation. DO : densité optique.

Afin de mieux cerner la nature du cycle infectieux de SIRV2, j'ai réalisé des observations en microscopie électronique de cellules infectées par ce virus (Figure 17). Dix heures après infection, des agrégats de virions en formation étaient visibles à l'intérieur des cellules (Figure 17 B2, B4, B6). De plus, ces observations ont montré que l'infection virale provoquait la mort des cellules hôtes, et que la densité optique résiduelle s'expliquait par le fait que les enveloppes cellulaires, vidées de leur contenu et adoptant une forme relaxée, demeuraient dans le milieu de culture après la sortie des virions (Figure 17 C1, C2). Enfin, elles ont mis en évidence un mécanisme de sortie des virions encore inconnu : des structures pyramidales à base heptagonale, distinctes des virions eux-mêmes, se forment au niveau de l'enveloppe cellulaire (Figure 17 B1-B5), et s'ouvrent par leur sommet en fin de cycle infectieux, permettant alors la sortie des virions dans le milieu extracellulaire (Figure 17 C2-C4). Ces structures ont été nommées VAPS (*Virus-Associated Pyramids*).

Les observations en microscopie électronique ont été complétées par des analyses de cytométrie en flux, en collaboration avec le laboratoire de Rolf Bernander (Suède), afin de quantifier les évolutions d'ADN total intracellulaire au cours du cycle infectieux, à l'échelle de cellules individuelles. Le détail des résultats se trouve dans l'article correspondant (Bize *et al.*, 2009). Les quantités d'ADN intracellulaire augmentaient dans un premier temps, ce que l'on pouvait supposer être lié à la réplication de l'ADN viral, puis elles chutaient brutalement, lors de la lyse cellulaire, en cohérence avec les observations de microscopie. Des expériences d'hybridation d'ADN (*Southern*), également visibles dans l'article, ont permis d'évaluer les quantités moyennes d'ADN intracellulaire, sur un grand nombre de cellules, tout en discriminant entre l'ADN de SIRV2 et l'ADN cellulaire, grâce à l'emploi de sondes radioactives spécifiques. Ces analyses ont conforté les conclusions concernant la nature lytique de SIRV2, et la durée du cycle infectieux. Elles ont montré que la dégradation d'ADN cellulaire débutait en partie avant la lyse, suggérant un possible recyclage de l'ADN cellulaire par le virus, comme cela est le cas pour les bacteriophages T chez *E. coli* (ex : (Hershey *et al.*, 1953)).
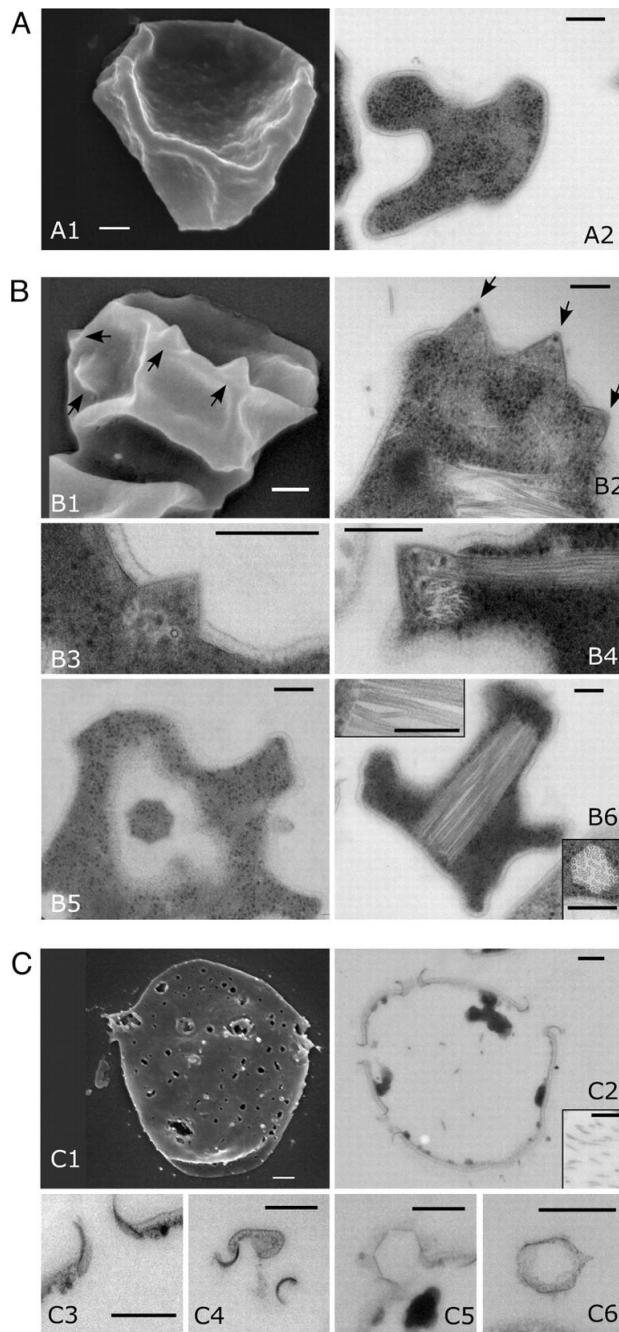
A1, B1, C1 : images obtenues par microscopie à balayage. Les autres images ont été obtenues par microscopie électronique en transmission (coupes) A. Cellules non infectées. B. Cellules 10 heures après infection. Les flèches indiquent des VAPs. On note la présence d'agrégats de particules virales en formation, à l'intérieur des cellules. En B5, la section perpendiculaire à la base d'une VAP met en évidence sa base heptagonale. C. Cellules 26 heures après infection. On note la présence de VAPs partiellement détruites. Barres : 200 nm.

Notons qu'un mécanisme similaire de sortie des virions, impliquant des structures pyramidales, a été découvert par une autre équipe à la même période, de manière indépendante, pour le virus icosédrique *Sulfolobus islandicus* turreted icosahedral virus 1 (STIV1) (Brumfield Susan *et al.*, 2009). SIRV2 et STIV1 partagent peu de gènes similaires, en particulier du fait qu'ils appartiennent à des familles différentes, *Rudiviridae* pour le premier et *Turriviridae* pour le second. Cela a permis de limiter à 3 le nombre de gènes candidats que l'on pouvait dans un premier temps supposer être impliqués dans ce mécanisme de sortie : SIRV2gp49 (ORF98), similaire à C92 chez STIV1, SIRV2gp35 (ORF121), similaire à A58 chez STIV1, et enfin SIRV2gp27 (ORF114), similaire à B116 chez STIV1. A la suite de mes

travaux de thèse, Tessa Quax a réalisé son doctorat dans le même laboratoire à l'Institut Pasteur, et elle a notamment poursuivi la caractérisation de ces structures pyramidales, identifiant P98 (SIRV2gp49, ORF98) comme leur principal constituant (Quax *et al.*, 2010, Quax *et al.*, 2011). SIRV2 s'est de plus développé comme virus modèle pour les archées, avec en particulier la mise au point d'une méthode d'édition de son génome basée sur CRISPR Cas, par l'équipe de Xu Peng (Mayo-Muñoz *et al.*, 2018, Alfastsen *et al.*, 2021), et le séquençage du génome de son hôte cellulaire (Jaubert *et al.*, 2013).

En conclusion, ma thèse s'est déroulée à une période où la recherche sur les virus d'archées était encore relativement jeune. Depuis, ces recherches se sont bien développées, bénéficiant particulièrement des nouvelles technologies telles que l'édition de génome avec CRISPR-Cas et le RNA-seq. Lorsque j'ai débuté ma thèse, les communautés scientifiques étudiant respectivement les virus d'archées et ceux de bactéries interagissaient relativement peu, ce qui se reflétait en particulier dans des différences terminologiques : *carrier state* / chronique, virus / bactériophage (…), mais également des différences de méthodes d'isolation des virus. Les liens ont depuis été renforcés, avec la création en 2010 de la conférence *Viruses of microbes*, qui rapproche les communautés scientifiques étudiant les virus d'archées, de bactéries, et d'eucaryotes unicellulaires. De tels échanges me semblent essentiels pour enrichir les points de vue et les approches.

IV.II   Identification des facteurs qui influencent la composition en k-mers des plasmides, virus et leurs hôtes, chez les archées

Quelques années après mon arrivée au Cemagref (devenu Irstea en 2012 puis INRAE en 2020), j'ai souhaité, comme évoqué précédemment, développer des recherches en écologie virale, avec l'analyse de métaviromes. Ce dernier aspect était nouveau pour moi. Certains outils d'analyse de métaviromes reposent sur les compositions en k-mers des génomes, en particulier pour la prédiction d'hôtes (Galiez *et al.*, 2017). Ayant prévu d'utiliser de tels outils, j'ai décidé de mettre à jour et de publier une analyse que j'avais débutée lors de ma thèse, qui consistait à explorer les compositions en k-mer au sein des archées, leurs virus et leurs plasmides. Cela me permettrait d'avoir une meilleure vision des facteurs modelant les profils de composition en k-mers, et donc de mieux percevoir le potentiel et les limites des outils de métagénomiques exploitant ce type information. A l'époque, j'avais observé pour l'ordre Sulfolobales que chaque famille d'élément extrachromosomique avait sa propre signature k-mer.

Pour mettre à jour cette étude, j'ai travaillé en collaboration avec Patrick Forterre (Institut Pasteur) et Violette Da Cunha (désormais au Genoscope). J'ai rassemblé la plupart des séquences de génomes de virus et plasmides d'archées disponibles dans les bases de données publiques, puis j'ai inclus les séquences des génomes d'archées appartenant aux mêmes ordres que leurs hôtes. Par exemple, si un virus infectant une archée de l'ordre Sulfolobales était présent dans le jeu de données, tous les génomes cellulaires de l'ordre Sulfolobales étaient intégrés à l'étude, qu'il s'agisse ou non de la souche hôte précise de ce virus. Cette étude est assez systématique et descriptive, j'ai donc choisi de l'illustrer par une sélection de résultats représentatifs, plutôt que par une présentation exhaustive. L'ensemble des résultats est visible dans la publication issue de ces travaux (Bize *et al.*, 2021).

L'étude portait sur un total de 589 séquences de cellules, plasmides, virus et provirus d'archées, associés à 11 ordres taxonomiques d'archées différents. Il s'agissait principalement d'halophiles, d'acidothermophiles, d'hyperthermophiles et de méthanogènes, avec également quelques génomes de *Marine Group II*, le groupe d'archées planctoniques le plus abondant à la surface des océans (Rinke *et al.*, 2019). A l'époque, toutes les archées incluses dans l'étude appartenaient aux phylums Crenarchaeota ou Euryarchaeota, ce qui reste encore vrai, à l'exception éventuelle de *Marine Group II*

([Tableau 3](#)). Il s'agit pour l'essentiel d'archées cultivées, qui sont loin de représenter la totalité de la diversité des archées.

Tableau 3. Ordres et phylums d'appartenance des archées incluses dans l'étude.

Ordre	Phylum
Desulfurococcales	Crenarchaeaota (TACK group)
Halobacteriales	Euryarchaeota
Haloferacales	Euryarchaeota
Natrialbales	Euryarchaeota
Marine Group II **(Candidatus Poseidoniales)**	Euryarchaeaota / Candidatus Thermoplasmatota
Methanobacteriales	Euryarchaeota
Methanoccocales	Euryarchaeota
Methanosarcinales	Euryarchaeota
Sulfolobales	Crenarchaeaota (TACK group)
Thermococcales	Euryarchaeota
Thermoproteales	Crenarchaeaota (TACK group)

Les profils ont été établis à partir des 1 024 mots de taille 5 (5-mers, 1 024 = 4^5), correspondant à un compromis entre spécificité et représentativité : les plasmides et virus ayant des génomes de petite taille, il est préférable d'utiliser des k-mers courts pour éviter d'avoir trop de zéros dans les matrices de comptage, ce qui introduirait des biais. Le dendrogramme de regroupement hiérarchique (*clustering*) présenté en [Figure 18](#) rassemble un sous-jeu de donné sélectionné aléatoirement, pour assurer la lisibilité de l'illustration. On observe tout d'abord un fort lien entre le contenu en GC des génomes et le pattern de clustering. En effet, le dendrogramme comporte 2 clusters principaux, l'un avec essentiellement des séquences riches en GC, et l'autre pauvres en GC. Concernant les génomes cellulaires, ce dendrogramme ne reflète pas fidèlement la phylogénie. Par exemple, les génomes des ordres Methanoccales et Methanosarcinales (phylum Euryarchaeaota), ne sont pas dans le même grand cluster que les génomes d'haloarchées (phylum Euryarchaeaota également, cluster a). En revanche, ils sont dans le même grand cluster que l'essentiel des génomes de Sulfolobales, qui appartiennent pourtant à un autre phylum, Crenarchaeaota (clusters b et d).

A l'échelle globale, il apparaît de plus que les virus et plasmides ne forment pas de cluster distinct des cellules. Ils ont tendance à être groupés avec des archées de même taxonomie que leurs hôtes. Cependant, le rang taxonomique varie selon les cas. Pour les archées halophiles, c'est au niveau de la classe, Halobacteria, qui rassemble ici les ordres Halobacteriales (orange), Haloferacales (jaune) et Natrialbales (vert), qu'on observait une bonne cohérence : cellules, plasmides et virus étaient rassemblés au sein du cluster a ([Figure 18](#)), à quelques exceptions près. Le même type de pattern était retrouvé, au rang taxonomique de l'ordre, pour Sulfolobales (rouge, principalement dans le cluster b), Thermococcales (vert foncé, principalement dans le cluster c) et Methanococcales (bleu vert, principalement dans le cluster d). Cela était moins clair pour les ordres Methanobacteriales, Thermoproteales, Desulfurococcales, et *Marine Group II*, dont les génomes étaient plus dispersés en diverses positions du dendrogramme. Certaines associations entre hôtes et éléments extrachromosomiques étaient néanmoins visibles au sein de petits clusters isolés, par exemple pour les ordres Methanobacteriales (gris clair) au sein du cluster e et Desulfurococcales (rose) au sein du cluster f.

Bien qu'imparfaite, cette association entre les éléments mobiles et leurs hôtes reflète l'effet de la co-évolution sur leur composition génomique en k-mers courts. On peut supposer que cet effet provient en grande partie d'une adaptation des éléments extrachromosomiques à l'usage des codons de l'hôte.

Cet effet semble dans la plupart des cas être cloisonné au rang taxonomique de l'ordre. Dans le cas de archées halophiles, cela se produit au niveau de la classe, comme déjà évoqué. Ceci peut certainement s'expliquer par une évolution convergente des compositions en k-mer de l'ADN, due à la forte pression de sélection exercée par la haute salinité (Paul *et al.*, 2008).

Au sein des 4 groupes pour lesquels le pattern d'association entre hôte et éléments extrachromosomiques était le plus cohérent (classe Halobacteria, ordres Sulfolobales, Thermococcales et Methanococcales), les positions des cellules, et celles des éléments extrachromosomiques dans le dendrogramme, n'étaient pas totalement entremêlées. Il était au contraire possible d'observer des clusters riches soit en cellules, soit en éléments extrachromosomiques. Cela indique qu'à un niveau plus fin, au-delà de la forte influence de la co-évolution avec leur hôte, les éléments extrachromosomiques conservent une composante spécifique en termes de composition en k-mers courts, probablement liée à leur nature différente.
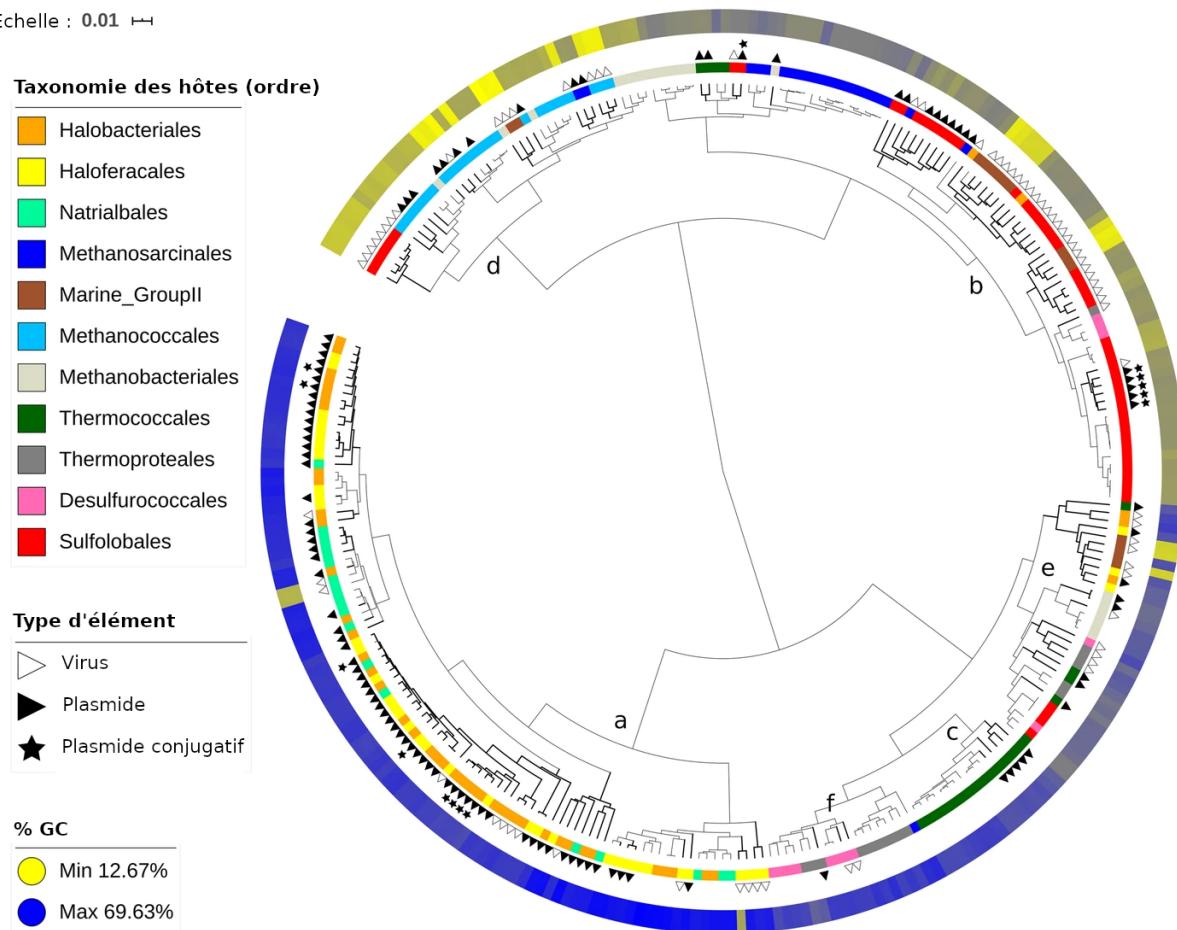
Figure 18. Dendrogramme basé sur les fréquences en 5-mers pour un sous-ensemble de archées et leurs éléments mobiles. Les comptages de 5-mers ont été obtenus avec Jellyfish 2.2.6 en utilisant de cluster de calcul de la plateforme bioinformatique INRAE MIGALE (https://migale.inrae.fr/) puis convertis en fréquences. Le clustering hiérarchique a été réalisé avec la fonction hclust de R, appliquée à la matrice de distances euclidiennes, avec la méthode Ward.D2. La visualisation a été générée avec iTOL (Letunic & Bork, 2007).

Cela était particulièrement bien illustré par le cas de Sulfolobales (Figure 18, lettre b), que j'ai examiné plus en profondeur, car il comporte une variété importante de familles virales et plasmidiques (Figure 19). Sans trop entrer dans les détails, on peut observer un *pattern* de *clustering* très cohérent pour les cellules, regroupées en 2 clusters distincts (couleur noir, Figure 19) : genres *Sulfolobus* et *Acidianus* d'une part (codes débutant par s et a), genre *Methallosphaera* d'autre part (codes débutant par m). Les génomes de ce dernier genre sont plus riches en GC, ce qui pourrait expliquer cette partition. De

façon similaire, le résultat du clustering était globalement très cohérent pour les familles virales *Rudiviridae* (turquoise), *Fuselloviridae* (rouge), *Lipothrixviridae* (vert clair), *Ampullaviridae* (vert foncé) et *Turriviridae* (beige clair), ainsi que pour les plasmides conjugatifs de la famille pNOB8 (orange) et les plasmides cryptiques de la famille pRN (magenta). De manière intéressante, les plasmides pRN étaient beaucoup plus distants du principal cluster d'hôte que les plasmides pNOB8, ces derniers formant le cluster le plus proche des hôtes. Ceci suggère que les plasmides de grande taille ont des compositions en k-mers courts similaires à celles des cellules, ce qui est moins le cas des plasmides courts et des virus. Cela pourrait s'expliquer par des fréquences d'échanges génétiques avec les hôtes plus faibles chez ces derniers.

Un réseau de gènes partagés, construit avec le même ensemble de génomes liés à l'ordre Sulfolobales, était en accord avec les principales conclusions observées sur la base des profils de composition en 5-mers (Figure 19 B). En particulier, la cohérence par familles, ainsi que la proximité entre les plasmides conjugatifs pNOB8 et leurs hôtes ont été retrouvées. Une exception notable est la proximité entre *Lipothrixviridae* et *Rudiviridae* : ces 2 familles ont des liens évolutifs, formant l'ordre *Ligamenvirales*. Si ces liens étaient bien visibles dans le réseau de gènes partagés, ce n'était pas le cas dans le dendrogramme basé sur les signatures génomiques, ce qui pourrait en partie s'expliquer par le très bas contenu en GC des membres de *Rudiviridae* (28,25% ± 6,17 en moyenne).

En utilisant ce réseau de gènes partagés, le nombre de liaisons entre différents types d'éléments a été calculé, en le pondérant par le nombre d'éléments présents dans les groupes considérés (Figure 19 C). Il apparaît un nombre plus élevé de liaisons entre cellules et plasmides qu'entre cellules et virus. De plus, en distinguant les deux familles de plasmides distinctes, pNOB8 et pRN, il apparaît que c'est avec la famille plasmide pNOB8 que le nombre de liaison est très élevé, tandis qu'il est très faible avec la famille de pRN. Ces observations sont tout-à-fait cohérentes avec les hypothèses formulées à partir des analyses de composition en 5-mers, à savoir une fréquence d'échanges génétiques plus élevée entre cellules et plasmides conjugatifs.

Cette étude conforte la possibilité d'utiliser les signatures génomiques pour prédire les hôtes des éléments mobiles, tel que le permettent certains outils, comme WiSH (Galiez *et al.*, 2017) ou PlasFlow (Krawczyk *et al.*, 2018). Il existe cependant des limites importantes. Typiquement, prédire un hôte précis pour un élément mobile halophile peut sembler hors de portée en utilisant des approches basées sur les k-mers. De plus, en dépit du *pattern* global de co-évolution observé, de nombreux éléments mobiles ont des compositions en 5-mer atypiques. Cela pourrait s'expliquer par la présence de gènes codants pour des ARN de transferts, comme évoqué par d'autres auteurs (Galiez *et al.*, 2017). Dans notre étude, nous n'avons pas noté de lien particulier entre composition atypique et présence de gènes d'ARNt, mais il est à noter que nous avons utilisé uniquement les annotations disponibles dans les bases de données publiques, et qu'il aurait pu être intéressant de mener spécifiquement une détection de ces gènes, car certains génomes n'avaient peut-être pas encore été analysés sous cet angle. Ainsi, la prédiction d'hôte à partir des signatures génomiques semble possible mais à considérer avec prudence, et son degré de précision peut varier selon les taxons d'hôtes. L'incohérence du dendrogramme avec la phylogénie des archées, et d'autres observations de détails que je n'expose pas ici, conduisent de plus à penser que la composition en k-mer des génomes évolue relativement rapidement, et qu'il peut donc être risqué d'utiliser les signatures k-mers pour effectuer de la reconstruction phylogénétique, tout particulièrement pour des événements ancestraux.
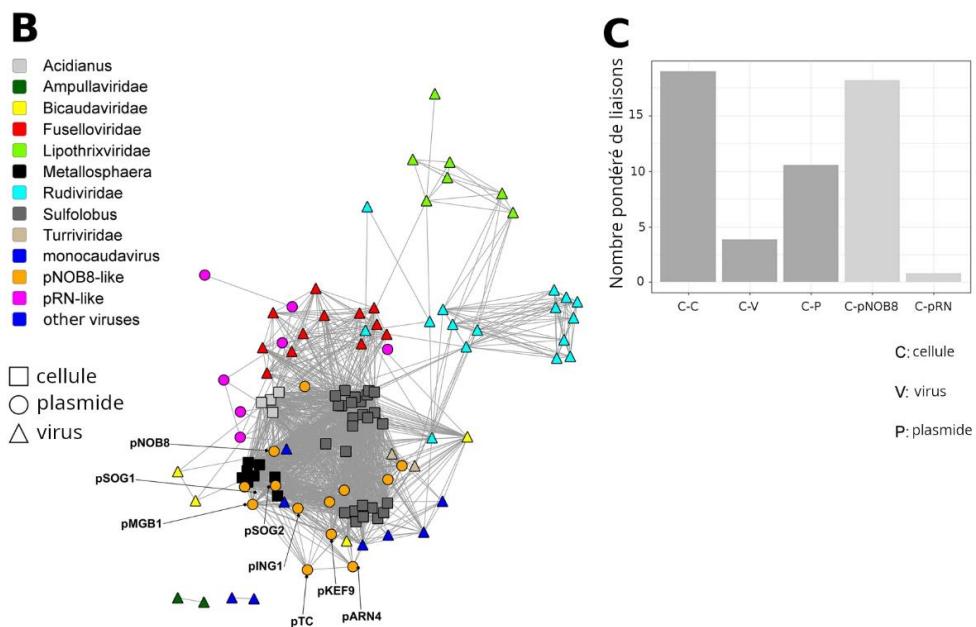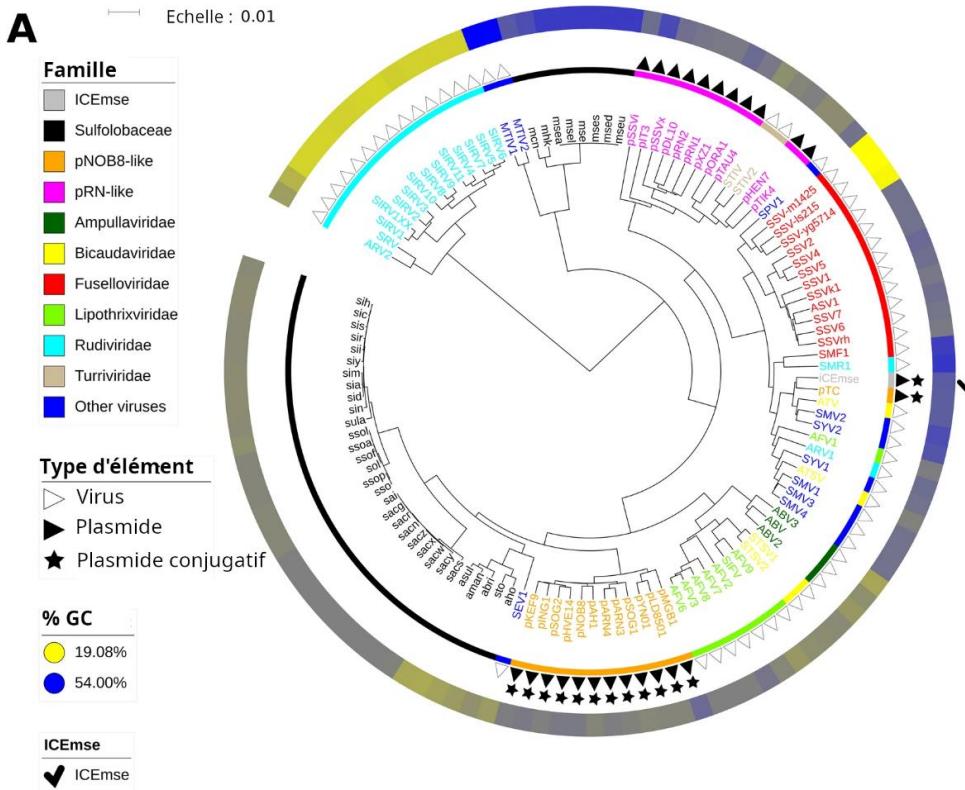
Figure 19. Aperçu des éléments mobiles de l'ordre Sulfolobales.

A. Dendrogramme basé sur les fréquences de 5-mers pour les membres de l'ordre Sulfolobales et leurs plasmides et virus. B. Réseau de gènes partagés basé sur le nombre normalisé de gènes partagés. Pour chaque paire d'éléments, le nombre de gènes partagé a été divisé par la longueur du génome le plus court des deux. De plus, les liaisons (*edges*) dont les valeurs normalisées était inférieures à 0.1 ne sont pas montrées, afin d'éliminer les interactions faibles. C. Graphique en barres du nombre de liaisons au sein du réseau, selon différentes catégories d'éléments. Les valeurs ont été normalisées pour prendre en compte le nombre d'éléments dans les catégories considérées. ICEmse correspond à un plasmide conjugatif de *Metallosphaera* identifié au cours de l'étude.

IV.III Diversité génomique des virus infectant les archées méthanogènes au sein de microcosmes de méthanisation du formate

Pour mon premier projet de recherche sur les virus de microorganismes présents dans les méthaniseurs, j'ai choisi de me concentrer sur les virus d'archées méthanogènes, à la croisée de différents questionnements fondamentaux et appliqués. En effet, les archées méthanogènes représentent un groupe fonctionnel clé du processus méthanisation, en catalysant l'ultime étape, qui génère le méthane : la méthanogenèse. Il est donc pertinent de mieux connaître leurs virus et les effets qu'ils peuvent avoir au sein des digesteurs. Par ailleurs, le méthane est un gaz à effet de serre puissant, émis dans différents environnements (sédiments, rizières, ruminants, fonte du permafrost…). Mieux connaître les virus d'archées et les fonctions biologiques qu'ils encodent revêt donc un intérêt pour la compréhension des écosystèmes environnementaux, et pour la mitigation des émissions de méthane. Enfin, parmi les virus d'archées, ceux infectant les méthanogènes sont encore relativement peu caractérisés. Une dizaine d'entre eux a été isolée jusqu'à présent (Tableau 4), ce qui reste limité considérant la grande diversité phylogénétique des archées méthanogènes. Ces virus ont été isolés en deux vagues successives. Une première a eu lieu vers les années 1990, alors que le domaine du vivant *Archaea* devenait bien établi (Woese *et al.*, 1990). La seconde a débuté en 2017, après une interruption de plus de 10 ans dont l'origine n'est pas claire. Elle pourrait s'expliquer par la moindre diversité apparente des virus d'archées méthanogènes, dominés par les caudovirus, qui aurait pu susciter moins d'intérêt dans un premier temps. La relative difficulté de cultivation des méthanogènes, par rapport aux modèles aérobies que sont les archées halophiles ou acidothermophiles, aurait pu renforcer ce désintérêt passager. Revenant aux virus de méthanogènes, on notera qu'à l'exception d'un virus issu d'une cheminée hydrothermale, tous les autres ont été isolés à partir de boues de méthaniseurs mésophiles ou thermophiles (Tableau 4), qui doivent relativement accessibles aux prélèvements. Outre ces 10 virus isolés, plusieurs provirus ont également été identifiés au sein de génomes d'archées méthanogènes. Ils ne seront pas décrits en détail ici.

Dans le cadre du projet ANR JCJC VIRAME (2017-2023), nous nous sommes concentrés sur la diversité des virus infectant des archées méthanogènes mésophiles, au sein de méthaniseurs. Ces travaux ont été réalisés essentiellement par Hoang Ngo, doctorant dont j'ai été l'encadrante principale (directeur de thèse : Théodore Bouchez), de janvier 2019 à juin 2022. Une collaboration a de plus été établie avec Mart Krupovic (Institut Pasteur), pour son expertise sur les virus d'archées, ainsi que François Enault (Université Clermont-Auvergne, LMGE), bioinformaticien dont les recherches portent spécifiquement les métaviromes. L'objectif était de caractériser cette diversité *in situ*, au sein de microcosmes de méthanisation mésophiles alimentés avec des substrats de méthanogenèse, en combinant marquage isotopique (*Stable Isotope Probing, SIP*) et séquençage métagénomique *shotgun*. Connaissant les limites des approches métagénomiques, j'ai souhaité utiliser le *SIP* en complément, pour pouvoir établir un lien direct plus direct avec l'activité des microorganismes.

Tableau 4. Virus d'archées méthanogènes (et pseudovirus) isolés jusqu'à présent.

Nom	Morphologie et famille	Genre de l'hôte	Origine	Génome à ADN	Référence
ΨM1	Tête-queue *Leisingerviridae*	*Methanothermobacter*	DA thermophile	30,4 kb - linéaire	(Meile *et al.*, 1989)
ΨM2*	Tête-queue *Leisingerviridae*	*Methanothermobacter*	DA thermophile	26,1 kb - linéaire	(Pfister *et al.*, 1998)
ΦF1	Tête-queue	*Methanobacterium*	DA thermophile	85 kb - linéaire	(Nölling *et al.*, 1993)
ΦF3	Tête-queue *Anaeroviridae*	*Methanobacterium*	DA thermophile	36 kb – linéaire ou circulaire	(Nölling *et al.*, 1993)
Drs3	Tête-queue *Anaeroviridae*	*Methanobacterium*	DA mésophile	37 kb - linéaire	(Wolf *et al.*, 2019)
Blf4	Tête-queue *Pungoviridae* ?	*Methanoculleus*	DA mésophile	37 kb - potentiellement circulaire	(Weidenbach *et al.*, 2021)
MetSV	Sphérique	*Methanosarcina*	DA mésophile	11 kb - linéaire	(Weidenbach *et al.*, 2017)
A3-VLP	Oblat	*Methanococcus*	DA mésophile	23 kb - circulaire	(Wood *et al.*, 1989)
MFTV1*	Tête-queue *Fervensviridae* ?	*Methanocaldococcus*	cheminée hydrothermale	31 kb -linéaire	(Thiroux *et al.*, 2021)

* : indique que le génome viral contient un gène d'intégrase – DA : digestion anaérobie.

Les archées méthanogènes sont généralement sous-dominantes dans les procédés de méthanisation, représentant de l'ordre de 5 à 15% de la communauté microbienne catalytique (Nelson *et al.*, 2011). Afin de pouvoir en étudier les virus plus aisément, nous avons utilisé du formate, l'un des substrats possibles de la méthanogenèse (Figure 20), pour enrichir la communauté microbienne en archées méthanogènes. De plus, nous avons employé du formate marqué au carbone 13 pour pouvoir séparer les ADN enrichis en ^{13}C, et ainsi différencier les archées méthanogènes actives, ayant assimilé le substrat marqué, des archées méthanogènes inactives. Il est à noter que le formate n'est pas un substrat spécifique aux archées méthanogènes. Il peut également être utilisé comme source de carbone par des bactéries formatotrophes, selon différentes voies métaboliques. La voie réductrice de l'acétyl-CoA est par exemple montrée en Figure 20. Des bactéries formatotrophes pouvaient donc également être sélectionnées par le substrat formate, et devenir abondantes.
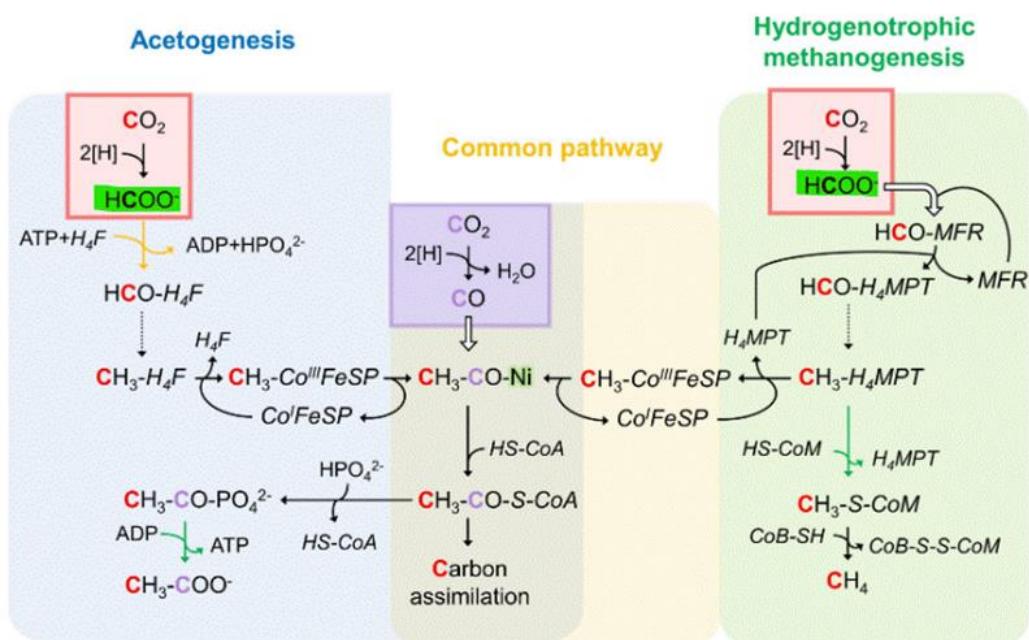
Figure 20. Voies réductrices de l'acétyl-CoA chez les bactéries et les archées méthanogènes.
Illustration d'après (Lemaire *et al.*, 2020). La voie réductrice de l'acétyl-CoA est également appelée voie de Wood-Ljungdahl. A gauche : branche du méthyl – Au milieu : branche du carbonyle – A droite : méthanogenèse hydrogénotrophe. Les flèches vertes indiquent des réactions couplées à la conservation d'énergie, et les oranges à l'hydrolyse d'ATP.

Nous avons établi 6 microcosmes mésophiles et discontinus (*batch*) de méthanisation du formate, 3 contenant du formate ^{13}C et les 3 autres contenant du formate non-marqué. Nous avons employé un ratio inoculum/substrat relativement bas pour limiter l'apport en carbone sous une autre forme que le formate. La quantité et la composition de biogaz produit ont été mesurées, ainsi que les valeurs de pH et les concentrations en acides gras volatils. Ceci a permis de vérifier la consommation du formate et la bonne mise en place de production de biogaz riche en méthane. A trois points de temps différents (8, 13 et 17 jours d'incubation), des paires de microcosmes, l'un alimenté au ^{13}C-formate, l'autre alimenté au formate non-marqué et constituant un contrôle, ont été sacrifiées afin de collecter les particules virales et d'analyser les ADN cellulaires et viraux. Ces points de temps correspondaient aux début, milieu et fin de la phase active de méthanogenèse.

Les ADN cellulaires ont été séparés selon leur densité, dans un gradient de Chlorure de Cesium, et les fractions collectées (fraction lourde, légère et intermédiaire) ont été séquencées par métabarcoding ADNr 16S. Ces premières analyses ont confirmé l'enrichissement de la communauté microbienne en archées méthanogènes au cours du temps, leur abondance relative passant de 1-3% au jour 0, à 26-

30% au jour 17. Il est de plus apparu que les archées étaient suffisamment enrichies en ^{13}C uniquement au jour 17, lorsque la production de biogaz s'achevait. Leur abondance relative atteignait alors 56% dans la fraction lourde. Ces archées étaient dominées par des méthanogènes hydrogénotrophes, *Methanobacterium* et dans une moindre mesure, *Methanobrevibacter* (Berghuis *et al.*, 2019), ce qui était cohérent avec l'emploi de formate. Ces tendances ont été confirmées par les résultats de séquençage métagénomique *shotgun*. Ainsi, la stratégie expérimentale choisie pour pouvoir étudier les virus d'archées méthanogènes consommant le formate fonctionnait à ce stade, nous avons donc poursuivi l'étude avec l'analyse des virus.

Les observations en microscopie électronique en transmission, réalisées à la plateforme INRAE MIMA2 de Jouy-en-Josas, ont suggéré une abondance relativement élevée de particules virales, et ont mis en évidence leur grande diversité (Figure 21). En particulier, des virions fusiformes, d'abondance modérée, ont été observés de manière répétée. Il s'agit d'un morphotype spécifique aux virus d'archées. Des virus en forme de baguette étaient également présents mais en abondance moindre : ils pourraient provenir d'archées ou de plantes, et seraient dans ce dernier cas vraisemblablement apportés par l'inoculum, contenant des résidus de biodéchets.
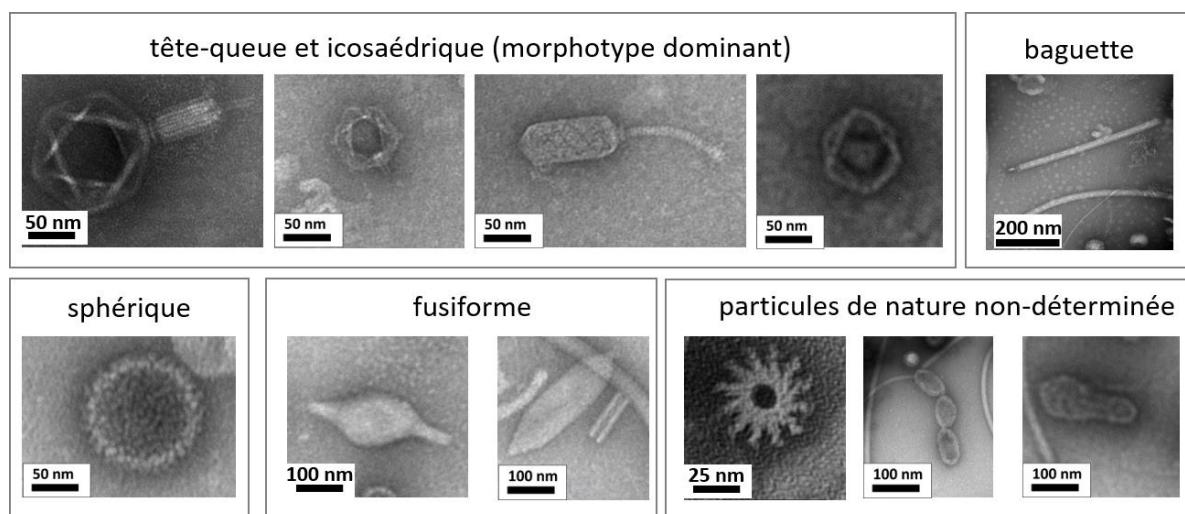
Figure 21. Aperçu de la diversité des particules virales dans les microcosmes.
Observations en microscopie électronique à transmission (plate-forme INRAE MIMA2, Jouy-en-Josas).

Une diversité intéressante de morphotypes viraux ayant notamment été observée aux jours 13 et 17, les métaviromes de ces deux points de temps ont été séquencés avec la technologie Illumina, soit 2 métaviromes. En complément, les métagénomes cellulaires des jours 13 et 17, ainsi que les 3 fractions de différentes densités du jour 17, ont également été séquencés, soit 5 métagénomes cellulaires. Ces métagénomes cellulaires ont été exploités pour améliorer la prédiction d'hôte des contigs viraux, par deux approches complémentaires. La première approche était basée sur la détection de *protospacers* dans les contigs viraux, en utilisant une base de données de *spacers* publique et, également, une base de donnée issue de la détection de *spacers* dans nos contigs cellulaires. La seconde approche reposait sur les signatures génomiques, grâce à l'outil WIsH (Galiez *et al.*, 2017), en employant, pour base de donnée d'hôtes potentiels, 81 *metagenomes assembled genomes* (MAGs) obtenus à partir de nos contigs cellulaires. Ces derniers incluaient 17 MAGs d'archées.

Ces informations ont été combinées aux annotations taxonomiques et fonctionnelles des gènes des métaviromes, ainsi qu'à l'annotation taxonomique des contigs viraux, pour prédire au mieux les hôtes des contigs viraux. Par ailleurs, trois outils de prédiction de génomes viraux, VirSorter2 (Guo *et al.*,

2021), CheckV (Nayfach *et al.*, 2021) et VIBRANT (Kieft *et al.*, 2020) ont été employés pour valider la nature virale des contigs étudiés. En intégrant l'ensemble de ces résultats d'analyses, des filtres automatiques et une validation experte ont permis de retenir 39 contigs viraux susceptibles de provenir de virus d'archées méthanogènes, sur un total de 5 571 contigs de plus de 3 kb obtenus par le co-assemblage des séquences de métaviromes.

Afin d'accéder à un aperçu global de la diversité de ces 39 contigs viraux, un réseau bipartite a été construit, permettant de visualiser les protéines partagées entre ces contigs ainsi qu'avec les génomes des virus d'archées déjà connus (Figure 22). Deux zones distinctes étaient visibles. La partie supérieure du réseau était dominée par des virus de la classe *Caudoviricetes*, c'est-à-dire de morphotype tête-queue (magrovirus, siphovirus, myovirus). La partie inférieure du réseau était riche en virus spécifiques aux archées (virus pléiomorphes *Pleolipoviridae*, virus en forme de filaments ou de baguettes *Tokiviricetes*, virus fusiformes, etc). Cette topologie est cohérente avec un réseau bipartite de virus d'archées précédemment publié (Iranzo *et al.*, 2016). Les 39 contigs viraux de l'étude étaient principalement situés dans la zone des caudovirus, tout en étant pour certains très distants, c'est-à-dire avec un nombre limité de connexions. Une minorité de contigs viraux, de l'ordre de 5, étaient situés dans la zone des familles virales spécifiques aux archées.

Cet aperçu est conforme avec la vision actuelle basée que nous avons de la diversité des virus d'archées méthanogènes, basée sur les quelques virus d'archées méthanogènes isolés jusqu'à présent : ces derniers sont dominés par les caudovirus (Tableau 4). Cela montre aussi qu'une fraction sous-dominante des virus d'archées méthanogènes pourrait appartenir à des familles spécifiques aux virus d'archées, moins cosmopolites que les caudovirus. Enfin, cela suggère que l'essentiel de cette diversité reste à découvrir. En effet, un seul virus d'archée méthanogène isolé jusqu'à présent n'est pas un caudovirus. Il s'agit de MetSV, virus sphérique de *Methanosarcina* (Tableau 4). De plus, considérant le faible nombre total de virus de méthanogènes isolés, il est certain que de nombreux caudovirus sont encore inconnus, en cohérence avec le nombre important de contigs viraux qui apparaissent faiblement connectés aux caudovirus, dans notre étude. Cette dernière observation est à nuancer du fait que la plupart de ces contigs viraux représentent des génomes incomplets, ce qui pourrait introduire un biais, à la baisse, sur le nombre de connexions détectées entre 2 contigs/génomes viraux.

Par la suite, nous nous sommes focalisés sur 4 contigs viraux présentant un intérêt particulier, soit parce qu'il s'agissait de génomes complets, soit parce qu'il pourrait s'agir de nouvelles familles virales spécifiques au archées (indiqués par une étoile rouge dans la Figure 22). Je présente ici deux d'entre eux (Figure 23), qui illustrent bien le potentiel de l'approche combinant SIP et métagénomique pour différencier les virus infectant des archées actives ou inactives.
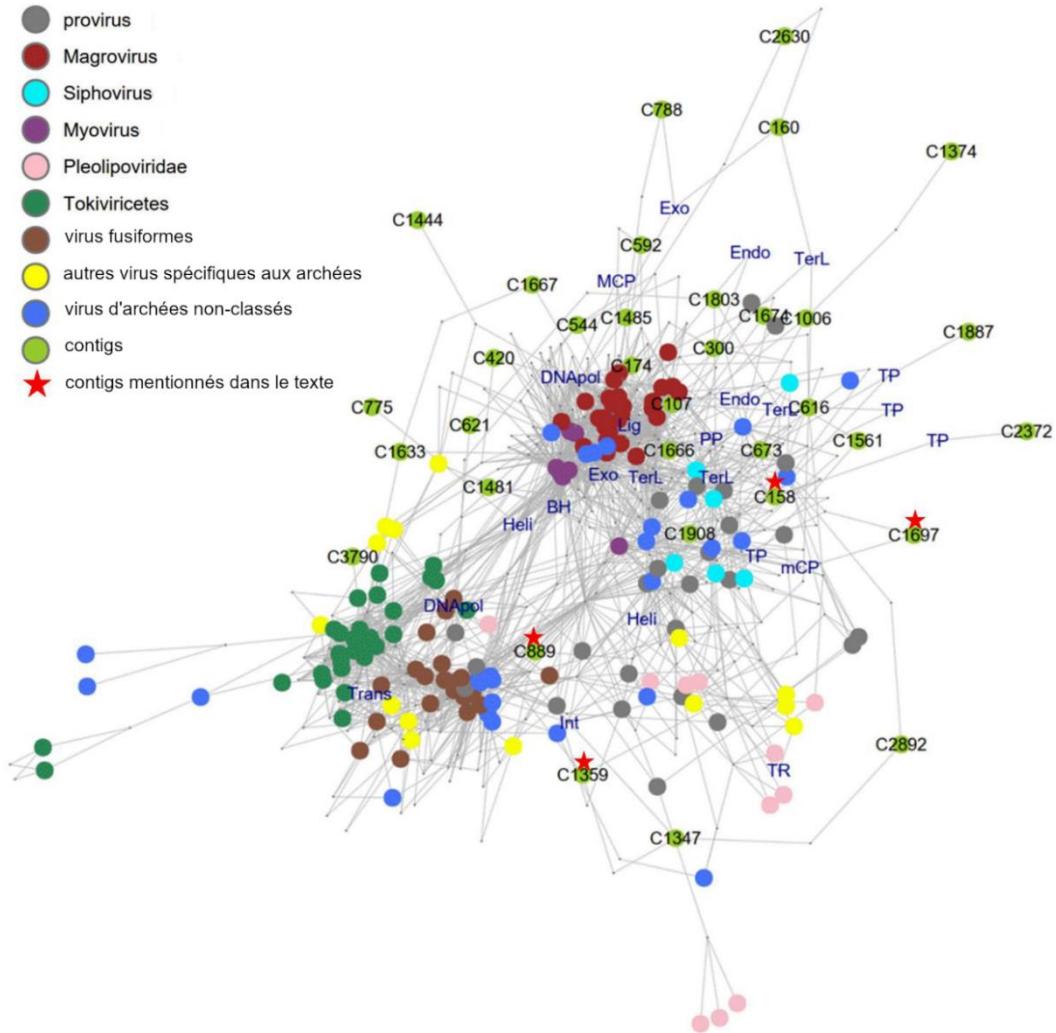
Les contigs viraux sont connectés *via* les groupes de protéines orthologues qu'ils partagent. Les génomes et les contigs sont représentés par des ronds colorés, et les groupes de protéines par les intersections des arêtes. Certaines annotations fonctionnelles de protéines sont indiquées. TerL: grande sous-unité de la terminase, MCP : protéine majeure de capside, mCP : protéine mineure de capside, Exo : exonucléase, TP : protéine de queue, PP : protéine portal, Lig : ligase, DNApol : ADN polymérase, BH : protéine de plaque basale, Heli : hélicase, Trans : transposase, TR : régulateur transcriptionnel.

Le premier, C158 (Figure 23 A), illustre bien le potentiel notre approche pour identifier des virus infectant un hôte actif. En effet, sur la base de son affiliation taxonomique et de sa signature génomique, son hôte prédit est *Methanobacterium*, qui était l'archée la plus abondante à la fois au cours de l'incubation, et dans la fraction lourde d'ADN cellulaire. De façon cohérente, C158 était le plus abondant parmi les 39 contigs viraux sélectionnés : son abondance normalisée atteignait 468 et 229 aux jours 13 et 17, respectivement, en RPKM (*Reads per Kilobase per Million Reads*, nombre de lectures par kilobase de séquence et par million de lectures). C158 est le génome complet d'un virus de la classe Caudoviricetes, dont la morphologie est probablement de type siphovirus, c'est-à-dire qu'il possède une longue queue non contractile. La présence d'une intégrase dans le génome de C158 suggère un mode de vie tempéré. Considérant la faible similarité de C158 avec des virus déjà connu, nous avons proposé la création d'une nouvelle famille virale dont il serait le premier représentant, *Speroviridae*.

(A)

(B)
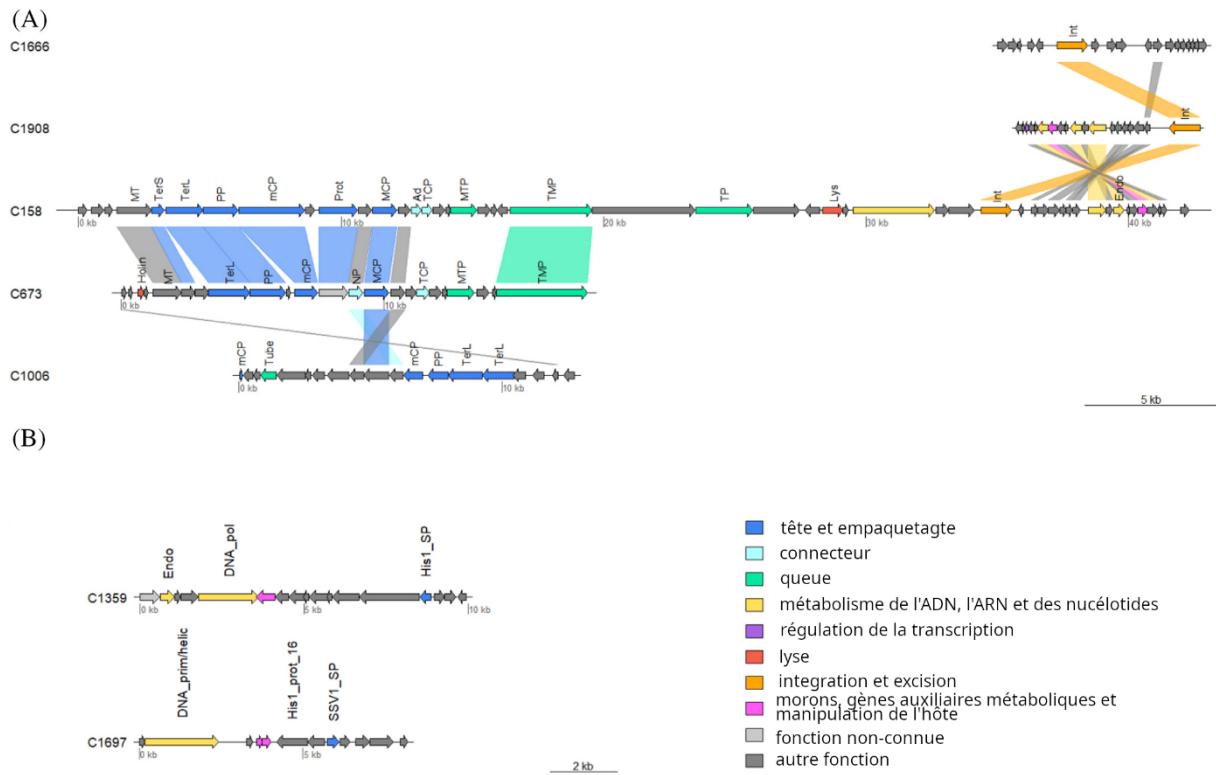
Figure 23. Carte des gènes pour une sélection de contigs viraux d'intérêt.

(A) Contigs d'un cluster viral (CV10) affilié à la classe Caudoviricetes, incluant C158 (nouvelle famille proposée : *Speroviridae*). (B) Contigs viraux de morphologie prédite comme fusiforme. Abbréviations pour les principales protéines virales : Ad, protéine adaptateur ; BP, protéine de plaque basale ; BH, protéine basale (*hub*) ; BS : protéine basale de spicules (*spike*) ; BW : protéine basale (*wedge*) ; DNA pol : ADN polymérase ; DNA prim : ADN primase ; DNA prim/helic : ADN primase/helicase ; Endo : endonucléase ; His1 prot 16 : similaire à la protéine 16 du virus 1 d'*Haloarcula hispanica* (His1) ; His1 SP : protéine majeure de capside de His1 ; holin : holine ; Int : intégrase ; Lys : endolysine ; mCP : protéine mineure de capside ; MCP : protéine majeure de capside ; MT : méthyltransférase ; MTP : protéine majeure de queue ; NP : protéines d'appendice du collier ; NT : ARNt nucléotidyltransférase ; PP : protéine portale ; Prot : protéase de maturation de la capside ; SSV1 SP : similaire à la protéine majeure de capside de SSV1 (*Sulfolobus* spindle-shaped virus 1) ; TCP : protéine de complétion de la queue ; TerL : grande sous-unité de la terminase ; TerS : petite sous-unité de la terminase ; TMP : protéine « mètre ruban » de la queue (*tail tape measure*) ; TP : protéine de queue ; Tube : protéine de tube.

Le second, C1697, représente un génome incomplet issu d'un virus minoritaire, et infectant une archée méthanogène n'ayant pas assimilé de formate marqué (Figure 23 B). Son hôte prédit, sur la base de son affiliation taxonomique et de sa composition en k-mer, est l'archée méthanogène *Methanosarcina*, dont l'abondance relative était très faible à l'ensemble des points de temps étudiés. Elle n'a en particulier pas été détectée dans les fractions intermédiaire et lourde de l'ADN cellulaire. A nouveau, les abondances de C1697 étaient cohérentes avec celle de l'hôte prédit, atteignant seulement 4 et 8 RPKM aux jours 13 et 17, respectivement. De façon intéressante, ce contig viral codait pour une protéine présentant une forte similarité avec la protéine majeure de capside de SSV1 (*Sulfolobus* Spindle-Shaped Virus 1), un virus fusiforme de la famille *Fuselloviridae* qui infecte des archées acidothermophiles (Pina *et al.*, 2011) : cela indiquait un morphotype fusiforme. C1697 codait par ailleurs pour une ADN polymérase de la famille B (à amorce protéique), ce qui suggérait un génome linéaire. Ce type de polymérase est présent chez deux autres virus fusiformes déjà connus, His1 (*Halspiviridae*) (Bath & Dyall-Smith Michael, 1998) et NSV1 (*Thaspiviridae*) (Kim *et al.*, 2019).

40

L'identification d'un virus fusiforme dans nos réacteurs était cohérente avec les observations en microscopie électronique, bien qu'il ne soit pas possible d'établir de lien précis entre un morphotype observé au microscope et le contig obtenu. Un autre génome incomplet de virus fusiforme infectant une archée méthanogène a été identifié pendant l'étude, C1359 (Figure 23). Son hôte précis n'a pas pu être prédit et il ne présentait pas de similarité avec C1697. Aussi, nos résultats suggèrent que les virus fusiformes pourraient être assez courants chez certaines archées méthanogènes mésophiles, bien qu'aucun d'entre eux n'ait été isolé à ce jour. Ceci est en particulier cohérent avec l'étude de (Calusinska *et al.*, 2016), qui rapporte l'observation de particules virales fusiformes au sein de méthaniseurs pleine échelle. Des virions fusiformes ont également été observés dans des sédiments profonds anoxiques du lac Pavin (France) et il a été suggéré qu'ils pourraient infecter des archées du genre *Methanosaeta* (Borrel et al., 2012). Enfin, des virus fusiformes infectant des archées mésophiles, bien que non méthanogènes, ont déjà été isolés par le passé (Kim *et al.*, 2019). Il s'agit des virus NSVs, infectant des thaumarchées du genre *Nitrosopumilus*, oxydant l'ammoniac.

Cette étude représente à ce jour la preuve la plus solide de l'existence de virus fusiformes infectant des archées méthanogènes, et renforce, pour les virus fusiformes, la notion de large distribution géographique et de grande diversité phylogénétique de leurs hôtes archéens. Elle a également montré l'intérêt du marquage isotopique pour étudier spécifiquement les virus infectant des hôtes impliqués dans un métabolisme particulier. Une approche similaire a été employée avec succès par d'autres auteurs pour étudier les virus de méthanotrophes du sol (Lee *et al.*, 2021), et ceux des archées oxydant l'ammoniac (Lee *et al.*, 2022).

Suite à notre première utilisation du SIP combiné à l'analyse de métaviromes, nous sommes actuellement en train de préparer une publication basée sur une approche très similaire, où nous nous intéressons à la fois aux virus d'archées et de bactéries ayant assimilé le carbone du formate, puisque nous avons constaté que des bactéries enrichies en ^{13}C étaient également abondantes au cours de ce type d'incubation. Plusieurs dizaines de contigs viraux de bactéries et archées sont actuellement en cours d'analyse. Nous nous concentrerons tout particulièrement sur la présence possible de gènes métaboliques auxiliaires.

Une direction assez naturelle à l'issue de ces résultats serait d'isoler les virus d'archées méthanogènes identifiés. En effet, cela pourrait permettre d'isoler un premier virus fusiforme d'archée méthanogène, et d'étudier son cycle infectieux, sa biologie. Des premières tentatives en ce sens ont été effectuées en collaboration avec Diana P. Baquero et Mart Krupovic (Institut Pasteur), mais elles ont malheureusement été infructueuses. Il me semblerait important de persévérer dans cette voie, car l'interprétation biologique et écologique sur la seule base de séquences de contigs viraux peut être fortement limitée, et sujette à caution.

C. Bilan et perspectives de recherche

Arrivée au Cemagref à une période où les méthodes méta-omiques perçaient tout juste dans le domaine des biotechnologies environnementales, j'ai été heureuse de participer à l'élan de développement, appropriation et mise en œuvre, de ces méthodes à haut-débit. En biologie environnementale d'une manière générale, ces dernières ont permis d'accéder à une vision bien plus approfondie de la diversité et des fonctions associées aux communautés microbiennes. Sur le plan fondamental, cette exploration a permis la découverte de nombreux groupes microbiens, notamment

chez les archées, et a abouti à un retour en force du débat sur l'origine des eucaryotes et la place des archées dans l'histoire évolutive, suite à la découverte du groupe Asgard : les archées sont-elles sœurs ou mères des eucaryotes ? Aujourd'hui, c'est cette seconde hypothèse qui semble convaincre le plus de scientifiques. Ce débat a percolé auprès du grand public, avec par exemple la publication d'un article le 3 janvier 2023 dans Le Monde : « Un chaînon manquant éclaire l'origine des cellules nucléées », Eva-Desvigne-Hansch. Ces dernières décennies d'exploration des communautés microbiennes environnementales ont également indiscutablement mis en lumière leur rôle majeur dans les cycles biogéochimiques de la planète, aspect tout aussi primordial et peut-être moins médiatisé.

Cependant, des verrous demeurent encore pour aller d'une écologie principalement descriptive vers une écologie plus prédictive, nécessitant non-seulement des connaissances mais également une compréhension approfondie des phénomènes (Widder *et al.*, 2016, Prosser, 2020). C'est typiquement le cas dans le domaine des biotechnologies environnementales, où les études méta-omiques s'accumulent sans pour l'instant déboucher sur des outils opérationnels concrets de *management* microbien. Des projets de recherche ont actuellement lieu dans différents laboratoires pour lever ce verrou. A PROSE, Olivier Chapleur et Laurent Mazéas mènent des études en ce sens. Des retombées concrètes pourraient commencer à émerger dans quelques années.

Ainsi, les approches méta-omiques ont beaucoup apporté, mais leurs limites sont aujourd'hui aussi bien identifiées : les fonctions prédites sont généralement seulement hypothétiques, et de nombreux gènes n'ont pas fonction connue. La qualité des annotations dépend fortement du contenu des bases de données. Enfin, il s'agit de données complexes, dont l'interprétation connaît inévitablement des limites. Mettre en œuvre des approches complémentaires est indispensable, tel qu'employer d'autres outils d'écologie moléculaire, ou encore isoler et caractériser des souches pures, ce qui peut passer par la culturomique.

A l'avenir, je souhaite que l'écologie virale et les approches méta-omiques continuent à occuper une place importante dans mes activités de recherches. Je perçois ces méthodes haut-débit comme un outil désormais incontournable. Je pense à court et moyen terme travailler dans la continuité de ce que j'ai développé ces quelques dernières années, en m'attachant à mieux cerner l'importance et les effets des virus microbiens au sein des bioprocédés anaérobies. Je mène en parallèle une réflexion sur la possibilité de m'orienter, à plus long terme, vers des bioprocédés dont les communautés microbiennes ont une diversité moins élevée, et sont moins caractérisées, que celles de la méthanisation. Il me semble en effet que le potentiel d'amélioration de tels procédés moins matures, sur la base de connaissance générées en écologie microbienne, y est peut-être plus élevé. De plus, un modèle avec une diversité microbienne plus limitée pourrait faciliter la démarche de compréhension des interactions écologiques et de la structuration des communautés.

C.I Comprendre la structuration et les effets des populations de virus au sein des digesteurs anaérobies, en étudiant l'effet de facteurs abiotiques

L'écologie virale des méthaniseurs émerge actuellement comme un sujet de fort intérêt. Plusieurs études ont été dédiées à l'exploration de la diversité des métaviromes dans les installations de méthanisations (Park *et al.*, 2007, Wu & Liu, 2009, Chien *et al.*, 2013, Calusinska *et al.*, 2016, Willenbücher *et al.*, 2022), mais encore relativement peu cherchent à comprendre les dynamiques (Zhang *et al.*, 2017) et à identifier les facteurs susceptibles de les affecter, ou encore à déterminer l'impact des virus sur le fonctionnement des procédés de méthanisation (Rossi *et al.*, 2022). Or il est établi pour d'autres écosystèmes que les virus peuvent avoir des effets divers et nombreux, comme cela est illustré ici pour le cas des bactériophages, les virus de bactéries (Figure 24).
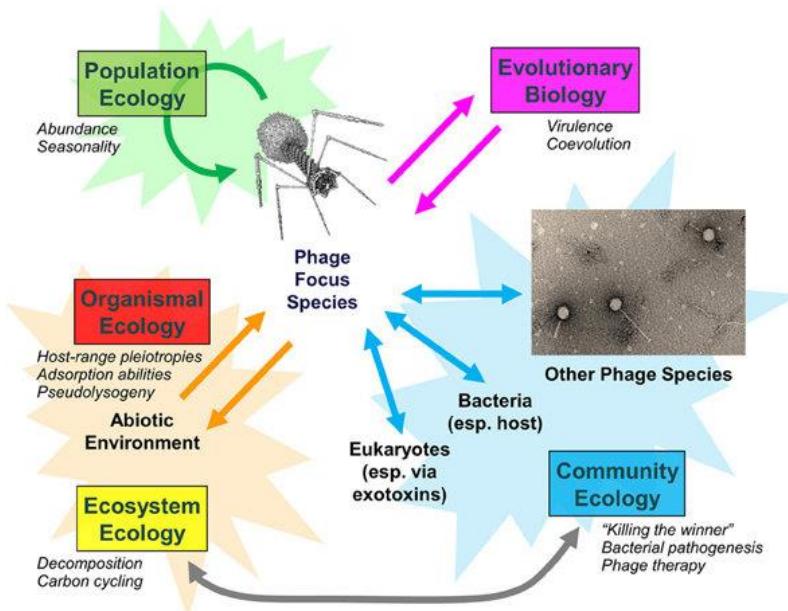
Mon premier projet, tourné vers l'exploration de la diversité des virus d'archées méthanogènes, a mis en évidence la diversité et la relative abondance des virus infectant les archées méthanogènes dans le cas de la digestion de formate, renforçant l'idée que les virus pourraient avoir des effets importants au sein des digesteurs. Nous avons en particulier identifié certains virus qui semblent tempérés (présence de gène d'intégrase) et d'autres virulents (absence de gènes caractéristiques de virus tempérés). Cela suggère que les virus d'archées méthanogènes, et plus largement les virus de bactéries, pourraient avoir des effets sur les digesteurs par différents mécanismes : mortalité des hôtes, mais aussi transfert ou régulation de gènes, effets sur la valeur sélective des hôtes (*fitness*), etc. Par la suite, je souhaiterais donc mettre en œuvre des projets visant à étudier les effets possibles des virus au sein des communautés microbiennes, et à en décrypter les mécanismes. Je pense me focaliser sur l'effet de certains facteurs abiotiques susceptibles d'affecter la composition des métaviromes, afin de mieux cerner le rôle des virus au sein des méthaniseurs.

Une bourse de thèse de l'école doctorale ABIES a été obtenue pour Marion Covès, qui a débuté ses travaux en octobre 2020 (co-encadrement : Laurent Mazéas). Son projet est encore en cours et s'intitule : « Effet d'inhibitions abiotiques de la digestion anaérobie sur les virus des écosystèmes de méthanisation ». L'objectif est d'évaluer, en microcosmes de digestion anaérobie mésophiles discontinus (*batch*), l'effet de différents composés chimiques connus pour inhiber la méthanisation, sur la composition des métaviromes. L'hypothèse sous-jacente est que parmi ces facteurs abiotiques, certains pourraient activer le cycle lytique de provirus. Trois inhibiteurs de la méthanisation déjà étudiés au sein de l'unité INRAE PROSE ont été retenus, afin de capitaliser sur l'expérience déjà acquise : le phénol, l'ammonium (sous la forme NH_4Cl) et le sel (NaCl). En outre, la mytomicine C a été employée comme agent bien connu d'induction des provirus.

Ce questionnement est très proche de celui de (Rossi *et al.*, 2022), dont les travaux ont été initiés de manière indépendante et publiés très récemment. En microcosmes de méthanisation, les auteurs ont étudié l'effet de différents facteurs de stress connus pour stimuler l'induction de prophages : oxygène,

température, pH, salinité, et surcharge organique. La mitomycine C y a également été employée comme contrôle. Dans le cas de (Rossi *et al.*, 2022), comme dans le nôtre, les éléments disponibles ne suggèrent pas de phénomène d'induction massive de provirus. Ainsi, la composition des métaviromes des méthaniseurs semble être dans ces deux études principalement pilotée par la dynamique des hôtes cellulaires. Les virus ne semblent pas causer d'aggravation significative des dysfonctionnements des méthaniseurs. Dans un autre type d'environnement, les cheminées hydrothermales en eaux peu profondes, il a été considéré que les virus tempérés pourraient être des acteurs majeurs du cycle du carbone (Rastelli *et al.*, 2017). Dans le cas de la méthanisation, les hôtes présentent une grande diversité et une certaine redondance fonctionnelle ; il semble donc plausible que les virus, étant abondants, aient un effet sur les flux élémentaires, sans toutefois causer de dysfonctionnement. Concernant la méthanisation, ce sont les toutes premières études sur le sujet et il n'existe pas de certitude à ce jour.

La force ionique m'est apparue comme un autre paramètre d'intérêt en lien avec les dynamiques hôtes-virus dans les méthaniseurs. Il est en effet rapporté dans la littérature que les particules virales sont plus abondantes dans les environnements salins, et que la force ionique a une influence sur l'adsorption des particules virales aux cellules hôtes (Kukkaro & Bamford, 2009) ou au sol (Kimura *et al.*, 2008). Au sein des méthaniseurs, la force ionique est susceptible de varier selon la nature de l'alimentation et le mode d'opération (voie humide ou voie sèche). Il pourrait donc être intéressant d'évaluer, en microcosmes de méthanisation, si la force ionique a un effet sur les dynamiques hôtes-virus et l'abondance des particules virales. Pour cela, je déposerai en 2023 une lettre d'intention à l'ANR (projet VIRALSONG), sous la forme d'un projet collaboratif. PROSE étudierait donc, en microcosmes, l'effet de la force ionique. Les partenaires du projet aborderaient des questions complémentaires : INRAE LBE et INRAE MICALIS examineraient l'effet du taux de siccité sur l'abondance et la dynamique des virus, tandis que l'Institut Pasteur (Archaeal virology) isolerait des virus d'archées méthanogènes. L'université de Padova, partenaire non-financé, renforcerait les compétences en analyses métagénomiques. Ce cadre collaboratif permettrait d'avancer sur plusieurs fronts complémentaires, pour développer de manière significative les connaissances en écologie virale des méthaniseurs.

C.II Travailler à l'échelle de virus isolés, pour mieux comprendre leur biologie

La démarche d'isolation et caractérisation de virus me paraît indispensable en complément des approches métagénomiques. Comme évoqué précédemment, la métagénomique employée seule ne permet pas d'obtenir des conclusions très solides concernant la biologie des virus présents. De plus, dans ces approches *shotgun*, de nombreuses séquences de génomes viraux demeurent incomplètes, lorsque l'on utilise la technologie Illumina seule.

D'après les résultats obtenus lors des projets récents (VIRAME, doctorat Marion Covès), de nombreux virus identifiés n'ont pas de lien de parenté significatif avec les familles virales connues, bien qu'appartenant à la classe *Caudoviricetes* (virus tête-queue). D'autres présentent des similarités avec des virus du microbiote intestinal, ce qui semble cohérent (environnement de fermentation anaérobie). Cependant, ces similarités restent limitées, il s'agit dans la plupart des cas de familles virales différentes. La diversité virale des méthaniseurs semble donc spécifique, et ceci confirme l'intérêt d'étudier leurs virus de manière plus approfondie.

Dans le cadre du projet VIRALSONG qui sera déposé à l'automne prochain, il est prévu comme évoqué ci-dessus que des virus d'archées méthanogènes soient isolés à l'Institut Pasteur. Les virus d'archées méthanogènes constituent une cible de choix car très peu sont encore caractérisés (Tableau 4), alors que les archées méthanogènes sont elles-mêmes très diverses, couvrant au moins 7 ordres

phylogénétiques distincts. De plus, les archées méthanogènes représentent un groupe fonctionnel clé des procédés de méthanisation, catalysant l'étape ultime de méthanogenèse. Enfin, le contrôle des archées méthanogènes représente un enjeu plus large en lien avec la problématique du réchauffement climatique, le méthane étant un puissant gaz à effet de serre : la caractérisation de virus d'archées méthanogènes pourrait contribuer à l'établissement de bases scientifiques pour la mise en place de (bio)contrôle des archées méthanogènes, par exemple dans le cas des tubes digestifs des ruminants.

INRAE PROSE n'est pas spécialisée dans la culture de souches pures. Fonctionner par collaboration est donc intéressant, au moins dans un premier temps. A terme, je n'exclus pas la possibilité d'isoler des virus à PROSE, dans la mesure où cette activité est très complémentaire des analyses de métaviromes et qu'elle est souvent menée de pair (exemples : INRAE MICALIS, INRAE SAYFOOD, et CAU Institute for General Microbiology à Kiel en Allemagne). Cela permet de tirer tout le bénéfice de cette complémentarité métagénomique/cultures de souches pures.

C.III   Explorer d'autres types de procédés, avec des approches d'écologie synthétique

J'ai débuté très récemment une réflexion pour m'orienter, à plus long terme, vers des approches d'écologie synthétique. Ce volet de l'écologie microbienne est en plein essor et apparaît comme une piste prometteuse pour lever des verrous bien identifiés dans le domaine, tels qu'atteindre une meilleure compréhension des mécanismes de structuration des communautés microbiennes, et ainsi développer une écologie microbienne plus prédictive (Großkopf & Soyer, 2014, Widder *et al.*, 2016). Etant consciente des limites des méthodes méta-omiques et de leur insuffisance, je suis désireuse de pouvoir mettre en œuvre des approches complémentaires et favorisant la compréhension du fonctionnement des communautés microbiennes.

Les systèmes de biocathodes, qui permettent la biosynthèse de composés d'intérêt par réduction de substrats comme le CO_2, pourraient constituer un modèle intéressant d'un point de vue fondamental et appliqué, mais mes réflexions sont encore à approfondir à ce stade. Tout d'abord, les procédés de biotechnologies environnementales reposant sur des systèmes bioélectrochimiques ne sont, pour la plupart, pas encore matures : les recherches en écologie microbienne sur ces systèmes ont peut-être un plus fort potentiel de retombées opérationnelles que pour des technologies plus matures. D'un point de vue scientifique, la diversité microbienne aux biocathodes, alimentées par des substrats simples, peut être supposée plus faible que dans les méthaniseurs, traitant des biodéchets complexes. Ceci pourrait être un avantage pour décrypter plus aisément les interactions et favoriser la compréhension de la communauté microbienne, au-delà de la description. Il pourrait en particulier être intéressant de s'y intéresser aux interactions hôtes-virus, en lien avec mes précédentes recherches.

Une approche dite *top-down* (simplifier les communautés microbiennes par dilution) semblerait adaptée dans un premier temps, permettant de traiter des questions pertinentes d'un point de vue opérationnel et écologique, et s'intégrant bien au cadre de l'unité INRAE PROSE. Nous y menons en premier lieu des recherches sur les communautés microbiennes et ne sommes pas spécialiste de la caractérisation et la manipulation de souches pures. Il serait en particulier pertinent de déterminer s'il existe un niveau optimal de diversité microbienne pour la stabilité et les performances du procédé. Une autre question à considérer avant cela est l'intérêt, d'un point de vue opérationnel, d'utiliser un consortium microbien ou un cocktail de souches, plutôt qu'une souche unique optimisée, étant donné la simplicité du substrat utilisé. On peut imaginer que l'utilisation d'un consortium performant pourrait permettre de travailler en conditions non-stériles, ce qui autoriserait des procédés moins coûteux. Du point de vue de la performance, l'étude de (Pande *et al.*, 2014) est particulièrement intéressante. Elle a montré que deux souches d'*Escherichia coli* modifiées génétiquement et dépendantes l'une de

l'autres par *cross-feeding* étaient généralement plus performantes qu'une souche non-déficiente sans partenaire. Ceci pourrait s'expliquer par une division de la charge métabolique : le coût de valeur sélective (*fitness*) lié à la surproduction de certains acides aminés était inférieur au bénéfce de ne pas avoir à en produire d'autres, lorsqu'ils étaient fournis par le partenaire. Cela suggère que même pour la bioconversion d'un substrat simple, un certain niveau de diversité pourrait présenter des avantages.

D'un point de vue pratique, il y a peu de croissance de biomasse aux cathodes, et ces systèmes peuvent difficilement être opérés en conditions stériles, ce qui peut poser des questions de faisabilité. Il est donc nécessaire d'évaluer de manière plus approfondie la pertinence d'un tel projet. Un stage co-encadré par Théodore Bouchez et moi-même est en cours à INRAE PROSE, pour la mise en place de biocathodes : cela pourrait être l'occasion d'évaluer la faisabilité et le potentiel de cette approche.

D. Conclusions

Ces 3 années passées en thèse à l'Institut Pasteur sur l'étude de virus d'archées acidothermophiles, puis ces 14 années en poste à INRAE (ex-Irstea, ex-Cemagref) dédiées à la caractérisation des communautés microbiennes de la méthanisation, m'ont permis d'aborder différentes questions de recherche liées aux aspects de diversité microbienne et virale, et de déterminisme de la structuration des communautés microbiennes. Mon intégration au sein d'une unité multidisciplinaire, sur des questions d'écologie microbienne et de développement d'approches méta-omiques représentait pour moi un défi, bien que très stimulant. Dans cette situation, je me suis d'abord orientée vers des thématiques relativement bien établies au sein de l'unité (méthanisation de la cellulose), tout en y apportant de nouveaux éléments, avec le développement de nouvelles collaborations, par exemple pour caractériser des substrats cellulosiques manufacturés. J'ai employé une démarche que je qualifierais de principalement exploratoire et descriptive, ce qui était plus aisé dans cette phase de développement et d'appropriation des méthodes métaprotéomiques et métagénomiques au sein de l'unité, à laquelle j'ai contribué de façon substantielle. Ces approches constituent aujourd'hui l'un des éléments forts de l'identité de notre unité. J'ai par la suite pu développer une thématique de recherche plus personnelle, en écologie virale de la méthanisation, et j'ai en parallèle établi des collaborations pour développer des recherches plus directement liées à des questions opérationnelles.

En cohérence avec mon parcours professionnel récent, je souhaiterais aujourd'hui approfondir les aspects d'écologie virale des procédés de biotechnologies environnementales anaérobies, afin notamment de pouvoir aller au-delà des aspects purement descriptifs et de contribuer à développer une meilleure compréhension des effets des virus au sein de ce type d'écosystèmes. Ces questions sont encore très peu abordées au sein de notre communauté scientifique, bien qu'une tendance à l'augmentation soit perceptible. L'écologie virale est déjà bien développée pour d'autres types d'écosystèmes, ce qui devrait faciliter et inspirer ce type de recherches dans le domaine des biotechnologies environnementales.

Toujours dans l'optique d'aller vers plus de compréhension des écosystèmes, je souhaiterais à moyen et long terme m'orienter vers des communautés microbiennes de procédés de moindre diversité, et également moins connues, en employant potentiellement des approches d'écologie synthétique. Le recours à des modèles réductionnistes est une tendance actuelle forte en écologie microbienne, considérant que des systèmes plus simples peuvent favoriser une meilleure compréhension et que travailler à différentes échelles de diversité apporte des complémentarités bénéfiques pour avancer vers une écologie plus prédictive.

Références bibliographiques

Adney WS, Rivard CJ, Shiang M & Himmel ME (1991) Anaerobic digestion of lignocellulosic biomass and wastes. *Applied Biochemistry and Biotechnology* **30**: 165-183.

Alfastsen L, Peng X & Bhoobalan-Chitty Y (2021) Genome editing in archaeal viruses and endogenous viral protein purification. *STAR Protocols* **2**: 100791.

Allen H & Abedon S (2013) That's disturbing! An exploration of the bacteriophage biology of change. *Frontiers in Microbiology* **4**.

Badalato N, Guillot A, Sabarly V*, et al.* (2017) Whole Proteome Analyses on Ruminiclostridium cellulolyticum Show a Modulation of the Cellulolysis Machinery in Response to Cellulosic Materials with Subtle Differences in Chemical and Structural Properties. *PLOS ONE* **12**: e0170524.

Baquero DP, Liu Y, Wang F, Egelman EH, Prangishvili D & Krupovic M (2020) Chapter Four - Structure and assembly of archaeal viruses. *Advances in Virus Research,* Vol. 108 (Kielian M, Mettenleiter TC & Roossinck MJ, eds.), p.^pp. 127-164. Academic Press.

Bath C & Dyall-Smith Michael L (1998) His1, an Archaeal Virus of theFuselloviridae Family That Infects Haloarcula hispanica. *Journal of Virology* **72**: 9392-9395.

Berghuis BA, Yu FB, Schulz F, Blainey PC, Woyke T & Quake SR (2019) Hydrogenotrophic methanogenesis in archaeal phylum Verstraetearchaeota reveals the shared ancestry of all methanogens. *Proceedings of the National Academy of Sciences* **116**: 5037-5044.

Bettstetter M, Peng X, Garrett RA & Prangishvili D (2003) AFV1, a novel virus infecting hyperthermophilic archaea of the genus acidianus. *Virology* **315**: 68-79.

Bize A, Midoux C, Mariadassou M, Schbath S, Forterre P & Da Cunha V (2021) Exploring short k-mer profiles in cells and mobile elements from Archaea highlights the major influence of both the ecological niche and evolutionary history. *BMC genomics* **22**: 1-22.

Bize A, Karlsson EA, Ekefjärd K, Quax TEF, Pina M, Prevost M-C, Forterre P, Tenaillon O, Bernander R & Prangishvili D (2009) A unique virus release mechanism in the Archaea. *Proceedings of the National Academy of Sciences* **106**: 11306-11311.

Bize A, Cardona L, Desmond-Le Quéméner E, Battimelli A, Badalato N, Bureau C, Madigou C, Chevret D, Guillot A & Monnet V (2015) Shotgun metaproteomic profiling of biomimetic anaerobic digestion processes treating sewage sludge. *Proteomics* **15**: 3532-3543.

Borrel G, Colombet J, Robin A, Lehours A-C, Prangishvili D & Sime-Ngando T (2012) Unexpected and novel putative viruses in the sediments of a deep-dark permanently anoxic freshwater habitat. *The ISME Journal* **6**: 2119-2127.

Brumfield Susan K, Ortmann Alice C, Ruigrok V, Suci P, Douglas T & Young Mark J (2009) Particle Assembly and Ultrastructural Features Associated with Replication of the Lytic Archaeal Virus Sulfolobus Turreted Icosahedral Virus. *Journal of Virology* **83**: 5964-5970.

Brussaard CPD, Wilhelm SW, Thingstad F*, et al.* (2008) Global-scale processes with a nanoscale drive: the role of marine viruses. *The ISME Journal* **2**: 575-578.

Calusinska M, Marynowska M, Goux X, Lentzen E & Delfosse P (2016) Analysis of dsDNA and RNA viromes in methanogenic digesters reveals novel viral genetic diversity. *Environmental Microbiology* **18**: 1162-1175.

Camargo FP, Sakamoto IK, Bize A, Duarte ICS, Silva EL & Varesche MBA (2021) Screening design of nutritional and physicochemical parameters on bio-hydrogen and volatile fatty acids production from Citrus Peel Waste in batch reactors. *International Journal of Hydrogen Energy* **46**: 7794-7809.

Carballa M, Regueiro L & Lema JM (2015) Microbial management of anaerobic digestion: exploiting the microbiome-functionality nexus. *Current Opinion in Biotechnology* **33**: 103-111.

Chapleur O, Bize A, Serain T, Mazéas L & Bouchez T (2014) Co-inoculating ruminal content neither provides active hydrolytic microbes nor improves methanization of 13C-cellulose in batch digesters. *FEMS Microbiology Ecology* **87**: 616-629.

Chien IC, Meschke JS, Gough HL & Ferguson JF (2013) Characterization of Persistent Virus-Like Particles in Two Acetate-Fed Methanogenic Reactors. *PLOS ONE* **8**: e81040.

Conrad R (2005) Quantification of methanogenic pathways using stable carbon isotopic signatures: a review and a proposal. *Organic Geochemistry* **36**: 739-752.

Delforno TP, Macedo TZ, Midoux C, Lacerda Jr GV, Rué O, Mariadassou M, Loux V, Varesche MBA, Bouchez T & Bize A (2019) Comparative metatranscriptomic analysis of anaerobic digesters treating anionic surfactant contaminated wastewater. *Science of the total environment* **649**: 482-494.

Desvaux M, Guedon E & Petitdemange H (2000) Cellulose Catabolism by Clostridium cellulolyticum Growing in Batch Culture on Defined Medium. *Applied and Environmental Microbiology* **66**: 2461-2470.

Desvaux M, Guedon E & Petitdemange H (2001) Carbon Flux Distribution and Kinetics of Cellulose Fermentation in Steady-State Continuous Cultures of Clostridium cellulolyticum on a Chemically Defined Medium. *Journal of Bacteriology* **183**: 119-130.

Drula E, Garron M-L, Dogan S, Lombard V, Henrissat B & Terrapon N (2022) The carbohydrate-active enzyme database: functions and literature. *Nucleic Acids Research* **50**: D571-D577.

Escudié F, Auer L, Bernard M, Mariadassou M, Cauquil L, Vidal K, Maman S, Hernandez-Raquet G, Combes S & Pascal G (2018) FROGS: Find, Rapidly, OTUs with Galaxy Solution. *Bioinformatics* **34**: 1287-1294.

Fuhrman JA (1999) Marine viruses and their biogeochemical and ecological effects. *Nature* **399**: 541-548.

Gal L, Pages S, Gaudin C, Belaich A, Reverbel-Leroy C, Tardif C & Belaich JP (1997) Characterization of the cellulolytic complex (cellulosome) produced by Clostridium cellulolyticum. *Applied and Environmental Microbiology* **63**: 903-909.

Galiez C, Siebert M, Enault F, Vincent J & Söding J (2017) WIsH: who is the host? Predicting prokaryotic hosts from metagenomic phage contigs. *Bioinformatics* **33**: 3113-3114.

Gehin A, Gelhaye E & Petitdemange H (1996) Adhesion of Clostridium cellulolyticum spores to filter paper. *Journal of Applied Bacteriology* **80**: 187-190.

Großkopf T & Soyer OS (2014) Synthetic microbial communities. *Current Opinion in Microbiology* **18**: 72-77.

Guérin-Rechdaoui S, Bize A, Levesque-Ninio C, Janvier A, Lacroix C, Le Brizoual F, Barbier J, Amsaleg CR, Azimi S & Rocher V (2022) Fate of SARS-CoV-2 coronavirus in wastewater treatment sludge during storage and thermophilic anaerobic digestion. *Environmental Research* **214**: 114057.

Guo J, Bolduc B, Zayed AA*, et al.* (2021) VirSorter2: a multi-classifier, expert-guided approach to detect diverse DNA and RNA viruses. *Microbiome* **9**: 37.

Hanreich A, Heyer R, Benndorf D, Rapp E, Pioch M, Reichl U & Klocke M (2012) Metaproteome analysis to determine the metabolically active part of a thermophilic microbial community producing biogas from agricultural biomass. *Canadian Journal of Microbiology* **58**: 917-922.

Hantula J, Kurki A, Vuoriranta P & Bamford DH (1991) Ecology of bacteriophages infecting activated sludge bacteria. *Applied and Environmental Microbiology* **57**: 2147-2151.

Häring M, Peng X, Brügger K, Rachel R, Stetter KO, Garrett RA & Prangishvili D (2004) Morphology and genome organization of the virus PSV of the hyperthermophilic archaeal genera Pyrobaculum and Thermoproteus: a novel virus family, the Globuloviridae. *Virology* **323**: 233-242.

Hershey AD, Dixon J & Chase M (1953) Nucleic acid economy in bacteria infected with bacteriophage T2. I. Purine and pyrimidine composition. *The Journal of general physiology* **36**: 777-789.

Iranzo J, Krupovic M & Koonin Eugene V (2016) The Double-Stranded DNA Virosphere as a Modular Hierarchical Network of Gene Sharing. *mBio* **7**: e00978-00916.

Iranzo J, Koonin Eugene V, Prangishvili D & Krupovic M (2016) Bipartite Network Analysis of the Archaeal Virosphere: Evolutionary Connections between Viruses and Capsidless Mobile Elements. *Journal of Virology* **90**: 11043-11055.

Iranzo J, Koonin EV, Prangishvili D & Krupovic M (2016) Bipartite network analysis of the archaeal virosphere: evolutionary connections between viruses and capsidless mobile elements. *Journal of virology* **90**: 11043-11055.

Jaubert C, Danioux C, Oberto J, Cortez D, Bize A, Krupovic M, She Q, Forterre P, Prangishvili D & Sezonov G (2013) Genomics and genetics of Sulfolobus islandicus LAL14/1, a model hyperthermophilic archaeon. *Open biology* **3**: 130010.

Jensen PD, Hardin MT & Clarke WP (2008) Measurement and quantification of sessile and planktonic microbial populations during the anaerobic digestion of cellulose. *Water Science and Technology* **57**: 465-469.

Jensen PD, Hardin MT & Clarke WP (2009) Effect of biomass concentration and inoculum source on the rate of anaerobic cellulose solubilization. *Bioresource Technology* **100**: 5219-5225.

Kieft K, Zhou Z & Anantharaman K (2020) VIBRANT: automated recovery, annotation and curation of microbial viruses, and evaluation of viral community function from genomic sequences. *Microbiome* **8**: 90.

Kim J-G, Kim S-J, Cvirkaite-Krupovic V, Yu W-J, Gwak J-H, López-Pérez M, Rodriguez-Valera F, Krupovic M, Cho J-C & Rhee S-K (2019) Spindle-shaped viruses infect marine ammonia-oxidizing thaumarchaea. *Proceedings of the National Academy of Sciences* **116**: 15645-15650.

Kimura M, Jia Z-J, Nakayama N & Asakawa S (2008) Ecology of viruses in soils: Past, present and future perspectives. *Soil Science and Plant Nutrition* **54**: 1-32.

Koonin Eugene V, Dolja Valerian V, Krupovic M, Varsani A, Wolf Yuri I, Yutin N, Zerbini FM & Kuhn Jens H (2020) Global Organization and Proposed Megataxonomy of the Virus World. *Microbiology and Molecular Biology Reviews* **84**: e00061-00019.

Koonin EV, Dolja VV & Krupovic M (2015) Origins and evolution of viruses of eukaryotes: The ultimate modularity. *Virology* **479-480**: 2-25.

Krawczyk PS, Lipinski L & Dziembowski A (2018) PlasFlow: predicting plasmid sequences in metagenomic data using genome signatures. *Nucleic Acids Research* **46**: e35-e35.

Krupovic M, Dolja VV & Koonin EV (2019) Origin of viruses: primordial replicators recruiting capsids from hosts. *Nature Reviews Microbiology* **17**: 449-458.

Krupovic M, Quemin ERJ, Bamford DH, Forterre P & Prangishvili D (2014) Unification of the globally distributed spindle-shaped viruses of the Archaea. *Journal of Virology* **88**: 2354-2358.

Krupovic M, Cvirkaite-Krupovic V, Iranzo J, Prangishvili D & Koonin EV (2018) Viruses of archaea: Structural, functional, environmental and evolutionary genomics. *Virus Research* **244**: 181-193.

Kukkaro P & Bamford DH (2009) Virus–host interactions in environments with a wide range of ionic strengths. *Environmental Microbiology Reports* **1**: 71-77.

Kunin V, He S, Warnecke F*, et al.* (2008) A bacterial metapopulation adapts locally to phage predation despite global dispersal. *Genome Research* **18**: 293-297.

Lai TE, Nopharatana A, Pullammanappallil PC & Clarke WP (2001) Cellulolytic activity in leachate during leach-bed anaerobic digestion of municipal solid waste. *Bioresource Technology* **80**: 205-210.

Lee S-H, Otawa K, Onuki M, Satoh H & Mino T (2007) Population dynamics of phage-host system of Microlunatus phosphovorus indigenous in activated sludge. *Journal of microbiology and biotechnology* **17**: 1704-1707.

Lee S, Sieradzki ET, Nicol GW & Hazard C (2022) Propagation of viral genomes by replicating ammonia-oxidising archaea during soil nitrification. *The ISME Journal*.

Lee S, Sieradzki ET, Nicolas AM, Walker RL, Firestone MK, Hazard C & Nicol GW (2021) Methane-derived carbon flows into host–virus networks at different trophic levels in soil. *Proceedings of the National Academy of Sciences* **118**: e2105124118.

Lemaire ON, Jespersen M & Wagner T (2020) CO2-Fixation Strategies in Energy Extremophiles: What Can We Learn From Acetogens? *Frontiers in Microbiology* **11**.

Letunic I & Bork P (2007) Interactive Tree Of Life (iTOL): an online tool for phylogenetic tree display and annotation. *Bioinformatics* **23**: 127-128.

Li T, Mazéas L, Sghir A, Leblon G & Bouchez T (2009) Insights into networks of functional microbes catalysing methanization of cellulose under mesophilic conditions. *Environmental Microbiology* **11**: 889-904.

Liu J, Cvirkaite-Krupovic V, Baquero DP, Yang Y, Zhang Q, Shen Y & Krupovic M (2021) Virus-induced cell gigantism and asymmetric cell division in archaea. *Proceedings of the National Academy of Sciences* **118**: e2022578118.

Lü F, Bize A, Guillot A, Monnet V, Madigou C, Chapleur O, Mazéas L, He P & Bouchez T (2014) Metaproteomics of cellulose methanisation under thermophilic conditions reveals a surprisingly high proteolytic activity. *The ISME Journal* **8**: 88-102.

Lynd Lee R, Weimer Paul J, van Zyl Willem H & Pretorius Isak S (2002) Microbial Cellulose Utilization: Fundamentals and Biotechnology. *Microbiology and Molecular Biology Reviews* **66**: 506-577.

Mahé F, Rognes T, Quince C, de Vargas C & Dunthorn M (2014) Swarm: robust and fast clustering method for amplicon-based studies. *PeerJ* **2**: e593.

Martin A, Yeats S, Janekovic D, Reiter W-D, Aicher W & Zillig W (1984) SAV 1, a temperate u.v.-inducible DNA virus-like particle from the archaebacterium Sulfolobus acidocaldarius isolate B12. *The EMBO Journal* **3**: 2165-2168.

Mayo-Muñoz D, He F, Jørgensen JB, Madsen PK, Bhoobalan-Chitty Y & Peng X (2018) Anti-CRISPR-Based and CRISPR-Based Genome Editing of Sulfolobus islandicus Rod-Shaped Virus 2. Vol. 10 p.^pp.

Medvedeva S, Sun J, Yutin N, Koonin EV, Nunoura T, Rinke C & Krupovic M (2022) Three families of Asgard archaeal viruses identified in metagenome-assembled genomes. *Nature Microbiology* **7**: 962-973.

Meile L, Jenal U, Studer D, Jordan M & Leisinger T (1989) Characterization of ψM1, a virulent phage of Methanobacterium thermoautotrophicum Marburg. *Archives of Microbiology* **152**: 105-110.

Moletta R (2015) *La méthanisation (3e éd.)*. Lavoisier.

Nayfach S, Camargo AP, Schulz F, Eloe-Fadrosh E, Roux S & Kyrpides NC (2021) CheckV assesses the quality and completeness of metagenome-assembled viral genomes. *Nature Biotechnology* **39**: 578-585.

Nelson MC, Morrison M & Yu Z (2011) A meta-analysis of the microbial diversity observed in anaerobic digesters. *Bioresource Technology* **102**: 3730-3739.

Noike T, Endo G, Chang J-E, Yaguchi J-I & Matsumoto J-I (1985) Characteristics of carbohydrate degradation and the rate-limiting step in anaerobic digestion. *Biotechnology and Bioengineering* **27**: 1482-1489.

Nölling J, Groffen A & de Vos WM (1993) ɸ F1 and ɸF3, two novel virulent, archaeal phages infecting different thermophilic strains of the genus Methanobacterium. *Microbiology* **139**: 2511-2516.

O'Sullivan CA, Burrell PC, Clarke WP & Blackall LL (2005) Structure of a cellulose degrading bacterial community during anaerobic digestion. *Biotechnology and Bioengineering* **92**: 871-878.

O'Sullivan CA, Burrell PC, Clarke WP & Blackall LL (2006) Comparison of cellulose solubilisation rates in rumen and landfill leachate inoculated reactors. *Bioresource Technology* **97**: 2356-2363.

Otawa K, Lee SH, Yamazoe A, Onuki M, Satoh H & Mino T (2007) Abundance, Diversity, and Dynamics of Viruses on Microorganisms in Activated Sludge Processes. *Microbial Ecology* **53**: 143-152.

Pande S, Merker H, Bohl K, Reichelt M, Schuster S, de Figueiredo LF, Kaleta C & Kost C (2014) Fitness and stability of obligate cross-feeding interactions that emerge upon gene loss in bacteria. *The ISME Journal* **8**: 953-962.

Park MO, Ikenaga H & Watanabe K (2007) Phage Diversity in a Methanogenic Digester. *Microbial Ecology* **53**: 98-103.

Parsiegla G, Juy M, Reverbel-Leroy C, Tardif C, Belaïch J-P, Driguez H & Haser R (1998) The crystal structure of the processive endocellulase CelF of Clostridium cellulolyticum in complex with a thiooligosaccharide inhibitor at 2.0 Å resolution. *The EMBO Journal* **17**: 5551-5562.

Parsley Larissa C, Consuegra Erin J, Thomas Stephen J*, et al.* (2010) Census of the Viral Metagenome within an Activated Sludge Microbial Assemblage. *Applied and Environmental Microbiology* **76**: 2673-2677.

Paul S, Bag SK, Das S, Harvill ET & Dutta C (2008) Molecular signature of hypersaline adaptation: insights from genome and proteome composition of halophilic prokaryotes. *Genome Biology* **9**: R70.

Pfister P, Wasserfallen A, Stettler R & Leisinger T (1998) Molecular analysis of Methanobacterium phage ΨM2. *Molecular Microbiology* **30**: 233-244.

Pina M, Bize A, Forterre P & Prangishvili D (2011) The archeoviruses. *FEMS Microbiology Reviews* **35**: 1035-1054.

Prangishvili D, Forterre P & Garrett RA (2006) Viruses of the Archaea: a unifying view. *Nature Reviews Microbiology* **4**: 837-848.

Prangishvili D, Arnold HP, Götz D, Ziese U, Holz I, Kristjansson JK & Zillig W (1999) A Novel Virus Family, the Rudiviridae: Structure, Virus-Host Interactions and Genome Variability of the Sulfolobus Viruses SIRV1 and SIRV2. *Genetics* **152**: 1387-1396.

Prosser JI (2020) Putting science back into microbial ecology: a question of approach. *Philosophical Transactions of the Royal Society B: Biological Sciences* **375**: 20190240.

Qu X, Vavilin VA, Mazéas L, Lemunier M, Duquennoi C, He PJ & Bouchez T (2009) Anaerobic biodegradation of cellulosic material: Batch experiments and modelling based on isotopic data and focusing on aceticlastic and non-aceticlastic methanogenesis. *Waste Management* **29**: 1828-1837.

Quax TEF, Krupovič M, Lucas S, Forterre P & Prangishvili D (2010) The Sulfolobus rod-shaped virus 2 encodes a prominent structural component of the unique virion release system in Archaea. *Virology* **404**: 1-4.

Quax TEF, Lucas S, Reimann J, Pehau-Arnaudet G, Prevost M-C, Forterre P, Albers S-V & Prangishvili D (2011) Simple and elegant design of a virion egress structure in Archaea. *Proceedings of the National Academy of Sciences* **108**: 3354-3359.

Rastelli E, Corinaldesi C, Dell'Anno A, Tangherlini M, Martorelli E, Ingrassia M, Chiocci FL, Lo Martire M & Danovaro R (2017) High potential for temperate viruses to drive carbon cycling in chemoautotrophy-dominated shallow-water hydrothermal vents. *Environmental Microbiology* **19**: 4432-4446.

Rinke C, Rubino F, Messer LF*, et al.* (2019) A phylogenomic and ecological analysis of the globally abundant Marine Group II archaea (Ca. Poseidoniales ord. nov.). *The ISME Journal* **13**: 663-675.

Rossi A, Morlino MS, Gaspari M, Basile A, Kougias P, Treu L & Campanaro S (2022) Analysis of the anaerobic digestion metagenome under environmental stresses stimulating prophage induction. *Microbiome* **10**: 125.

Sauterey B, Charnay B, Affholder A, Mazevet S & Ferrière R (2020) Co-evolution of primitive methane-cycling ecosystems and early Earth's atmosphere and climate. *Nature Communications* **11**: 2705.

Schleper C, Kubo K & Zillig W (1992) The particle SSV1 from the extremely thermophilic archaeon Sulfolobus is a virus: demonstration of infectivity and of transfection with viral DNA. *Proceedings of the National Academy of Sciences* **89**: 7645-7649.

Schlüter A, Bekel T, Diaz NN*, et al.* (2008) The metagenome of a biogas-producing microbial community of a production-scale biogas plant fermenter analysed by the 454-pyrosequencing technology. *Journal of Biotechnology* **136**: 77-90.

Shapiro OH, Kushmaro A & Brenner A (2010) Bacteriophage predation regulates microbial abundance and diversity in a full-scale bioreactor treating industrial wastewater. *The ISME Journal* **4**: 327-336.

Shoham Y, Lamed R & Bayer EA (1999) The cellulosome concept as an efficient microbial strategy for the degradation of insoluble polysaccharides. *Trends in Microbiology* **7**: 275-281.

Song H, Clarke WP & Blackall LL (2005) Concurrent microscopic observations and activity measurements of cellulose hydrolyzing and methanogenic populations during the batch anaerobic digestion of crystalline cellulose. *Biotechnology and Bioengineering* **91**: 369-378.

Stent G (1966) Phage and the origins of molecular biology.

Suttle CA (2005) Viruses in the sea. *Nature* **437**: 356-361.

Thiroux S, Dupont S, Nesbø CL, Bienvenu N, Krupovic M, L'Haridon S, Marie D, Forterre P, Godfroy A & Geslin C (2021) The first head-tailed virus, MFTV1, infecting hyperthermophilic methanogenic deep-sea archaea. *Environmental Microbiology* **23**: 3614-3626.

Weidenbach K, Wolf S, Kupczok A, Kern T, Fischer MA, Reetz J, Urbańska N, Künzel S, Schmitz RA & Rother M (2021) Characterization of Blf4, an Archaeal Lytic Virus Targeting a Member of the Methanomicrobiales. Vol. 13 p.^pp.

Weidenbach K, Nickel L, Neve H*, et al.* (2017) Methanosarcina Spherical Virus, a Novel Archaeal Lytic Virus Targeting Methanosarcina Strains. *Journal of Virology* **91**: e00955-00917.

Weinbauer MG (2004) Ecology of prokaryotic viruses. *FEMS Microbiology Reviews* **28**: 127-181.

Widder S, Allen RJ, Pfeiffer T*, et al.* (2016) Challenges in microbial ecology: building predictive understanding of community function and dynamics. *The ISME Journal* **10**: 2557-2568.

Willenbücher K, Wibberg D, Huang L*, et al.* (2022) Phage Genome Diversity in a Biogas-Producing Microbiome Analyzed by Illumina and Nanopore GridION Sequencing. Vol. 10 p.^pp.

Withey S, Cartmell E, Avery LM & Stephenson T (2005) Bacteriophages—potential for application in wastewater treatment processes. *Science of The Total Environment* **339**: 1-18.

Woese CR, Magrum LJ & Fox GE (1978) Archaebacteria. *Journal of Molecular Evolution* **11**: 245-252.

Woese CR, Kandler O & Wheelis ML (1990) Towards a natural system of organisms: proposal for the domains Archaea, Bacteria, and Eucarya. *Proceedings of the National Academy of Sciences* **87**: 4576-4579.

Wolf S, Fischer MA, Kupczok A, Reetz J, Kern T, Schmitz RA & Rother M (2019) Characterization of the lytic archaeal virus Drs3 infecting Methanobacterium formicicum. *Archives of Virology* **164**: 667-674.

Wood AG, Whitman WB & Konisky J (1989) Isolation and characterization of an archaebacterial viruslike particle from Methanococcus voltae A3. *Journal of Bacteriology* **171**: 93-98.

Wu Q & Liu W-T (2009) Determination of virus abundance, diversity and distribution in a municipal wastewater treatment plant. *Water Research* **43**: 1101-1109.

Zhang J, Gao Q, Zhang Q*, et al.* (2017) Bacteriophage–prokaryote dynamics and interaction within anaerobic digestion processes across time and space. *Microbiome* **5**: 57.

Zillig W, Kletzin A, Schleper C, Holz I, Janekovic D, Hain J, Lanzendörfer M & Kristjansson JK (1993) Screening for Sulfolobales, their Plasmids and their Viruses in Icelandic Solfataras. *Systematic and Applied Microbiology* **16**: 609-628.

Zillig W, Arnold HP, Holz I, Prangishvili D, Schweier A, Stedman K, She Q, Phan H, Garrett R & Kristjansson JK (1998) Genetic elements in the extremely thermophilic archaeon Sulfolobus. *Extremophiles* **2**: 131-140.

Expérience professionnelle

Depuis 2008, j'occupe une poste de chercheuse en écologie microbienne à Antony (92), avec un statut d'IPEF en position normale d'activité. J'ai rejoint ce poste en 2008. Cependant les contours de l'unité, le nom de l'institut et son périmètre ont changé à plusieurs reprises. Mes principaux sujets d'intérêt sont les suivants :

- Ecologie microbienne des procédés de biotechnologie environnementale
- Diversité et dynamique des virus dans les écosystèmes de digestion anaérobie (depuis 2017)
- Hydrolyse et digestion anaérobie des déchets lignocellulosiques (depuis 2010)
- Mise en œuvre de méthodes méta-omiques pour comprendre la structuration et le fonctionnement des écosystèmes microbiens
- Développement d'un système d'information pour les données méta-omiques issues de procédés de biotechnologie environnementale (DeepOmics)

Depuis 2020 **Université Paris-Saclay, INRAE PROSE** (1461, Procédés biotechnologiques au service de l'environnement), Centre Île-de-France - Jouy-en-Josas-Antony, Université Paris-Saclay, chercheuse en écologie microbienne.

2019-2020 **Irstea PROSE**, chercheuse en écologie microbienne

2012-2019 **Irstea HBAN**, chercheuse en écologie microbienne

2008-2012 **Cemagref HBAN**, chercheuse en écologie microbienne

2005-2008 **Doctorat en microbiologie** sur les virus d'archées acidothermophiles, sous la direction de David Prangishvili. Institut Pasteur, BMGE (Biologie Moléculaire du Gène chez les Extrêmophiles), Paris. Ecole doctorale Ville et Environnement, Ecole Nationale des Ponts et Chaussées, Université Paris-Est.

Formation

2004-2005 **Master 2 Approches interdisciplinaires du vivant (AIV)**, Ecole Normale Supérieure Paris et Université Paris 7

2002-2004 **Ecole Nationale des Ponts et Chaussées**, Champs-sur-Marne - Corps des Ponts et Chaussées, Voie Ville, Environnement, Transport

1999-2002 **Ecole Polytechnique, Palaiseau** - Majeures Biologie et Economie

Enseignement

Depuis 2017 **Ecole des Ponts ParisTech**, enseignement pour les élèves de 1ère année Cours d'Introduction aux Sciences du Vivant, ~ 6 h/an Encadrement d'un projet d'initiation à la recherche en écologie microbienne, ~ 10 h/an, 6 élèves

Depuis 2012 **AgroParisTech**, UE Biotechnologie Microbienne pour l'Environnement, approches méta-omiques pour les procédés de biotechnologie environnementale, ~ 3 h/an

Depuis 2018 **Ecole Polytechnique**, Environmental Engineering and Sustainability Management Master, Approches méta-omiques pour les procédés de biotechnologie environnementale, ~ 2 h/an

Depuis 2022 **Sorbonne université**, Master 1 de Biologie, UE Diversité et fonctions des microorganismes : du microbiote humain et animal aux écosystèmes naturels, écologie microbienne de la méthanisation, ~ 1h30/an

Encadrement

Stagiaires de Master ou d'école d'ingénieur
11 stagiaires en tant qu'encadrante ou co-encadrante principale

2023 **Baptiste Rousseau, 6 mois**, Université de Paris Cité, Master 2 de Bioinformatique, *Développement d'un connecteur et d'une interface logicielle entre DeepOmics, un système d'information de données biologiques de la méthanisation, et les outils de soumission des données de séquençage à l'European Nucleotide Archive (ENA) de l'Institut Français de Bioinformatique (IFB)*

2022 **Caroline Talleu, 6 mois**, co-accréditation Université Bretagne Occidentale (UBO) et Université de Rennes 1 (UR1), Master 2 Microbiologie Fondamentale et Appliquée (MFA), *A la recherche de nouveaux virus d'archées grâce à des approches combinant isotopie et analyses bioinformatiques de métaviromes*

2021 **Sleheddine Kastalli, 6 mois**, Université Paris Diderot, Master 2 Biologie Informatique, *Analyse statistique des communautés microbiennes dans les systèmes bioélectrochimiques*

2021 **Lucia Ripol Malo, 6 mois**, Université de Pau (UPPA), Master 2 Biologie Moléculaire et Microbiologie de l'Environnement (BME). University of Pau et *University of Oviedo* (Espagne), Biotechnology of Environment and Health, *Diversité des communautés microbiennes et des particules virales au sein de méthaniseurs de déchets organiques en Ile-de-France*

2020 **Ferran Colomies, 6 mois**, Sorbonne Université, Master 2 de Microbiologie Environnement Santé, *Etudier les virus d'archées méthanogènes au sein des communautés microbiennes complexes de méthaniseurs : exploration d'approches* in silico *et expérimentales*

2019 **Maximilien Sotomski, 6 mois**, Sorbonne Université, Master 2 Microbiologie-Environnement-Santé, *Vers une nouvelle méthode moléculaire pour établir le lien entre les virus d'archées et leur hôte au sein d'écosystèmes de digestion anaérobie*

2019 **Oumar Telly Diallo, 2 mois** Université Paris-Saclay, Master 1 Biodiversité et Génome, *Démonstration de fouille de données de biotechnologies environnementales - Effets de l'inhibition par le phénol ou l'ammoniac sur les communautés microbiennes de réacteurs de méthanisation*

2018 **Maxime Allioux, 6 mois**, Université Paris-Saclay, Master 2 Génomique et Environnement, *Vers une nouvelle approche basée sur le Stable Isotope Probing pour mettre en évidence les associations entre les virus et leur hôte*

2017 **Lilia Ahmed Zaid, 6 mois**, Université de Bordeaux, Master 2 Bioinformatique, *Analyse du lien entre puissance métabolique et entropie du génome chez les micro-organismes*

2013 **Elodie Perrin, 6 mois**, Université Pierre et Marie Curie, Master 2 Microbiologie Appliquée à l'Environnement et à la Santé, *Diversité de virus présents dans les bioprocédés de méthanisati*on

2010 **Amandine Barket, 6 mois**, Université Paris-Sud 11, Master 2 Microbiologie Appliquée et Génie Biologique, *Colonisation de la cellulose et performances de sa dégradation anaérobie : comparaison d'inoculums issus de différentes installations de traitements des déchets*

6 autres stagiaires co-encadrés

2023   Julien Gordonnat, 5 mois, Université Bretagne Sud, Master 1 en Biotechnologies, *Etude des dynamiques microbiennes au sein d'un microméthaniseur en vue de l'identification de nouvelles stratégies opératoires*

2023   Louise Rigaud, 6 mois, Université Technologique de Compiègne (5ème année), *Préparation et caractérisation de consortia microbiens électrotrophes*

2020   Chloé Soulard, 6 mois, Université Technologique de Compiègne (2ème année), *Etude de l'inhibition de la digestion anaérobie des boues urbaines par l'acide propionique*

2013   Mathilde Lagesse, 6 mois, Chimie ParisTech (2ème année), *Mise en place d'un protocole de détection in situ des gènes impliqués dans l'hydrolyse de la cellulose. Application à l'observation et au suivi de la colonisation de déchets lignocellulosiques modèles lors de leur méthanisation.*

2012   Romain Guyonnet, **6 mois,** Université Paris Diderot, Master 2 Microbiologie Fondamentale, *Etude du prétraitement mécanique par broyage de déchets lignocellulosiques modèles : Influence sur les dynamiques de colonisation et de dégradation de* Clostridium cellulolyticum

2010   Carolina Hoyos, 6 mois, Université Paris-Est Créteil (UPEC), Master 2 Ingénierie biologique pour l'environnement (IBE), *Potentiel de dégradation du S-métolachlore dans les zones humides artificielles : une évaluation en microcosmes.*

Doctorants
3 doctorants en tant que co-encadrante principale

2020-2022   Marion Covès, Ecole doctorale ABIES, Université Paris-Saclay, *Effet d'inhibitions abiotiques de la digestion anaérobie sur les virus des écosystèmes de méthanisation.* Direction de thèse : Ariane Bize (INRAE PROSE, ADR).

2019-2021   Hoang Ngo, Ecole doctorale ABIES, Université Paris-Saclay. *Caractérisation génomique de virus d'archées et de bactéries au sein de procédés de digestion anaérobie par le couplage d'approches isotopiques et métagénomiques.* Direction de thèse : Théodore Bouchez (INRAE PROSE).

2010-2014   Nelly Badalato, Ecole doctorale ABIES, AgroParisTech, *Structure de déchets lignocellulosiques : effets sur la colonisation, les communautés microbiennes et les performances de méthanisation, caractérisés par des approches fonctionnelles et haut-débit.* Direction de thèse : Jean-Jacques Pernelle (Irstea HBAN).

4 autres doctorants en co-encadrement

2017-2021   Franciele P Camargo, Université de São Paulo, Brésil, *Obtention de biogaz et d'autres composes d'intérêt biotechnologique à partir du traitement de résidus d'agrumes*.

2017-2020   Lays P Leonel, Université de Campinas, Brésil, *Effets de la réutilisation d'eaux usées traitées en agriculture*.

2009-2013   Andreia F Salvador, Université de Minho, Portugal, *Analyse fonctionnelle d'écosystèmes microbiens dégradant les acides gras à longue chaîne par syntrophie*.

2009-2013   Liping Hao, Université de Tongji, Chine, Irstea, France, *Structure microbienne dans des granules anaérobies thermophiles, à différentes concentrations d'azote ammoniacal.*

Post-doctorants
En co-encadrement

2017-2018   Tiago P Delforno, Université de Campinas, Brésil – Irstea, France, *Fouille de données méta-omiques pour les procédés biotechnologiques de digestion anaérobies*

2014-2015   Sabine Podmirseg, Université d'Innsbruk, Autriche – Irstea, France, *Métatranscriptomique et métagénomiques des champignons anaérobies*.

2013-2015   Liping Hao, *Déploiement des approches omiques afin de diagnostiquer les causes de dysfonctionnement de digesteurs et proposer de nouveaux modes de gestion*.

2009-2013   Lü Fan, Université de Tongji, Chine – Irstea, France, Protéogénomique de l'hydrolyse de la cellulose dans des écosystèmes microbiens anaérobies naturels et de procédés

Projets de recherche

(Sauf indication contraire, le budget total de chaque projet est mentionné).

2023   programme EC2CO INSU-CNRS, projet DEPICTO, 7,8 k€, *Déchiffrer par epicPCR des interactions cellules-virus au sein de méthaniseurs de déchets organiques* – Coordinatrice scientifique.

2023-(2027)   action du programme MOCOPEE, 10 k€/an, *Etudier les relations entre la diversité microbienne et le fonctionnement des systèmes industriels, pour une application opérationnelle des outils de biologie moléculaire dans la conduite des procédés de traitement anaérobie*- Coordinatrice scientifique

2023-2026   HORIZON-INFRA-2022-TECH-01, BIOINDUSTRY 4.0, total de 10 M€, (101094287, Research and Innovations Actions)**,** *RI services to promote deep digitalization of Industrial Biotechnology - towards smart biomanufacturing* - Partenaire scientifique

2022-2023   INRAE, Département MICA, 30 k€, MISTRAL, *Compréhension du déterminisme de la dynamique taxonomique au sein des niches écologiques d'un microméthaniseur, en vue de l'identification de nouvelles stratégies opératoires de pilotage de la digestion anaérobie* - Partenaire scientifique

2022-2025   3BCAR projet de ressourcement MEMOS, total de 197 k€, *Entrepôts facilitant la fouille de données méta-omiques dans le domaine des biotechnologies environnementales, afin de favoriser la maîtrise des systèmes microbiens complexes*, Coordinatrice scientifique

2020-2021   INRAE, Département MICA, 30 k€, financement pour la poursuite du développement de DeepOmics (Digital Environmental Engineering Platform for Omics data) - Coordinatrice scientifique

2018-2022   action du programme MOCOPEE, 50 k€, *Systèmes de traitement anaérobies des boues d'épuration urbaines : mieux comprendre le lien entre communautés microbiennes et fonctionnement* – Partenaire scientifique (projet coordonné par l'Université de Rennes I, ECOBIO, France).

2017-2021   ANR JCJC VIRAME, 248 k€ (ANR-17-CE05-0011-01) incluant le financement d'une thèse sur "Caractérisation in situ du contenu génomique de virus d'archées méthanogènes au sein de bioprocédés de fermentation de déchets organiques" (https://virame.inrae.fr/) – Coordinatrice scientifique

2014-2019   Irstea, ~50 personnes.mois d'ingénieur en informatique (appel d'offre interne compétitif), projet en informatique scientifique concernant le développement d'un entrepôt pour les données méta-omiques de procédés de biotechnologies environnementales (DeepOmics, Digital Environmental Engineering Platform for Omics data) – Chef fonctionnel de projet

2010-2013   Région Île-de-France, 177 k€ (R2DS 2010-08) incluant une bourse de doctorat sur la co-digestion de déchets organiques – Coordinatrice scientifique

2010-2013   Irstea, demi-bourse de doctorat (appel d'offre interne compétitif) sur les liens entre la colonisation microbienne de la lignocellulose et les performances de sa dégradation en conditions anaérobies – Encadrante de la doctorante

Jury et comité de thèse

2022-2024   **membre du comité de thèse de Julia Gendre**, Université Paris Saclay, ED ABIES, *Caractérisation de l'impact des bactériophages sur la conduite de fermentations alimentaires*, (INRAE SayFood, Palaiseau, co-encadrants : Claire Le-Hénaff-Le-Marrec (INRAE Oenologie) et Eric Dugat-Bony, directrice de thèse : Sophie Landaud)

2019   **membre externe du jury de thèse de Sarah Thiroux**, Université de Bretagne Occidentale, EDSM, *Etudes des interactions entre virus et hôtes archéens hydrothermaux hyperthermophiles* (UBO LM2E, Brest – directeurs de thèse : Claire Geslin and Anne Godfroy)

2018   **membre externe du jury de thèse de Jeffrey Cornuault**, Université Paris-Saclay, ED ABIES, *Impact des phages tempérés sur la stabilité du microbiote intestinal : la lysogénie n'est pas un long fleuve tranquille* (équipe Phage, INRAE MICALIS, Jouy-en-Josas – encadrante : Marianne de Paepe, directrice de thèse : Marie-Agnès Petit)

2010-2013   **membre du comité de thèse de Charles Motte**, Université de Montpellier II, ED SPSA, *Digestion anaérobie par voie sèche de résidus lignocellulosiques : Etude dynamique des relations entre paramètres de procédés, caractéristiques du substrat et écosystème microbien* (INRAE LBE, Narbonne – encadrante : Claire Dumas, directeurs de thèse : Jean-Philippe Delgenes, Nicolas Bernet)

Autres comités, expertise

2022   **Présidente du jury pour le concours externe INRAE** de recrutement de techniciens de recherche en génomique (n°TRA09)

2021-2024   **Membre élue du Conseil Scientifique du Département MICA d'INRAE** (Microbiologie de la chaîne alimentaire)

2018   **Membre du comité de sélection pour le recrutement d'un maître de conférences** en microbiologie/bactériologie à Paris Sorbonne Université (n°65MCFl054 (0017)

2014-2016   **Participation à l'analyse stratégique collective** (Asco) « *Biologie de synthèse pour la chimie, l'énergie et l'environnement* » pour AllEnvi, alliance nationale de recherche pour l'environnement

Publications

24 publications dans des revues internationales à comité de lecture

Articles de recherche

1.   Viveros ML, Azimi S, Pichon E, Roose-Amsaleg C, **Bize A**, Durandet F, Rocher V**. 2022**. Wild type and variants of SARS-COV-2 in Parisian sewage: presence in raw water and through processes in wastewater treatment plants. *Environmental Science and Pollution Research,* 29:67442-67449, https://doi.org/10.1007/s11356-022-22665-x.

2.   Ngo VQH, Enault F, Midoux C, Mariadassou M, Chapleur O, Mazéas L, Loux V, Bouchez T, Krupovic M, **Bize A. 2022**. Diversity of novel archaeal viruses infecting methanogens discovered through coupling of stable isotope probing and metagenomics. *Environmental Microbiology,* 24:4853-486 https://doi.org/10.1111/1462-2920.16120.

3.   Leonel LP, **Bize A**, Mariadassou M, Midoux C, Schneider J, Tonetti AL**. 2022**. Impacts of disinfected wastewater irrigation on soil characteristics, microbial community composition, and crop yield. *Blue-Green Systems* 4**:**247-271. https://doi.org/10.2166/bgs.2022.126.

4.   Guérin-Rechdaoui S, **Bize A**, Levesque-Ninio C, Janvier A, Lacroix C, Le Brizoual F, Barbier J, Amsaleg CR, Azimi S, Rocher V**. 2022**. Fate of SARS-CoV-2 coronavirus in wastewater treatment

sludge during storage and thermophilic anaerobic digestion. *Environmental Research* 214:114057, https://doi.org/10.1016/j.envres.2022.114057.

5. Hao L, Fan L, Chapleur O, Guenne A, **Bize A**, Bureau C, Lü F, He P, Bouchez T, Mazéas L. **2021**. Gradual development of ammonia-induced syntrophic acetate-oxidizing activities under mesophilic and thermophilic conditions quantitatively tracked using multiple isotopic approaches. *Water Research,* 204: 117586, https://doi.org/10.1016/j.watres.2021.117586.

6. Godon J-J, **Bize A**, Ngo H, Cauquil L, Almeida M, Petit M-A, Zemb O. **2021**. Bacterial Consumption of T4 Phages. *Microorganisms,* 9 :1852. https://doi.org/10.3390/microorganisms9091852.

7. Camargo FP, Sakamoto IK, Delforno TP, Mariadassou M, Loux V, Midoux C, Duarte ICS, Silva EL, **Bize A**, Varesche MBA**. 2021**. Microbial and functional characterization of an allochthonous consortium applied to hydrogen production from Citrus Peel Waste in batch reactor in optimized conditions. *Journal of Environmental Management* 291:112631, https://doi.org/10.1016/j.jenvman.2021.112631.

8. Camargo FP, Sakamoto IK, **Bize A**, Duarte ICS, Silva EL, Varesche MBA**. 2021**. Screening design of nutritional and physicochemical parameters on bio-hydrogen and volatile fatty acids production from Citrus Peel Waste in batch reactors. *International Journal of Hydrogen Energy* 46**:**7794-7809, https://doi.org/10.1016/j.ijhydene.2020.06.084https://www.sciencedirect.com/science/article/pii/S0360319920322217.

9. **Bize A**, Midoux C, Mariadassou M, Schbath S, Forterre P, Da Cunha V**. 2021**. Exploring short k-mer profiles in cells and mobile elements from Archaea highlights the major influence of both the ecological niche and evolutionary history. *BMC Genomics* 22**:**186, https://doi.org/10.1186/s12864-021-07471-y.

10. Delforno TP, Macedo TZ, Midoux C, Lacerda GV, Rué O, Mariadassou M, Loux V, Varesche MBA, Bouchez T, **Bize A**, Oliveira VM**. 2019**. Comparative metatranscriptomic analysis of anaerobic digesters treating anionic surfactant contaminated wastewater. *Science of The Total Environment* 649**:**482-494, https://doi.org/10.1016/j.scitotenv.2018.08.328.

11. Tian J-H, Pourcher A-M, **Bize A**, Wazeri A, Peu P**. 2017**. Impact of wet aerobic pretreatments on cellulose accessibility and bacterial communities in rape straw. *Bioresource Technology* 237**:**31-38, https://doi.org/10.1016/j.biortech.2017.03.142.

12. Badalato N, Guillot A, Sabarly V, Dubois M, Pourette N, Pontoire B, Robert P, Bridier A, Monnet V, Sousa DZ, Durand S, Mazéas L, Buléon A, Bouchez T, Mortha G, **Bize A. 2017**. Whole Proteome Analyses on Ruminiclostridium cellulolyticum Show a Modulation of the Cellulolysis Machinery in Response to Cellulosic Materials with Subtle Differences in Chemical and Structural Properties. *PloS one* 12**:**e0170524-e0170524. https://doi.org/10.1371/journal.pone.0170524.

13. Poirier S, **Bize A**, Bureau C, Bouchez T, Chapleur O. **2016**. Community shifts within anaerobic digestion microbiota facing phenol inhibition: Towards early warning microbial indicators? *Water Research* 100**:**296-305. https://doi.org/10.1016/j.watres.2016.05.041.

14. Hao L, **Bize A**, Conteau D, Chapleur O, Courtois S, Kroff P, Desmond-Le Quéméner E, Bouchez T, Mazéas L**. 2016**. New insights into the key microbial phylotypes of anaerobic sludge digesters under different operational conditions. *Water Research* 102**:**158-169, https://doi.org/10.1016/j.watres.2016.06.014.

15. **Bize A,** Cardona L, Desmond-Le Quéméner E, Battimelli A, Badalato N, Bureau C, Madigou C, Chevret D, Guillot A, Monnet V, Godon J-J, Bouchez T**. 2015**. Shotgun metaproteomic profiling of biomimetic anaerobic digestion processes treating sewage sludge. *PROTEOMICS* 15**:**3532-3543. https://doi.org/10.1002/pmic.201500041.

16. Lü F, **Bize A**, Guillot A, Monnet V, Madigou C, Chapleur O, Mazéas L, He P, Bouchez T**. 2014**. Metaproteomics of cellulose methanisation under thermophilic conditions reveals a surprisingly high proteolytic activity. *The ISME Journal* 8**:**88-102. https://doi.org/10.1038/ismej.2013.120.

17. Chapleur O, **Bize A**, Serain T, Mazéas L, Bouchez T**. 2014**. Co-inoculating ruminal content neither provides active hydrolytic microbes nor improves methanization of 13C-cellulose in batch digesters. *FEMS Microbiology Ecology* 87**:**616-629. https://doi.org/10.1111/1574-6941.12249.

18. Jaubert C, Danioux C, Oberto J, Cortez D, **Bize A**, Krupovic M, She Q, Forterre P, Prangishvili D, Sezonov G**. 2013**. Genomics and genetics of Sulfolobus islandicus LAL14/1, a model hyperthermophilic archaeon. *Open Biology* 3**:**130010. https://doi.org/10.1098/rsob.130010.

19. **Bize A,** Karlsson EA, Ekefjärd K, Quax TEF, Pina M, Prevost M-C, Forterre P, Tenaillon O, Bernander R, Prangishvili D**. 2009**. A unique virus release mechanism in the Archaea. *Proceedings of the National Academy of Sciences* 106**:**11306. https://doi.org/10.1073/pnas.0901238106.

20. Vestergaard G, Shah SA, **Bize A**, Reitberger W, Reuter M, Phan H, Briegel A, Rachel R, Garrett RA, Prangishvili D. **2008**. Stygiolobus Rod-Shaped Virus and the Interplay of Crenarchaeal Rudiviruses with the CRISPR Antiviral System. *Journal of Bacteriology* 190**:**6837. https://doi.org/10.1128/JB.00795-08.

21. Steinmetz NF, **Bize A**, Findlay KC, Lomonossoff GP, Manchester M, Evans DJ, Prangishvili D. **2008**. Site-specific and Spatially Controlled Addressability of a New Viral Nanobuilding Block: Sulfolobus islandicus Rod-shaped Virus 2. *Advanced Functional Materials* 18**:**3478-3486. https://doi.org/10.1002/adfm.200800711.

22. **Bize A**, Peng X, Prokofeva M, MacLellan K, Lucas S, Forterre P, Garrett RA, Bonch-Osmolovskaya EA, Prangishvili D. **2008**. Viruses in acidic geothermal environments of the Kamchatka Peninsula. *Research in Microbiology* 159**:**358-366. https://doi.org/10.1016/j.resmic.2008.04.009.

23. Guyon J, **Bize A**, Paul G, Stewart E, Delmas J-F, Taddéi F. **2005**. Statistical Study of Cellular Aging. *ESAIM: Proc* 14:100-114. https://doi.org/10.1051/proc:2005009.

Revue

24. Pina M, **Bize A**, Forterre P, Prangishvili D**. 2011**. The archeoviruses. *FEMS Microbiology Reviews* 35**:**1035-1054. https://doi.org/10.1111/j.1574-6976.2011.00280.x.

Vulgarisation / semi-vulgarisation

25. **Bize A**, Sezonov G, Prangishvili D**. 2013**. Énigmatiques virus d'archées. *Biologie Aujourd'hui* 207**:**169-179. https://doi.org/10.1051/jbio/2013015.

26. **Bize A,** Sezonov G. **2010**. Les virus d'archées: de la morphologie aux nanotechnologies: Les virus de microbes. *Biofutur* (Puteaux)**:**41-45.

27. **Bize A,** Forterre P, Prangishvili D**. 2010**. Les archéovirus. *Virologie* 14**:**101-117.

Fiche technique

28 Bouchez T, **Bize A**. **2016**. Fiche application déchets 1. « Diagnostic de l'hydrolyse et de la méthanisation de déchets cellulosiques à l'aide d'approches métaprotéomiques » (2 pages) (2016) Fiche application dans « *La microbiologie moléculaire au service du diagnostic environnemental* », ADEME Observatoire des sols vivants.

Annexe : sélection de publications

1. **Lü F, Bize A, Guillot A, Monnet V, Madigou C, Chapleur O, Mazéas L, He P, Bouchez T.** 2014. Metaproteomics of cellulose methanisation under thermophilic conditions reveals a surprisingly high proteolytic activity. The ISME Journal **8:**88-102. doi:10.1038/ismej.2013.120. https://doi.org/10.1038/ismej.2013.120.

2. **Badalato N, Guillot A, Sabarly V, Dubois M, Pourette N, Pontoire B, Robert P, Bridier A, Monnet V, Sousa DZ, Durand S, Mazéas L, Buléon A, Bouchez T, Mortha G, Bize A.** 2017. Whole Proteome Analyses on Ruminiclostridium cellulolyticum Show a Modulation of the Cellulolysis Machinery in Response to Cellulosic Materials with Subtle Differences in Chemical and Structural Properties. PLOS ONE **12:**e0170524. doi:10.1371/journal.pone.0170524. https://doi.org/10.1371/journal.pone.0170524.

3. **Bize A, Karlsson EA, Ekefjärd K, Quax TEF, Pina M, Prevost M-C, Forterre P, Tenaillon O, Bernander R, Prangishvili D.** 2009. A unique virus release mechanism in the Archaea. Proceedings of the National Academy of Sciences **106:**11306-11311. doi:10.1073/pnas.0901238106. https://doi.org/10.1073/pnas.0901238106.

4. **Bize A, Midoux C, Mariadassou M, Schbath S, Forterre P, Da Cunha V.** 2021. Exploring short k-mer profiles in cells and mobile elements from Archaea highlights the major influence of both the ecological niche and evolutionary history. BMC genomics **22:**1-22.

5. **Ngo VQH, Enault F, Midoux C, Mariadassou M, Chapleur O, Mazéas L, Loux V, Bouchez T, Krupovic M, Bize A.** 2022. Diversity of novel archaeal viruses infecting methanogens discovered through coupling of stable isotope probing and metagenomics. Environmental Microbiology **24:**4853-4868. doi:https://doi.org/10.1111/1462-2920.16120. https://doi.org/10.1111/1462-2920.16120.

npg

ORIGINAL ARTICLE

Metaproteomics of cellulose methanisation under thermophilic conditions reveals a surprisingly high proteolytic activity

Fan Lü[1,2], Ariane Bize[1], Alain Guillot[3], Véronique Monnet[3], Céline Madigou[1], Olivier Chapleur[1], Laurent Mazéas[1], Pinjing He[2] and Théodore Bouchez[1]

[1]Irstea, UR HBAN, F-92761, Antony, France; [2]State Key Laboratory of Pollution Control and Resource Reuse, Tongji University, Shanghai, China and [3]INRA, UMR1319 MICALIS, PAPPSO, Jouy-en-Josas, France

Cellulose is the most abundant biopolymer on Earth. Optimising energy recovery from this renewable but recalcitrant material is a key issue. The metaproteome expressed by thermophilic communities during cellulose anaerobic digestion was investigated in microcosms. By multiplying the analytical replicates (65 protein fractions analysed by MS/MS) and relying solely on public protein databases, more than 500 non-redundant protein functions were identified. The taxonomic community structure as inferred from the metaproteomic data set was in good overall agreement with 16S rRNA gene tag pyrosequencing and fluorescent _in situ_ hybridisation analyses. Numerous functions related to cellulose and hemicellulose hydrolysis and fermentation catalysed by bacteria related to _Caldicellulosiruptor_ spp. and _Clostridium thermocellum_ were retrieved, indicating their key role in the cellulose-degradation process and also suggesting their complementary action. Despite the abundance of acetate as a major fermentation product, key methanogenesis enzymes from the acetoclastic pathway were not detected. In contrast, enzymes from the hydrogenotrophic pathway affiliated to _Methanothermobacter_ were almost exclusively identified for methanogenesis, suggesting a syntrophic acetate oxidation process coupled to hydrogenotrophic methanogenesis. Isotopic analyses confirmed the high dominance of the hydrogenotrophic methanogenesis. Very surprising was the identification of an abundant proteolytic activity from _Coprothermobacter proteolyticus_ strains, probably acting as scavenger and/or predator performing proteolysis and fermentation. Metaproteomics thus appeared as an efficient tool to unravel and characterise metabolic networks as well as ecological interactions during methanisation bioprocesses. More generally, metaproteomics provides direct functional insights at a limited cost, and its attractiveness should increase in the future as sequence databases are growing exponentially.

Introduction

Lignocellulosic materials including paper, textile, wood, yard trimmings and crop straws are dominant in municipal solid waste (MSW), agricultural waste and energy crops. Lignocellulose is the most abundant biochemical renewable energy source on Earth, because the bioenergy it contains can be recovered by cost-effective and robust anaerobic digestion technologies that are now widely applied. These processes involve sophisticated self-assembled and largely uncultured microbial communities comprising tens to hundreds of operational taxonomy units (Chouari _et al._, 2005; Krause _et al._, 2008; Kröber _et al._, 2009; Riviere _et al._, 2009). However, various biotechnological barriers still hinder a fully optimised process operation. Major issues include the polysaccharide hydrolysis efficiency, the methanogenesis stability and the sensitivity to inhibitions and disturbances (Chen _et al._, 2008; Ward _et al._, 2008; Ganidi _et al._, 2009; Holm-Nielsen _et al._, 2009). To better pilot anaerobic bioprocesses, additional functional insight into the catalysis of complex organic substrate digestion by the anaerobic microbial communities is required.

Significant progress has been achieved over the last few years. Fluorescent _in situ_ hybridisation (FISH) approaches contributed to the identification of cellulolytic bacteria in various anaerobic environments (O'Sullivan _et al._, 2007). Isotopic labelled

substrates were exploited to identify the functional microbial groups catalysing the methanisation of cellulose using stable isotope probing (Li et al., 2009). Several metagenomic studies shed light on the identity of the functional groups and provided an extended catalogue of the catalytic potential (Krause et al., 2008; Schlüter et al., 2008; Jaenicke et al., 2011; Rademacher et al., 2012). However, these approaches do not generate direct information on the expressed genes and associated metabolic processes. Recently, metatranscriptome sequencing provided insight into the metabolically active communities of a mesophilic biogas plant (Zakrzewski et al., 2012). Two metaproteomic approaches were implemented on similar anaerobic systems, and the feasibility of such approaches was proved (Abram et al., 2011; Hanreich et al., 2012), and more specifically on the complex matrix of anaerobic lignocellulose-degrading communities (Hanreich et al., 2012). Both based on a two-dimensional gel separation, they enabled the identification of a few dozen functions in the case of psychrophilic anaerobic digestion of glucose-fed wastewater (Abram et al., 2011) and of a dozen functions for methanisation of agricultural biomass in thermophilic conditions (Hanreich et al., 2012).

As thermophilic methanisation is currently emerging as a promising process (van Lier et al., 2001), we investigated the metaproteome obtained from a microcosm containing office paper and inoculated with stabilised digestate from a thermophilic MSW anaerobic digester, with the objective of achieving a comprehensive view. To increase the identification depth, a combination of various separation techniques was employed, and a high number of protein fractions were analysed. The possibility of obtaining valuable information based solely on public protein databases was questioned. For this, identification results using the UniprotKB database were compared with analyses of the microcosm generated independently by a polyphasic approach (16S rRNA gene pyrosequencing, FISH, stable isotopic fractioning signatures of methanogenesis processes). Encouragingly, more than 500 non-redundant protein functions were identified, and the complementary approaches were in overall good agreement.

After presenting an overview of the metaproteomic data set and assessing its validity, functional insights are presented with a focus on each methanisation step (hydrolysis, fermentation, acetogenesis and methanogenesis) and on unexpectedly high proteolytic activities.

Materials and methods

Anaerobic incubations
Sludge was sampled from a 21-m^3 thermophilic anaerobic industrial pilot digester located in France, fed with the organic fraction of MSW. The sample was sieved, stabilised at 55 °C, centrifuged at 13 100 g at 4 °C, aliquoted and stored at -80 °C to serve as inoculum. In each of five 1-l bottle (replicates A–E), 5 g unprinted office paper, 10 g inoculum (wet mass) and 500 g Biochemical Methane Potential buffer (EN-ISO-11734, 1998) were added (see also Supplementary Materials Section S0). Three similar microcosms without paper were set up as control. The bottles were rubber-sealed, the headspace was purged with N_2 and all microcosms were incubated in anaerobic and thermophilic conditions (55 ± 0.5 °C).

Chemical analyses
The degradation dynamics was assessed by measuring the biogas production and composition, pH, concentrations of total organic carbon, total inorganic carbon and volatile fatty acids in the liquid phase. The biogas isotopic composition was analysed by determining $\delta13CH_4$ and $\delta13CO_2$ values and calculating the apparent fractionation factor α_C (Qu et al., 2009). Biogas production and composition were assessed as described in Qu et al. (2009). The detected gas included CO_2, CH_4, H_2S, N_2, H_2 and O_2. The liquid samples were recovered with a syringe and needle through the rubber septum. They were centrifuged briefly at maximum speed in a bench centrifuge to separate the liquid phase from the cell-containing pellet. The pH was measured on the liquid phase just after sampling. The supernatants and pellets were stored separately at -80 °C for further chemical and biological analyses.

Protein preparation and mass spectrometry analyses
Proteins were extracted and purified from 50 ml samples of the paper anaerobic incubation (replicate A, day 60) using a protocol from Wilmes and Bond (2004), modified to handle samples from a MSW digester containing much debris. Briefly, cells were disrupted by bead-beating in the lysis buffer, and the proteins were purified from the obtained supernatant by trichloroacetic acid–acetone precipitation. Protein concentrations were assessed using the 2-D Quant Kit (GE Healthcare, Aulnay sous Bois, France) and the High Sensitivity Protein 250 Kit (Agilent, Les Ulis, France). The purified proteins were further processed for subsequent MS/MS analyses according to three different strategies detailed below. For every strategy, fixation of each fraction in SDS-PAGE gel (NuPAGE Novex 4–12% Bis-Tris Gel 1.0 mm; Invitrogen, Saint Aubin, France) was the last step, with or without prior separation. Strategy 1 was the separation according to the molecular weight, resulting in 26 fractions: a protein aliquot was migrated by SDS-PAGE and the gel lane was cut into 26 sections. Strategy 2, performed in three technical replicates, was the separation into 12

fractions according to the pI, resulting in a total of 36 fractions: for each protein aliquot, 12 liquid fractions were generated by off-gel isoelectric focusing (OFFGEL-IEF, Agilent 3100 OFFGEL Fractionator, low-resolution kit, pH 3–10, 12 cm Immobilised pH Gradient strip), and each obtained fraction was fixed into SDS-PAGE gel by a very short-duration migration and gel excision. Strategy 3, performed in three technical replicates, was the absence of separation and generated 3 fractions: the protein aliquot was simply fixed in an SDS-PAGE gel fragment by a very short-duration migration followed by gel excision. As a result, a total of 65 fractions $(26+36+3)$ fixed in SDS-PAGE gel fragments were separately submitted to in-gel tryptic digestion followed by shotgun analyses by nanoLC-MS/MS (LTQ-Orbitrap, Thermo Fisher, Waltham, MA, USA; PAPPSO proteomic platform, INRA, Jouy-en-Josas). The detailed procedures are supplied in the Supplementary Section S1.

Peptide identification and data processing
The mass spectrometry data set produced for each fraction was analysed for peptide and protein identification using X!Tandem software (http://www.thegpm.org/tandem/) and the UniProtKB database with nearly 20 million entries (version January 2012, http://uniprot.org). The 65 X!Tandem result sets were imported into the software Scaffold 2.0 (Proteome Software, Inc., Portland, OR, USA) to combine, compare and validate identified proteins based on peptide and protein probability. The filtering thresholds were the protein probability >95%, at least two unique peptides per protein, and the peptide probability of at least one unique peptide >90%. The potential contaminant proteins, such as keratin and trypsin, were excluded from the analysis.

The obtained identified redundant proteins were grouped according to two different methods: first, non-redundant protein groups were obtained in Scaffold 2.0 based on the presence of shared identified peptides; second, and independently, another grouping was performed based on belonging to shared UniRef clusters with various identity thresholds (for example, 50%, 90% and 100%). The latter methods provided a simple and convenient insight into the taxonomic specificity of the identified redundant proteins. For instance, the case where an identified protein belonged to a UniRef50 cluster containing solely this protein was a strong indication that the taxonomic assignment and the inferred function were very specific because no other closely related protein sequence was present in the database. The possible presence of a signal peptide within protein sequences was analysed with SignalP (Petersen *et al.*, 2011). The detailed procedures are supplied in the Supplementary Materials (Supplementary Section S1).

DNA analyses
DNA was extracted using the PowerSoil DNA Isolation Kit (MO BIO Laboratories, Inc., Carlsbad, CA, USA) according to the manufacturer's instructions. A series of 16S rRNA-based techniques were conducted to describe the microbial community, including automated ribosomal intergenic spacer analyses (ARISA, Supplementary Figure S12), 16S rRNA gene pyrosequencing and FISH. Pyrosequences were obtained for the raw digestate directly sampled from the industrial facility, the inoculum (sieved and stabilised digestate, see section *Anaerobic incubations*), the replicate A (day 0 and 60) and the replicate B (day 60 and 73). They were deposited in the National Center for Biotechnology Information Short Read Archive as BioProject PRJNA182049. More detailed procedures are supplied in the Supplementary Section S2.

Results and Discussion

Office paper degradation dynamics and general overview of the metaproteome data set
Office paper was selected as a model cellulosic substance, because it is a major component in MSW and has a relatively fixed composition of 70% cellulose and 30% hemicellulose. Batch anaerobic incubations of office paper were conducted at 55 °C in five replicated microcosms labelled from A to E. At day 60, the replicate A was used for metaproteomic analyses, whereas replicates B–E were further incubated until day 120 (Supplementary Materials Section S0). The replicates exhibited a good level of reproducibility, and classical degradation trends were observed (Supplementary Figure S2). The rapid onset of a hydrolytic and acidogenic activity led to the accumulation of volatile fatty acid, reaching a concentration of 360 mg carbon per liter at day 18 and mainly corresponding to acetate. This induced a pH decrease from 7.5 to 5.8. After a phase of slow methane production, the methanogenic activity increased gradually from day 20 onwards until it reached a plateau around day 60. At this stage, 62% of the carbon initially introduced as paper had been degraded, and volatile fatty acid concentrations in the liquid phase were low and mainly corresponded to acetate, propionate and lactate.

All metaproteomic analyses were performed on samples from replicate A, day 60: 65 protein sample fractions were prepared by three different separation strategies and analysed by MS/MS (Materials and methods) to favour a sufficient analysis depth (Figure 1a). Strategy 2 based on OFFGEL-IEF generated the highest number of identified non-redundant protein groups (266, 358 and 360 for each technical replicate, respectively), followed by strategy 1 based on SDS-PAGE (212 protein groups) and by strategy 3 (no separation, 54, 90 and 120 protein groups for each technical
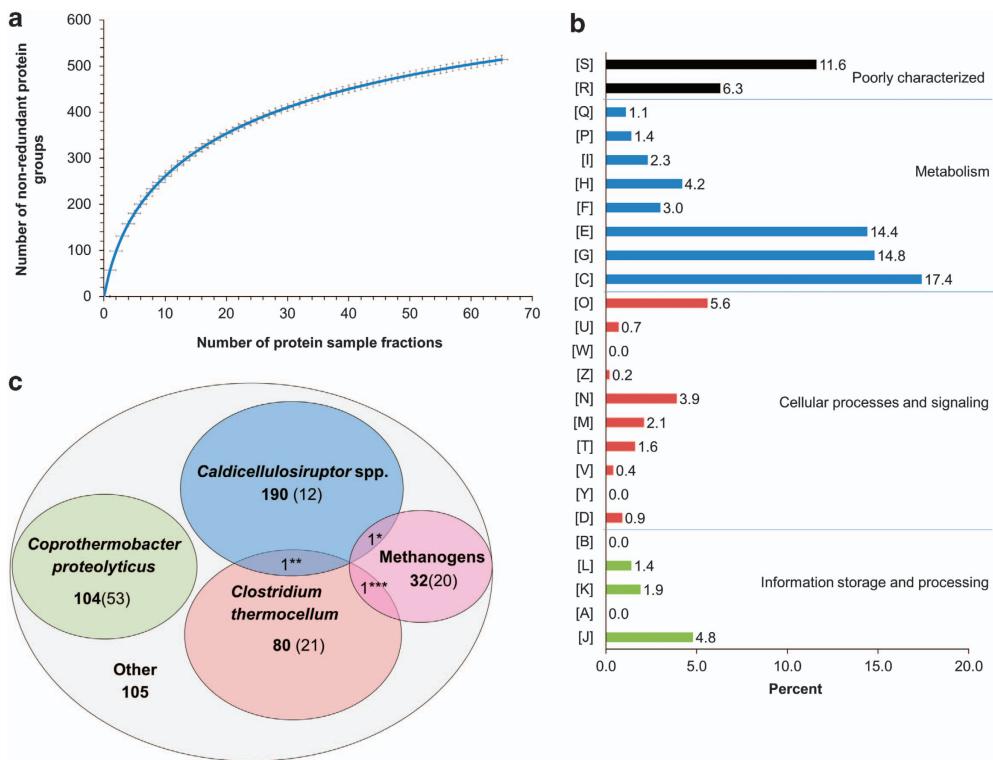
Figure 1 Overview of the metaproteomics results. (**a**) Rarefaction curve. Number of identified non-redundant protein groups in function of the number of protein sample fractions included in the analyses. The fractions were generated by several technical replicates and various separation procedures (Materials and methods). A total of 65 protein fractions were included. (**b**) Distribution of the identified non-redundant protein groups into Clusters of Orthologous Groups. [S] Function unknown, [R] General function prediction only, [Q] Secondary metabolites biosynthesis, transport and catabolism, [P] Inorganic ion transport and metabolism, [I] Lipid transport and metabolism, [H] Coenzyme transport and metabolism, [F] Nucleotide transport and metabolism, [E] Amino acid transport and metabolism, [G] Carbohydrate transport and metabolism, [C] Energy production and conversion, [O] Post-translational modification, protein turnover and chaperon functions, [U] Intracellular trafficking, secretion and vesicular transport, [W] Extracellular structures, [Z] Cytoskeleton, [N] Cell motility, [M] Cell wall/membrane/envelope biogenesis, [T] Signal transduction mechanisms, [V] Defense mechanisms and [Y] Nuclear structure. (**c**) Taxonomic distribution of the identified non-redundant protein groups. Bold number: number of non-redundant protein groups. Number in parenthesis: number of non-redundant protein groups belonging to UniRef50 clusters specific for the considered taxonomic or functional group (*C. proteolyticus* species, *C. thermocellum* species, *Caldicellulosiruptor* genus, *Methanothermobacter* genus and other methanogens, respectively) (see Materials and methods for more details about UniRef50). *Putative uncharacterised protein A4XIB5_CALS8 from *Caldicellulosiruptor*, and/or A7I471_METB6 from Methanogens **RNA polymerase sigma factor A4XHW4_CALS8 from *Caldicellulosiruptor* and/or A3DDV0_CLOTH from *C. thermocellum* ***Phenylacetate-CoA ligase A3DD21_CLOTH from *C. thermocellum* and/or A6VIF0_METM7 from Methanogens.

replicate, respectively) (Supplementary Table S4, Supplementary Figure S5). Both the separation step and the amount of starting protein material appeared as important factors to favour identification (Supplementary Table S3). All the data sets were then combined and analysed together. A significant number of proteins were identified using the UniprotKB database (full content). Among the 7 17 065 spectra obtained, 40 818 ($\approx 6\%$) were assigned to peptides by X!Tandem search. As a reference, 25% of the spectra are typically assigned when studying 10 µg proteins from a pure culture of *Lactococcus lactis* (Beganovic *et al.*, 2010). After filtering with Scaffold software, 13 090 peptides corresponding to 2541 potentially redundant proteins were retained. The latter corresponded to 514 non-redundant protein groups and also to 497 distinct UniRef50 clusters (Materials and method). Except for a few protein groups that were highly redundant (for example actin, pyridoxine, RNA

polymerase sigma factor and some oxidoreductases), the redundancy was overall limited and distributed throughout the various protein groups.

According to the classification into the 25 Clusters of Orthologous Groups (Figure 1b), the category amino acids [E] seemed unexpectedly abundant, given that the initial substrate was only composed of carbohydrate. Indeed, the dominant categories were related to energy [C], carbohydrate [G] and amino acids [E], together accounting for 46.6% of the 514 non-redundant protein groups. Post-translation [O], translation [J] and coenzyme [H] were the next dominant functional categories, representing 14.6% of the total. Most proteins related to coenzyme [H] were assigned to archaea.

The taxonomic distribution of the identified proteins based on the non-redundant protein groups (Figure 1c, Supplementary Table S17) suggested the presence of a few dominant groups, possibly linked to the thermophilic conditions and the presence of

one major substrate for growth (office paper). The retrieved taxa were consistent with the anaerobic thermophilic conditions. *Caldicellulosiruptor* species ($\approx 41\%$), *Coprothermobacter proteolyticus* ($\approx 20\%$) and *Clostridium thermocellum* strains ($\approx 16\%$) were the most represented, suggesting the dominance of members of the order Thermoanaerobacterales followed by Clostridiales. *Methanothermobacter thermoautotrophicus* was the dominant archaea ($\approx 5\%$).

Validity and representativeness of the identified proteins

Polyphasic experiments and *in silico* approaches were implemented to evaluate the validity of the metaproteomic data before developing more detailed biological interpretation. This approach suggested that invaluable functional insight could be gained for most of the dominant microbial groups.

The accuracy of the taxonomic distribution resulting from the metaproteomic data set was evaluated by 16S rRNA gene tag pyrosequencing (Figure 2, Supplementary Figure S13). On the same sample (Figure 2, A60) and on samples from replicate B at similar time points (Figure 2, B60 and B73), the most dominant genera were consistently retrieved (*Caldicellulosiruptor*, *Coprothermobacter* and *Methanothermobacter*), and the dominance of members of the orders Thermoanaerobacterales followed by Clostridiales was confirmed. These dominant strains were not abundant in the inoculum (Figure 2, Inoc, A0; Supplementary Figure S12), and they thus developed in the course of the incubation, which further supports their active role during the office paper digestion. FISH analyses, reflecting the abundance of rRNAs in the specifically targeted cells, were performed for the replicate A at day 60. They provided additional elements of proof for the activity and abundance of the three above-mentioned bacterial genera (Supplementary Figure S11). The genera *Gelria* (Firmicutes phylum) and Tta-b61 (Firmicutes phylum) were poorly represented in the metaproteomic data set, although they appeared to be of significant importance based on the pyrosequencing results (Figure 2b, operational taxonomic units 489 and 356–383, respectively). These observations were probably mainly linked to the absence of closely related sequenced genomes in the database (Table 1). In addition, in contrast to the metaproteomic analyses, a low proportion of sequences were attributed to *C. thermocellum* and other Ruminococcaceae cellulolytic strains including *Clostridium cellulosi* (Figure 2b, A60). One possible reason could be that the classification of Firmicutes is very complex and that the taxonomic assignment results obtained on partial 16S rDNA sequences must be interpreted cautiously (average read lengths ~ 300–$400\,bp$). PCR bias on the sample from microcosm A, day 60, could also be an explanation for this discrepancy because
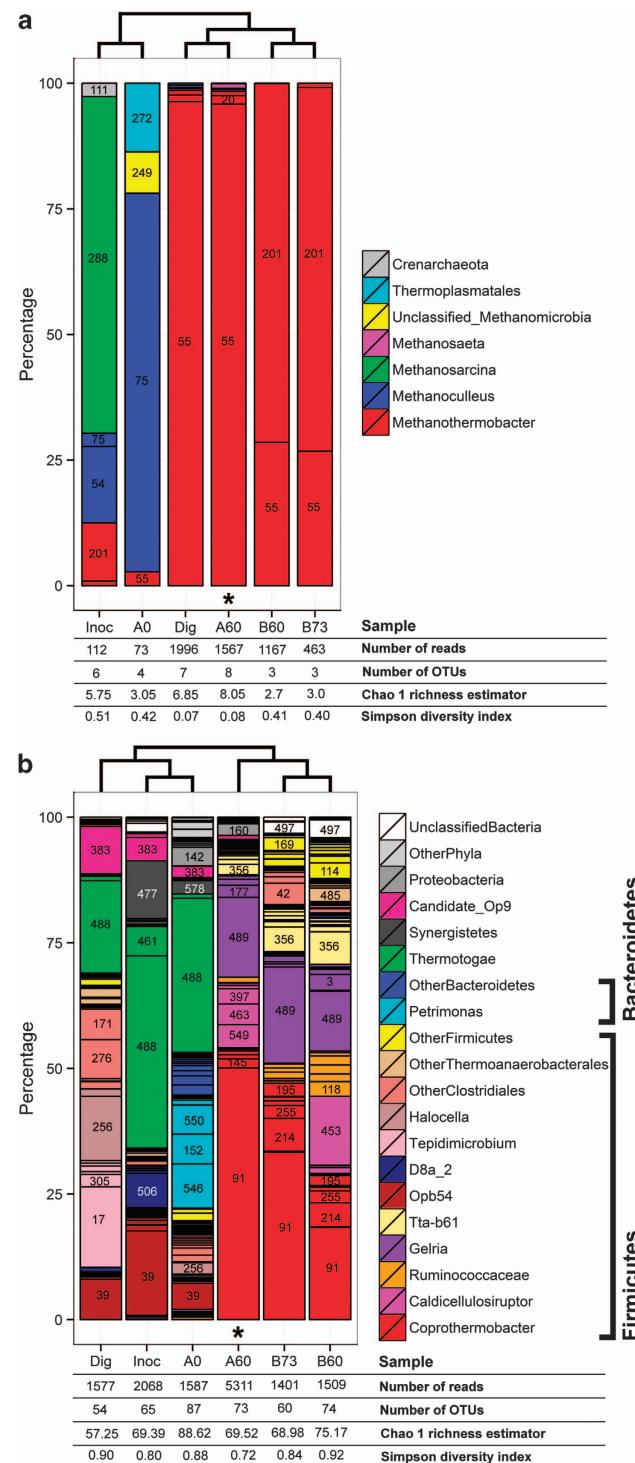
Figure 2 Taxonomic distributions obtained by 16S rRNA gene pyrosequencing: (**a**) with the archaeal primer set and (**b**) with the bacterial primer set. Dig: collected raw thermophilic digestion sludge; Inoc: thermophilic inoculum (sieved and stabilised digestion sludge); A0, A60, B60 and B73: microcosm and incubation day; * sample also analysed by metaproteomics (A60). For the most abundant groups, the arbitrary operational taxonomic unit (OTU) number is indicated. The cladogram is based on genus frequencies for archaea (**a**) and on OTU frequencies for bacteria (**b**). See Supplementary Section S2 and Supplementary Table S1 for more details about the procedures. Rarefaction curves are shown in Supplementary Figure S13.

Table 1 Overview of protein detection for some of the abundant bacterial taxonomic groups

	C. proteolyticus	C. thermocellum and Ruminoccaceae	Caldicellulosiruptor species	Gelria	Tta-b61
% in sample from microcosm A, day 60[a]	43–65	1–2	10–16	20–25	7–12
Closest fully sequenced microbial genome(s) (number of putative proteins encoded in the genomes)	C. proteolyticus (1482)	C. thermocellum (2911–3173)	Caldicellulosiruptor species (2147–2682)	Natranaerobius thermophilus, Desulfotomaculum kuznetsovii, Pelotomaculum thermopropionicum (2882, 3398, 2919)	Thermoanaerobacter pseudethanolicus (2239)
% 16S ID[b]	97–100	87–100	95–97	88	88
% NAI[c]	60–100	43–100	50–60	45	45
% AAI[d]	95–100	52–100	70–100	53	53
Protein identification probability[e]	0.62–1.00	0.02–1.00	0.21–1.00	0.02	0.02
% of the total number (514) of identified non-redundant proteins	20% (104 proteins)	16% (82 proteins)	37% (192 proteins)	3% (13 proteins)	1% (6 proteins)

[a]Rough estimate based on the 16S rRNA gene pyrosequencing data values from the bacterial primer set ±20%, and neglecting the archaeal population.
[b]Percentage of sequence identity between the sample-related sequences and the 16S rRNA gene sequences from the most closely related fully sequenced microbial genome represented in UniprotKB.
[c]Nucleic acid identity. Estimation based on Zaneveld *et al.* (2010).
[d]Amino acid identity. Estimation based on Konstantinidis and Tiedje (2007).
[e]Estimation based on Denef *et al.* (2007).

abundant related strains were observed in the sample by FISH with probe UCL284 (Supplementary Table S1, Supplementary Figure S11).

The identification of peptides and proteins using the full UniprotKB database was highly specific. As pointed out in a study by Denef *et al.*, 2007, similar procedures enable cross-strain identification and avoid most cross-species false-positive hits. Indeed in the present case, the overlap between the dominant microbial groups after protein identification was limited to three non-redundant protein groups (Figure 1c). In addition, 24% of the 427 UniRef50 clusters affiliated to the dominant microbial groups were specific for the concerned group, respectively, *Caldicellulosiruptor* species, *C. proteolyticus*, *C. thermocellum* and methanogenic archaea of various orders (Figure 1c, Supplementary Figure S10). All this confirmed the specificity of the identification procedure, and thus strengthens the validity of the identified functions.

In conclusion, the presence in the database of proteomes from fully sequenced genomes closely related to the sample's strains together with the specificity of the identification procedure suggested that interesting functional insights could be gained for most abundant groups (*C. proteolyticus*, *C. thermocellum*-related strains, *Caldicellulosiruptor* species, *Methanothermobacter* species). More limited information was expected for *Gelria* and Tta-b61 groups, which could have been significantly present in the sample.

Ruminococcaceae strains and Caldicellulosiruptor species are complementary key actors for polysaccharide hydrolysis

Consistent with paper being the main substrate for the incubations, a large number of proteins potentially involved in cellulose and hemicellulose binding and hydrolysis or in oligosaccharide metabolism were identified (Table 2). Ruminococcaceae strains including *C. thermocellum*, and *Caldicellulosiruptor* members, appeared as key actors for paper hydrolysis. Moreover, the details of the identified functions suggested a synergetic action of both actors. The former could have been hydrolysing the cellulosic part of the substrate, including the crystalline part, whereas the latter could have been specifically involved in hemicellulose hydrolysis in addition to cellulose hydrolysis.

More specifically, 11 non-redundant protein groups related to polysaccharide hydrolysis were attributed to *C. thermocellum*. This widely studied thermophilic cellulolytic species is known for efficiently hydrolysing the cellulose, including in its crystalline shape (reviewed in Maki *et al.*, 2009). Its cellulosome has been extensively characterised (reviewed in Bayer *et al.*, 2008), and the presence of dockerin/cohesin domains within a protein is a strong indication that they are cellulosomal subunits. The identified proteins included some major cellulosomal structural and catalytic components. In particular, CelS and CelJ proteins are among the most upregulated enzymes for *C. thermocellum* Avicel-grown cells compared with cellobiose-grown cells (Gold and Martin, 2007). Hence, even if a limited number of the over 30 cellulosomal proteins from *C. thermocellum* (Gold and Martin, 2007) were detected, the information was sufficient to confirm the cellulolytic activity of strains closely related to *C. thermocellum*.

Seven non-redundant protein groups related to polysaccharide hydrolysis were affiliated to the *Caldicellulosiruptor* genus (Table 2). This taxon contains cellulolytic members, such as *C. obsidiansis* or *C. bescii* (formerly *Anaerocellum thermophilum*). These do not possess cellulosomes, their cellulolytic system being based on secreted multifunctional

Table 2 Identified proteins putatively related to polysaccharide hydrolysis

Identified proteins[a]	Uniprot entry (gene)[b] [Reference][c]	Information based on public databases and publications				Mass spectrometry	
		UniRef50 clusters: ID name taxonomic range	CAZy[d]	Domains[e]	Putative functions or activities[d]	Number of peptide ions[f]	Coverage (%)[g]
Affiliated mainly to C. thermocellum strains (cellulolytic system based on a cellulosome)							
O86999 CLOTM (slpA) *A3DHX1_CLOTH*	UniRef50 O86999 S-layer protein C. thermocellum			SLH signal	S-layer protein; functional partners: carbohydrate-binding protein A3DET8_CLOTH (STRING 9 network http://string-db.org/)	32	42
Q06851_CLOTH (cipA) [1] *D1NHZ0_CLOTM E6U1J82_CLOTL*	UniRef50 Q06851 Cellulosomal-scaffolding protein A *C. thermocellum*	CBM3	Dockerin cohesin signal	Cellulosomal-scaffolding protein	20	15
A3DD30_CLOTH *C7HDM6_CLOTM D1NI43_CLOTM E6USK3_CLOTL* P71140_CLOTM (celJ) [1]	UniRef50 C7HDM6 Glycoside hydrolase family 9 domain protein, Ig domain protein cellulolytic Clostridiales	CBM30 CBM44 GH9 GH44	Dockerin Ig-like_fold signal	GH44: endoglucanase, xyloglucanase, active on many substances (GH9: endoglucanase)	11	9.3
A3DH67_CLOTH (celS) [1] *P0C2S5_CLOTM (celS) C7HDZ8_CLOTM E6US41_CLOTL D1NQ73_CLOTM*	UniRef50 P22534 Endoglucanase A mainly cellulolytic Clostridia	GH48	Dockerin signal	GH48: processive exoglucanase acting on reducing end, endo-processive cellulose, key component for enzymatic synergy, most abundant enzyme subunit	7	6.9
Q06852_CLOTH (olpB) [1] *D1NHZ1_CLOTM E6U1J83_CLOTL C7HDC4_CLOTM E6U1RK4_CLOTL*	UniRef50 Q06852 Cell surface glycoprotein 1 mainly *C. thermocellum*		Cohesin SLH signal	Cellulosome anchoring protein	6	4.8
	UniRef50 D1NNW4 Type 3a cellulose-binding domain protein *C. thermocellum*	CBM3	signal	CBM3a	5	6.7
A3DBG8_CLOTH *C7HH13_CLOTM E6UN62_CLOTL D1NKR6_CLOTM*	UniRef50 A3DBG8 Ig domain protein group 2 domain protein mainly *C. thermocellum*		SLH Big_2 signal	Functional partners: carbohydrate-binding protein A3DET8 (STRING 9 network http://string-db.org/)	5	1.4
A3DC35 CLOTH *C7HGL0_CLOTM E6ULX2_CLOTL*	UniRef50 D2Q7C4 Cellobiose phosphorylase mainly cellulolytic Clostridiales, Actinobacteridae	CBM GH94		GH94: cellobiose phosphorylase	3	3.3
A3DDD5_CLOTH *C7HD15_CLOTM D1NMI1_CLOTM E6URU4_CLOTL*	UniRef50 C7HD15 Cellulosome protein dockerin type I mainly *C. thermocellum*		Dockerin signal	Cellulosome protein dockerin type I	2	3.4
A3DJQ6_CLOTH *C7HEH9_CLOTM D1NRB6_CLOTM E6UTK1_CLOTL* Q93HT8_CLOTM (cdp-ym4)	UniRef50 D9SRI5 Glycosyltransferase 36(h) mainly *C. thermocellum*	GH94		GH94: cellodextrin phosphorylase	2	2.8
C7HJV5_CLOTM *D1NR31_CLOTM D1NRP9_CLOTM* O52779_CLOTM (xynV) *E6UT15_CLOTL* O52780_CLOTM (xynU) O87118_CLOTM (xynB) O87119_CLOTM (xynA) [1] *E6UT14_CLOTL* A3DJP0_CLOTH Q8GJ44_CLOSR (xynA)	UniRef50 A3DJP0 Glycoside hydrolase family 11 mainly cellulolytic Clostridiales	CBM6 GH11 CE4	Dockerin signal	GH11: endo-β-1,4-xylanase	2	4.2
Affiliated mainly to Caldicellulosiruptor species (cellulolytic system based on multifunctional enzymes)							
E4SCW4_CALK2	UniRef50 Q9KWY5 Beta-mannanase mainly thermophilic bacteria and archaea, including many Clostridia	GH3		GH3: removes single glycosyl residues from the non-reducing end, dual or broad substrate specificities, β-glucosidase	7	16
A4XM78_CALS8	UniRef50 F8F4V8 Esterase mainly Clostridia and Bacilli	CE7		CE7: acetyl xylan esterase	4	11
P22534_CALSA (celA) [2] *A4XIF5_CALS8*	UniRef50 P22534 Endoglucanase A mainly Clostridia and other bacteria	CBM3 GH9 GH48	Signal	GH48: processive exoglucanase acting on reducing end, endo-processive cellulose, key component for enzymatic synergy, most abundant enzyme subunit	3	1.3

Table 2 (Continued)

Identified proteins[a]	Information based on public databases and publications				Mass spectrometry	
Uniprot entry (gene)[b] [Reference][c]	UniRef50 clusters: ID name taxonomic range	CAZy[d]	Domains[e]	Putative functions or activities[d]	Number of peptide ions[f]	Coverage (%)[g]
E4S6B0_CALKI G2PWX9_9FIRM [2]	UniRef50_A4XIG8 Glycosyltransferase 36[h] Caldicellulosiruptor	GH94 CBM		GH94: cellobiose phosphorylase, membrane protein	2	3.9
G2PWE2_9FIRM	UniRef50_E4Q5G9 Endo-1,4-beta-xylanase mainly Caldicellulosiruptor	CBM9 CenC-like CBM GH10	SLH Signal	GH10: endo-beta-1,4-xylanase, membrane protein	2	1.3
A4XN28_CALS8 B9MLB5_ANATD D9TFV3_CALOO E4Q4G6_CALOW E4QE34_CALH1 E4S9M6_CALKI E4SEW2_CALK2 F7KFZ8_Lachno	UniRef50_P39130 Putative glucosidase lplD mainly Clostridia and Bacillales	GH4		GH4: distinct substrate specificities, unusual mechanism involving NAD+	2	6.7
B9MN93_ANATD D9TGX7_CALOO E4Q6A9_CALOW E4QAK4_CALH1	UniRef50_Q9KWY5 Beta-mannanase mainly thermophilic bacteria and archaea including many Clostridia	GH3		GH3: remove single glycosyl residues from the non-reducing end, dual or broad substrate specificities, β-glucosidase	2	7.1
Affiliated to Coprothermobacter proteolyticus strains (non-cellulolytic organisms)						
B5Y6P8_COPPD	UniRef50_B5Y6P8 Amylopullulanase C. proteolyticus	GH13		GH13:amylopullulanase, α-amylase, pullulanase (etc), wide range of different preferred substrates and products	26	21
B5Y7R5_COPPD	UniRef50_B5Y7R5 Endoxylanase C. proteolyticus		SLH signal	Endoxylanase, pectin lyase	2	3.3

[a]Redundant protein list: all possibilities corresponding to a given set of peptides are shown. *Italic font*: entry showing 100% sequence identity with the above entry. The affiliation to distinct UniRef90 clusters is materialised by distinct paragraphs.

[b]*Clostridium* species: CLOTH, *C. thermocellum* strain ATCC 27405/DSM 1237; CLOTM, *C. thermocellum* strain YS; CLOTL, *C. thermocellum* strain DSM 1313/LMG 6656/LQ8; CLOSR, *C. stercorarium*; *Lachnospiraceae* species: Lachno, *L. bacterium*; *Caldicellulsiruptor* species: 9FIRM, *C. lactoaceticus*; ANATD, *C. beccsii*; CALH1, *C. hydrothermalis*; CALK2, *C. kronotskyensis*; CALKI, *C. kristjanssonii*; CALS8, *C. saccharolyticus*; CALSA, *C. saccharolyticus*; CALOO, *C. obsidiansi*; CALOW, *C. owensensis*; and *Coprothermobacter* species: COPPD, *C. proteolyticus*.

[c]Reference indicating experimental evidence for the expression and/or activity of the corresponding protein or a closely related one. [1] = Gold and Martin, 2007; [2] = Lochner et al., 2011.

[d]More information at http://www.cazy.org/ and http://www.cazypedia.org (Cantarel et al., 2009). CBM, carbohydrate-binding module; GH, glycoside hydrolase; CE, carbohydrate esterase.

[e]SLH, S-layer homology domain; Big, bacterial immunoglobulin-like, group 2; Ig, immunoglobulin-like; signal, signal sequence (secreted or membrane protein).

[f]Number of unique parent peptide ions matched (including different charge states, modifications).

[g]Percentage of amino acid coverage to the matched protein.

[h]Glycosyltransferase 36 was recently reclassified as glycoside hydrolase 94 (GH94) according to CAZy.

enzymes; they usually utilise a broad range of plant materials, including crystalline cellulose, cellulose, hemicelllose, starch and pectin, with a very high hydrogen yield (van de Werken *et al.*, 2008; VanFossen *et al.*, 2009; Yang *et al.*, 2009; Hamilton-Brehm *et al.*, 2010; Lochner *et al.*, 2011). Among the non-redundant protein groups identified for *Caldicellulosiruptor*, four were probably related to hemicellulose degradation (two beta-mannanases UniRef50_Q9KWY5, one acetyl xylan esterase UniRef50_F8F4V8 and one endoxylanase UniRef50_E4Q5G9). This suggested that *Caldicellulosiruptor* members were also present in the incubations and partly specialised in hemicellulose degradation.

To the best of our knowledge, results on cellulo-lytic species functions and interactions have not been reported in such detail for cellulose methanisation by complex microbial communities.

Identified proteins related to central carbon metabolism
The identified proteins related to central carbon metabolism (Supplementary Tables S6–S8, Supplementary Figures S7–S9) reinforced the conclusions concerning the major contribution of *Caldicellulosiruptor* species (see also Supplementary Table S20) and *C. thermocellum* (see also Supplementary Table S19) to saccharification and fermentation during the incubation, with the former more oriented towards hemicellulolysis compared with the latter, which is more specialised for cellulolysis. In addition, these results highlighted the important contribution of *C. proteolyticus* as fermenting microorganism (see also Supplementary Table S18) and indicated the possible activity of other non-dominant species. Finally, they suggested that each of these groups produced a distinct combination of metabolic end products and that acetate was produced by most of the identified fermentative bacterial groups.

More specifically, after polysaccharide and oligosaccharide hydrolysis, monosaccharides are channelled to the central catabolic pathways to generate pyruvate. None of the enzymes from the Entner-Doudoroff pathway were detected. In contrast, proteins from the Embden–Meyerhof pathway (glycolysis) corresponding to 24 distinct UniRef50 clusters were identified (Supplementary Table S6, Supplementary Figure S7). Mainly attributed to strains from the *C. thermocellum*, *C. proteolyticus* and *Caldicellulosiruptor* genera, they confirmed the importance of these Gram-positive members of the class Clostridia in the studied community. Remarkably, considering together the identified proteins affiliated to the *Caldicellulosiruptor* genus, only the phosphofructo-kinase was not detected; all other nine enzymes from the Embden–Meyerhof pathway were retrieved for this genus. Proteins from the non-oxidative arm of the pentose phosphate pathway were present as

well (Supplementary Table S6, Supplementary Figure S7, 10 distinct UniRef50 clusters) and overall attributed to the same three Clostridia taxa. In particular, enzymes processing the xylose—a major building block from hemicellulose—to the pentose phosphate pathway non-oxidative branch were identified and attributed to *Caldicellulosiruptor* species (Supplementary Table S6, Supplementary Figure S7, for example, xylose isomerase, step 28, and xylulokinase, step 29). This was consistent with the above-mentioned role of *Caldicellulosiruptor* species in hemicellulolyse.

Proteins associated with pyruvate metabolism and corresponding to 23 distinct UniRef50 clusters were also identified (Supplementary Table S7, Supplementary Figure S8), pyruvate being the glycolytic pathway end product. They were, however, distributed among various taxa, mainly *C. thermocellum*, *C. proteolyticus* and *Caldicellulosiruptor* species (as above), as well as *Pelotomaculum thermopropionicum* (Clostridia class). The latter is a known anaerobic thermophilic, syntrophic, propionate-oxidising bacterium (Imachi *et al.*, 2002).

Acetyl-CoA and other tricarboxylic acid cycle intermediates are further catalysed into different fermentation products. The identified proteins (Supplementary Table S8, Supplementary Figure S9) corresponded to 21 distinct UniRef50 clusters and were mainly affiliated to *Caldicellulosiruptor* species and *C. proteolyticus*, followed by *C. thermocellum* and *Thermosinus carboxydivorans*. The latter is an anaerobic thermophilic hydrogen-producing bacterium (Sokolova *et al.*, 2004). According to the identified enzymes, a variety of fermentation products could have been generated. *C. thermocellum* strains could have been mainly generating lactate, ethanol and acetate as metabolic end products (Supplementary Table S8, (11), (13), (17)). *Caldicellulosiruptor* species could have been producing mainly lactate, propanoate and acetate (Supplementary Table S8, (11), (13), (15), (16)). *C. proteolyticus* strains could have been producing formate, butanol, butanoate and acetate (Supplementary Table S8, (1), (7), (11), (12)).

However, it is difficult to draw definite conclusions concerning the nature and number of the fermentation products because of incomplete information.

Syntrophic acetate oxidation and hydrogenotropic methanogenesis are the dominant pathways for methane production
At the metabolic level, most methanogens (including representatives of the *Methanothermobacter* genus) perform hydrogenotrophic methanogenesis exclusively; only members of the order Methanosarci-nales perform acetoclastic methanogenesis or both methanogenesis pathways (Thauer *et al.*, 2008). The data generated by pyrosequencing, metaproteomic approaches and isotopic analyses together strongly
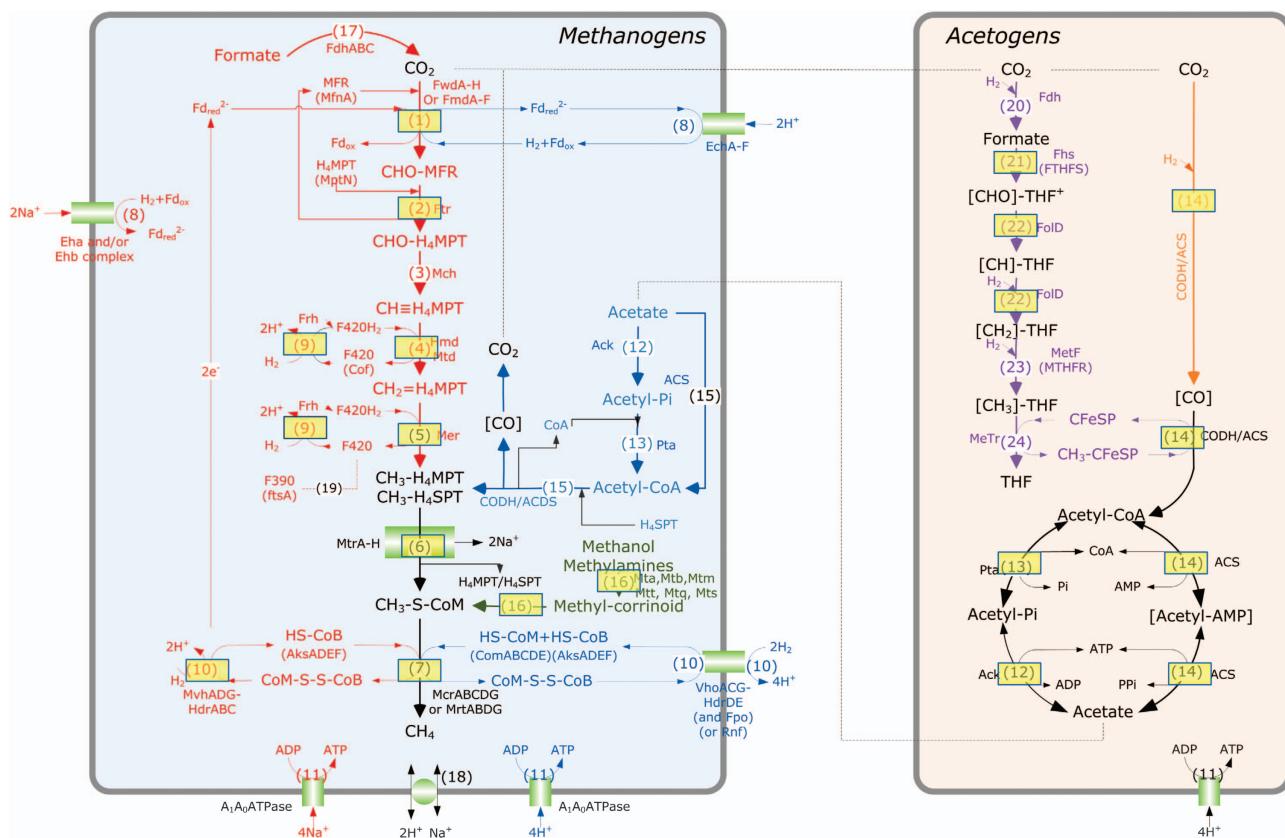
Figure 3 Identified enzymes involved in the methanogenesis and Acetyl-CoA pathways mapped over generic pathways. ▢: proteins identified during the study. Red: hydrogenotrophic methanogenesis. Blue: acetoclastic methanogenesis. Green: methylotrophic methanogenesis. Purple: Eastern branch of the acetyl-CoA pathway. Orange: Western branch of the acetyl-CoA pathway. ▢: membrane proteins. (Cofactors): key enzymes for the biosynthesis of cofactor. (1) Formylmethanofuran dehydrogenase, FwdA-F or FmdA-F; (2) Formylmethanofuran—tetrahydromethanopterin N-formyltransferase, Ftr; (3) Methenyltetrahydromethanopterin cyclohydrolase, Mch; (4) Methylenetetrahydromethanopterin dehydrogenase, Hmd/mtd; (5) Coenzyme F420-dependent N5,N10-methenyltetrahydromethanopterin reductase, Mer; (6) Tetrahydromethanopterin S-methyltransferase, Mtr; (7) Methyl-coenzyme M reductase, MrtABG or McrABG; (8) Membrane hydrogenase, Ech, Eha or Ehb, Rnf, Fpo; (9) Coenzyme F420 hydrogenase, Frh, Fru, Frc; (10) F420 non-reducing hydrogenase/heterodisulfide reductase complex, MvhADG-HdrABC or VhoACG-HdrDE; (11) ATPase, AhaA-K; (12) Acetate kinase, Ack; (13) Phosphotransacetylase, Pta; (14) Carbon monoxide dehydrogenase/acetyl-coA synthase complex, CODH/ACS; (15) Carbon monoxide dehydrogenase/acetyl-coA synthase/decarbonlyase (CODH/ACDS) complex; (16) Methylcobamide:CoM metyltransferase, MtaABC, Mtb, Mtm, Mtq, Mts or Mtt; (17) Formate dehydrogenase, Fdh; (18) Na + /H + antiporter; (19) Coenzyme F390; (20) Formyltetrahydrofolate (CHO-THF) synthetase, Fhs or FTHFS; (21) Biofunctional protein—Methenyltetrahydrofolate (CH-THF) cyclohydrolase and methylenetetrahydrofolate (CH2-THF) dehydrogenase complex, FolD; (22) Methylenetetrahydrofolate (CH2-THF) reductase, MetF or MTHFR; (23) THF:Fe-S-Co Methyltransferase, MeTr. See also Supplementary Table S7.

supported the production of methane through the hydrogenotrophic pathway by strains of the genus *Methanothermobacter*. Consistent with these results, thermophilic conditions are known to favour the hydrogenotrophic methanogenesis pathway (Schink, 1997; Hattori, 2008), but this is not systematically the case (for example, Hanreich *et al.*, 2012). More precisely, among the 21 non-redundant protein groups identified as enzyme subunits required for hydrogenotrophic methanogenesis (Figure 3, Methanogens, steps 1–10 framed by a rectangle; Supplementary Table S9), three were specific for the hydrogenotrophic pathway, catalysing the reduction of CoM–S–S–CoB by H_2 (Figure 3, Methanogens, hdrA, hdrC and mvhD genes, step 10, red). Moreover, none of the enzymes specific for acetoclastic methanogenesis were identified (Figure 3, Methanogens, steps 12, 13 and

15, blue). The identified proteins related to methanogenesis were mainly affiliated to *Methanothermobacter* members (Supplementary Table S9), consistent with a hydrogenotrophic pathway and with the pyrosequencing analyses (see above). Finally, the dominance of the hydrogenotrophic pathway was also supported by the apparent isotopic fractionation α_C values determined over time (Figure 4, Supplementary Figure S3) (Conrad, 2005).

As acetate first accumulated during the degradation (Supplementary Figure S2) and was then consumed during the phase of biogas isotopic enrichment, it was reasonable to assume that acetate was transformed into H_2 and CO_2 by syntrophic acetate oxidation (SAO). The presence of syntrophic hydrogen suppliers in the community was therefore questioned (reviewed in Schink, 1997).
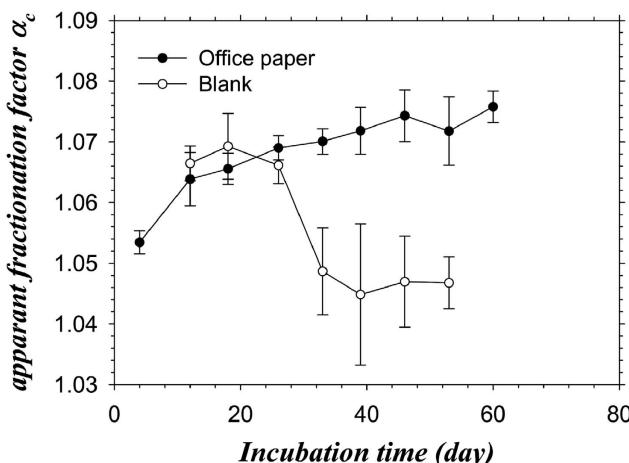
Figure 4 Temporal evolution of the apparent fractionation factor α_C calculated from the ^{13}C isotopic signal of methane (δCH_4) and CO_2 (δCO_2) in the biogas. For substrates of natural isotopic composition, such as office paper in the present case, α_C values > 1.065 are associated with hydrogenotrophic methanogenesis (Conrad, 2005). Measures were performed on the five replicates and on two controls. The mean values and s.d. are shown.

Metaproteomic together with the pyrosequencing data brought some interesting insights, but definitive conclusions could not be drawn owing to the lack of known protein specific for the SAO pathways and to the still limited knowledge about thermophilic SAO microorganisms compared with mesophilic ones (reviewed in Hattori, 2008; Westerholm *et al.*, 2011; and other publications from Schnürer's group).

In the present case, the majority of the key bacterial enzymes involved in the Acetyl-CoA pathway were identified (Figure 3, Acetogens, steps 12–14), suggesting the presence of acetogenic bacteria. They represented eight non-redundant protein groups with an average of four peptides per group and 12% protein coverage (Supplementary Table S9). They included subunits from the carbon monoxide dehydrogenase/acetyl-coA synthase/decarbonlyase complex (Supplementary Table S9, step 14) assigned to *Moorella thermoacetica* (acsC gene), *Carboxydothermus hydrogenoformans* and *Thermodesulfatator indicus* (cooSI gene). Four formyltetrahydrofolate synthetase proteins (regarded as the key and specific enzymes for the oxidative Acetyl-CoA pathway, also named Eastern branch of the Wood-Ljungdahl pathway) were retrieved and assigned to *Caldicellulosiruptor lactoaceticus* (two of them), *Desulfatibacillum alkenivorans* and *C. proteolyticus*, respectively. Finally, a serine hydroxymethyltransferase was assigned to *C. proteolyticus* (Supplementary Table S9, step 22).

Comparing these results to the pyrosequencing data, *C. proteolyticus* and *Gelria* bacteria can be proposed as important H_2 producers in the system and might have established efficient syntrophy with *Methanothermobacter* archaea (Plugge *et al.*, 2002; Sasaki *et al.*, 2011). However, according to the

literature (Plugge *et al.*, 2002; Sasaki *et al.*, 2011) and to their protein expression profile, they corresponded to typical fermenting microorganisms rather than to obligate SAO. Based on the pyrosequencing data, other good candidates for SAO could be other members of the order Thermoanaerobacterales (Figure 2b). For instance, the family Thermoanaerobacteraceae comprises thermophilic and strictly anaerobic acetogenic species such as *Thermoacetogenium phaeum* (SAO, Hattori, 2008), *M. thermoacetica* (homoacetogen, Pierce *et al.*, 2008) and *C. hydrogenoformans* (hydrogenogenic bacterium producing the carbon monoxide dehydrogenase/acetyl-coA synthase/decarbonlyase complex with the corresponding acs operon closely related to that of *M. thermoacetica*; Wu *et al.*, 2005). The moderate proportion of pyrosequencing sequences attributed to the Thermoanaerobacterales order ($\sim 0.4\%$ without *Gelria*) seems similar to proportions observed for SAO bacteria in other systems (for example, Wersterholm *et al.*, 2011). Strains belonging to the Clostridiales order may also have been performing SAO.

In conclusion, the microorganisms acting as SAO in syntrophy with the hydrogenotrophic methanogens could not be clearly identified, but the data indicated the presence of several strains closely related to known homoacetogens or SAO microorganisms. As presented above, the limited number of identified proteins related to SAO candidates was probably linked to the lack of closely related entries in the protein database.

Nitrogen metabolism and recycling: the abundance of proteolytic strains from the species C. proteolyticus
Among the numerous identified proteins putatively involved in ammonia assimilation and amino acid biosynthesis (Supplementary Table S10) and in peptidase activities (Supplementary Table S11), 22 were surprisingly affiliated to *C. proteolyticus* (11 in each of Supplementary Tables S10 and S11), indicating that *C. proteolyticus* strains could have been exerting an intensive proteolytic activity during the methanisation (Cai *et al.*, 2011).

The extracellular protease activity from *C. proteolyticus* strains was first supported by the identification of a putative extracellular cell wall-attached protease (B5Y6Q5, MEROPS S8A). Based on MEROPS information (Rawlings *et al.*, 2012), this protease could have a role in nutrition and it could have been abundantly produced because 13 peptides were detected by MS/MS analyses, the highest level among identified peptidases (Supplementary Table S10). In addition, a putative peptidase T from *C. proteolyticus* was possibly involved in general protein turnover (B5Y9V8, MEROPS M20). Very interestingly, three of the putative peptidases identified from *C. proteolyticus* were potentially involved in microcin synthesis (bacterial toxin composed of few peptides) (Duquesne *et al.*, 2007;

Cai *et al.*, 2011). Indeed, they belonged to the three distinct UniRef50 clusters UniRef50_B5Y6N5 Tltd/PmbA, UniRef50_B5Y6N6 LmbIH and UniRef50_B5Y9Y2 TldD, and to the MEROPS family U62. This suggested that *C. proteolyticus* could have been actively predating other microorganisms in addition to simply scavenging extracellular proteinaceous material.

This proteolytic activity was further supported by other retrieved protein functions. In particular, seven ABC transporter clusters were attributed to *C. proteolyticus* including three related to peptide transport (B5Y898, B56YV2-B5Y6V3 and B5Y897), suggesting an important peptide import/export activity (Supplementary Table S12). Several proteins from *C. proteolyticus* related to oxidative stress response or to virulence factor were also identified, but their role remains unclear.

Finally, the important activity of *C. proteolyticus* members in the present anaerobic system was generally supported by the present body of data. One-hundred and four non-redundant protein groups were affiliated to *C. proteolyticus*

(Figure 1c, Supplementary Table S18), among which 53 belonged to UniRef50 clusters specific to the species. About 30% of the 16S rRNA gene pyrotags were attributed to the species (Figure 2b), and a significant fraction of the bacterial community was hybridised with a probe targeting *Corprothermobacter* species (Supplementary Figure S11).

The proteolytic activity of *Coprothermobacter* members is clearly supported by the literature, and such strains were detected a number of times during anaerobic digestion of protein-rich waste (for example, Ollivier *et al.*, 1985; Kersters *et al.*, 1994; Etchebehere *et al.*, 1998; Sasaki *et al.*, 2007, 2011). It was furthermore reported that their ability to ferment carbohydrates was far inferior compared with that to ferment proteins (Ollivier *et al.*, 1985; Kersters *et al.*, 1994; Etchebehere *et al.*, 1998). Their abundance in a system fed with paper as the sole exogenous substrate is surprising and suggests that extracellular proteinaceous material was abundant. Several distinct and non-exclusive protein sources can be suggested: the protein constituents of the EPS (up to 40% based on Liu
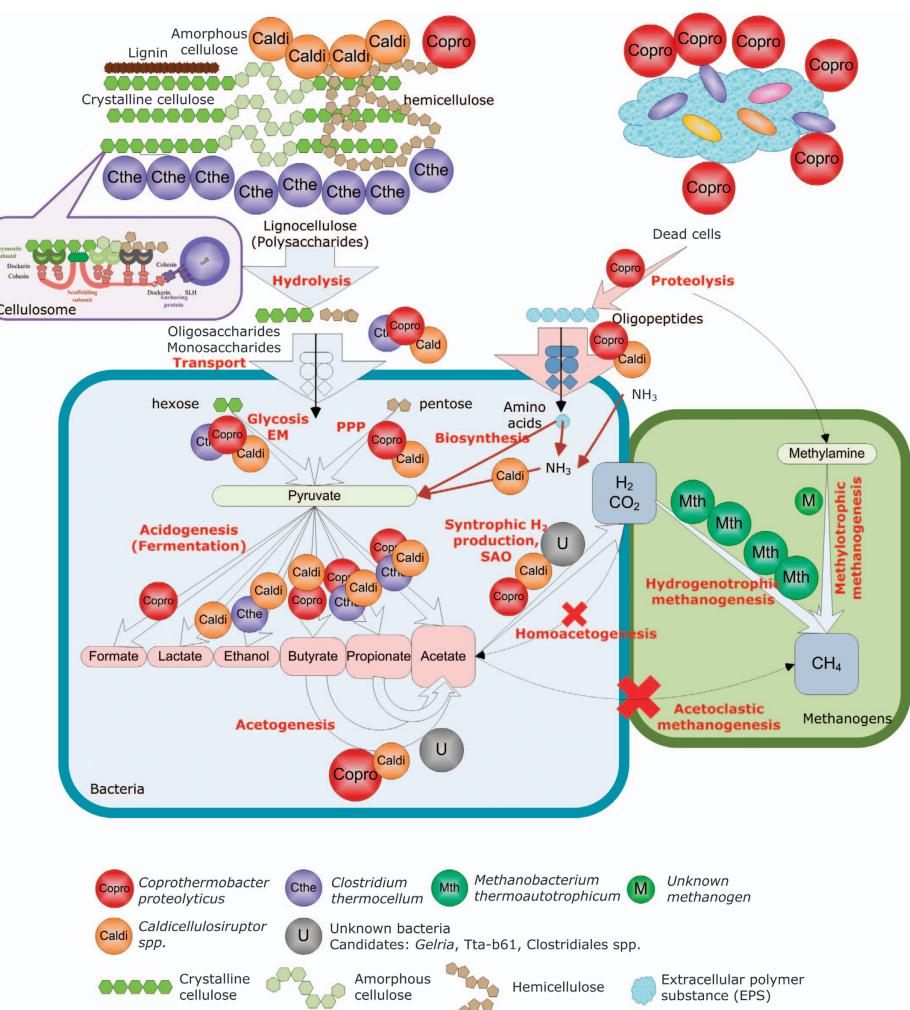
Figure 5 Functional model for the anaerobic digestion of lignocellulose by complex thermophilic communities.

and Fang, 2002); the abundant extracellular enzymes, in particular cellulosomes/cellulases or other proteins secreted by *C. thermocellum* (Ellis *et al.*, 2012) and other cellulolytic strains; dead cell material present in the initial inoculum; and finally dead cell material generated during the incubation by the counter-selection of poorly competitive strains, by various stresses or by predation. Possible predators could be *C. proteolyticus* members, as discussed above, as well as viruses, whose presence is compatible with the four identified CRISPR-associated protein groups (Supplementary Table S15) (Bhaya *et al.*, 2011).

Concluding remarks

Combining metaproteomic analyses and isotopic, chemical and molecular approaches, the present study provides one of the most comprehensive views of expressed biological functions during cellulosic waste methanisation, complementing the information gained from previous metagenomic or metatranscriptomic studies (for example, Krause *et al.*, 2008; Schluter *et al.*, 2008; Jaenicke *et al.*, 2011; Rademacher *et al.*, 2012; Zarkzewski *et al.*, 2012). Using the complete UniprotKB database, over 500 protein functions were identified. The novel information gained on the microbial catalysts highlights the importance of ecological interactions between microbial groups, especially cooperation, for efficient methanisation. Based on the results, a model for the studied ecosystem is suggested (Figure 5). Its main features are as follows: the complementarity of distinct cellulolytic microbial groups to hydrolyse the recalcitrant substrate; the absence of acetoclastic methanogenesis despite the abundance of acetate as fermentation product; the dominance of hydrogenotrophic methanogenesis in association with syntrophic microorganisms; and finally the abundance of proteolytic fermentative *Coprothermobacter* strains. For a given microbial group, the identification level strongly relied on the availability of genome sequences from closely related strains. Consequently, the picture is still incomplete, and additional information could probably be gained concerning the function of *Gelria* and Tta-b61 with a more specific database that could typically be obtained by complementary metagenomic approaches.

To our knowledge, the abundant proteolytic activity has never been reported for similar polysaccharide-fed bioprocesses. Further investigation of proteolytic activity in large-scale methanisation plants treating lignocellulosic waste could help to better understand the carbon and nitrogen fluxes during such processes. The probable complementarity of distinct cellulolytic strains is also an important aspect, and whether it is a general feature in similar systems remains an open question.

Metaproteomics appears as an attractive tool for providing direct and cost-limited access to functional information (reviewed in Wilmes and Bond, 2009; Schneider and Riedel, 2010). The exponential increase in sequence database size should reinforce the attractiveness of this approach in the future.

Conflict of Interest

The authors declare no conflict of interest.

Acknowledgements

References

Abram F, Enright AM, O'Reilly J, Botting CH, Collins G, O'Flaherty V. (2011). A metaproteomic approach gives functional insights into anaerobic digestion. *J Appl Microbiol* **110**: 1550–1560.

Bayer EA, Lamed R, White BA, Flint HJ. (2008). From cellulosomes to cellulosomics. *Chem Rec* **8**: 364–377.

Beganovic J, Guillot A, van de Guchte M, Jouan A, Gitton C, Loux V *et al.* (2010). Characterization of the insoluble proteome of *Lactococcus lactis* by SDS-PAGE LC-MS/MS leads to the identification of new markers of adaptation of the bacteria to the mouse digestive tract. *J Proteome Res* **9**: 677–688.

Bhaya D, Davison M, Barrangou R. (2011). CRISPR-Cas systems in bacteria and archaea: versatile small RNAs for adaptive defense and regulation. *Annu Rev Genet* **45**: 273–297.

Cai H, Gu J, Wang Y. (2011). Protease complement of the thermophilic bacterium *Coprothermobacter proteolyticus*. In: Arabina HR, Tran Q-N (eds) *Proceeding of the International Conference on Bioinformatics and Computational Biology BIOCOMP'11*. ISBN: Las Vegas, Nevada, USA, pp 18–21.

Cantarel BL, Coutinho PM, Rancurel C, Bernard T, Lombard V, Henrissat B. (2009). The carbohydrate-active enzymes database (CAZy): an expert resource for glycogenomics. *Nucleic Acids Res* **37**: D233–D238.

Chen Y, Cheng JJ, Creamer KS. (2008). Inhibition of anaerobic digestion process: a review. *Biores Technol* **99**: 4044–4064.

Chouari R, Le Paslier D, Daegelen P, Ginestet P, Weissenbach J, Sghir A. (2005). Novel predominant archaeal and bacterial groups revealed by molecular analysis of an anaerobic sludge digester. *Environ Microbiol* **7**: 1104–1115.

Conrad R. (2005). Quantification of methanogenic pathways using stable carbon isotopic signatures: a review and a proposal. *Org Geochem* **36**: 739–752.

Denef VJ, Shah MB, VerBerkmoes NC, Hettich RL, Banfield JF. (2007). Implications of strain- and species-level sequence divergence for community and isolate shotgun proteomic analysis. *J Proteome Res* **6**: 3152–3161.

Duquesne S, Destoumieux-Garzon D, Peduzzi J, Rebuffat S. (2007). Microcins, gene-encoded antibacterial peptides from enterobacteria. *Nat Prod Rep* **24**: 708–734.

Ellis LD, Holwerda EK, Hogsett D, Rogers S, Shao X, Tschaplinski T et al. (2012). Closing the carbon balance for fermentation by *Clostridium thermocellum* (ATCC 27405). *Biores Technol* **103**: 293–299.

Etchebehere C, Pavan ME, Zorzopulos J, Soubes M, Muxi L. (1998). *Coprothermobacter platensis* sp. nov., a new anaerobic proteolytic thermophilic bacterium isolated from an anaerobic mesophilic sludge. *Int J Syst Bacteriol* **48**: 1297–1304.

Ganidi N, Tyrrel S, Cartmell E. (2009). Anaerobic digestion foaming causes—a review. *Biores Technol* **100**: 5546–5554.

Gold ND, Martin VJJ. (2007). Global view of the *Clostridium thermocellum* cellulosome revealed by quantitative proteomic analysis. *J Bacteriol* **189**: 6787–6795.

Hamilton-Brehm SD, Mosher JJ, Vishnivetskaya T, Podar M, Carroll S, Allman S et al. (2010). *Caldicellulosiruptor obsidiansis* sp. nov., an anaerobic, extremely thermophilic, cellulolytic bacterium isolated from Obsidian Pool, Yellowstone National Park. *Appl Environ Microbiol* **76**: 1014–1020.

Hanreich A, Heyer R, Benndorf D, Rapp E, Pioch M, Reichl U et al. (2012). Metaproteome analysis to determine the metabolically active part of a thermophilic microbial community producing biogas from agricultural biomass. *Can J Microbiol* **58**: 917–922.

Hattori S. (2008). Syntrophic acetate-oxidizing microbes in methanogenic environments. *Microbes Environ* **23**: 118–127.

Holm-Nielsen JB, Al Seadi T, Oleskowicz-Popiel P. (2009). The future of anaerobic digestion and biogas utilization. *Biores Technol* **100**: 5478–5484.

Imachi H, Sekiguchi Y, Kamagata Y, Hanada S, Ohashi A, Harada H. (2002). *Pelotomaculum thermopropionicum* gen. nov., sp. nov., an anaerobic, thermophilic, syntrophic propionate-oxidizing bacterium. *Int J Syst Evol Microbiol* **52**: 1729–1735.

Jaenicke S, Ander C, Bekel T, Bisdorf R, Dröge M, Gartemann KH et al. (2011). Comparative and joint analysis of two metagenomic datasets from a biogas fermenter obtained by 454-pyrosequencing. *PLoS One* **6**: e14519.

Kersters I, Maestrojuan GM, Torck U, Vancanneyt M, Kersters K, Verstraete W. (1994). Isolation of *Coprothermobacter proteolyticus* from an anaerobic digest and further characterization of the species. *Sys Appl Microbiol* **17**: 289–295.

Konstantinidis KT, Tiedje JM. (2007). Prokaryotic taxonomy and phylogeny in the genomic era: advancements and challenges ahead. *Curr Opin Microbiol* **10**: 504–509.

Krause L, Diaz NN, Edwards RA, Gartemann K-H, Krömeke H, Neuwger H et al. (2008). Taxonomic composition and gene content of a methane producing microbial community isolated from a biogas reactor. *J Biotechnol* **136**: 91–101.

Kröber M, Bekel T, Diaz NN, Goesmann A, Jaenicke S, Krause L et al. (2009). Phylogenetic characterization of a biogas plant microbial community integrating clone library 16S-rDNA sequences and metagenome sequence data obtained by 454-pyrosequencing. *J Biotechnol* **142**: 38–49.

Li T, Mazeas L, Sghir A, Leblon G, Bouchez T. (2009). Insights into networks of functional microbes catalysing methanization of cellulose under mesophilic conditions. *Environ Microbiol* **11**: 889–904.

Liu H, Fang HH. (2002). Extraction of extracellular polymeric substances (EPS) of sludges. *J Biotechnol* **95**: 249–256.

Lochner A, Giannone RJ, Keller M, Antranikian G, Graham DE, Hettich RL. (2011). Label-free quantitative proteomics for the extremely thermophilic bacterium *Caldicellulosiruptor obsidiansis* reveal distinct abundance patterns upon growth on cellobiose, crystalline cellulose, and switchgrass. *J Proteome Res* **10**: 5302–5314.

Maki M, Leung KT, Qin W. (2009). The prospects of cellulase-producing bacteria for the bioconversion of lignocellulosic biomass. *Int J Biol Sci* **5**: 500–516.

Ollivier BM, Mah RA, Ferguson TJ, Boone DR, Garcia JL, Robinson R. (1985). Emendation of the genus Thermobacteroides—Thermobacteroides-Proteolyticus sp-nov, a proteolytic acetogen from a methanogenic enrichment. *Int J Syst Bacteriol* **35**: 425–428.

O'Sullivan C, Burrell PC, Clarke WP, Blackall LL. (2007). A survey of the relative abundance of specific groups of cellulose degrading bacteria in anaerobic environments using fluorescence in situ hybridization. *J Appl Microbiol* **103**: 1332–1343.

Petersen TN, Brunak S, von Heijne G, Nielsen H. (2011). SignalP 4.0: discriminating signal peptides from transmembrane regions. *Nat Methods* **8**: 785–786.

Pierce E, Xie G, Barabote RD, Saunders E, Han CS, Detter JC et al. (2008). The complete genome sequence of *Moorella thermoacetica* (f. *Clostridium thermoaceticum*). *Environ Microbiol* **10**: 2550–2573.

Plugge CM, Balk M, Zoetendal EG, Stams AJM. (2002). *Gelria glutamica* gen. nov., sp. nov., a thermophilic, obligately syntrophic, glutamate-degrading anaerobe. *Int J Syst Evol Microbiol* **52**: 401–407.

Qu X, Mazéas L, Vavilin VA, Epissard J, Lemunier M, Mouchel J-M et al. (2009). Combined monitoring of changes in delta13CH4 and archaeal community structure during mesophilic methanization of municipal solid waste. *FEMS Microbiol Ecol* **68**: 236–245.

Rademacher A, Zakrzewski M, Schlüter A, Schönberg M, Szczepanowski R, Goesmann A et al. (2012). Characterization of microbial biofilms in a thermophilic biogas system by high-throughput metagenome sequencing. *FEMS Microbiol Ecol* **79**: 785–799.

Rawlings ND, Barrett AJ, Bateman A. (2012). MEROPS: the database of proteolytic enzymes, their substrates and inhibitors. *Nucleic Acids Res* **40**: D343–D350.

Riviere D, Desvignes V, Pelletier E, Chaussonnerie S, Guermazi S, Weissenbach J et al. (2009). Towards the definition of a core of microorganisms involved in anaerobic digestion of sludge. *ISME J* **3**: 700–714.

Sasaki K, Haruta S, Ueno Y, Ishii M, Igarashi Y. (2007). Microbial population in the biomass adhering to supporting material in a packed-bed reactor degrading organic solid waste. *Appl Microbiol Biotechnol* **75**: 941–952.

Sasaki K, Morita M, Sasaki D, Nagaoka J, Matsumoto N, Ohmura N *et al.* (2011). Syntrophic degradation of proteinaceous materials by the thermophilic strains *Coprothermobacter proteolyticus* and *Methanothermobacter thermautotrophicus*. *J Biosci Bioeng* **112**: 469–472.

Schink B. (1997). Energetics of syntrophic cooperation in methanogenic degradation. *Microbiol Mol Biol Rev* **61**: 262–280.

Schlüter A, Bekel T, Diaz NN, Dondrup M, Eichenlaub R, Gartemann KH *et al.* (2008). The metagenome of a biogas-producing microbial community of a production-scale biogas plant fermenter analysed by the 454-pyrosequencing technology. *J Biotechnol* **136**: 77–90.

Schneider T, Riedel K. (2010). Environmental proteomics: analysis of structure and function of microbial communities. *Proteomics* **10**: 785–798.

Sokolova TG, Gonzalez JM, Kostrikina NA, Chernyh NA, Slepova TV, Bonch-Osmolovskaya EA *et al.* (2004). *Thermosinus carboxydivorans* gen. nov., sp. nov., a new anaerobic, thermophilic, carbon-monoxide-oxidizing, hydrogenogenic bacterium from a hot pool of Yellowstone National Park. *Int J Syst Evol Microbiol* **54**: 2353–2359.

Thauer RK, Kaster A-K, Seedorf H, Buckel W, Hedderich R. (2008). Methanogenic archaea: ecologically relevant differences in energy conservation. *Nat Rev Micro* **6**: 579–591.

van de Werken HJG, Verhaart MRA, VanFossen AL, Willquist K, Lewis DL, Nichols JD *et al.* (2008). Hydrogenomics of the extremely thermophilic bacterium *Caldicellulosiruptor saccharolyticus*. *Appl Environ Microbiol* **74**: 6720–6729.

van Lier JB, Tilche A, Ahring BK, Macarie H, Moletta R, Dohanyos M *et al.* (2001). New perspectives in anaerobic digestion. *Water Sci Technol* **43**: 1–18.

VanFossen AL, Verhaart MRA, Kengen SMW, Kelly RM. (2009). Carbohydrate utilization patterns for the extremely thermophilic bacterium *Caldicellulosiruptor saccharolyticus* reveal broad growth substrate preferences. *Appl Environ Microbiol* **75**: 7718–7724.

Ward AJ, Hobbs PJ, Holliman PJ, Jones DL. (2008). Optimisation of the anaerobic digestion of agricultural resources. *Biores Technol* **99**: 7928–7940.

Westerholm M, Dolfing J, Sherry A, Gray ND, Head IM, Schnürer A. (2011). Quantification of syntrophic acetate-oxidising microbial communities in biogas processes. *Environ Microbiol Rep* **3**: 500–505.

Wilmes P, Bond PL. (2004). The application of two-dimensional polyacrylamide gel electrophoresis and downstream analyses to a mixed community of prokaryotic microorganisms. *Environ Microbiol* **6**: 911–920.

Wilmes P, Bond PL. (2009). Microbial community proteomics: elucidating the catalysts and metabolic mechanisms that drive the Earth's biogeochemical cycles. *Curr Opin Microbiol* **12**: 310–317.

Wu M, Ren Q, Durkin AS, Daugherty SC, Brinkac LM, Dodson RJ *et al.* (2005). Life in hot carbon monoxide: the complete genome sequence of *Carboxydothermus hydrogenoformans* Z-2901. *PLoS Genet* **1**: e65.

Yang SJ, Kataeva I, Hamilton-Brehm SD, Engle NL, Tschaplinski TJ, Doeppke C *et al.* (2009). Efficient degradation of lignocellulosic plant biomass, without pretreatment, by the thermophilic anaerobe 'Anaerocellum thermophilum' DSM 6725. *Appl Environ Microbiol* **75**: 4762–4769.

Zakrzewski M, Goesmann A, Jaenicke S, Jünemann S, Eikmeyer F, Szczepanowski R *et al.* (2012). Profiling of the metabolically active community from a production-scale biogas plant by means of high-throughput metatranscriptome sequencing. *J Biotechnol* **158**: 248–258.

Zaneveld JR, Lozupone C, Gordon JI, Knight R. (2010). Ribosomal RNA diversity predicts genome diversity in gut bacteria and their relatives. *Nucleic Acids Res* **38**: 3869–3879.

Supplementary Information accompanies this paper on the ISME Journal website (http://www.nature.com/ismej)

Whole Proteome Analyses on *Ruminiclostridium cellulolyticum* Show a Modulation of the Cellulolysis Machinery in Response to Cellulosic Materials with Subtle Differences in Chemical and Structural Properties

Nelly Badalato[1], **Alain Guillot**[2], **Victor Sabarly**[3], **Marc Dubois**[3], **Nina Pourette**[1], **Bruno Pontoire**[4], **Paul Robert**[4], **Arnaud Bridier**[1], **Véronique Monnet**[2], **Diana Z. Sousa**[5,6], **Sylvie Durand**[4], **Laurent Mazéas**[1], **Alain Buléon**[4], **Théodore Bouchez**[1], **Gérard Mortha**[7], **Ariane Bize**[1] *

1 UR HBAN, Irstea, Antony, France, 2 UMR 1319 MICALIS, PAPPSO, INRA, Jouy-en-Josas, France, 3 Omics Services, Paris, France, 4 UR 1268 BIA, INRA, Nantes, France, 5 Centre of Biological Engineering, University of Minho, Braga, Portugal, 6 Laboratory of Microbiology, Wageningen University, Wageningen, The Netherlands, 7 LGP2, UMR CNRS 5518, Grenoble INP-Pagora, Saint Martin d'Hères, France

* ariane.bize@irstea.fr

Abstract

Lignocellulosic materials from municipal solid waste emerge as attractive resources for anaerobic digestion biorefinery. To increase the knowledge required for establishing efficient bioprocesses, dynamics of batch fermentation by the cellulolytic bacterium *Ruminiclostridium cellulolyticum* were compared using three cellulosic materials, paper handkerchief, cotton discs and Whatman filter paper. Fermentation of paper handkerchief occurred the fastest and resulted in a specific metabolic profile: it resulted in the lowest acetate-to-lactate and acetate-to-ethanol ratios. By shotgun proteomic analyses of paper handkerchief and Whatman paper incubations, 151 proteins with significantly different levels were detected, including 20 of the 65 cellulosomal components, 8 non-cellulosomal CAZymes and 44 distinct extracytoplasmic proteins. Consistent with the specific metabolic profile observed, many enzymes from the central carbon catabolic pathways had higher levels in paper handkerchief incubations. Among the quantified CAZymes and cellulosomal components, 10 endoglucanases mainly from the GH9 families and 7 other cellulosomal subunits had lower levels in paper handkerchief incubations. An in-depth characterization of the materials used showed that the lower levels of endoglucanases in paper handkerchief incubations could hypothetically result from its lower crystallinity index (50%) and degree of polymerization (970). By contrast, the higher hemicellulose rate in paper handkerchief (13.87%) did not result in the enhanced expression of enzyme with xylanase as primary activity, including enzymes from the "*xyl-doc*" cluster. It suggests the absence, in this material, of molecular structures that specifically lead to xylanase induction. The integrated approach developed in this work shows that subtle differences among cellulosic materials regarding

chemical and structural characteristics have significant effects on expressed bacterial functions, in particular the cellulolysis machinery, resulting in different metabolic patterns and degradation dynamics.

Introduction

Conversion of cellulose during the degradation of biomass residues and agricultural waste and products has been extensively studied in the context of biofuel production [1–3]. Other sources of lignocellulosic materials, such as waste, are currently emerging as attractive options for biorefinery based on anaerobic digestion [4–6]. The cellulosic fraction in municipal solid waste (MSW) accounts for up to 50% weight in developed countries. Using this resource for biofuel or synthon production can potentially cuts down emissions of greenhouse gases while improving resource efficiency. In contrast with native biomass, this lignocellulosic fraction mainly contains diverse manufactured products with heterogeneous properties such as sanitary textiles, papers, or cardboards.

The optimal strategies to efficiently recover energy or added-value molecules from these specific waste materials are not fully established yet. Approaches relying on fermentation by pure strain cultures, similar to current bioprocesses for bioethanol or biofuel production, could be considered, such as various processes based on fermentation by yeasts [6] or consolidated bioprocessing with various microorganisms such as the bacterium *Lachnoclostridium phytofermentans* or the fungus *Trichoderma reesei* [7, 8]). Alternatively, bioprocesses based on the action of complex microbial communities, such as those classically used for organic waste treatment and valorization (e.g. methanization) could also be invaluable options [5, 9].

Characterization and understanding of the fermentation process of lignocellulosic manufactured materials are needed to establish the scientific bases required for the development of bioprocesses efficiently exploiting their potential. In this respect, a limited number of such studies have been published so far [9–12]. The present work focuses on three cellulosic materials containing no lignin, cotton discs, paper handkerchief and Whatman filter paper, which will be referred as "Cotton", "Tissue" and "Whatman paper", respectively. These substrates are rather homogeneous compared to the variety of lignocellulosic waste materials and their bioconversion has been only little studied so far [9–12].

To characterize their anaerobic fermentation dynamics and mechanisms in simple model conditions, *Ruminiclostridium cellulolyticum*, formerly known as *Clostridium cellulolyticum*, was selected as model species. Although the wild-type strain does not produce high concentrations of ethanol nor other biofuel or platform molecules (acetate being the main metabolic end-product), the species is currently considered as a model organism for consolidated bioprocessing through metabolic engineering as exemplified by a recently engineered strain producing isobutanol directly from cellulose [13]. Moreover, closely related members of *R. cellulolyticum* have been detected in anaerobic digesters treating waste with high cellulose content [14] and the species has recently been shown to improve wheat straw methanization by bioaugmentation [15]. Finally, the wild-type *R. cellulolyticum* bacterium is an important biological model of mesophilic anaerobic cellulolytic bacterium, so that a robust knowledge framework is available for data interpretation including knowledge on the cellulolysis machinery and on its metabolism upon growth on cellulose and its derivatives [16–18]. In particular, detailed studies of its metabolism upon growth on cellobiose *versus* cellulose showed key metabolic nodes in the central metabolic pathways [18, 19]. Its cellulolysis machinery relies both on

cellulosomal proteins and non-cellulosomal secreted enzymes [20]. Cellulosomes are complex extracellular muti-enzyme machineries produced by numerous cellulolytic microorganisms. The modulation of *R. cellulolyticum* cellulosome composition at the protein level according to the carbohydrate growth substrate has been described in details by targeted approaches [16, 17]. Recently, a transcriptomic and proteomic study of *R. cellulolyticum* grown on a variety of substrates (glucose, xylose, cellobiose, cellulose, xylan or corn stover) showed that core cellulases are regulated by carbon catabolite repression, while most of the accessory CAZymes and their associated transporters are regulated by the Two-Component Systems [21].

To achieve a global insight into the bioconversion dynamics and mechanisms of the three studied cellulosic materials by *R. cellulolyticum*, a combined approach was developed. The fermentation dynamics was monitored in batch microcosms and the biodegradation mechanisms were analyzed at the protein level for Tissue and Whatman Paper through label-free quantitative shotgun proteomics. The in-depth characterization of the three materials was moreover conducted to elaborate precise hypotheses regarding the origin of the differences in metabolic end-product concentrations and in protein levels. The present work is, to our knowledge, the first global shotgun proteomic study of *R. cellulolyticum*. The obtained data provide evidence that Tissue is degraded the fastest by *R. cellulolyticum* and is associated to a distinct metabolic pattern compared to both other materials. When comparing Tissue and Whatman Paper, the data show a clear influence of the substrates, even though they are rather similar, on protein levels from the cellulolysis machinery and the central carbon metabolism. Based on the material characteristics, it is postulated that the crystallinity rate and the degree of polymerization had a preponderant influence on the cellulosome composition here, compared to the hemicellulose content.

Materials and Methods

Bacterial strain and culture conditions

R. cellulolyticum H10 ATCC 35319 (DSM 5812) was grown anaerobically at 37°C, as indicated on ATCC website (www.lgcstandards-atcc.org/Products/Cells_and_Microorganisms/Bacteria/Alphanumeric_Genus__Species/35319.aspx#culturemethod). The basal medium (initial pH 7.1) contained, per liter: Na_2HPO_4, 0.4 g; KH_2PO_4, 0.4 g; NH_4Cl, 0.3 g; NaCl, 0.3 g; $MgCl_2$, 0.1 g; $CaCl_2$, 0.1 g; $NaHCO_3$, 4.0 g; $Na_2S.9H_2O$, 0.2 g. The medium was supplemented with 0.2 mL/L of vitamin solution and 1 mL/L of acid and alkaline trace element solutions (each) [22]. 5 mg/L resazurin were added to the medium as a redox indicator. Inoculation was realized with 10% (v/v) of a pre-adapted culture grown on 1 g/L Sigmacell microcrystalline cellulose in 125 mL flasks with 50 mL working volume. Three cellulosic materials were used separately as sole carbon substrates: cotton pads ("Cotton", Leader Price, Disques à Démaquiller—Simplement, Duo face, 100% cotton), Whatman qualitative filter paper, Grade 1 ("Whatman Paper", 1001–125, 11 μm, 125 mm diameter) and paper handkerchief ("Tissue", Lotus Classic, large handkerchiefs, made from pure cellulose fibers, pure virgin pulp). These substrates were cut in bands (~1.5 cm x 5 cm) and added to each flask to a final concentration of 2.5 g/L. For each cellulosic substrate separately, triplicate flasks were dedicated to physico-chemical monitoring of the degradation dynamics. For Tissue and Whatman paper separately, six additional replicate flasks by substrate were operated and sacrificed at 2 different time points for proteomic analyses during the incubation.

Physico-chemical analyses of the incubation samples

At each sampling time, fresh samples (2 mL) were recovered from the 125 mL flasks and centrifuged at 10 000g for 10 min at 4°C. The obtained pellets and supernatants were stored

separately at -80˚C. Volatile fatty acid concentrations (including lactate and acetate) were measured using a DX 120 Ion Chromatograph (Dionex) with an IonPAc ICE-AS1 column. Ethanol concentrations were quantified by headspace gas chromatography—mass spectrometry (GC-MS) (Trace GC Ultra and DSQ II from Thermo with a TR-WAX column (30 m length, 0.25 mm intern diameter, 0.25 μm thick polyethylene glycol film).

DNA extraction and quantification

DNA was extracted from the pellets with the PowerSoil DNA Isolation Kit (MO BIO Laboratories, Inc., Carlsbad, CA, USA) according to the manufacturer's instructions and was quantified with the fluorescence-based Qubit dsDNA HS assay (Life Technologies). The obtained values were multiplied by a constant factor specific of each DNA extraction series to take into account the extraction yields. Based on *R. cellulolyticum* genome size, it was estimated that 1 ng of DNA corresponded to 227,703 genome copies.

Physico-chemical characterization of the substrates

The total solids and volatile solids in each substrate were calculated from the moisture and ash content, determined by oven-drying, following instructions described in EN ISO 12879 and EN ISO 12880. Material elemental composition was analyzed using a Vario EL III (Elementar Analysensysteme GmbH, Hanau, Germany). Chemical oxygen demand was measured on solid substrates with the LCK 514 kit (Hach Lange) according to the manufacturer's instructions.

To determine the crystallinity index, diffraction diagrams were monitored by recording X-ray diffraction diagrams every 10 min on a Bruker D8 Discover diffractometer. Cu Kα1 radiation (Cu Kα1 = 1.5405 Å), produced in a sealed tube at 40 kV and 40 mA, was selected and parallelized using a Gobël mirror parallel optics system and collimated to produce a 500 μm beam diameter. Crystallinity index was calculated based on [23], as follows:

$$Crystallinity\ index\ (\%) = \frac{\sum_{2\theta}|U - A|_{2\theta}}{\sum_{2\theta}|C - A|_{2\theta}} \times 100$$

With A the value obtained for the amorphous standard, C the value obtained for crystalline standard and U the value obtained for the sample.

To determine the molar mass distribution of the cellulosic substrates, the latter were dissolved in N,N-dimethylacetamide (DMAc) and derivatized by tri-carbanilation using phenylisocyanate as reactant (reaction time 5 days at 40˚C)[24]. The reaction was quenched with methanol and direct samples from the obtained solutions were analyzed after dilution in tetrahydrofurane (THF) in a size-exclusion chromatographic system (Viscotek TDA-302 apparatus) equipped with 3 Varian PLGel Mixed B columns (7.8×300) with a guard column. The coupled detection was UV at 260 nm, DRI, RALS/LALS/RI at 670 nm (laser 3 mW, 670 nm) and viscometer detector. DRI was used as concentration detector and UV as control. The chromatographic solvent was THF, injected concentrations were 1 mg/mL (injection volume 100 microliter), and the dn/dC of cellulose tricarbanilate in THF was taken at a predetermined value of 0.165. Data were treated by the OmniSec™ (4.5.6. version) program (Malvern Co.).

The lignocellulose sugar content of the samples was determined by the commercial facility of Celignis Analytical (http://www.celignis.com, Analysis Package P7, substrate hydrolysis followed by ion chromatography). Van Soest fractionation of the substrates was performed by the commercial facility of INRA Transfert Environnement as in [25] (https://www6.montpellier.inra.fr/it-e).

Proteomic analyses

Total proteome and exoproteome extraction and shotgun MS/MS analyses. After 46 h and 70 h of incubation, 3 flasks for each substrate were sacrificed at each time point and sampled for proteomic analyses. The total volume, except 2 mL used for chemical analyses, was supplemented with 1 mM Phenylmethylsulfonyl Fluoride to inhibit protease activity and centrifuged at 6 000g for 20 min. The collected supernatant was further centrifuged at 13 000g for 15 min. Pellets from the first centrifugation round and supernatants from the second one were frozen in liquid nitrogen immediately and stored at -80˚C. Proteins from the pellets were extracted as described in [26] with minor modifications. Briefly, cell disruption was performed by bead-beating of the pellets resuspended in 1X PBS using 0.1 mm silica and glass beads, and by a subsequent ultrasonication step. Proteins were extracted using liquid phenol and precipitated with ammonium acetate in methanol. After several washing steps, the protein pellets were resuspended in buffer (7 M urea, 2 M thiourea, 4% (w/v) CHAPS) as described in [9] and stored at -80˚C. Culture supernatants were filtered through 0.22 μm PVDF membrane filters and proteins were precipitated and resuspended as described above. Protein concentrations were measured using an Agilent 2100 Bioanalyzer with the High Sensitivity Protein 250 kit following the manufacturer's instructions.

For each sample, 5μg of each whole proteome were classically purified by SDS-PAGE, digested by trypsin in gel and analyzed by LC-MS/MS on a LTQ-Orbitrap Discovery mass spectrometer (Thermo Fisher, USA). Peptide separation was realized with an Ultimate 3000 RSLCnano system (Dionex, Voisins le Bretonneux, France) using a long gradient and a C18 column (Pepmap100, 0.075 x 50 cm, 100 Å, 2 μm, Thermo) during 188 min to enhance the resolution and the sensitivity of peptide detection by mass spectrometry. A detailed protocol is provided S1 File.

Protein database search and label free quantification. The data processing pipeline was designed using the TOPPAS software [27], part of the OpenMS project [28]. X!Tandem (directly provided in the OpenMS archive) was used to perform database searches in batch mode. Peak lists were created by an OpenMS dedicated tool with an additional processing step. Indeed, a precursor recalculation was computed for each tandem spectrum to improve the number of matches between a spectrum and a peptide sequence. While proteins were digested with trypsin, the analysis program allowed for 2 missed trypsin cleavage sites. Cysteine carbamidomethylation was set as a fixed modification; methionine oxidation and protein N-terminal acetylation were set as variable modifications for all X!Tandem searches. The mass tolerances in MS and MS/MS were set to 10 ppm and 0.6 Da respectively. Data were searched against a target/decoy concatenated database to obtain a false discovery rate (FDR) value at the peptide level. All identifications were validated with a final peptide FDR of 5% and a calculated protein FDR of 1%.

For the relative quantification based on eXtracted Ion Chromatogram (XIC), peaks were detected in each sample using the « PeakPickerCentroid » algorithm (OpenMS software). Validated identification data were matched with detected peaks. Peaks which were assigned to the same peptide sequence in different samples were used as anchors for retention time alignment between those samples. A Protein Abundance Index (PAI) was calculated and defined as the average of XIC area values from the three most intense peptides identified for a given protein.

Statistical analyses of the label-free quantification proteome datasets. All statistical analyses were performed using the programming language R on $\log_{10}$-transformed and quantile normalized PAI. The substrate, incubation time and substrate-by-incubation time interaction effects were assessed by Tobit models with a censoring threshold equal to the minimum non-zero value for a given protein. With y^*_{ijkl} the normalized $\log_{10}$ PAI of protein l, for the

substrate S_i, incubation time T_j and replicate k, the model used for each protein is the following:

$$y_{ijkl} = \left\{ \begin{array}{c} \min_{i'j'k':\ y^*_{i'j'k'l}>0} (y^*_{i'j'k'l}) \text{ if } y^*_{ijkl} = 0 \text{ and } \exists k' : y^*_{ijk'l} \neq 0 \\ \\ y^*_{ijkl} \text{ otherwise} \end{array} \right.$$

$$\text{with } y^*_{ijkl} = \mu_l + \alpha_{il}S_i + \beta_{jl}T_j + \gamma_{ijl}(S,T)_{ij} + \varepsilon_{ijkl}$$

$$\text{and } \varepsilon_{ijkl} \sim \mathcal{N}(0, \sigma_l^2)$$

Tobit regressions were computed using the Markov chain Monte Carlo algorithm from the R package MCMCpack [29]. Runs that did not pass the Heidelberger and Welch's convergence diagnostic [30, 31] were discarded. Substrate (α_{il}) and interaction (γ_{ijl}) effect coefficients as well as p-values were determined from the posterior distributions. P-values were adjusted for multiple comparisons using Benjamini and Hochberg's false discovery rate [32]. The significance threshold used corresponds to a false discovery rate of 1%.

For the supernatants, the proteins with at least one statistically significant effect were filtered to select only proteins likely to be released in the extracellular milieu, cellulosomal proteins and secreted CAZYmes (see next section). The other proteins from the supernatants (cytoplasm/cell wall) were considered as originating from cell lysis.

In silico predictions of protein sub-cellular localization

Prediction of sub-cellular localization was obtained from LocateP database (www.cmbi.ru.nl/locatep-db/cgi-bin/locatepdb.py) and by analyzing *R. cellulolyticum* protein sequences with the SurfG+ program 1.02 [33] with default parameter values with a local Galaxy instance (migale.jouy.inra.fr/galaxy/). The protein sequences were moreover analyzed with the last available version of specific tools: SignalP 4.1 server (www.cbs.dtu.dk/services/SignalP/), LipoP 1.0 server (www.cbs.dtu.dk/services/LipoP/), SecretomeP 2.0 server (www.cbs.dtu.dk/services/SecretomeP/), TMHMM server v. 2.0 (www.cbs.dtu.dk/services/TMHMM/), selecting the "Gram-positive" option whenever available. The results were manually examined and confronted to the knowledge on *R. cellulolyticum* proteins, in particular on cellulosomal proteins and cellulases [16, 21]. For ABC transporters, the subfamily name was retrieved from the Archaeal and Bacterial ABC Systems database (ABCdb database www-abcdb.biotoul.fr/) and for peptidases, the Peptidase database identity (MEROPS ID) was indicated (merops.sanger.ac.uk/).

The mass spectrometry proteomic data are available via ProteomeXchange, identifier PXD001051 and DOI 10.6019/PXD001051.

Results and Discussion
Growth and fermentation patterns

Fermentation dynamics of the three studied materials by *R. cellulolyticum* was characterized by incubating them separately in anaerobic, mesophilic conditions. For all incubations and as was expected [18, 20], acetate was the major end-product, followed by ethanol and lactate (Fig 1A–1C). Degradation occurred at a faster rate for Tissue than for both other substrates during the 94 first hours of incubation, as shown by the faster accumulation of acetate, ethanol and lactate in Tissue incubations over this time period (Fig 1A–1C, S1 Fig, panel A for the total dissolved organic carbon). A distinct metabolic profile was moreover observed for Tissue incubations,
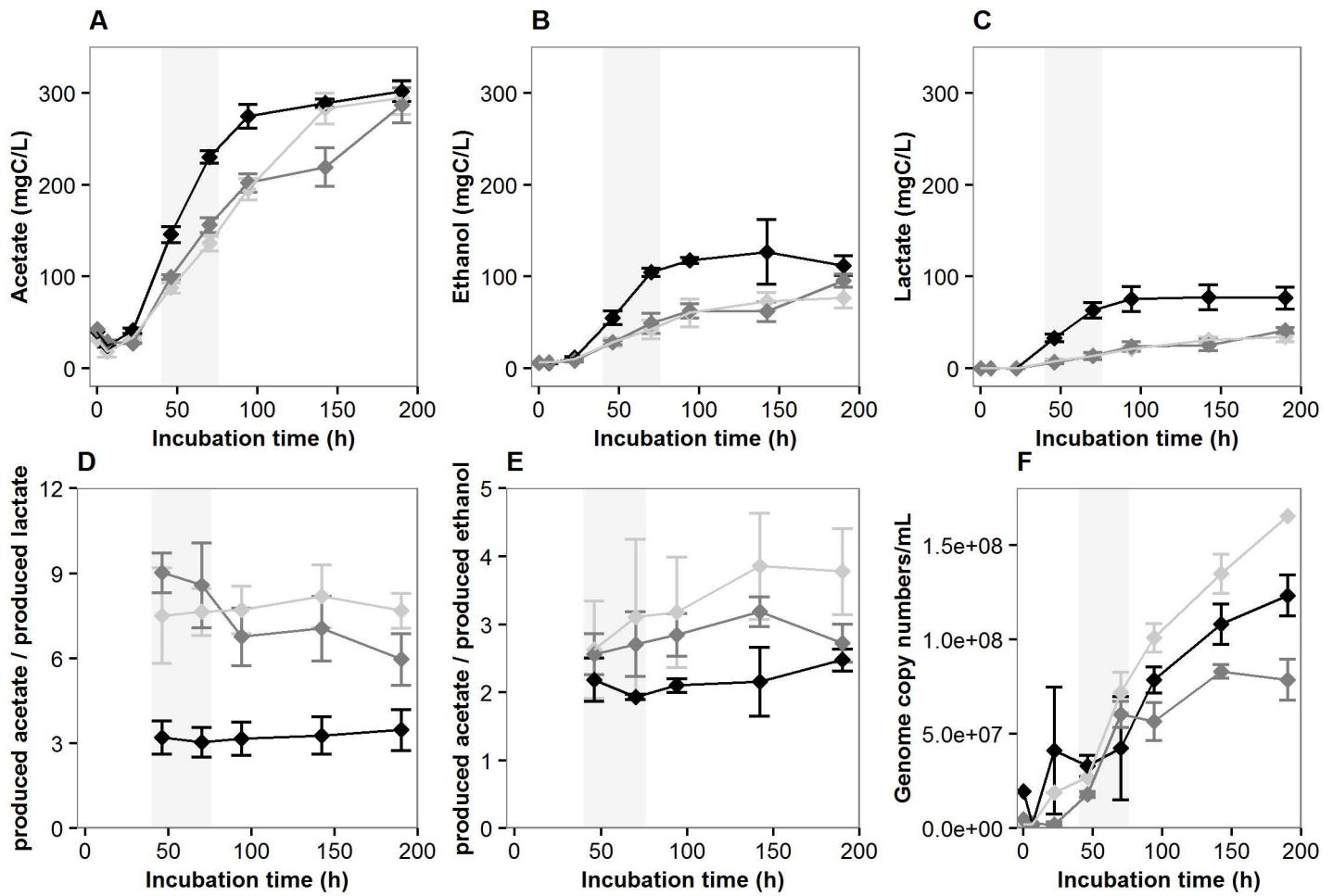
Fig 1. Growth and fermentation dynamics of *R. cellulolyticum* on Tissue (black symbols), Whatman Paper (grey symbols) and Cotton (light grey symbols). Acetate (A), ethanol (B) and lactate (C) are the three most abundant fermentation products and their concentration ratios are shown in (D-E). Genome copy numbers estimated from the amount of total extracted DNA are shown in (F). Error bars indicate standard deviations calculated from triplicate samples, except in F (duplicate samples). Light grey areas indicate the time points selected for subsequent proteomic analyses.

doi:10.1371/journal.pone.0170524.g001

with lower acetate-to-lactate concentration ratios (Fig 1D) as well as lower acetate-to-ethanol concentration ratios compared to both other substrates (Fig 1E). This specific metabolic profile likely results from the higher sugar influx in the cells and the faster pH decrease in the milieu over time, from pH 7.1 to pH ~6.3 (S1 Fig, panel B). Indeed, it has previously been shown for *R. cellulolyticum* that the carbon flux partition between acetate, lactate and ethanol is greatly influenced by pH and entering carbon flows [20, 34, 35].

At the end of the 190 hours of incubation, the substrates were only partly degraded, as shown by the degradation yields estimated from carbon mass distributions (S1 File, S2 Fig and S1 Table), but significantly altered, disrupting easily during sample handling. Such a partial degradation is expected in batch microcosms where suboptimal growth conditions emerge over time, typically with the accumulation of H_2 (values at the final time point shown in S1 Fig, panel C) or other fermentation products. Independently, colonization of the cellulosic substrates by the bacterial cells was qualitatively followed in microplate incubations over the incubation period, by wet mount microscopy and scanning electron microscopy (S1 File and S3 Fig). The observations were in good agreement with the degradation dynamics, since

Table 1. Detailed characteristics of Tissue, Whatman Paper and Cotton.

	Tissue	W. Paper	Cotton
Total Solids/Volatil Solids (%/%)	94.9/94.2	96.0/95.8	96.5/96.4
Carbon/Nitrogen (%/%)	41.9/0.07	43.1/0.03	41.1/0.05
Chemical Oxygen Demand (g/g)	1.06	1.14	1.11
Crystallinity Index (%)	50	94	74
Degree of Polymerization*	~970	~1300	~2730
Total sugar content (% Dry Matter)	**97.23**	**99.22**	**99.47**
Hexoses (% Dry Matter)	**83.35**	**98.67**	**98.94**
Glucose	81.54	98.58	98.69
Galactose	0.14	0.03	0.14
Mannose	1.66	0.05	0.07
Rhamnose	0.02	0.01	0.03
Pentoses (% Dry Matter)	**13.87**	**0.55**	**0.54**
Xylose	13.72	0.51	0.45
Arabinose	0.15	0.04	0.09
Van Soest fractionation			
Neutral Detergent Soluble fraction (%)	0.05	0.01	5.3
Acid Detergent Soluble fraction (%)	14.89	4.01	43.6
Sulfuric Acid Soluble fraction (%)	85.06	84.84	49.8
Insoluble Volatile Solids fraction (%)	0	11.13	1.2

W. Paper: Whatman Paper.

*The DP values correspond to the MW values of the peak of individual (i.e. non-aggregated) cellulose chains from the molar mass distribution plots (S5 Fig) divided by the mass of the tricarbanilated anhydroglucose unit (519 Da) (see Materials and Methods).

doi:10.1371/journal.pone.0170524.t001

colonization occurred the fastest for Tissue and the slowest for Cotton. Scanning electron microscopy provided insight into the cell aggregate structure and suggested the absence of thick biofilm (S1 File and S3 Fig).

Substrate characterization

To identify substrate's properties likely to explain the differences observed during their fermentation by *R. cellulolyticum*, the three cellulosic substrates were characterized in details (Table 1). Measurement values of total solids, volatile solids, carbon, nitrogen and organic contents (Table 1) were each highly similar among the three substrates and consistent with cellulose being their major component. Moreover, the dominant chemical functional groups were the same in all substrates and typical of cellulosic materials since primary and secondary alcohols and glycosidic bonds were identified by Fourier transform infrared spectroscopy (FTIR, S1 File, S4 Fig).

Important differences between the substrates concerned the crystallinity index (CI) (Table 1), the molar mass distributions (Table 1, S1 File and S5 Fig) and the hemicellulose content (Table 1). CI is commonly measured to estimate the amount of crystalline regions in cellulose, less easily degradable compared to amorphous regions. Based on the calculated CI, cellulose was the most amorphous in Tissue and the most crystalline in Whatman Paper (Table 1). The average degree of polymerization (DP) (Table 1, S1 File and S5 Fig) was the lowest for Tissue and the highest for Cotton. Both Whatman Paper and Cotton were composed almost exclusively of glucose (98.58% and 98.69% respectively, Table 1), whereas Tissue contained a significant proportion (13.87%) of pentoses (mainly xylose, 13.72%) in addition to the

dominant glucose (81.54%) (Table 1). The composition data were consistent with the molecular weight distribution analyses (S1 File and S5 Fig).

Material characterization highlighted differences among the substrates both in terms of composition and structure. Based on these characteristics, the faster bioconversion observed for Tissue compared to both other substrates could arise from its low CI and its low average DP. Moreover, the presence of hemicelluloses in Tissue is likely to increase its enzyme accessibility and/or its hydrophilicity at the supramolecular level since networks of hemicelluloses and cellulose are less ordered and crystalline than networks of pure cellulose, which is thus likely to favor a faster degradation [36].

By contrast, Van Soest fractionation indicated that Whatman Paper is overall the substrate the less readily solubilized by chemical solutions (total of 4.02% solubilization by the first two detergents), followed by Tissue (total of 14.94% solubilization by the first two detergents) and Cotton (total of 48.9% solubilization by the first two detergents) (Table 1). These results highlight that great differences exist between chemical and biological reactivity for the studied cellulosic substrates.

Comparative proteome-wide label-free quantification

To investigate which biological functions could predominantly be influenced by the substrate during fermentation, a sensitive shotgun proteomic approach based on XIC was implemented and served as basis for comparative quantitative analyses. Two substrates were selected for this approach, Tissue due to its specific metabolic profile and Whatman Paper as a reference since cellulose colonization by *R. cellulolyticum* has previously been studied on this substrate [37, 38]. Two time points were selected to be able to discriminate between the effect of the sole substrate and the effect of time and substrate interaction. Illustrative examples of such effects are provided in S6 Fig. Proteins were extracted from the pellets and supernatants separately at incubation times 46h and 70h (S7 Fig, panel A) and analyzed by LC-MS/MS. At the selected time points, fermentation products were just starting to significantly accumulate (Fig 1A–1C) and the biomass was actively growing (Fig 1F), limiting the possible effects of inhibitors accumulating in batch microcosms.

A total of 1194 proteins were quantified by the XIC approach (S1 Dataset). A good reproducibility was obtained (S7 Fig, panel B) and the identified functions were consistent with cellulose fermentation (S8 Fig, S2 Table). Comparative statistical analyses were conducted for each protein by adjusting models accounting for the influence of substrate, of time and of the interaction of substrate and time. In total, 151 proteins showing significantly different levels (Q-value < = 0.01) were identified and validated (S1 Dataset), including 132 with an effect in the pellets exclusively, 16 with an effect in the supernatants exclusively and 3 with an effect in both. They besides included 20 cellulosomal components (providing information about substrate hydrolysis mechanisms), 8 enzymes from the central carbon catabolism (providing information about the response to intracellular carbon fluxes) and 44 extracytoplasmic proteins (providing information about substrate transport and other specific functions).

Carbohydrate-active enzymes (CAZymes) and cellulosomal proteins. Considering their essential role in substrate deconstruction and catabolism, proteins related to cellulolysis and CAZymes were specifically examined. *R. cellulolyticum* cellulosomes include structural subunits and numerous different catalytic subunits [16, 17, 20] encoded by 65 distinct genes [21]. Their average protein composition is highly modulated according to the nature of the carbohydrate growth substrate [16, 21]. *R. cellulolyticum* genome encodes up to 149 CAZymes according to [21] and only a subset of them are extracellular components involved in lignocellulose deconstruction since CAZymes participate to a variety of other biological processes. For

the present study, a total of 153 proteins from *R. cellulolyticum* were considered (S1 Dataset) corresponding to its CAZymes and to its known non-CAZyme cellulosomal structural subunits. Among these 153 proteins, 103 were quantified in at least one sample and 28 showed significantly different levels when comparing growth on Tissue and Whatman Paper (Tables 2–4). The present approach appears as complementary to specific proteomic approaches targeting cellulosomal components (such as in [16]) since the sensitivity is comparable and additional functions can also be detected. Indeed, out of the 52 cellulosomal components and CAZymes detected in [16], 47 were quantified in the present study and, for instance, 7 additional proteins with a dockerin-module (cellulosomal components), not detected in [16], were quantified here.

As expected, the 28 proteins with significantly different levels were dominated by cellulosomal subunits (Tables 2 and 3, total of 20 proteins). The levels of an important proportion of the 65 cellulosomal proteins were thus significantly influenced although both cellulosic substrates are rather similar, highlighting the sensitivity of cellulosome composition to subtle substrate differences. Among the 20 cellulosomal proteins with significant effects, 17 had lower levels or levels that decreased faster when comparing growth on Tissue and on Whatman Paper (Tables 2 and 3, at least one negative log fold change), encompassing 6 glycoside hydrolase (GH) families and 2 other CAzyme families (Tables 2 and 3). This observation is at first sight unexpected: minor sugar components being more abundant in Tissue compared to Whatman, especially hemicelluloses (Table 1), it could have been anticipated that a variety of cellusome enzymes could have higher levels in the Tissue incubations. In the model proposed by [21], core cellulosomal genes are activated, or not repressed, when intracellular levels of glycolytic intermediates are low (carbon catabolite repression) [21]. Since less carbon flowed through glycolysis during growth on Whatman paper compared to Tissue, results obtained here reflect that hydrolysis of the most recalcitrant substrate requires more diverse cellulosomal enzymes during a longer time.

Hemicelluloses from Tissue were very likely at least partly fermented during the incubations since 3 intracellular proteins involved in xylose or xylose oligomer catabolism (encoded by Ccel_0203, Ccel_3438, Table 4, and Ccel_3429) had higher levels in Tissue incubations, suggesting a higher xylodextrin intracellular influx therein. It is however unclear whether the higher abundance of xylose in Tissue specifically induced the higher expression of certain xylanases. Indeed, cellulosomal proteins encoded by the "*xyl-doc*" cluster (14 genes more specifically oriented towards hemicellulolysis [16]) did not show any significant differences, although 8 of them were quantified in the whole dataset (S1 Dataset). The cellulosomal endoglucanase Ccel_0429, presenting higher levels in Tissue incubations, could have participated to hemicellulose deconstruction since it exerts xyloglucan depolymerization as a secondary activity [39]; however its regulation mechanisms are unknown. Finally, the other cellulosomal subunits with a known xylanase activity and statistically significant differences had lower levels in Tissue incubations (Ccel_0755 in Table 2, Ccel_0931 in Table 3).

The lack of specific activation of the "*xyl-doc*" cluster has previously been reported [16, 21] during growth of *R. cellulolyticum* on oat spelt xylan, as well as its activation during growth on wheat straw [16] or corn stover [21]. Together with the present work, these observations suggest that natural lignocellulosic substrates, in which hemicellulose is associated to lignin within complex entangled structures, are more likely to induce the expression of the "*xyl-doc*" cluster than more simple or engineered materials where the growth substrates are more readily available.

The subset of cellulosomal proteins with significantly different levels was enriched in endoglucanases (12 out of 20 cellulosomal proteins with different levels, Table 2, compared to 17 endoglucanases in total among the 65 cellulosomal proteins annotated in *R. cellulolyticum*

Table 2. Cellulosomal endoglucanases with significantly different levels when comparing Tissue and Whatman Paper incubations.

Gene ID	Protein/Gene name	Pellet Substrate[a]	Pellet Interaction[b]	Supernatant Substrate[a]	Supernatant Interaction[b]	Modular structure[c]	Localization[d]	Protein function or name[e]
Ccel_1648	β-glucanase R, Cel9R	-0.24			+0.29	S-GH9-CBM3-UNK-CBM3-DOC1	cellulosome	*Endoglucanase Cel9R*
Ccel_0732*	Cel9E	-0.26				S-UNK-CBM4-UNK-GH9-UNK-DOC1	cellulosome	**Endoglucanase/ cellobiohydrolase Cel9E**
Ccel_0734*	Cel9H	-0.18				S-UNK-GH9-UNK-CBM3-DOC1	cellulosome	*Endoglucanase Cel9H*
Ccel_1249	β-glucanase T, Cel9T	-0.25				S-GH9-CBM3-DOC1	cellulosome	*Endoglucanase Cel9T*
Ccel_0735*	Cel9J	-0.42	-0.38			S-GH9-CBM3-DOC1	cellulosome	*Endoglucanase Cel9J*
Ccel_0740*	Cel5N	-0.28	-0.25			S-GH5-DOC1	cellulosome	Endoglucanase Cel5N
Ccel_0753	Cel9P, P90	-0.22	-0.28			S-GH9-CBM3-UNK-DOC1	cellulosome	*Endoglucanase Cel9P*
Ccel_0755	cellulase U / cellulase S, Cel9U		-0.13			S-UNK-GH9-DOC1	cellulosome	*Endoglucanase Cel9U*
Ccel_2392	Cel9V		-0.21			S-UNK-CBM4-UNK-GH9-DOC1	cellulosome	*Endoglucanase/ cellobiohydrolase Cel9V*
Ccel_1099	celCCA, Cel5A, Cca	+0.28		-3.63	-3.63	S-GH5-DOC1	cellulosome	**Endoglucanase/ endoxylanase Cel5A**
Ccel_2337	CMCase, P66					S-GH5-DOC1	cellulosome	Endoglucanase
Ccel_0429	P99			+0.40		S-GH44-DOC1-UNK-CBM44	cellulosome	Endoglucanase Cel44O (PKD domain containing protein)

The proteins are listed according to the observed effects and to their function. The statistical models take into account the replicates and their variability.

* in the first left column ("Gene ID") indicate genes encoded in the "*cip-cel*" gene cluster, which codes for 12 key cellulosomal components.

a) The log10 fold change values are indicated for the proteins with statistically significant substrate effects (Q-value < = 0.01). Tissue incubations are used as a reference (positive values when the protein levels are higher in the Tissue incubations).

b) The log10 fold change values are indicated for the proteins with statistically significant substrate-by-time interaction effects (Q-value < = 0.01). Tissue incubations are used as a reference (positive values when the protein levels increase faster or decrease slower in the Tissue incubations).

c) According to [16, 21] and the present study: S: signal sequence; GH: family of glycoside hydrolase; PL: family of pectate lyase; CE: family of carbohydrate esterase; GT: family of glycosyl transferase; CBM: family of carbohydrate-binding module; DOC1: dockerin type 1 module; COH: cohesin type I module; LNK: linker sequence; SLH: surface-layer homology sequence; COG: clusters of orthologous groups; UNK: unknown function module or sequence; TSP_C: thrombospondin C-terminal region; fn3: fibronectin type III domain.

d) Known (in bold) or predicted localization.

e) Predicted or characterized (in bold) activities or protein names according to [16] and to UniprotKB database.

doi:10.1371/journal.pone.0170524.t002

Table 3. Other cellulosomal proteins with significantly different levels when comparing Tissue and Whatman Paper incubations.

| Gene ID | Protein/ Gene name | Pellet | | Supernatant | | Modular structure | Localization | Protein function or name |
		Substrate	Interaction	Substrate	Interaction			
Ccel_0931	P41a, xyn10A	-0.31				S-GH10-DOC1	**cellulosome**	Xylanase Xyn10A
Ccel_2162	P42	-0.27				S-DOC1-CE2	**cellulosome**	Acetyl-xylan esterase
Ccel_1655		-0.22				S-DOC1-UNK	**cellulosome**	Unknown (cellulosome protein dockerin type I)
Ccel_1060		-4.09	-3.78			S-COG2755 / COG2845-DOC1	**cellulosome**	SGNH-hydrolase
Ccel_2243		-0.23	-0.22			S-PL1-UNK-DOC1-UNK	**cellulosome**	Pectate lyase
Ccel_0379	P76		-0.20			S-GH5-LNK-CBM32-DOC1	**cellulosome**	Mannanase
Ccel_1597	P50		-0.15			S-GH27-UNK-DOC1	**cellulosome**	α-Galactosidase
Ccel_1543				+0.60		S-TSP_C–(fn3)4-CBM	**cellulosome**	Cellulosome anchoring protein cohesin region

The column titles are identical to those from Table 2.

genome). Among these endoglucanases, 10 had significantly lower levels in the Tissue incubations, particularly enzymes from the GH5 and GH9 families (Table 2). This result is consistent with the presence of shorter cellulose chains in this Tissue as well as its lower CI. In the closely related *Ruminiclostridium thermocellum* [40] and *Ruminiclostridium clariflavum* [41], it has indeed been shown that the levels GH9 endoglucanases, which are also very common in their cellulosomes, are influenced by the substrate's nature, with increased levels in the presence of crystalline cellulose.

Overall, CAZyme expression appears to be influenced here more by substrate structure than by its carbohydrate composition, although the present experiments do not provide a fully formal proof.

Central carbon metabolism. To determine whether the level of sugar influx affected the expression of enzymes from the central carbon metabolism, enzymes from the glucose and

Table 4. Non-cellulosomal CAZymes with at least one statistically significant effect when comparing incubations on Tissue and Whatman Paper.

| Gene ID | Protein/ Gene name | Pellet | | Supernatant | | Modular structure | Localization | Protein function or name |
		Substrate	Interaction	Substrate	Interaction			
Ccel_0428	Cel5I			-3.85	+3.83	S-GH5-CBM17-CBM28-(SLH)3	cell wall	Endoglucanase Cel5I
Ccel_2417				+3.86	+3.86	GT39	cell wall	Glycosyl transferase family 39
Ccel_0881				-0.60		S-CBM16-UNK	secreted	Unknown (Carbohydrate-binding, CenC-like protein)
Ccel_1036				-0.27		GH51-UNK	secreted	α-Arabinofuranosidase
Ccel_2893				-3.64	+3.64	S-GH18-UNK	secreted	β-Glycosidase
Ccel_1139		+0.30				UNK-GH3-UNK	intracellular	β-Glucosidase
Ccel_0203			+0.31			GH3-UNK	intracellular	β-Xylosidase
Ccel_3438			+0.34			GH43-UNK	intracellular	β-Xylosidase/a-arabinofuranosidase

The column titles are identical to those from Table 2.

xylose catabolic pathways of *R. cellulolyticum* (Fig 2 and S1 Dataset) were specifically examined. Among them, 23 were successfully quantified in at least one of the pellet samples and statistical models could be adjusted for 18 of them. Finally, 8 proteins showed statistically significant effects: the glucose-6-phosphate isomerase (product from Ccel_1445, pgi), the ATP-dependent 6-phosphofructokinase (Ccel_2612, pfkA), the xylose isomerase (Ccel_3429, xylA), the glyceraldehyde-3-phosphate dehydrogenase (Ccel_2275), the phosphoglycerate kinase (Ccel_2260), the phosphoglycerate mutase (Ccel_0619), the 2,3-bisphosphoglycerate-independent phosphoglycerate mutase (Ccel_2259, gpmI) and the phosphopyruvate hydratase, also known as enolase (Ccel_2254, eno). All 8 of them are enzymes from upstream of the pyruvate node and they are distributed over the Embden-Meyerhof-Parnas (EMP) and xylose-utilization pathways (Fig 2). Except for the enolase, they all showed only positive effects, indicating higher protein levels and/or lower protein level decreases in Tissue incubations compared to Whatman Paper incubations.

This result shows that the higher sugar influx during Tissue fermentation leads to an overall enhanced expression of enzymes from the carbon catabolic pathways, which is shown for the first time here for *R. cellulolyticum* and contrasts to a previous report on the closely related *Ruminiclostridium termitidis* [42].

Other extracytoplasmic proteins. The nature of the cellulosic substrate most certainly directly or indirectly influences the expression levels of numerous extracytoplasmic proteins since the latter are involved in a variety of biological processes in bacteria, including substrate colonization, substrate uptake, or cell-cell interactions, which are all relevant for the present study. Predicting the subcellular localization of proteins is a complex issue [43]. To identify extracytoplasmic proteins in the present dataset, the 115 proteins showing at least one statistically significant effect when comparing growth on Tissue and Whatman Paper (other than those already described in the result section on CAZymes and cellulosomal proteins) were specifically analyzed *in silico*. Among the 115 manually examined proteins, 44 had a predicted extracytoplasmic localization and the detailed results are shown in S3 Table. For 36 of these 44 proteins, the prediction can be considered as very robust since at least both SurfG+ and LocateP indicate an extracytoplasmic localization (with however sometimes distinct predicted subcellular localizations).

Noticeably, the 44 selected proteins include 8 components of ATP-binding Cassette (ABC) transporters (Table 5) and 4 of them, encoded by Ccel_1987, Ccel_2997, Ccel_0998 and Ccel_1133, are likely involved in sugar transport based on blastp analyses against the manually curated part of ABCdb (www-abcdb.biotoul.fr, CleanDb). They indeed show significant similarity (e-value $< 10^{-5}$) with the sequence of several proteins annotated as sugar binding proteins and with at least one experimentally well-characterized sugar-transporter component such as the D-xylose transporter subunit encoded by the xylF gene from *Escherichia coli* (Ccel_1987) or the multiple sugar-binding ABC transporter, sugar-binding protein precursor MsmE encoded by msmE gene from *Streptococcus mutans* (Ccel_2997, Ccel_0998, Ccel_1133). The trends regarding their expression are contrasted, since higher or lower levels are observed in Tissue incubations according to the protein (Table 5). These results show the modulation of the sugar ABC transporter profile according to the nature of the cellulosic substrate and consistently, genes Ccel_1987, Ccel_0998 and Ccel_1133 belong to genomic regions regulated by Two-Components Systems responding to the availability of specific extracellular soluble sugars, as described in [21].

Interestingly, a Fibronectin type III domain protein (encoded by Ccel_0648) predicted to be released in the extracellular milieu by both SurfG+ and LocateP was associated to significant negative effects in the pellets, indicating higher levels in the Whatman paper incubations (S3 Table). A Fibronectin type III-like repeat from the *Ruminiclostridium thermocellum*
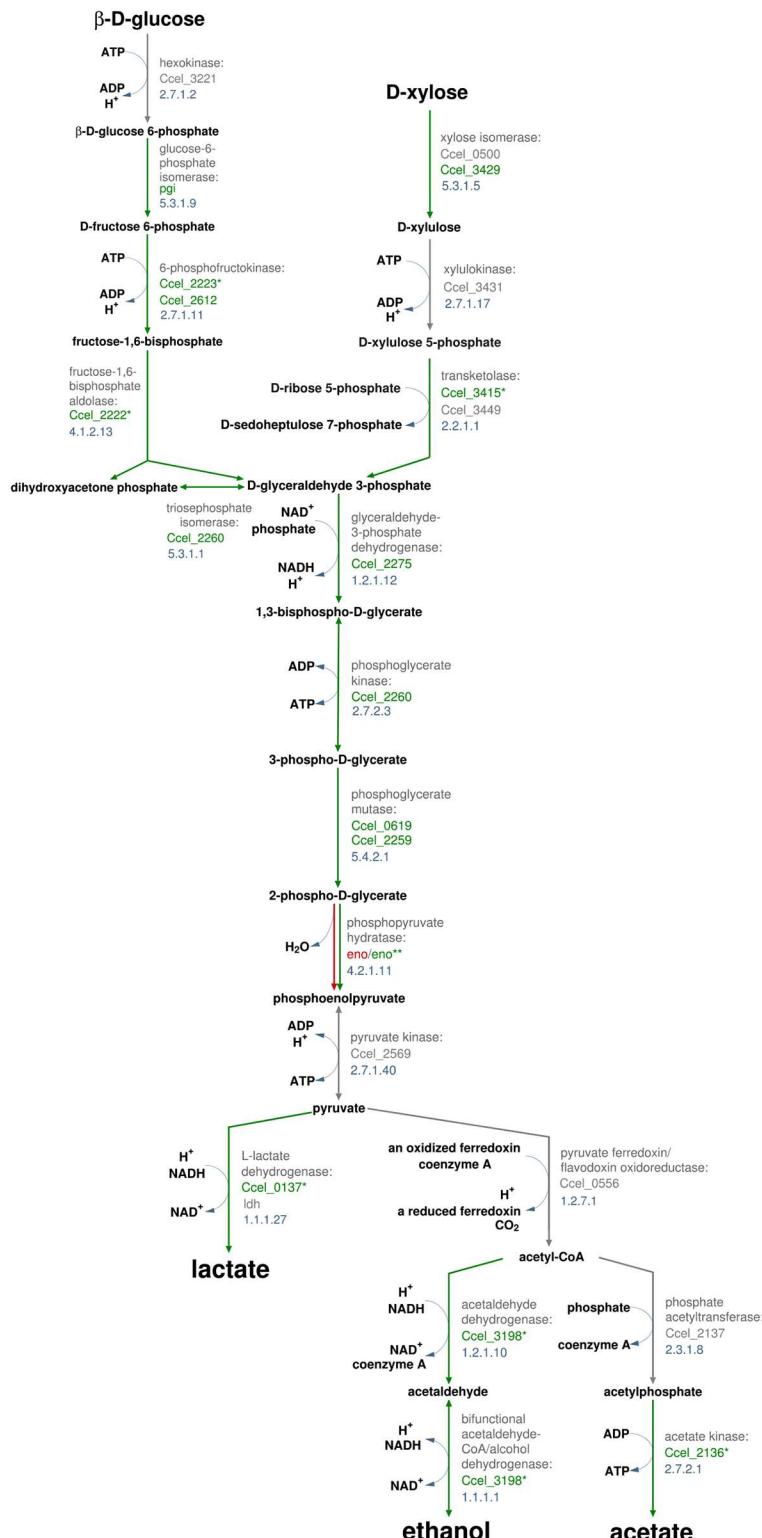
Fig 2. Proteins with significant effects when comparing growth on Tissue and Whatman Paper mapped over *R. cellulolyticum* glucose and xylose catabolic pathways. Statistically significant substrate and substrate-by-time interaction effects were considered. The green color indicates positive effects while the red color indicates negative effects. Positive effects correspond to quantified protein levels higher in Tissue incubations than in Whatman Paper incubations (substrate effect) and to quantified protein levels increasing

more or decreasing less in Tissue incubations than in Whatman Paper incubations (substrate-by-time interaction effect). The statistical models (see Materials and Methods) take into account the replicates and their variability. Protein names and EC numbers are indicated in grey. * indicate effects not significant when adjusting for multiple comparisons (Q-values > 0.01) but still supporting the overall trend (p-values < = 0.05). ** indicate a significant negative substrate effect (q-value ≤ 0.01) and a positive interaction effect (q-value > 0.01 but p-value ≤ 0.05). Pathways were adapted from the Biocyc website (http://biocyc.org/).

cellobiohydrolase CbhA was previously shown to promote hydrolysis of cellulose by modifying its surface [44]. If the protein encoded by Ccel_0648 participates to a similar function, its higher concentration levels in Whatman Paper incubations could favor the hydrolysis of this more recalcitrant substrate compared to Tissue.

Other extracytoplasmic proteins with significantly different levels could be interesting, such as those containing SLH domains; however, a fine interpretation is overall hindered by the currently limited knowledge on non-cellulolytic extracytoplasmic proteins in *R. cellulolyticum*. For instance, the ABC transporter specificities are poorly described for this species. Strikingly, among the 16 proteins of unknown function present in the dataset of the 151 significant proteins, 13 correspond to predicted extracytoplasmic proteins (S3 Table). Better characterizing transporters and other extracytoplasmic proteins from *R. cellulolyticum* thus appears of great

Table 5. ABC transporter proteins with significantly different levels when comparing Tissue and Whatman Paper incubations.

Gene ID	Protein name	Pellets		Supernatants		SurfG+[c]	LocateP[d]
		Sub. [a]	Inter. [b]	Sub. [a]	Inter. [b]		
Ccel_1987	Putative solute-binding component of ABC transporter (S_1ab)	+0.37				PSE	Lipid anch.
Ccel_1133	Extracellular solute-binding protein family 1 (S_5ab)	+0.55				PSE	Lipid anch.
Ccel_1768	Extracellular solute-binding protein family 5 (S_2a)		+0.27			PSE	Lipid anch.
Ccel_2997	Extracellular solute-binding protein family 1 (S_5ab)	-7.65		NA	NA	PSE	N-ter anch.
Ccel_0967	Transport permease protein (M_7a)	-0.15		NA	NA	MB	Memb.
Ccel_0998	Extracellular solute-binding protein family 1 (S_5ab)	-3.66	+3.66	NA	NA	PSE	Lipid anch.
Ccel_1156	Periplasmic solute binding protein (S_8b)	-3.68	+3.68	NA	NA	PSE	Lipid anch.

The proteins are listed according to the observed effects and to the subcellular localization predicted by SurfG+.

[a] The log10 fold change values are indicated for proteins with statistically significant substrate effects ("Sub., Q-value < = 0.01). Tissue incubations are used as a reference (positive values when the protein levels are higher in the Tissue incubations). The statistical models take into account the replicates and their variability.

[b] The log10 fold change values are indicated for proteins with statistically significant substrate-by-time interaction effects ("Inter.", Q-value < = 0.01). Tissue incubations are used as a reference (positive values when the protein levels increase faster or decrease slower in the Tissue incubations). The statistical models take into account the replicates and their variability.

[c] Subcellular localization predicted by SurfG+. PSE: potentially surface exposed; MB: membrane.

[d] Subcellular localization from LocateP database. Lipid anch.: Lipid anchored; Memb: Multi-transmembrane;. N-ter anch.: N-terminally anchored (No cleavage site).

importance to better understand its physiology during cellulose degradation and to be able to implement global approaches such as systems metabolic engineering.

Conclusions

Fermentation by model cellulolytic bacteria of engineered materials has been little studied so far. In this study, fermentation by *R. cellulolyticum* of three cellulosic substrates containing no lignin, paper handkerchief, cotton discs and Whatman filter paper was considered. Paper handkerchief was fermented the fastest and 151 proteins had significantly different levels when comparing paper handkerchief and Whatman filter paper incubations, including 8 enzymes from the central carbon metabolic pathways and 44 distinct extracytoplasmic proteins. They moreover comprised 20 out of the 65 cellulosomal components and 4 non-cellulosomal extracytoplasmic CAZymes potentially involved in cellulolysis, highlighting the sensitivity of the cellulolysis machinery to subtle differences in substrate properties. In particular, ten cellulosomal endoglucanases, mainly from GH5 and GH9 families, had lower levels during fermentation of paper handkerchief when comparing with fermentation of Whatman paper. This observation hypothetically results from the lower crystallinity rate and degree of polymerization of cellulose in paper handkerchief. Paper handkerchief exhibited higher hemicellulose content and the enhanced level of intracellular xylose isomerase suggested that the hemicellulose was at least partly metabolized. However, regarding hemicellulose hydrolysis, none of the known extracytoplasmic enzymes with xylanolysis as primary activity had significantly higher levels in the Tissue incubations. It appears that natural lignocellulosic substrates, in which hemicellulose is associated to lignin within complex entangled structures, could be more likely to induce the expression of the "*xyl-doc*" cluster or other specialized xylanases than more simple or engineered materials where the growth substrates are more readily available. Similar to differences occurring among Tissue and Whatman paper incubations, there could be significant differences on protein levels among Whatman paper and Cotton incubations, especially regarding the cellulolysis machinery, since these substrates have different crystallinity index and degrees of polymerization. Addressing this question would require further proteomic analyses. The present study provides, to our knowledge, the first whole-proteome analysis on the model cellulolytic bacterium *R. cellulolyticum* and expands the knowledge on the proteome response of this bacterium to cellulosic substrates.

Supporting Information

S1 Fig. Data on fermentation of Tissue (black symbols), Whatman Paper (grey symbols) and Cotton (light grey symbols) by *R. cellulolyticum*. (A) Evolution over time of the total Dissolved Organic Content (DOC). (B) Evolution over time of pH. (C) Cumulated gas production at the final incubation time point. Error bars indicate standard deviations calculated from triplicate samples. Light grey areas in (A) and (B) indicate the time points selected for subsequent proteomic analyses. *R. cellulolyticum* was grown in 50 mL batch fermentation microcosms on 2.5 g/L cellulosic substrate.
(TIF)

S2 Fig. Average carbon mass distributions in the microcosms at the initial and final incubation time points. The carbon masses are in mg. Sampled: carbon mass removed from the microcosms through sampling of the liquid phase—CO2 gas: carbon mass in CO_2 in the headspace—DIC: inorganic carbon mass in the liquid phase (Dissolved Inorganic Carbon)–DOC: organic carbon mass in the liquid phase (Dissolved Organic Carbon)–Cellulosic material: estimated carbon mass in the substrate (contained either in Tissue, Whatman Paper or Cotton).

The percent values next to each substrate name indicate the estimated average degradation yield (percentage of carbon from the substrate that was degraded). Details on the calculation method are available in S1 File.
(TIF)

S3 Fig. Illustrative microscopy images of substrate colonization by *R. cellulolyticum* during growth in microplates on Tissue (left column), Whatman Paper (middle column) and Cotton (right column) at a final concentration of 5 g/l. Confocal Laser Scanning Microscopy images were acquired from Tissue (A-D), Whatman Paper (F-I) and Cotton (K-N) incubations, on wet mount samples stained with a cellular esterase activity marker (green) after removal of the planktonic cells. Scale bars are 200 μm. A total of 75 representative images were acquired. Scanning Electron Microscopy images were acquired from Tissue (E), Whatman Paper (J) and Cotton (O) incubations, on samples collected after 48 h of incubation. Scale bars are 3 μm. A total of 137 images of scanning electron microscopic were acquired. Details on the methods are available in S1 File.
(TIF)

S4 Fig. Mid-infrared absorption spectra obtained for Tissue, Whatman Paper and Cotton. W. Paper: Whatman Paper. The 5 peaks annotated with arrows on the spectra are related to the presence of cellulose. Details on the method are available in S1 File.
(TIFF)

S5 Fig. Molar mass distribution curves of the cellulose and hemicellulose chains from Tissue (black lines), Whatman Paper (grey lines) and Cotton (light grey lines). M stands for Molar Mass. Vertical colored lines indicate the positions of the M values corresponding to the peak of individual (i.e. non-aggregated) cellulose chains (see Table 1). Grey area A1: hemicellulose distribution peak observed for Tissue, originating from bleached wood pulp. Grey area A2: peaks corresponding to very high molecular weight polymers and, more likely, to cellulose chain aggregates. Details on the method are available in S1 File.
(TIFF)

S6 Fig. Illustrative examples of statistically significant substrate and substrate-by-time interaction effects on the protein levels. The illustrative examples are selected from the dataset of the pellet proteins. On each dot plot, the values are shown for the incubation times 46h, 70h and for the blank. The shown values correspond to the log-transformed and normalized data. The red color corresponds to Tissue incubations and the green color to Whatman Paper incubations. The "+" and "-"signs indicate the sign of the considered effect (substrate or substrate-by-time interaction). From left to right and from top to bottom: Ccel_1139 encodes a β-Glucosidase (see also Table 4); Ccel_0734 encodes the endoglucanase Cel9H (see also Table 2); Ccel_0203 codes for a β-Xylosidase (see also Table 4); Ccel_2392 codes for the endoglucanase/cellobiohydrolase Cel9V (see also Table 2); Ccel_1570 encodes a putative uncharacterized protein; Ccel_0735 encodes the endoglucanase Cel9J (see also Table 2); Ccel_3392 encodes a putative uncharacterized protein; Ccel_0428 encodes the endoglucanase Cel5I (see also Table 4).
(TIF)

S7 Fig. Quantification of total extracted proteins and principal component analyses of the proteomes based on the individual protein quantification data. A) Total amounts of proteins extracted from the Tissue (black) and Whatman Paper (grey) incubations, after 46h and 70h of incubation, from the pellets and supernatants respectively. B) Principal component analysis of the samples based on the label-free quantitative proteomic data (XIC approach).
(TIF)

S8 Fig. Functional profiles for all proteins encoded in *R. cellulolyticum* genome and for the proteins showing significantly different levels in the presence of Tissue compared to Whatman Paper. A selection of 32 Gene Ontology (GO) terms is shown, corresponding to the categories with highest percentages of annotations and to the most enriched or depleted categories when comparing the dataset of proteins with significantly different levels (after removal of categories with less than 3 proteins with significantly different levels) and all genome-encoded proteins. The GO terms are shown from the most enriched to the most depleted, from top to bottom. *R. cellulolyticum* genome encodes 3290 proteins, of which 2081 have GO annotations in UniprotKB, corresponding to a total of 3674 GO annotations. 151 proteins showed significantly different levels, of which 116 have GO annotations in UniprotKB, corresponding to a total of 385 annotations. Numeric values and additional details are shown in S2 Table.
(TIF)

S1 Dataset. Summary of the quantitative results obtained for each protein of *R. cellulolyticum*.
(XLSX)

S1 Table. Carbon mass distributions in the microcosms at the initial and final incubation time points. The carbon masses are in mg. The average values (± one standard deviation where relevant) are shown. Cellulosic material: carbon mass in the substrate (contained either in Tissue, Whatman Paper or Cotton)–DOC: organic carbon mass in the liquid phase (Dissolved Organic Carbon)–DIC: inorganic carbon mass in the liquid phase (Dissolved Inorganic Carbon)–CO2 gas: carbon mass in CO2 in the headspace—Sampled: carbon mass removed from the microcosms through sampling of the liquid phase—Total: total carbon mass in the microcosms at the initial incubation time point—Degradation yield: estimated percentage of degraded carbon from the substrate. The carbon mass in the cellulosic material at the final time point is calculated by considering that the total carbon mass is identical at time points 0h and 190h in the system. To calculate the carbon mass removed through sampling of the liquid phase, two options were considered: no substrate particles were sampled (option 1), substrate particles at a concentration of 2.6 g/L (corresponding to the initial concentration) were sampled (option 2). Consequently, two different values were obtained for the carbon mass in the cellulosic material at time point 190h. More details on the method are available in S1 File.
(PDF)

S2 Table. Functional profiles for all proteins encoded in *R. cellulolyticum* genome and for proteins showing significantly different levels in the presence of Tissue compared to Whatman Paper. a) BP: Biological Process—CC: Cellular Component—MF: Molecular Function. b) For each GO term, percentage of the GO term annotations among the significant proteins. c) For each GO term, percentage of the GO term annotations among all proteins encoded in *R. cellulolyticum* genome. d) Ratios of both percentages. The GO terms are presented by decreasing ratio values. See legend from S7 Fig for additional details.
(XLSX)

S3 Table. Extracytoplasmic proteins with at least one statistically significant effect when comparing incubations on Tissue and Whatman Paper. The proteins are listed according to the observed effects, to the subcellular localization predicted by SurfG+ and according to their function. a) The Gene IDs (Ccel_) are given according to UniprotKB database. b) The Protein names are given according to UniprotKB database. c)—d) Statistically significant effects for proteins quantified in Pellets and Supernatants respectively. Substrate effects (Q-value < = 0.01) are indicated with + for positive effects (quantified protein levels higher in Tissue incubations than in Whatman Paper incubations) and—for negative effects (quantified protein levels

lower in Tissue incubations than in Whatman Paper incubations). Substrate-by-time interaction effects (Q-value < = 0.01) are indicated with + for positive effects (quantified protein levels in Tissue incubations increase more or decrease less than in Whatman Paper incubations) and—for negative effects (quantified protein levels in Tissue incubations increase less or decrease more than in Whatman Paper incubations).The statistical models take into account the replicates and their variability. e) Subcellular localization predicted by SurfG+. PSE: potentially surface exposed; EXT: extracellular milieu; CYTO: cytoplasm; MB: membrane. f) Subcellular localization from LocateP database. Lipid anch.: Lipid anchored; Released: Secretory (released) (with cleavage site); Memb: Multi-transmembrane;. N-ter anch.: N-terminally anchored (No cleavage site); Intracell.: Intracellular. g) SignalP predictions concerning the presence of a signal peptide. Y: yes; N: no. h) SecretomeP predictions concerning the secretion by a non-classical pathway (without signal peptide). Y: yes; N: no. i) LipoP predictions concerning lipoproteins and signal peptides. SpII: lipoprotein signal peptide (signal peptidase II); SpI: signal peptide (signal peptidase I); TMH: n-terminal transmembrane helix (this is generally not a very reliable prediction according to LipoP website); CYT: cytoplasmic (all others). j) TMHMM predictions: number of predicted transmembrane helixes. k) TMHMM predictions: possible presence of a signal peptide (as indicated on THMM server website, "predicted TM segments in the N-terminal region sometime turn out to be signal peptides").
(XLSX)

S1 File. Supplementary Materials and Methods and References.
(PDF)

Acknowledgments

Author Contributions

Conceptualization: NB TB A. Bize.

Formal analysis: NB MD VS A. Bize.

Funding acquisition: A. Bize.

Investigation: NB AG NP BP SD.

Methodology: NB AG A. Bridier DZS.

Project administration: A. Bize.

Resources: PR VM DZS LM A. Buléon TB GM.

Software: MD VS.

Supervision: TB A. Bize.

Visualization: NB VS SD GM A. Bize.

Writing – original draft: NB GM A. Bize.

Writing – review & editing: NB AG VS VM DZS TB A. Bize.

References

1. Hansen MA, Kristensen JB, Felby C, Jorgensen H. Pretreatment and enzymatic hydrolysis of wheat straw (Triticum aestivum L.)—the impact of lignin relocation and plant tissues on enzymatic accessibility. Bioresource technology. 2011; 102(3):2804–11. Epub 2010/11/03. doi: 10.1016/j.biortech.2010.10.030 PMID: 21036603

2. Himmel ME, Bayer EA. Lignocellulose conversion to biofuels: current challenges, global perspectives. Current opinion in biotechnology. 2009; 20(3):316–7. Epub 2009/06/16. doi: 10.1016/j.copbio.2009.05.005 PMID: 19523813

3. Ogeda TL, Silva IB, Fidale LC, El Seoud OA, Petri DF. Effect of cellulose physical characteristics, especially the water sorption value, on the efficiency of its hydrolysis catalyzed by free or immobilized cellulase. Journal of biotechnology. 2012; 157(1):246–52. Epub 2011/12/08. doi: 10.1016/j.jbiotec.2011.11.018 PMID: 22146618

4. Kalogo Y, Habibi S, MacLean HL, Joshi SV. Environmental implications of municipal solid waste-derived ethanol. Environmental science & technology. 2007; 41(1):35–41. Epub 2007/02/03.

5. Marshall CW, LaBelle EV, May HD. Production of fuels and chemicals from waste by microbiomes. Current opinion in biotechnology. 2013; 24(3):391–7. Epub 2013/04/17. doi: 10.1016/j.copbio.2013.03.016 PMID: 23587964

6. Buruiana C-T, Garrote G, Vizireanu C. Bioethanol production from residual lignocellulosic materials: A review-Part 2. Annals of the University Dunarea de Jos of Galati, Fascicle VI: Food Technology. 2013; 37(1):25–38.

7. Hasunuma T, Okazaki F, Okai N, Hara KY, Ishii J, Kondo A. A review of enzymes and microbes for lignocellulosic biorefinery and the possibility of their application to consolidated bioprocessing technology. Bioresource technology. 2013; 135:513–22. Epub 2012/12/01. doi: 10.1016/j.biortech.2012.10.047 PMID: 23195654

8. Olson DG, McBride JE, Shaw AJ, Lynd LR. Recent progress in consolidated bioprocessing. Current opinion in biotechnology. 2012; 23(3):396–405. Epub 2011/12/20. doi: 10.1016/j.copbio.2011.11.026 PMID: 22176748

9. Lu F, Bize A, Guillot A, Monnet V, Madigou C, Chapleur O, et al. Metaproteomics of cellulose methanisation under thermophilic conditions reveals a surprisingly high proteolytic activity. The ISME journal. 2014; 8(1):88–102. Epub 2013/08/21. doi: 10.1038/ismej.2013.120 PMID: 23949661

10. Chu KH, Feng X. Enzymatic conversion of newspaper and office paper to fermentable sugars. Process Safety and Environmental Protection. 2013; 91(1):123–30.

11. Sangkharak K. Optimization of enzymatic hydrolysis for ethanol production by simultaneous saccharification and fermentation of wastepaper. Waste management & research: the journal of the International Solid Wastes and Public Cleansing Association, ISWA. 2011; 29(11):1134–44. Epub 2011/01/19.

12. Lima D, Gouveia E, editors. Increase in bioethanol production from used office paper by Saccharomyces cerevisiae UFPEDA 1238. BMC Proceedings; 2014: BioMed Central Ltd.

13. Higashide W, Li Y, Yang Y, Liao JC. Metabolic engineering of Clostridium cellulolyticum for production of isobutanol from cellulose. Applied and environmental microbiology. 2011; 77(8):2727–33. Epub 2011/03/08. doi: 10.1128/AEM.02454-10 PMID: 21378054

14. Li T, Mazeas L, Sghir A, Leblon G, Bouchez T. Insights into networks of functional microbes catalysing methanization of cellulose under mesophilic conditions. Environmental microbiology. 2009; 11(4):889–904. Epub 2009/01/09. doi: 10.1111/j.1462-2920.2008.01810.x PMID: 19128320

15. Peng X, Borner RA, Nges IA, Liu J. Impact of bioaugmentation on biochemical methane potential for wheat straw with addition of Clostridium cellulolyticum. Bioresource technology. 2014; 152:567–71. Epub 2013/12/21. doi: 10.1016/j.biortech.2013.11.067 PMID: 24355075

16. Blouzard JC, Coutinho PM, Fierobe HP, Henrissat B, Lignon S, Tardif C, et al. Modulation of cellulosome composition in Clostridium cellulolyticum: adaptation to the polysaccharide environment revealed by proteomic and carbohydrate-active enzyme analyses. Proteomics. 2010; 10(3):541–54. Epub 2009/12/17. doi: 10.1002/pmic.200900311 PMID: 20013800

17. Ravachol J, Borne R, Tardif C, de Philip P, Fierobe HP. Characterization of all family-9 glycoside hydrolases synthesized by the cellulosome-producing bacterium Clostridium cellulolyticum. The Journal of biological chemistry. 2014; 289(11):7335–48. Epub 2014/01/24. doi: 10.1074/jbc.M113.545046 PMID: 24451379

18. Desvaux M. Unravelling carbon metabolism in anaerobic cellulolytic bacteria. Biotechnology progress. 2006; 22(5):1229–38. Epub 2006/10/07. doi: 10.1021/bp060016e PMID: 17022659

19. Desvaux M. Mapping of carbon flow distribution in the central metabolic pathways of Clostridium cellulolyticum: Direct comparison of bacterial metabolism with a soluble versus an insoluble carbon source. Journal of microbiology and biotechnology. 2004; 14(6):1200–10.

20. Desvaux M. Clostridium cellulolyticum: model organism of mesophilic cellulolytic clostridia. FEMS microbiology reviews. 2005; 29(4):741–64. Epub 2005/08/17. doi: 10.1016/j.femsre.2004.11.003 PMID: 16102601

21. Xu C, Huang R, Teng L, Wang D, Hemme CL, Borovok I, et al. Structure and regulation of the cellulose degradome in Clostridium cellulolyticum. Biotechnology for biofuels. 2013; 6(1):73. Epub 2013/05/10. doi: 10.1186/1754-6834-6-73 PMID: 23657055

22. Stams AJ, Van Dijk JB, Dijkema C, Plugge CM. Growth of syntrophic propionate-oxidizing bacteria with fumarate in the absence of methanogenic bacteria. Applied and environmental microbiology. 1993; 59 (4):1114–9. Epub 1993/04/01. PMID: 16348912

23. Wakelin JH, Virgin HS, Crystal E. Development and Comparison of Two X-Ray Methods for Determining the Crystallinity of Cotton Cellulose. Journal of Applied physics. 2004; 30(11):1654–62.

24. Henniges U, Kloser E, Patel A, Potthast A, Kosma P, Fischer M, et al. Studies on DMSO-containing carbanilation mixtures: chemistry, oxidations and cellulose integrity. Cellulose. 2007; 14(5):497–511.

25. Mottet A, François E, Latrille E, Steyer JP, Déléris S, Vedrenne F, et al. Estimating anaerobic biodegradability indicators for waste activated sludge. Chemical Engineering Journal. 2010; 160(2):488–96.

26. Hanreich A, Heyer R, Benndorf D, Rapp E, Pioch M, Reichl U, et al. Metaproteome analysis to determine the metabolically active part of a thermophilic microbial community producing biogas from agricultural biomass. Canadian journal of microbiology. 2012; 58(7):917–22. Epub 2012/06/14. doi: 10.1139/w2012-058 PMID: 22690648

27. Junker J, Bielow C, Bertsch A, Sturm M, Reinert K, Kohlbacher O. TOPPAS: a graphical workflow editor for the analysis of high-throughput proteomics data. Journal of proteome research. 2012; 11(7):3914–20. Epub 2012/05/16. doi: 10.1021/pr300187f PMID: 22583024

28. Sturm M, Bertsch A, Gropl C, Hildebrandt A, Hussong R, Lange E, et al. OpenMS—an open-source software framework for mass spectrometry. BMC bioinformatics. 2008; 9:163. Epub 2008/03/28. doi: 10.1186/1471-2105-9-163 PMID: 18366760

29. Martin AD, Quinn KM, Park JH. MCMCpack: Markov Chain Monte Carlo in R. Journal of Statistical Software. 2011; 42(9):1–21.

30. Heidelberger P, Welch PD. Simulation Run Length Control in the Presence of an Initial Transient. Operations Research. 1983; 31(6):1109–44.

31. Plummer M, Best N, Cowles K, Vines K. CODA: Convergence diagnosis and output analysis for MCMC. R News. 2006; 6(1):7–11.

32. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. Journal of the Royal Statistical Society Series B (Methodological). 1995:289–300.

33. Barinov A, Loux V, Hammani A, Nicolas P, Langella P, Ehrlich D, et al. Prediction of surface exposed proteins in Streptococcus pyogenes, with a potential application to other Gram-positive bacteria. Proteomics. 2009; 9(1):61–73. Epub 2008/12/05. doi: 10.1002/pmic.200800195 PMID: 19053137

34. Desvaux M, Guedon E, Petitdemange H. Cellulose catabolism by Clostridium cellulolyticum growing in batch culture on defined medium. Applied and environmental microbiology. 2000; 66(6):2461–70. Epub 2000/06/01. PMID: 10831425

35. Desvaux M, Guedon E, Petitdemange H. Metabolic flux in cellulose batch and cellulose-fed continuous cultures of Clostridium cellulolyticum in response to acidic environment. Microbiology (Reading, England). 2001; 147(Pt 6):1461–71. Epub 2001/06/08.

36. Krässig HA. Cellulose: structure, accessibility and reactivity. Gordon and Breach Science Publ.; 1993. p. 6–42, 167–323.

37. Gelhaye E, Gehin A, Petitdemange H. Colonization of Crystalline Cellulose by Clostridium cellulolyticum ATCC 35319. Applied and environmental microbiology. 1993; 59(9):3154–6. Epub 1993/09/01. PMID: 16349055

38. Gehin A, Gelhaye E, Petitdemange H. Adhesion of Clostridium cellulolyticum spores to filter paper. Journal of applied bacteriology. 1996; 80(2):187–90.

39. Ravachol J, de Philip P, Borne R, Mansuelle P, Maté MJ, Perret S, et al. Mechanisms involved in xyloglucan catabolism by the cellulosome-producing bacterium Ruminiclostridium cellulolyticum. Scientific reports. 2016; 6.

40. Raman B, Pan C, Hurst GB, Rodriguez M Jr, McKeown CK, Lankford PK, et al. Impact of pretreated switchgrass and biomass carbohydrates on Clostridium thermocellum ATCC 27405 cellulosome composition: a quantitative proteomic analysis. PloS one. 2009; 4(4):e5271. doi: 10.1371/journal.pone. 0005271 PMID: 19384422

41. Artzi L, Morag E, Barak Y, Lamed R, Bayer EA. Clostridium clariflavum: key cellulosome players are revealed by proteomic analysis. MBio. 2015; 6(3):e00411–15. doi: 10.1128/mBio.00411-15 PMID: 25991683

42. Munir RI, Spicer V, Krokhin OV, Shamshurin D, Zhang X, Taillefer M, et al. Transcriptomic and proteomic analyses of core metabolism in Clostridium termitidis CT1112 during growth on α-cellulose, xylan, cellobiose and xylose. BMC microbiology. 2016; 16(1):1.

43. Renier S, Micheau P, Talon R, Hebraud M, Desvaux M. Subcellular localization of extracytoplasmic proteins in monoderm bacteria: rational secretomics-based strategy for genomic and proteomic analyses. PloS one. 2012; 7(8):e42982. Epub 2012/08/23. doi: 10.1371/journal.pone.0042982 PMID: 22912771

44. Kataeva IA, Seidel RD 3rd, Shah A, West LT, Li XL, Ljungdahl LG. The fibronectin type 3-like repeat from the Clostridium thermocellum cellobiohydrolase CbhA promotes hydrolysis of cellulose by modifying its surface. Applied and environmental microbiology. 2002; 68(9):4292–300. Epub 2002/08/30. doi: 10.1128/AEM.68.9.4292-4300.2002 PMID: 12200278

A unique virus release mechanism in the Archaea

Ariane Bize[a], Erik A. Karlsson[b], Karin Ekefjärd[b], Tessa E. F. Quax[a], Mery Pina[a], Marie-Christine Prevost[c], Patrick Forterre[a], Olivier Tenaillon[d], Rolf Bernander[b], and David Prangishvili[a,1]

[a]Biologie Moléculaire du Gène chez les Extrêmophiles and [c]Plate-Forme de Microscopie Ultrastructurale, Institut Pasteur, 25 rue du Docteur Roux, 75724 Paris cedex 15, France; [b]Department of Molecular Evolution, Evolutionary Biology Center, Uppsala University, Norbyvägen 18C, SE-752 36 Uppsala, Sweden; and [d]Institut National de la Santé et de la Recherche Médicale, Unité 722, Faculté de médecine Xavier Bichat, Université Paris 7, 16 rue Henri Huchard, 75018 Paris, France

Little is known about the infection cycles of viruses infecting cells from Archaea, the third domain of life. Here, we demonstrate that the virions of the archaeal *Sulfolobus islandicus* rod-shaped virus 2 (SIRV2) are released from the host cell through a mechanism, involving the formation of specific cellular structures. Large pyramidal virus-induced protrusions transect the cell envelope at several positions, rupturing the S-layer; they eventually open out, thus creating large apertures through which virions escape the cell. We also demonstrate that massive degradation of the host chromosomes occurs because of virus infection, and that virion assembly occurs in the cytoplasm. Furthermore, intracellular viral DNA is visualized by flow cytometry. The results show that SIRV2 is a lytic virus, and that the host cell dies as a consequence of elaborated mechanisms orchestrated by the virus. The generation of specific cellular structures for a distinct step of virus life cycle is known in eukaryal virus-host systems but is unprecedented in cells from other domains.

lysis | virus factory | hyperthermophile | infection cycle

As for organisms belonging to the Bacteria and Eukarya, members of the domain Archaea are infected by specific viruses. The majority of archaeal viruses isolated so far contain dsDNA as the genetic material and infect hyperthermophilic hosts from the phylum Crenarchaeota (1). The diversity and uniqueness of these viruses at both the morphological and genetic levels are such that they have been classified into 7 viral families (2). The knowledge on the biology of this exceptional group of viruses is still limited, partly because of the unique genetic content: very few genes have detectable functions or homologs in the databases (3).

In particular, little is known about relationships of crenarchaeal viruses with their hosts. Except for a few isolated cases (4–6), it is generally presumed that these viruses persist in the host cell in a carrier state, a nonlytic relationship in which virions are continuously secreted by the still-dividing cells (7). However, the classification of crenarchaeal viruses as chronic is based on indirect experimental evidence, such as a lack of optical density (OD) decrease and absence of cellular debris in infected cultures (e.g., 8, 9). Detailed characterization of the infection cycle and the carrier state has not been specifically addressed in the scarce reports on crenarchaeal host-virus interactions (see e.g., 10).

To study the nature of host-virus relationships in crenarchaea, we selected the nonenveloped, rod-shaped virus SIRV2 and its hyperthermophilic and acidophilic host, *Sulfolobus islandicus*. SIRV2, originally described as a carrier state, nonlysogenic virus (11), belongs to a common crenarchaeal virus family, the *Rudiviridae* (9, 11, 12, 13, 14), and contains a linear 35.5-kb dsDNA genome (15). The host belongs to a well characterized crenarchaeal genus, *Sulfolobus* (16, 17), from which also other viruses are known (2). We describe detailed in vivo effects of the virus on its host and, unexpectedly, demonstrate that SIRV2 is a cytocidal, lytic virus. Remarkably, a unique virus release mechanism was encountered during the characterization, involving the generation of pyramidal structures that, by opening out, cause local disruption of the cell envelope and allow virion escape. In addition, intracellular viral DNA was visualized by flow cytometry, and the technique was also used to demonstrate chromosome degradation in infected cells.

Results

Growth Kinetics of SIRV2-Infected Cultures. OD and CFU values from uninfected and infected [multiplicity of infection (moi) ≈7] cultures of *S. islandicus* were monitored over time. The effects of the virus were visible already 1.5 h after infection (Fig. 1). Whereas uninfected cultures pursued normal growth with a generation time of ≈13 h, the OD in infected cultures remained constant for ≈60 h (Fig. 1 A and C), after which growth resumed (Fig. 1C). During this time period, the CFU values of uninfected controls remained constant or increased slightly. In contrast, the CFU values decreased dramatically in infected cultures, resulting in an ≈1,000-fold reduction at 6 h after infection (Fig. 1B, 10.5 h). The CFU values also revealed growth of a minor cell population in infected cultures starting at early time points (Fig. 1D, from 15 h). This growth was initially not detectable in the OD measurements (Fig. 1C), because of the low concentration of this cell population at early time points. Thus, infection by SIRV2 has a pronounced effect on the host cultures, preventing growth of a majority of the cells.

To exclude the possibility that the results were linked to the high moi used, or to the specific growth conditions, similar experiments were performed at low moi ($\approx 10^{-3}$), at different temperatures (70 °C, 75 °C, and 78 °C), pHs (3.0 and 3.5), medium richnesses (standard medium or 5-fold less rich medium), and with different host strains (*S. islandicus* strains KVEM10H3, HVE10/4, and LAL14/1). No significant differences were observed, indicating that the effects occurred independently of these parameters.

The cell population growing in the presence of SIRV2 consisted of cells completely resistant to SIRV2 infection, not producing any detectable infectious virions nor carrying the SIRV2 genome (*SI Text* and Fig. S1). This was consistent with the observation that the SIRV2 genome does not integrate into the host chromosome (11), and excluded the possibility that SIRV2 established a carrier state relationship with its host. The high initial proportion of resistant cells suggested that specific mechanisms could be involved in their generation, in addition to random mutations, such as CRISPR-related mechanisms (18).

Flow Cytometry Analysis of Infected Cells over Time. The cell size and intracellular DNA content in uninfected and SIRV2-infected cultures (moi ≈10) over time were monitored by flow cytometry (Fig. 2, Fig. S2, and Fig. S3).

The relative lengths of the *S. islandicus* cell cycle periods in the control cultures were found to be similar to those of other *Sulfolobus* species (16, 19), with the post-replicative phase occupying a large fraction of the generation time (68%, Fig. S4). Based on comparison of the flow cytometry fluorescence (DNA content)

MICROBIOLOGY
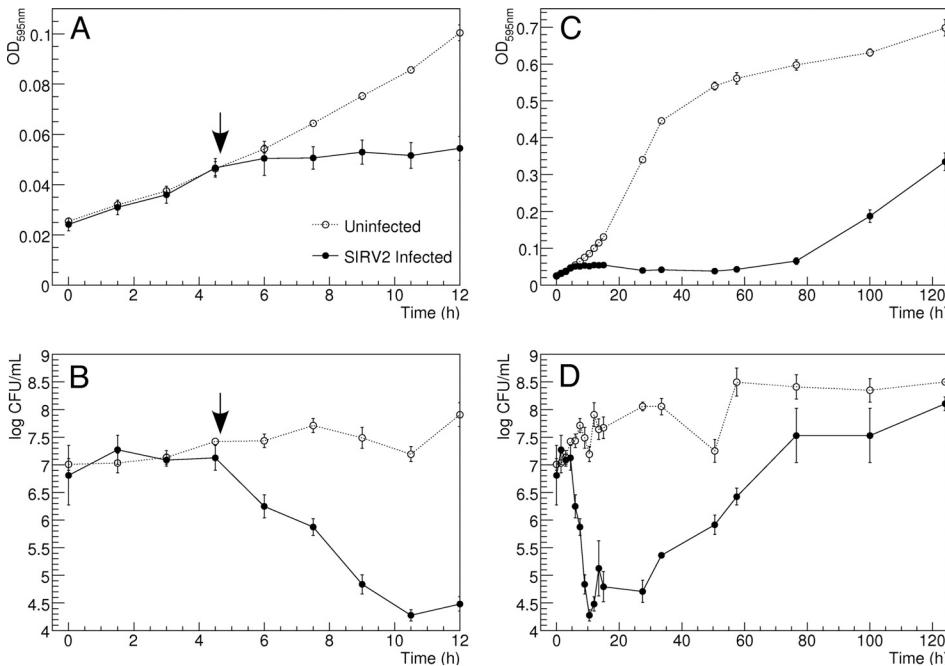
Fig. 1. Impact of SIRV2 infection on the growth kinetics of *S. islandicus* cultures. Cultures infected at a moi of ≈7 (filled circles, continuous line) and uninfected cultures (empty circles, dotted lines), were launched in triplicates. Averages of the replicates ± 1 SD are shown. The vertical arrows in *A* and *B* correspond to virus addition (4.5 h). (*A*) $OD_{595\ nm}$, detail of the first hours. (*B*) Log transformation of the CFU titres, detail of the first hours. (*C*) $OD_{595\ nm}$ over the entire time course. (*D*) Log transformation of the CFU titers over the entire time course.

signals with those from the sequenced genomes of *S. acidocaldarius* and *S. solfataricus*, the genome size was estimated to ≈2.6 Mb (Fig. S5). The average cell size (Fig. 2*A Left* and Fig. S3) progressively decreased when the cultures approached stationary phase. In the infected cultures (Fig. 2*B Left*), a cell size increase initially occurred in part of the cell population, evident as an extension of the distribution toward the right (6–8 h). Subsequently, the average cell size gradually decreased over time.

The DNA content distributions of the control cultures (Fig. 2*A Right* and Fig. S2) were typical for exponentially growing *Sulfolobus* cells (19), with a majority of the cells containing 2 chromosomes. In the infected cultures (Fig. 2*B Right*), cells with a very low DNA content (≪1 genome equivalent) started to appear at 0.5 h after infection and then increased in proportion over time whereas the proportion of cells containing 1–2 genome equivalents decreased. Thus, at 12 h, a large majority of the cell population contained no detectable intracellular DNA. The SIRV2 latent period is 8–10 h (below), and chromosome degradation, thus, occurred before virus release in a significant fraction of the cell population. Interestingly, the populations of chromosome-less cells and cells containing DNA were clearly separated and well defined (Fig. 2*B Right* and Fig. S6). Thus, for a given infected cell, chromosome degradation must have occurred within a brief time interval.

In parallel to chromosome degradation, an increase in the total DNA content occurred in part of the cell population, evident as an extension of the 2-genome equivalents peak toward the right (Fig. 2*B Right*, 2 h and onwards). The increase was (Fig. 2*C*, arrow) estimated to ≈0.5 genome equivalents per cell, or 1.3 Mb, on average 3 h after infection, and it corresponded to newly synthesized viral DNA (below).

The results demonstrate that infection by SIRV2 causes massive degradation of the host chromosome in virtually all infected cells during the first 12 h of infection, excluding the possibility that SIRV2 genomes are vertically transmitted between cell generations.

Links Between the Virus Infection Cycle, the Kinetics of Host Chromosome Degradation, and Cell Death. To discriminate between host chromosome and viral DNA, uninfected and infected cultures (moi ≈15) were monitored by dot blot hybridizations, in addition to flow cytometry. In an uninfected control culture, the percentage of chromosome-less cells did not exceed 5% (Fig. 3*A*) and tended to

decrease over time. In infected cultures, chromosome-less cells began to accumulate in the first hours, and after 5 h, the percentage was ≈40%, confirming that significant degradation occurred before virion release (at ≈8–10 h, see below) and, at 11 h, >80% of the cells were chromosome-less. Subsequent degradation occurred at a lower rate and finally reached 97%, confirming that genome degradation occurred in most cells.

The intracellular amount of SIRV2 DNA (Fig. 3 *B* and *D*) increased gradually and reached a maximum after ≈8 h, followed by a large decrease up until 14 h. The initial increase presumably corresponded to viral DNA replication, and the decrease to virus release, indicating a latent period of ≈8–10 h. Thus, a single round of infection occurred in the cultures at the high moi used. To relate viral DNA production to the dynamics of chromosome degradation, the percentage of DNA-less cells appearing between successive time points was superimposed (Fig. 3*B*). A small peak of degradation, visible at 0.5 h after infection, was most likely an artifact caused by the low signal-to-noise ratio for DNA-less cells in the very early time points. The major peak occurred at 11 h, in the middle of the virus release period. The use of a 16S rRNA gene probe combined with similar data analysis confirmed the chromosome degradation observed by flow cytometry (Fig. 3 *C* and *E*).

To confirm the latent period of 8–10 h and to estimate the burst size, a 1-step growth experiment was performed (*SI Text* and Fig. S7). Virus release was shown to begin at ≈8–10 h after infection, confirming that no infectious virions were released before this time point, consistent with all other data. The burst size was estimated to 30 ± 10 viruses per cell. Finally, a membrane potential-sensitive probe (*SI Text* and Fig. S8) was used to confirm that cell death occurred in connection to virus release.

In conclusion, a single round of infection occurred when a high moi was used, and the SIRV2 latent period was ≈8–10 h under the conditions used. Massive host chromosome degradation occurred throughout the infection cycle, starting from the early stage, and cell death took place concomitantly with virus release. Thus, SIRV2 is a lytic virus that kills the host cell during the process of virus production and release.

Identification of Cellular Ultrastructures Induced by SIRV2 Infection. To obtain insights into the details of the virus-host interaction, infected cells were analyzed by SEM and TEM. The cells were fixed
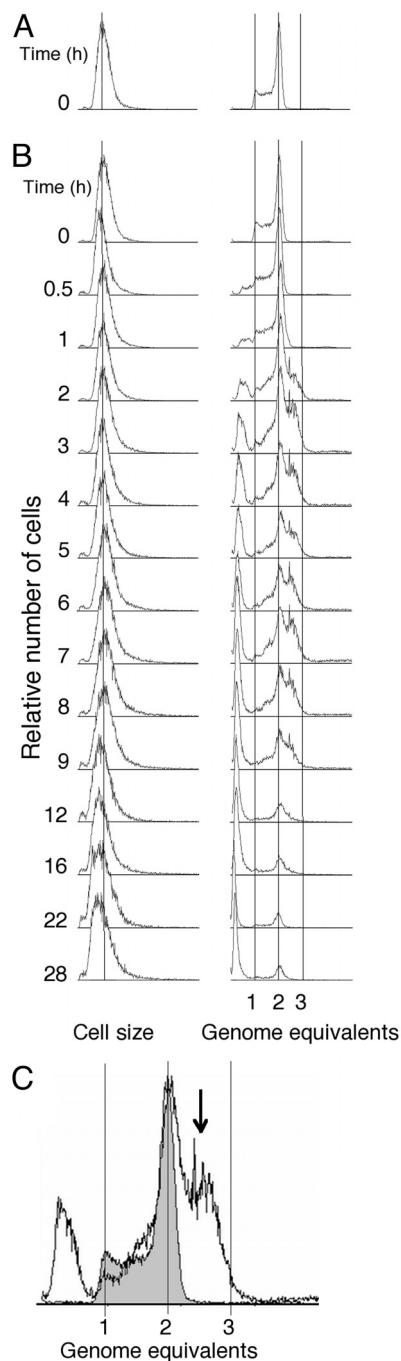
Fig. 2. Flow cytometry time-course analysis of *S. islandicus* cells infected by SIRV2. (*A*) Representative cell size and DNA content distributions from an uninfected culture. (*B*) Cell size and DNA content distributions from a culture infected with SIRV2 (moi ≈10). The virus was added just after time point 0 h. (*C*) Visualization of intracellular SIRV2 DNA by flow cytometry at 3 h after infection. The DNA content distribution from an infected *S. islandicus* culture is shown against the distribution from an uninfected culture (translucent gray). The arrow indicates additional DNA in infected cells.

at 10 h (just before virion release), 13 h (middle of release period), and 26 h (after release) after infection. Uninfected cells in midexponential growth phase were used as control. For analysis with TEM, ultrathin sections of samples were prepared.

The irregular coccoid morphology of uninfected cells was typical for *Sulfolobales* species, with the cell envelope consisting of a lipid membrane and an S-layer (Fig. 4*A1* and *A2*). At 10 h after infection,

multiple pyramidal protrusions were observed on the cell surface by SEM (Fig. 4*B1*, arrows), which were absent in uninfected control cells. In thin sections analyzed by TEM, these structures appeared as large angular protrusions associated with a local absence of S-layer on the cell envelope (Fig. 4 *B2–B4*). Both with SEM and TEM, several such virus-associated pyramids (VAPs) were usually visible per cell (Fig. 4 *B1* and *B2*). The pyramidal structure of the VAPs, suggested by SEM, was confirmed by TEM, showing a polygonal base in a plane parallel to the cell envelope (Fig. 4*B5*). In thin sections, the VAPs often contained regions producing a denser staining (Fig. 4*B2*, arrows), localized at the tip of the pyramidal structure.

Dense aggregates of virions were visible by TEM within numerous cells from the infected culture (examples in Fig. 4 *B4* and *B6*), showing that virion assembly occurred in the cytoplasm. Up to 3 densely packed aggregates, together containing up to ≈150 virions, were detected in the cell sections, and occupied a high fraction of the intracellular volume. The higher number of virions compared with the estimates from the 1-step growth experiment (above) could be due to that virions may still be aggregated after release.

At 13 h after infection, together with cells resembling the examples shown in Fig. 4 *B1* and *B2*, cells lacking VAPs and displaying numerous perforations on the cell surface were observed (similar to Fig. 4 *C1* and *C2*), and 26 h after infection almost all cells were perforated and empty (Fig. 4 *C1* and *C2*). The perforated cells exhibited spherical morphotype, different from the native phenotype, suggesting an altered intracellular organization. Thin section analysis of perforated cells displayed virion remains (Fig. 4*C2 Inset*) and disappearance of most of the cytoplasmic content (Fig. 4*C2*). The cell perforations were heterogeneous in size, and their majority visible in thin sections had a diameter in the range of 200 nm. TEM analysis revealed that the perforations were delimited by C-shaped structures (Fig. 4 *C2* and *C3*), and it is likely that these represented VAP remains. Apart from the perforations, the cell envelope appeared to be intact, with both the S-layer and the membrane visible (Fig. 4 *C2* and *C3*). Notably, the characteristic structures at the boundary of the perforations of the lysed cells were sometimes observed detached from the cell envelope (Fig. 4 *C4–C6*). The resemblance of polygonal shapes in Fig. 4 *B5*, *C5*, and *C6*, as well as the similarity of the structures in Fig. 4 *B3* and *C3*, supports the hypothesis that the structures in Fig. 4*C* represented remains of the VAPs shown in Fig. 4*B*. Thus, the VAPs were apparently involved in perforation of the cell envelope. Because ongoing virus release could not be detected, this must have occurred within a brief time interval.

Discussion

We report a detailed cellular study of the infection cycle of a crenarchaeal virus and demonstrate that SIRV2 is a lytic virus. The virions are assembled in the cytoplasm of the host cell and, 8–10 h after infection, start to be released through well defined apertures in the cell envelope. Remarkably, formation of these openings is preceded and facilitated by the generation of virus-induced cellular structures of pyramidal shape, VAPs, located at the cell envelope and pointing outwards. The VAPs perforate the membrane and S-layer, and after disruption leave behind apertures delimited by a ring structure of polygonal shape. After virion release, the cell envelope remains as a stable empty shell. Intracellular viral DNA was visualized by flow cytometry, and the same technique was used to show that host chromosome is completely degraded during the viral infection cycle. The combination of the data from 1-step growth experiment, flow cytometry, and TEM showed that chromosome degradation most likely occurred before virion release, in the majority of the cell population. Together, all of our results demonstrate that the host cells die as a consequence of specific and unique mechanisms orchestrated by the virus, rather than from general deleterious effects of the infection. The deduced viral life cycle is schematically illustrated in Fig. 5.
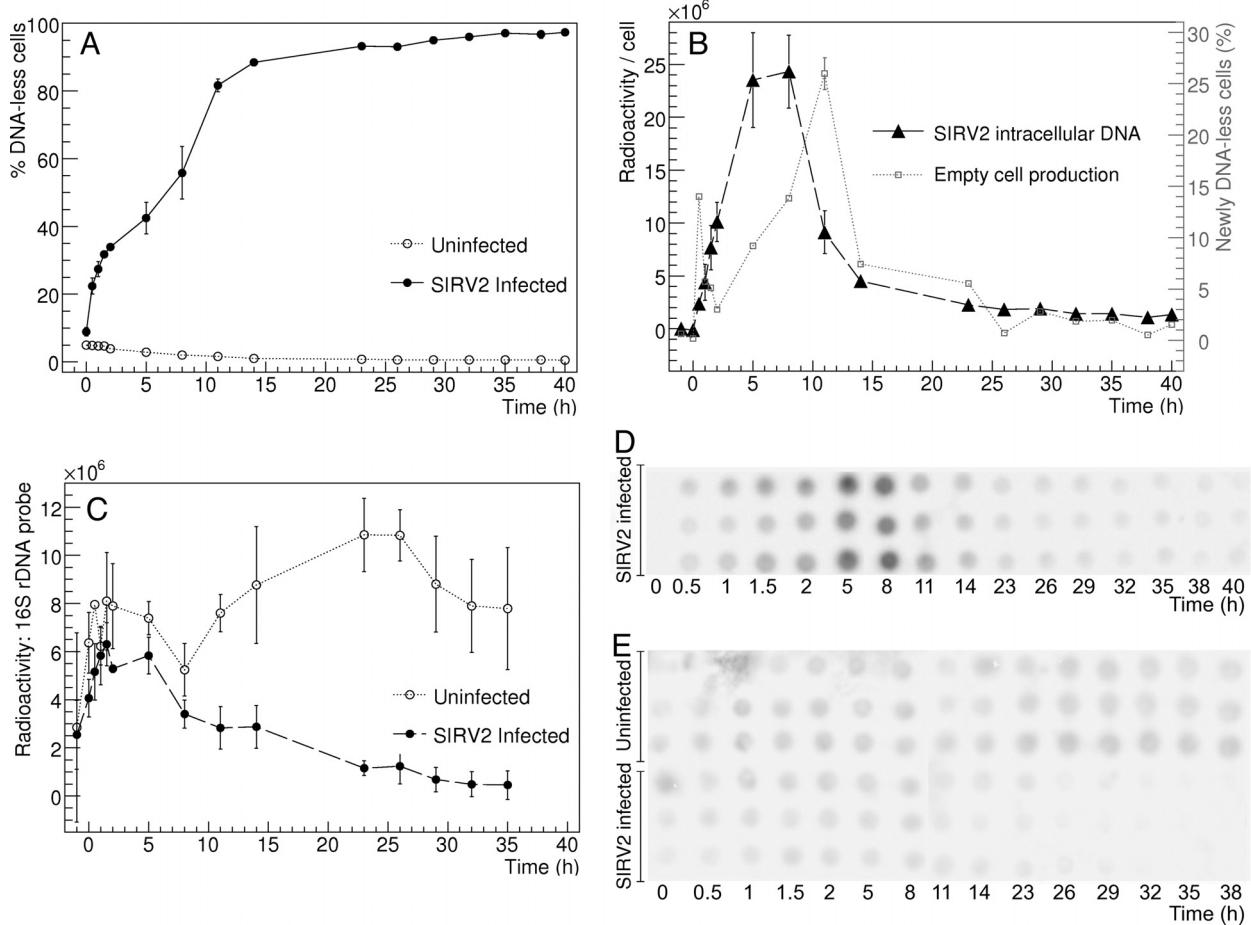
Fig. 3. Links between the kinetics of host chromosome degradation and the SIRV2 infection cycle. Infected cultures and uninfected cultures were launched in triplicates. SIRV2 was added (moi ≈ 15) just after time point 0 h. Averages of 3 infection replicates ± 1 SD are shown in *A–C*. (*A*) Percentage of DNA-less cells in uninfected and infected cultures. The values were obtained by flow cytometry analysis, using data from 2-parameter distributions, gating them as illustrated in Fig. S6B. (*B*) Radioactivity/cell (filled triangles, discontinuous line, left axis), indicative of SIRV2 intracellular DNA in infected cultures, over a time course. Values in arbitrary units were obtained by quantifying the hybridization signal from each spot in the image shown in *D*. The percentage of DNA-less cells appearing between 2 successive time points (empty circles, dotted line, right axis) was also plotted, using the data from *A*. (*C*) Radioactivity/cell indicative of intracellular 16S rDNA amounts in uninfected cultures (empty circles, dotted line) or infected cultures (filled circles, discontinuous line). Values in arbitrary units were obtained by quantifying the hybridization signal from each spot in the image shown in *E*. (*D*) Autoradiogram of hybridization of spots of cells sampled from infected cultures with a SIRV2-specific probe. Each spot corresponds to the same approximate number of cells, based on OD measurements. The time course corresponds to horizontal lines, with the 3 replicates shown vertically for each time point. (*E*) Autoradiogram of hybridization of spots of cells sampled from uninfected and infected cultures with a 16S rRNA gene-specific probe. See *D* for additional explanations.

It is likely that a set of viral genes must control the formation of the VAPs and the generation of the apertures through which the virions are released. The genes might either directly code for the proteins involved or modulate host-encoded mechanisms. The timing of VAP disruption and virus release must also be controlled by virus-encoded functions, such that cell lysis does not occur until the virions have been assembled, as for any lytic virus. Further, host chromosome degradation could also be an active mechanism, encoded by viral genes.

To our knowledge, the virus release mechanism identified here is unprecedented in virus biology. In lytic bacteriophages, the 2 main lysis strategies rely on the direct degradation of peptidogycan, for example, with the holin-endolysin system (20), or on the inhibition of cell wall synthesis (21). Both strategies result in complete cell disruption, and do not involve a modification of the cell envelope in several localized regions, as reported here. Also for eukaryotes there are no reports, to our knowledge, on generation of distinct structures for cell perforation and viral release. Modification of intracellular membranes (endoplasmic reticulum, Golgi complex) does occur as a result of infection with certain eukaryotic

RNA and DNA viruses, but this appears to be linked to viral replication rather than release (22). Recently, alteration of the *Sulfolobus* S-layer as a result of infection with the lytic icosahedral STIV virus was reported (23). It would be highly interesting if viruses that display little similarity in morphology and gene content would share a related mechanism for extrusion from the host cell.

The number and extent of elaborate modifications caused by SIRV2 on the host cell result in a radically transformed cell that can hardly be contemplated as the archaeon *Sulfolobus*. The whole infected cell rather appears to be converted into a complex viral factory, conceptually identical to those built by some eukaryotic viruses inside infected cells. In such cases, the structures of the factory are enclosed by a membrane to exclude cellular organelles. Ribosomes are, however, present, and the factory is dedicated to viral genome replication and virion assembly (24, 25). The eukaryotic viral factories were suggested to constitute the genuine identity of viruses (26), which thus might be considered as a specific type of living organisms (26, 27). A weakness of this concept was the failure to observe viral factories in cells from other domains. SIRV2, as described above, constitutes an example of archaeal virus producing

Fig. 5. Schematic representation of the major stages of SIRV2 infection cycle in the *Sulfolobus* host cell. Times after infection are indicated in hours. At 0 h, 2 chromosomes of *Sulfolobus* are shown in blue. Later between 0 and 8 h, they degrade concomitantly with viral DNA synthesis (gray helices). At 10 h, the VAPs (shown in red) and the intracellular clusters of assembled virions are shown. Finally, at time points between 10 and 14 h, the VAPs open (remains of VAPs shown in red), the cell lyses, and the virions are extruded. The gradual opening out of VAPs (at time points between 10 and 14 h) is illustrated in more details with fragments from the TEM of thin sections.

properties may have been overlooked in other crenarchaeal viruses. The original notion that the carrier state host-virus relationship is dominant in crenarchaea was consistent with the suggestion that this lifestyle would provide a durable intracellular refuge for the virus population in the harsh physico-chemical conditions at which cultured representatives of the Crenarchaeota thrive (T ≈60–90 °C, pH ≈3.0–6.0). In contrast, our findings imply that virus particles can persist in such extreme ecosystems long enough to encounter a new host cell. The SIRV2 virions are well adapted to harsh environments, being extremely stable in various solvents and other inhospitable conditions (14, 28), and almost as stable at 80 °C as phages of mesophilic bacteria are at 37 °C (29). Geothermal environments are extremely heterogeneous, due to a variety of gradients, dynamic movements and changes over time, and viruses may be trapped and preserved for long time periods in different environmental refuges in the absence of potential hosts. Finally, the fact that virus particles are apparently able to travel across the globe (30–32) also suggests that they are robust to variable environmental conditions and display stability over very extended time periods in a variety of biotopes.

Materials and Methods

Virus, Host Strains, and Cultures. Virus stocks were prepared by PEG precipitation of the virions from the culture supernatants, followed by concentration and purification on Cesium chloride density gradients, as described in ref. 33.

The cells of *S. islandicus* LAL14/1 were grown in shaken 50-mL flasks at 78 °C, pH 3.0, in rich medium as described in ref. 14. Colonies were obtained on Gelrite plates as described in ref. 14. To infect cultures, the appropriate volume of virus solution was dialyzed against medium or water on 0.025- or 0.05-μm MF membrane filters (Millipore) and directly added to the liquid cultures during the early exponential phase ($OD_{600\,nm}$ between 0.09 and 0.25). For the time-course experiments [growth kinetics, flow cytometry, dot blot hybridization, and DiBac$_4$ (3) staining], all conditions were tested in triplicates. Six identical 50-mL cultures were launched by dilution of a same preculture. After overnight growth, SIRV2 was added to 3 of them at the appropriate m.o.i.

Titrations, OD, and Fluorescence Measurements. To determine CFU values, culture samples were submitted to serial dilutions and 5 μL of each dilution were spotted on plates. After incubation, the colonies were counted in the last or last 2 positive spots.

To determine the PFU values, the same method was used, except that 5 μL of each dilution were spotted on a fresh cell lawn. When required, the cells were removed by centrifugation before spotting. The cell lawns were prepared as described in ref. 11, using a soft Gelrite overlay. After incubation, single plaques were counted in the last or last 2 positive spots.

ODs were measured in 96-well round-bottomed culture microplates (TPP) in a

Fig. 4. VAPs, detected by SEM and TEM on SIRV2 infected *Sulfolobus* cells. *A1*, *B1*, and *C1* micrographs were obtained by SEM, all other micrographs are TEM images from thin sections. (*A*) Uninfected cells. (*B*) Cells 10 h after infection. (*B2*, *B3*, *B4*, and *B6*) Thin sections in a plane perpendicular to the cell envelope. (*B5*) Thin section in a plane parallel to the cell envelope. (*B1* and *B2*) arrows indicate VAPs. (*B6 Insets*) Details of intracellular virion aggregates, sectioned according to a parallel (up) or perpendicular (down) plane. (*C*) Cells 26 h after infection. (*C2*, *C3*, and *C5*) Thin sections in planes perpendicular to the cell envelope. (*C5*) Disrupted VAP partly detached from cell envelope. (*C4* and *C6*) Thin sections of disrupted detached VAPs in different section planes. (*C2 Inset*) Virion remains inside a lysed cell. (Scale bars, 200 nm.)

a transient viral factory, consisting of the whole transformed infected cell.

Our results show that lytic cycles may be more common for crenarchaeal viruses than previously assumed (7) and that lytic

Multiskan Ascent microplate photometer (Thermo LabSystems) at 595 nm, using 200 μL of the culture.

Flow Cytometry. Sampling and flow cytometry were performed as described in ref. 19; the cells were fixed in 70% (vol/vol) ethanol and the intracellular DNA was stained with mithramycin A and Etd bromide. Samples were analyzed in a A40 Analyzer (Apogee, 25 mW solid-state laser, 405 nm wavelength). *S. islandicus* cell cycle was characterized preliminarily to the study of infected cultures (Fig. S4 and Fig. S5).

For the study of infected cultures, a high moi was used ($\approx$10–15) to obtain as synchronous an infection as possible. At each time point, $OD_{595\,nm}$ was measured and CFU titers were determined to control that the usual growth pattern was obtained.

The distinct cell populations were identified based on the cell size distributions, DNA content distributions and 2D diagrams of cell size and DNA content. The data were gated, and several contours tested, to ensure the robustness of the analysis and of the identified cell populations. The proportion of empty cells over time was computed by gating the 2-D diagrams, similar to what is shown in Fig. S6. In Fig. 3*A*, the total percentage of chromosome-less cells in the culture is shown. In Fig. 3*B*, for the curve related to chromosome-less cells, the difference between the values at time points T and T-1 is plotted, reflecting the production of empty cells between 2 successive time points.

Dot Blot Hybridization. Cells were washed once in cold medium, pelleted by low-speed centrifugation, and stored at −20 °C until further use. Cell pellets were resuspended in Tris-acetate pH 6.0 precooled at 4 °C. The suspension volume was adjusted for cell concentration to be roughly constant in all samples, on the basis of OD measurements. Four microliters of each sample were spotted on Hybond-*n* + nylon membranes (Amersham Biosciences). The membranes were further prepared as for colony hybridization (34).

The probes were generated by PCR. An $\approx$240bp SIRV2 DNA fragment was generated using primer combination [5'-ACATGAAAAGTTAGAGAGATA-CAAACG(3872) 5'-TGGTTACCACTAGCTTCGCTAC(4086)] and a 1,300-bp fragment of the 16S rDNA of *S. islandicus* LAL14/1 was generated by using primers 8aF and 1512uR (35). The probes were [^{32}P]-end-labeled with EasyTide [α-^{32}P]-dATP (PerkinElmer) using a random-primed DNA labeling Kit (Roche Applied Science), according to the manufacturer's instructions.

All hybridization steps were performed at 65 °C in prewarmed solutions. After a minimum of 2 h prehybridization followed by overnight hybridization, both performed in Church Buffer [7% SDS (wt/vol), 0.5 M sodium phosphate, pH 7.2, and 1 mM EDTA], membranes were washed 2 times for 15 min in a solution of 2$\times$ SSC and 0.1% SDS, and 2 times for 15 min in a solution of 0.5$\times$ SSC and 0.1% SDS.

Membranes were exposed on a GP Phosphor Screen (Amersham Biosciences). The screen was scanned in a Molecular Dynamics Storm 860 (Amersham Biosciences). The images were analyzed with the ImageQuantTL software (Amersham Biosciences). After contrast and brightness adjustment, the radioactivity of each spot on the membranes was quantified, using the background removal option (local average). The images of Fig. 3 *D* and *E* were processed with ImageQuantTL for contrast and brightness adjustment and with ImageJ software (http://rsbweb.nih.gov/ij/) for background removal, using the "sliding paraboloid" function.

Transmission Electron Microscopy. Cells were pelleted by low speed centrifugation. The cell pellet was fixed overnight at 4 °C with 2.5% (wt/vol) glutaraldehyde in 20 mM Tris-acetate, pH 6, buffer, postfixed for 1 h in 1% (wt/vol) OsO_4, and dehydrated in a graded series of ethanol dilutions (25% (v/v) to 100% (v/v)). The cells were embedded in an epoxy resin which was polymerized at 60 °C for 48 h. Ultrathin sections ($\approx$60 nm) were cut on a Leica Ultratruc UCT microtome and deposited on carbon-coated copper grids. They were stained for 30 min with 2% (wt/vol) uranyl acetate and for 5 min with 2.5% (wt/vol) lead citrate.

The grids were examined under a JEOL JEM-1010 transmission electron microscope operated at 80 kV. Images were recorded using an Eloise Keen View camera and the Analysis Pro software version 3.1 (Eloïse SARL).

Scanning Electron Microscopy. Cells were pelleted by low-speed centrifugation and fixed overnight at 4 °C with 2.5% (wt/vol) glutaraldehyde in 0,1 M Tris buffer, pH 6. Cells were adsorbed to polylysine-coated coverslips and postfixed 1 h in 1% (w/v) OsO_4 solution. Samples were dehydrated through a graded series of ethanol dilutions (25% (v/v) to 100% (v/v)) and critical point dried using a Leica EM CPD030 device. The dried coverslips were sputtered with 15-nm gold palladium in a GATAN Ion Beam Coater before examination with a JEOL JSM-6700F field emission scanning electron microscope operated at 5 kV. Images were acquired from the upper SE detector (SEI).

Note Added in Proof. At the final stage of preparation of the present publication, a detailed description of the findings reported in ref. 23 was published (36).

1. Prangishvili D, Forterre P, Garrett RA (2006) Viruses of the Archaea: A unifying view. *Nat Rev Microbiol* 4:837–848.
2. Prangishvili D, Basta T, Garrett RA (2008) Crenarchaeal viruses: Morphotypes and genomes. *Encyclopedia of Virology*, eds Mahy BWJ, van Regenmortel MHV (Elsevier, Oxford), pp 587–595.
3. Prangishvili D, Garrett RA, Koonin EV (2006) Evolutionary genomics of archaeal viruses: Unique viral genomes in the third domain of life. *Virus Res* 117:52–67.
4. Janekovic D, et al. (1983) TTV1, TTV2 and TTV3, a family of viruses of the extremely thermophilic, anaerobic, sulfur reducing archaebacterium *Thermoproteus tenax*. *Mol Gen Genet* 192:39–45.
5. Prangishvili D, et al. (2006) Structural and genomic properties of the hyperthermophilic archaeal virus ATV with an extracellular stage of the reproductive cycle. *J Mol Biol* 359:1203–1216.
6. Ortmann AC, et al. (2008) Transcriptome analysis of infection of the archaeon Sulfolobus solfataricus with Sulfolobus turreted icosahedral virus. *J Virol* 82:4874–4883.
7. Prangishvili D, Garrett RA (2005) Viruses of hyperthermophilic Crenarchaea. *Trends Microbiol* 13:535–542.
8. Schleper C, Kubo K, Zillig W (1992) The particle SSV1 from the extremely thermophilic archaeon Sulfolobus is a virus: Demonstration of infectivity and of transfection with viral DNA. *Proc Natl Acad Sci USA* 89:7645–7649.
9. Vestergaard G, et al. (2005) A novel rudivirus, ARV1, of the hyperthermophilic archaeal genus Acidianus. *Virology* 336:83–92.
10. Contursi P, et al. (2006) Characterization of the Sulfolobus host-SSV2 virus interaction. *Extremophiles* 10:615–627.
11. Prangishvili D, et al. (1999) A novel virus family, the Rudiviridae: Structure, virus-host interactions and genome variability of the sulfolobus viruses SIRV1 and SIRV2. *Genetics* 152:1387–1396.
12. Rice G, et al. (2001) Viruses from extreme thermal environments. *Proc Natl Acad Sci USA* 98:13341–13345.
13. Vestergaard G, et al. (2008) Stygiolobus rod-shaped virus and the interplay of crenarchaeal rudiviruses with the CRISPR antiviral system. *J Bacteriol* 190:6837–6845.
14. Zillig W, et al. (1994) Screening for *Sulfolobales*, their plasmids and their viruses in Icelandic solfataras. *System Appl Microbiol* 16:609–628.
15. Peng X, et al. (2001) Sequences and replication of genomes of the archaeal rudiviruses SIRV1 and SIRV2: Relationships to the archaeal lipothrixvirus SIFV and some eukaryal viruses. *Virology* 291:226–234.
16. Bernander R (2007) The cell cycle of Sulfolobus. *Mol Microbiol* 66:557–562.
17. Duggin IG, Bell SD (2006) The chromosome replication machinery of the archaeon Sulfolobus solfataricus. *J Biol Chem* 281:15029–15032.
18. Barrangou R, et al. (2007) CRISPR provides acquired resistance against viruses in prokaryotes. *Science* 315:1709–1712.
19. Bernander R, Poplawski A (1997) Cell cycle characteristics of thermophilic archaea. *J Bacteriol* 179:4963–4969.
20. Wang IN, Smith DL, Young R (2000) Holins: The protein clocks of bacteriophage infections. *Annu Rev Microbiol* 54:799–825.
21. Bernhardt TG, Wang IN, Struck DK, Young R (2002) Breaking free: ''Protein antibiotics'' and phage lysis. *Res Microbiol* 153:493–501.
22. Miller S, Krijnse-Locker J (2008) Modification of intracellular membrane structures for virus replication. *Nat Rev Microbiol* 6:363–374.
23. Fulton J, et al. (2009) Genetics, biochemistry and structure of the archaeal virus STIV. *Biochem Soc Trans* 37:114–117.
24. Novoa RR, et al. (2005) Virus factories: Associations of cell organelles for viral replication and morphogenesis. *Bio Cell* 97:146–172.
25. Fontana J, Lopez-Montero N, Elliott RM, Fernandez JJ, Risco C (2008) The unique architecture of Bunyamwera virus factories around the Golgi complex. *Cell Microbiol* 10:2012–2028.
26. Claverie JM (2006) Viruses take center stage in cellular evolution. *Genome Biol* 7:110.
27. Raoult D, Forterre P (2008) Redefining viruses: Lessons from Mimivirus. *Nat Rev Microbiol* 6:315–319.
28. Steinmetz NF, et al. (2008) Site-specific and spatially controlled addressability of a new viral nanobuilding block: *Sulfolobus islandicus* rod-shaped virus 2. *Adv Funct Mater* 18:1–9.
29. De Paepe M, Taddei F (2006) Viruses' life history: Towards a mechanistic basis of a trade-off between survival and reproduction among phages. *PLoS Biol* 4:e193.
30. Breitbart M, Miyake JH, Rohwer F (2004) Global distribution of nearly identical phage-encoded DNA sequences. *FEMS Microbiol Lett* 236:249–256.
31. Sano E, Carlson S, Wegley L, Rohwer F (2004) Movement of viruses between biomes. *Appl Environ Microbiol* 70:5842–5846.
32. Snyder JC, et al. (2007) Virus movement maintains local virus population diversity. *Proc Natl Acad Sci USA* 104:19102–19107.
33. Bettstetter M, Peng X, Garrett RA, Prangishvili D (2003) AFV1, a novel virus infecting hyperthermophilic archaea of the genus acidianus. *Virology* 315:68–79.
34. Sambrook J, Fritsch EF, Maniatis T (1989) *Molecular Cloning: A Laboratory Manual* (Cold Spring Harbor Lab Press, Plainview, NY), 2nd Ed.
35. Eder W, Ludwig W, Huber R (1999) Novel 16S rRNA gene sequences retrieved from highly saline brine sediments of kebrit deep, Red Sea. *Arch Microbiol* 172:213–218.
36. Brumfeld SK, et al. (2009) Particle assembly and ultrastructural features associated with replication of the lytic archaeal virus Sulfolobus turreted icosahedral virus. *J Virol* 83:5964–5970.

BMC Genomics

RESEARCH ARTICLE   **Open Access**

Exploring short k-mer profiles in cells and mobile elements from *Archaea* highlights the major influence of both the ecological niche and evolutionary history

Ariane Bize[1]* , Cédric Midoux[1,2,3], Mahendra Mariadassou[2,3], Sophie Schbath[2,3], Patrick Forterre[4,5]* and Violette Da Cunha[5]

Abstract

Background: K-mer-based methods have greatly advanced in recent years, largely driven by the realization of their biological significance and by the advent of next-generation sequencing. Their speed and their independence from the annotation process are major advantages. Their utility in the study of the mobilome has recently emerged and they seem a priori adapted to the patchy gene distribution and the lack of universal marker genes of viruses and plasmids.

To provide a framework for the interpretation of results from k-mer based methods applied to archaea or their mobilome, we analyzed the 5-mer DNA profiles of close to 600 archaeal cells, viruses and plasmids. *Archaea* is one of the three domains of life. Archaea seem enriched in extremophiles and are associated with a high diversity of viral and plasmid families, many of which are specific to this domain. We explored the dataset structure by multivariate and statistical analyses, seeking to identify the underlying factors.

Results: For cells, the 5-mer profiles were inconsistent with the phylogeny of archaea. At a finer taxonomic level, the influence of the taxonomy and the environmental constraints on 5-mer profiles was very strong. These two factors were interdependent to a significant extent, and the respective weights of their contributions varied according to the clade. A convergent adaptation was observed for the class *Halobacteria*, for which a strong 5-mer signature was identified. For mobile elements, coevolution with the host had a clear influence on their 5-mer profile. This enabled us to identify one previously known and one new case of recent host transfer based on the atypical composition of the mobile elements involved. Beyond the effect of coevolution, extrachromosomal elements strikingly retain the specific imprint of their own viral or plasmid taxonomic family in their 5-mer profile.

(Continued on next page)

* Correspondence: ariane.bize@inrae.fr; patrick.forterre@pasteur.fr
[1]Université Paris-Saclay, INRAE, PROSE, F-92761 Antony, France
[4]Institut Pasteur, Unité de Virologie des Archées, Département de Microbiologie, 25 Rue du Docteur Roux, 75015 Paris, France
Full list of author information is available at the end of the article

(Continued from previous page)

Conclusion: This specific imprint confirms that the evolution of extrachromosomal elements is driven by multiple parameters and is not restricted to host adaptation. In addition, we detected only recent host transfer events, suggesting the fast evolution of short k-mer profiles. This calls for caution when using k-mers for host prediction, metagenomic binning or phylogenetic reconstruction.

Keywords: Extrachromosomal element, Virus, Plasmid, 5-mer, Codon composition, Multivariate analysis, Signature, Halophily, Hyperthermophily, Host transfer

Background

In the field of nucleic acid sequence analysis, k-mer based methods have greatly advanced in recent years, supported by the advent of next-generation sequencing (reviewed in [1]). As the main advantages, they usually provide reasonable computation durations compared to most traditional alignment-based tools; they are also annotation-independent, and they enable the comparison of incomplete or nonhomologous sequences on a common basis. While they first emerged for practical purposes, their biological significance was subsequently established (reviewed in [2]). In particular, it appeared that the composition of short k-mers is conserved throughout the genome sequence, giving rise to the concept of a k-mer signature, originally based on dinucleotide composition [3]. This finding raised questions regarding the evolutionary significance of this concept and of the underlying mechanisms [4]. Meanwhile, a variety of k-mer-based applications started to proliferate. In the field of environmental microbiology, many k-mer-based tools are dedicated to metagenomic analysis. The k-mer composition of contigs can be used for binning, an important step in the reconstruction of metagenome-assembled genomes (MAGs) (e.g. [5, 6]). It is also used for the taxonomic assignation of sequences (e.g. [7–9]) and to compare different metagenomes by examining distances between k-mer profiles (e.g. [10, 11]). Quite recently, tools specifically dedicated to mobile elements have been developed, that seem a priori adapted to the patchy gene distribution and to the lack of universal marker genes of viruses and plasmids. They enable, for instance, the prediction of viral [12] or plasmid [13] sequences from metagenomes, the assignment of hosts to viruses [14] or plasmids [13], or the classification of viruses [15]. For the study of microbial diversity and evolution, the possibility of using k-mers for phylogenetic [16–19] or evolutionary network [20, 21] reconstruction is also being explored; its application to the detection of horizontal gene transfer (HGT) was proposed more than 10 years ago [22], and a tool for HGT detection within metagenomic data has been recently published [23].

Since these tools are generally based on statistical methods, the results may inevitably contain false or true positives. It is thus necessary to continue exploring k-mer signatures across the genomosphere to establish a framework for interpretation of results obtained with k-mer-based tools. In the present work, we focused specifically on the cells and mobile elements from *Archaea*, one of the three domains of life.

The diversity of viruses and plasmids in *Archaea* is high, with a great number of approved families compared to the relatively low number of isolated elements [24–26]. This provides an interesting case for comparing k-mer composition among hosts and viruses. In particular, viruses of extreme thermophilic crenarchaea are highly diverse. They often belong to *Archaea*-specific viral families, with unusual morphotypes. In the class *Halobacteria,* head-and-tail viruses belonging to *Caudovirales* are abundant and are predominant in hypersaline environments, which are dominated by haloarchaea [27]. While *Caudovirales* is a cosmopolitan order of viruses (the most abundant order infecting *Bacteria* [28]), *Halobacteria* members are also infected by *Archaea*-specific viral families, such as *Pleioipoviridae*. Many archaeal plasmids have not yet been classified into well-defined families; however, several families of plasmids have been defined according to plasmid size, replication mode, and genomic content (reviewed in [25]).

Among archaea, there are no known pathogens for humans, plants or animals, so there is no overrepresentation bias linked to pathogens in the databases. Other biases are, however, present: the mobile elements from several archaeal taxonomic groups (orders or even phyla,) are very poorly represented in public databases, so the view on global diversity remains incomplete. In addition to the diversity of their mobile elements, archaea constitute an interesting case in terms of adaptation or loss of adaptation to extreme environments, which has played an important role in their evolutionary history [29].

Several studies on k-mer signatures previously included archaeal genomes. For instance, in 1999, Campbell et al. [30] studied genome signatures across a wide phylogenetic range, encompassing bacteria, archaea, plasmids and mitochondrial DNA. This work highlighted the similarity of signatures between hosts and plasmids, the lack of consistent signatures among thermophiles and, finally, the high signature divergence among five archaeal genomes available at that time. In 2006, van Passel et al. [31] showed the difference in dinucleotide

composition between hosts and plasmids in *Archaea* and *Bacteria*. In 2008, Bohlin et al. [32] obtained a similar trend by using 4-mers and zero-order Markov models. The same authors studied the composition of bacterial and archaeal genomes in 2- to 8-mers, with 44 archaeal genomes among the 581 analyzed genomes. They observed a higher variability in AT-rich and host-associated genomes compared to GC rich or free-living archaea and bacteria [33].

Currently, the number of publicly available genomes has greatly increased, warranting a new study of signatures across the domain *Archaea*. Selecting close to 600 cellular, viral and plasmid genomes, we applied metrics based on short k-mer profiles to understand how mobile elements are distributed with respect to their hosts in the profile landscape. We used multivariate and statistical analyses to explore the dataset structure and identify some key structuring factors, namely, the taxonomic classification, the genomic GC content, the ecological niche and, for mobile elements, the taxonomy of the host. Moreover, we examined whether 5-mer profiles enable the detection of singular evolutionary trajectories, such as host transfers, among mobile elements. We also searched for 5-mer signatures for halophily and hyperthermophily in *Archaea*.

Results

The 5-mer profiles of archaeal genomes are influenced by the taxonomy and GC content

Before focusing on extrachromosomal elements, we first analyzed the 5-mer profile distribution of archaeal cellular genomes. We selected 239 archaeal genomes, focusing mainly on taxonomic groups for which many plasmids and/or viruses have already been classified into distinct families: *Halobacteria*, *Sulfolobales*, *Thermococcales* and a few other groups of *Euryarchaeota* and *Crenarchaeota*.

We first noticed from the dendrogram obtained by hierarchical clustering that the sequences were distributed into two main clusters according to GC content values, suggesting a major influence of the GC content on the k-mer distribution (Fig. 1a). The most GC-rich cluster (Fig. 1a, letter c) exclusively included *Halobacteria* members, consistent with the fact that *Halobacteria* have a high genomic GC-content, 63.28% ± 4.29 SD on average in our dataset. At the other extreme, the less GC-rich cluster (Fig. 1a, letter b) comprised only Group I methanogens (*Methanococcales* and *Methanobacteriales*), except for one Group II *Methanosarcinales* genome.

We also identified taxonomy as an important factor, and many clusters were dominated by a single taxonomic group (Fig. 1a). In particular, all members of the class *Halobacteria* were located in a single cluster (Fig. 1a, letters c) with only two exceptions, corresponding to

the two *Haloquadratum walsbyi* genomes (order *Haloferacales*). Similarly, 33 out of 37 members of the order *Methanosarcinales* were gathered in a single cluster (Fig. 1a, letter d). Members of the order *Sulfolobales* were divided into a major cluster (31 genomes out of 39) and a minor cluster (8 genomes out of 39) (Fig. 1a, letters e and f, respectively). The latter corresponded to the *Metallosphaera* genomes, which have a higher GC content than the other *Sulfolobales* genomes. The 17 members of the order *Methanococcales* were divided into two neighboingr clusters (Fig. 1a, within cluster b), which also included several *Methanobacteriales* members, which are Group I methanogens, similar to *Methanococcales* members.

We did not observe similar clustering for *Methanobacteriales*, *Thermococcales*, *Thermoproteales* and *Desulfurococcales.* In such cases, archaea belonging to the same order were distributed into several clusters, sometimes distant across the dendrogram. However, at the local scale, small- to medium-sized clusters enriched in one of these orders were still visible, such as a medium-sized cluster comprising exclusively *Thermococcales* members (23 genomes out of 39) (Fig. 1a, letter g).

To quantify the relative contribution of the taxonomy and of the GC content to the 5-mer composition, we performed a permutational multivariate analysis of variance (PERMANOVA) (Additional file 1). We applied PERMANOVA to the pairwise Euclidian distance matrix computed from the 5-mer profiles, which we will denote as D_{5_cells} hereafter. Among the three considered taxonomic levels (phylum, order, genus), order had the strongest influence; it alone explained 75.94% of the cell profile dissimilarity variance (model: D_{5_cells} ~ Genus), compared to 7.06% for phylum (D_{5_cells} ~ Phylum) and 17.74% for genus, when the effect of the phylum and order was first removed (D_{5_cells} ~ Phylum*Order*Genus).

Notably, the GC content alone contributed almost as much to the variance (69.10%, D_{5_cells} ~ GC%) as the taxonomic rank of the order (D_{5_cells} ~ order). These last two factors appeared to be highly dependent, explaining 56.71% of the cell dissimilarity variance (D_{5_cells} ~ order*GC%) in an indistinguishable manner.

Despite the strong influence of the taxonomy, the global topology of the dendrogram obtained by hierarchical clustering was inconsistent with the phylogeny of archaea. While *Sulfolobales* belongs to the *Crenarchaeota* phylum, its main cluster grouped with a cluster dominated by Group I methanogens from the *Euryarchaeota* phylum. Moreover, within the major *Halobacteria* cluster, archaea from the three orders *Haloferacales*, *Halobacteriales* and *Natrialbales* were interconnected (especially due to *Halobacteriales*), showing the blurring of phylogenetic information.

Fig. 1 Dendrograms based on 5-mer frequencies for archaeal cells and mobile elements. **a**. Archaeal cells. **b**. Archaeal viruses and plasmids
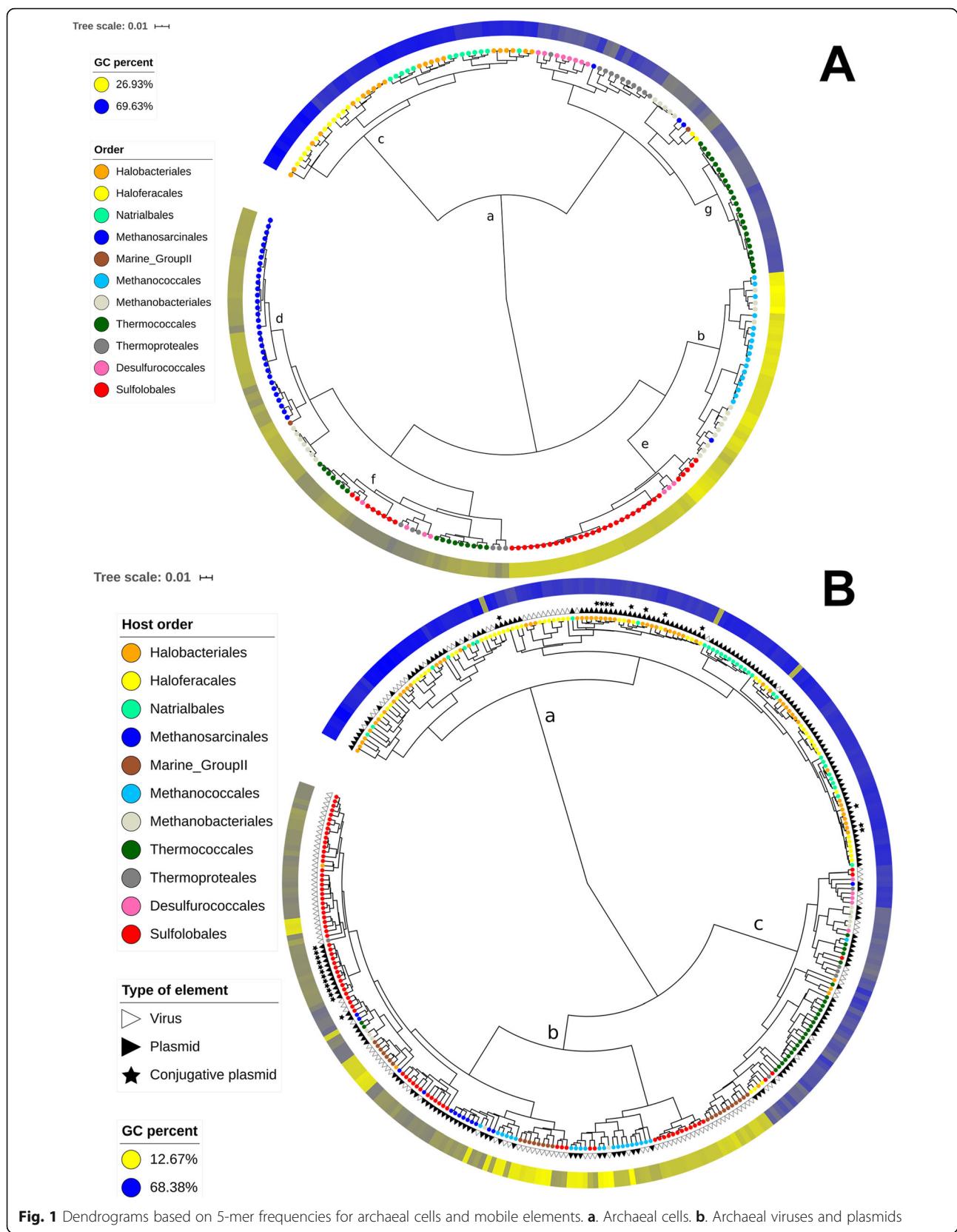
A strong link between the ecological niche and the 5-mer composition of archaeal cellular genomes

Many archaea thrive in extreme conditions, and adaptation to such specific environments has played an important role in their evolution [34, 35]. We therefore assumed that major properties of the environmental niches could be another important factor underlying the 5-mer composition among archaea. We focused on salinity and temperature and defined 8 "Niche" categories. All *Halobacteria* members were categorized as "halophile". The remaining archaea were labeled according to 7 qualitative growth temperature categories, ranging from "weak mesophile" to "extreme hyperthermophile" (Additional File 2), based on the BacDive database [36] and on the literature, e.g. [37].

The clustering pattern was clearly influenced by the "Niche" categories (Fig. 2 a). Among the 6 main clusters of the dendrogram for cells (Fig. 2 a, clusters a to f), cluster b was largely dominated by thermophiles to extreme hyperthermophiles. Cluster c was dominated by extreme thermophiles, corresponding mostly to *Sulfolobales* members. Cluster d comprised exclusively thermophiles to extreme hyperthermophiles. Finally, clusters e and f were dominated by weak mesophiles and mesophiles, although a small patch of hyperthermophiles was visible in cluster e. *Sulfolobales* comprises exclusively acidophilic members, which could explain their specific signature compared to other thermophilic/hyperthermophilic extrachromosomal elements. Indeed, cytoplasmic pH regulation does not fully compensate for the decrease in intracellular pH in acidic environments: the intracellular pH in acidophiles is higher by approximately 3 to 4 points than that of the surrounding acidic environment, but on the whole, it is still lower than that in neutrophiles [38]. It has previously been suggested that acidophilic archaea and bacteria have purine-poor codons in their long genes [39]; however, the effects of acidophily on compositional features seem to have been studied less than the adaptation to high temperatures.

Based on PERMANOVA, the "Niche" categories explained 64.17% of the dataset variance (D_{5_cells} ~ Niche). Although this percentage is lower than that explained by the taxonomic rank of order (namely, 75.94%), it is still very high. As anticipated, the GC content, taxonomic rank and "Niche" had a high level of dependency (Additional file 1, D_{5_cells} ~ Niche*Order*GC%). In particular, the last two factors explained 60.56% of the cell profile dissimilarity variance in an indistinguishable manner (D_{5_cells} ~ Order*Niche), consistent with the strong links between the ecological niche and the evolutionary history in *Archaea.* Finally, we noticed that a model combining the genomic GC content, ecological niche and taxonomy (order rank) explained almost all the cell dataset variance, namely, 95.48% (Additional file 1, D_{5_cells} ~

Niche*Order*GC%). Overall, a limited number of factors are therefore sufficient to explain the differences in 5-mer composition of the archaeal cell genomes included in our study.

The extrachromosomal element profiles are also influenced by the GC content and host taxonomy, with higher profile dispersion

We analyzed the 5-mer composition of archaeal plasmids and viruses (extrachromosomal elements) with a similar approach. The obtained dendrogram was divided into two major clusters. One of them (Fig. 1b, letter a), corresponded to elements with the highest GC contents, including nearly all 154 *Halobacteria* mobile elements, except for 9. The second cluster, with the lowest GC content, was divided into two subclusters (Fig. 1b, letters b and c). Subcluster b was dominated by *Sulfolobales* extrachromosomal elements but also included a significant number of extrachromosomal elements from *Methanococcales, Methanosarcinales* and *Marine Group II.* Subcluster c was dominated by *Thermococcales* extrachromosomal elements but also comprised significant numbers of extrachromosomal elements from *Marine Group II, Desulfurococcales, Thermoproteales* and *Methanobacteriales.*

Compared to the pattern obtained for cells, visual inspection showed that the extrachromosomal elements, categorized according to the taxonomy of their host, had a more intertwined distribution, except for viruses and plasmids of *Halobacteria.* Consistent with this observation, the taxonomy of the host at the order level explained only 57.36% of the extrachromosomal element dissimilarity variance (Additional File 3, D_{5_mobile} ~ Host order), compared to 75.94% for the cells. As in the case of cellular genomes, the rank of their hosts appeared more informative at the order level than at the phylum or genus level (Additional File 3, D_{5_mobile} ~ Host Phylum*Host Order*Host Genus).

The less consistent pattern obtained for extrachromosomal elements compared to cells could theoretically reflect more frequent genetic exchanges between extrachromosomal elements present in hosts belonging to different taxonomic groups. However, this does not seem to be the case. For instance, while several cases of host transfers between *Thermococcales* and *Methanococccales* plasmids have been previously documented [25], *Methanococcales* extrachromosomal elements clustered mostly with those of *Sulfolobales* rather than with those of *Thermococcales* in our analysis. Another hypothesis to explain such a complex pattern for extrachromosomal elements could be the influence of their GC content. Indeed, extrachromosomal element genomes harbor, in many cases, a distinct average GC content compared to their hosts (Additional File 4). We noticed that the extent and even

Fig. 2 Mapping of temperature and salinity-related growth conditions on the archaeal cell and mobile element dendrograms. **a**. Archaeal cells. **b**. Archaeal viruses and plasmids
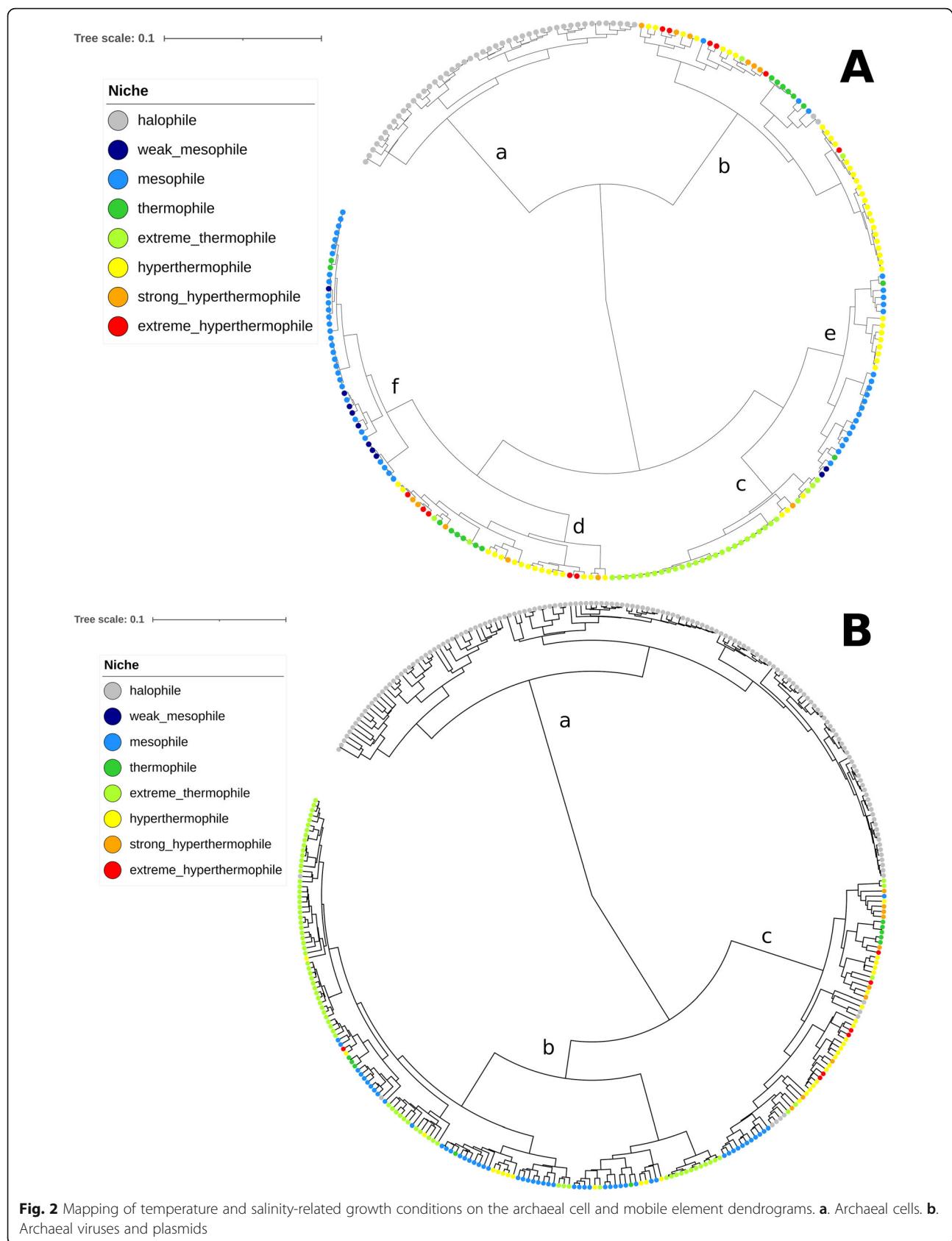
the direction of these shifts in GC content varied greatly according to the host's taxonomy (at the order level) and to the type of extrachromosomal element (Additional File 4). Since the GC content had a strong global influence on the obtained pattern (45.13% of the variance, Additional File 3, $D_{5\_mobile} \sim$ GC%), these shifts in GC content could greatly contribute to the more complex pattern obtained for archaeal extrachromosomal elements compared to that obtained for archaeal cells.

Similar to cells, the host taxonomy (at the order level) and the genomic GC-content were highly interdependent factors for extrachromosomal elements (Additional File 3): 39.71% of the dissimilarity variance was explained indistinguishably by these two factors ($D_{5\_mobile} \sim$ Host Order*GC% and $D_{5\_mobile} \sim$ GC% * Host Order). Interestingly, the taxonomic classification of viruses and plasmids was by far the most influential factor, alone explaining 68.30% of the extrachromosomal element dissimilarity variance (Additional File 3, $D_{5\_mobile} \sim$ Family). This could be due partly to the high number of viral and plasmid families in the dataset (60 compared to only 11 different host orders), which must support a better fit of the model. However, this finding also suggests that individual viral and plasmid families could have a specific 5-mer composition.

The extrachromosomal element family and the taxonomy of their hosts at the order level were strongly dependent, since 51.90% of the extrachromosomal element dissimilarity variance was explained indistinguishably by one of the factors (Additional File 3, $D_{5\_mobile} \sim$ Host Order*Family and $D_{5\_mobile} \sim$ Family*Host Order). This could reflect the fact that the host range of a given plasmid or viral family is limited. The fact that viruses and plasmids coevolved with their hosts and that they were not frequently transferred to new hosts from other orders could explain this limitation.

A significant but weaker influence of the ecological niche on the 5-mer composition of archaeal extrachromosomal elements

We used the same "Niche" categories and method to analyze plasmids and viruses of archaea (Fig. 2 b). As already identified above (Fig. 2 b), extrachromosomal elements from halophiles grouped together (cluster a), with a very limited number of exceptions. The viruses and plasmids from extreme thermophiles, corresponding mostly to *Sulfolobales*, tended to group with mesophilic extrachromosomal elements, in cluster b. By contrast, most other thermophilic to extremely hyperthermophilic extrachromosomal elements were in a separate group (cluster c).

The consistency of the 5-mer profile distribution with the "Niche" was lower than that for cells: the "Niche" explained 50.12% of the dissimilarity variance from the

extrachromosomal element profiles (Additional File 3, $D_{5\_mobile} \sim$ Niche). As we observed for cells, the information about the "Niche" was almost fully included in the host taxonomic classification, since the "Niche" explained only 1.16% of the extrachromosomal element dataset variance when the influence of host taxonomy was first removed (Additional File 3, $D_{5\_mobile} \sim$ Host Order*Niche). A statistical model combining the genomic GC content, the ecological niche and the taxonomy of the host explained 70.85% of the profile dissimilarity variance (Additional File 3, $D_{5\_mobile} \sim$ Niche*Host Order*GC%); adding the extrachromosomal element family as a variable to the model enabled us to reach 89.29% of explained variance (Additional File 3, $D_{5\_mobile} \sim$ Niche*Host Order*GC% and $D_{5\_mobile} \sim$ Niche*Host Order*Family*GC%).

A clear 5-mer signature for halophily and a weaker signature for hyperthermophily

Considering the strong association between the ecological niche and the 5-mer profile distribution, we decided to identify some of the most discriminant 5-mers between halophilic and nonhalophilic entities on the one hand, and between hyperthermophilic versus nonhyperthermophilic entities on the other. For this purpose, in each case, we applied partial least square discriminant analysis (PLS-DA) to archaeal cells and extrachromosomal element profiles separately. In each situation, we retained the ten most discriminant 5-mers (Table 1, Additional file 5).

For both cells and extrachromosomal elements, the separation according to the salinity-related growth properties was very strong, consistent with the hierarchical clustering results (principal component analysis (PCA) and PLS-DA, Additional files 6, 7, 8, 9). Consistent with this, the average frequency of the ten most discriminant 5-mers was significantly different between halophiles and nonhalophiles (Mann-Whitney-Wilcoxon test, $p < 0.01$, Additional files 10 and 11). Considering the marked separation between halophilic and nonhalophilic entities (Fig. 3, Additional Files 6, 7, 8, 9), many additional 5-mers likely have significantly different frequencies between both groups. The ten most discriminant 5-mers were more abundant in halophilic archaea or in their extrachromosomal elements, except for one 5-mer, which was more abundant in nonhalophilic archaea.

The signatures of halophilic cells and extrachromosomal elements were expected to be similar, since most *Halobacteria* extrachromosomal elements grouped with *Halobacteria* cells in a joint dendrogram (Fig. 3). Indeed, each of the ten discriminant 5-mers identified for the cells also had significantly different frequencies within extrachromosomal elements (Mann-Whitney-Wilcoxon test, $p < 0.01$). However, only 4 out of the 10 most

Table 1 Sets of 10 most discriminant 5-mers identified by PLS-DA

	Archaeal cells	Archaeal mobile elements
Halophiles high frequency 5-mers	**CGAAC**, **GTTCG**, **ACCGA**, GACCG, CGGTC, **TCGGT**, GTGAC, GTCAC, TCGAC	**GTTCG**, **ACCGA**, TTCGA, **CGAAC** TCGAA, **TCGGT**, TCGGA, CGAGT, TCCGA, ATCGA
Halophiles low frequency 5-mers	TGAAG	–
Hyperthermophiles high frequency 5-mers	TCAAC, GTTGA, **AGCTT**, **AAGCT**	TTTGG, GAGCT, AGCTC, **AAGCT**, **AGCTT**, TTGAG, (TTGGA), GCCAA, (TCCAA)
Non-hyperthermophiles low frequency 5-mers	TCAGA, TCTGA, TCAGT, ACTGA, CAGAT, ATCTG	CGAAT

Bold characters: in each table line, most discriminant 5-mers shared between cells and mobile elements, for a considered niche category. In parenthesis: statistically non-significant frequency differences based on a t-test ($p \geq 0.01$), in a considered niche category

discriminant 5-mers identified for halophiles were common between cells and mobile elements (Table 1, Additional file 5). The 10 most discriminant preferred 5-mers in haloarchaea were GC-rich, as expected (Table 1, Additional file 4).

To identify discriminant 5-mers according to the growth temperature, we removed all *Halobacteria* representatives from the dataset and classified the remaining elements into two categories: elements with growth temperatures below 80 °C (weak mesophiles to extreme thermophiles) and those with growth temperatures above 80 °C (hyperthermophiles to extreme hyperthermophiles).

For archaeal cells, hyperthermophiles and nonhyperthermophiles separated quite well based on PCA and PLS-DA (Additional files 12 and 13). The 10 most discriminant 5-mers identified by PLS-DA all had significantly different frequencies between the two groups (Mann-Whitney-Wilcoxon test, $p < 0.01$, Additional file 14). However, the differences were less pronounced than those for halophiles.

For the extrachromosomal elements, with the same defined categories, the separation between the two temperature groups was less clear, as assessed by PCA (Additional file 15); but the barycenters were
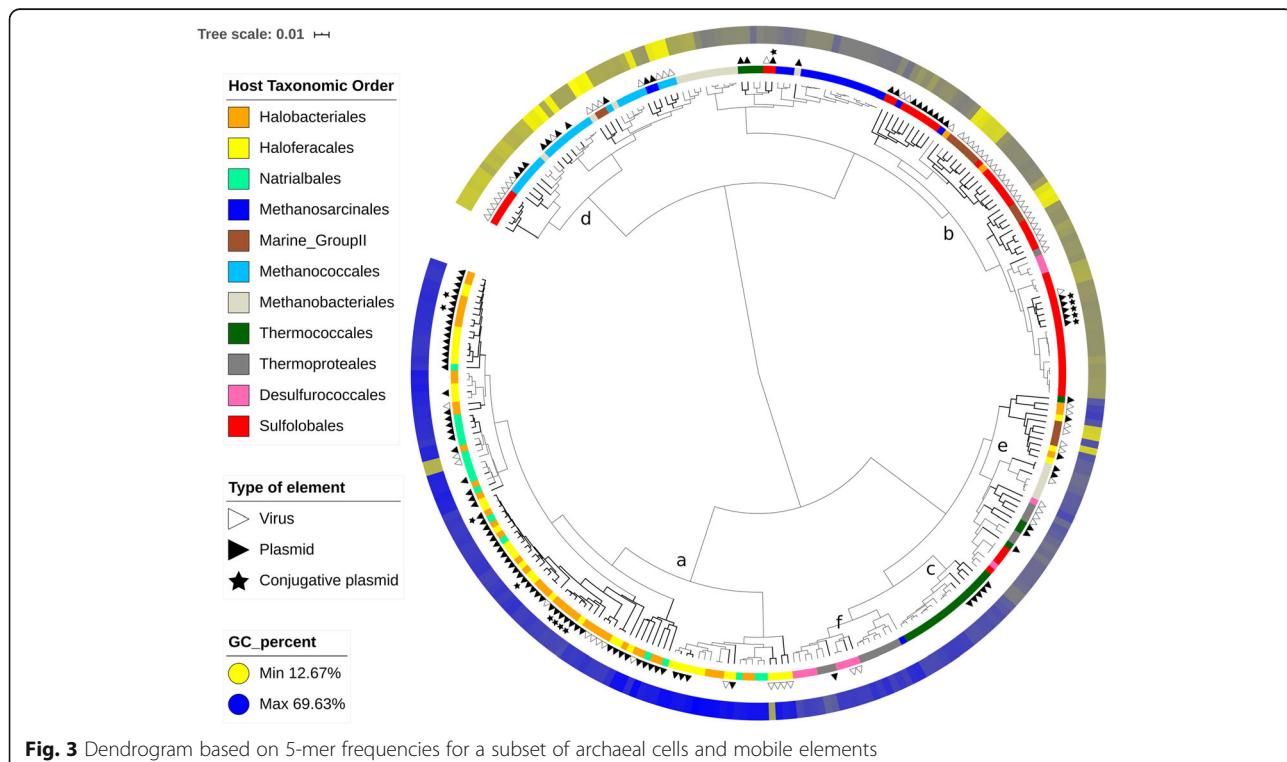
Fig. 3 Dendrogram based on 5-mer frequencies for a subset of archaeal cells and mobile elements

still quite distant from each other. Eight of the 10 most discriminant 5-mers identified by PLS-DA (Additional file 16) had significantly different frequencies between the two groups (Mann-Whitney-Wilcoxon test, $p < 0.01$, Additional File 17). Only two of them were shared with those identified for cells, with higher frequencies in hyperthermophiles than in the lower growth temperature group. Seven of the 10 most discriminant 5-mers identified for the cells also had significantly different levels in extrachromosomal elements (Additional file 18), indicating that the signatures of archaeal cells and extrachromosomal elements with respect to hyperthermophily are similar without being strictly identical.

The signal for hyperthermophily was much weaker overall than that for halophily. In addition, most hyperthermophiles in our dataset were from the orders *Desulfurococcales*, *Thermoproteales* and *Thermococcales*. The few others (e.g., some *Sulfolobales* and *Methanococcales* members) tended to be located within the lower-temperature group, as assessed by PCA. It is therefore not clear whether the identified discriminant 5-mers constitute a general signature for hyperthermophilic archaea.

Codon frequencies influence 3-mer and 5-mer profile distributions

It has been previously shown that amino acid usage and codon frequencies vary according to environmental conditions, particularly for archaea and extreme environments [29, 35, 40, 41]. Since the proportion of coding regions is high in archaeal genomes, it is likely that their 5-mer composition is somehow correlated with the codon frequencies. To evaluate this hypothesis, we focused only on the genomes for which the positions of coding regions were available in public databases, namely 238 out of 239 archaea and 288 out of 345 archaeal viruses and plasmids, in our dataset (Additional file 2).

We first compared, for halophiles and hyperthermophiles, the 10 most discriminant 3-mers of the whole-genome sequences to their 10 most discriminant codons (Table 2). In each case, several of the most discriminant codons were also present among the most discriminant 3-mers of the whole genome sequences (Table 2, underlined words), which supported, as expected, the link between codon frequencies and 3-mer composition in archaea and their extrachromosomal elements.

Table 2 Sets of 10 most discriminant codons and 3-mers identified by PLS-DA

	Discriminant codons with high frequency; corresponding amino acids	Discriminant codons with low frequency; corresponding amino acids	Discriminant 3-mers (whole genome) with high frequency	Discriminant 3-mers (whole genome) with low frequency
Halophilic archaea	**CGA**, GAC, **CGC**, GTC, CGT, CAC, CGG, **GCG**, **TCG**, **CCG** **R,** D, **R**, V, R, H, R, **A, S, P**	-	**GAC**, **GTC**, **CGA**, **TCG**, **ACG**, **CGT**	CTT, **AAG**, AGG, CCT
Halophilic mobile elements of archaea	**CGC**, **GCG**, **CCG**, **CGA**, **TCG** **R, A, P, R, S**	TAG, TTA, TAA, CTA, TTT Stop, L, Stop, L	**TCG**, **CGA**, CGT, **GTC**, **CCG**, **ACG**, **GAC**, CGG,	**AAG**, CTA
Hyperthermophilic archaea	**AGC**, GAG, **GCT**, (TCT), AGA **S**, E, **A**, (S), R	ATC, ACT, **CAG**, CTG, TTC I, T, **Q**, L, F	**AGC**, **GCT**, GAG	CAA, TTG, (CTG), (CAG), ATC, GAT, (**AAC**)
Hyperthermophilic mobile elements of archaea	**AGC**, (**TTG**), **GCT**, AGG **S**, (L), **A**, R	CAC, **CAG**, TAC, CAT, (TTA), **AAC** H, **Q**, Y, H, (L), N	TGG, **AGC**, GAG, **GCT**, CTC, (CAG), **TTG**	ACA, (**CAA**), **AAC**

Underlined: most discriminant words shared between codons and 3-mers in whole genomes, for a considered niche category. Bold characters: most discriminant words shared between cells and mobile elements, for a considered niche category. In parenthesis: statistically non-significant frequency differences based on a t-test ($p \geq 0.01$), in a considered niche category

The 10 most discriminant preferred codons in haloarchaea were GC rich, as expected (Table 2, Additional file 4). They encoded arginine (R) (through 4 different codons), aspartic acid (D), valine (V), histidine (H), alanine (A), serine (S) and proline (P). Contrary to previous results on amino acid composition [35, 41, 42], we did not detect preferred codons for glutamic acid (E) [35, 42, 43] and threonine (T) [35]. D and V have been repeatedly identified as preferred amino acids in halophiles [35, 41, 42]. A higher abundance of R in halophiles has been reported when comparing halophiles to thermophiles [42] or in specific cases [35, 43]; an increase in H has also been documented [41]. The enrichment in R probably compensates for the avoidance of K [35, 41–43]: this latter amino acid is similar to R, a basic, polar and positively charged amino acid; however, the side chains of R can bind more water molecules than those of K. In our study, the identification of 4 preferred codons coding for R could therefore partly result from a selection process operating at the protein level.

Our results on the most discriminant codons for hyperthermophilic archaea can be compared with those from [44], for the identification of differentially abundant codons between thermophilic and mesophilic archaea and bacteria. A limited number of codons identified in [44] were also retrieved in our analysis (Table 2): GAG (E), AGA (R) and AGG (R), which were more frequent in hyperthermophilic archaea or in their extrachromosomal elements; CAG (glutamine, Q), which was less frequent in both hyperthermophilic archaea and their extrachromosomal elements; and finally CAT (H), which was less frequent in hyperthermophilic extrachromosomal elements. However, the majority of the most discriminant codons for hyperthermophily that we identified (Table 2) were not detected as differentially abundant in [44]. In archaea and bacteria, the nature of the discriminant codons is likely influenced by proteomic adaptation to temperature [45]. In 2007, the amino acids isoleucine (I), V, tyrosine (Y), tryptophan (W), R, E and leucine (L) were proposed as universal markers for the optimal growth temperature in prokaryotes (IVYWREL) [45]. These amino acids were already identified to some extent prior to 2007 [44, 46, 47] . Although not present in the IVYWREL set, K was identified by other authors as a preferred amino acid [44, 47]. By contrast, thermophiles tend to be impoverished in at least Q, T and H [44, 46]. Our results on most discriminant codons showed a certain consistency with these established amino acid signatures, since 6 of them translated to one of these amino acids (Table 2, preferred codons translating to E or L and avoided codons translating to Q or H). In our analysis, some codons translating to S, R, and A appeared to be preferred in both hyperthermophilic archaea and their extrachromosomal elements. Finally, 3 avoided codons corresponded to the preferred amino acids I, L, and Y (Table 2), showing the difficulty of fully reconciling the signature at the codon level from this study to the amino acid signature from previous studies.

Examining the influence of codon frequency on the 5-mer profiles is less straightforward, since each 5-mer includes three overlapping 3-mers. We thus implemented a different approach to obtain a global estimate of this influence. We first established another type of 5-mer-based profile, taking into account the codon composition. For each element, this new profile was based on the concatenated coding regions. For each 5-mer, the profile value consisted of an exceptionality score, reflecting how unexpectedly frequent or rare this 5-mer is, considering the codon composition of the sequence. This other type of profile therefore does not necessarily highlight frequent 5-mers. Rather, it highlights 5-mers that have an unexpected frequency in the studied sequence, given the codon frequencies. After obtaining the profiles, we calculated the distance matrices ($D_{5\_cells\_e}$ and $D_{5\_mobile\_e}$) before applying PERMANOVA. The influence of the niche was much lower on this new type of profile, decreasing from 64.22 to 41.75% for archaeal cells ($D_{5\_cells} \sim$ Niche and $D_{5\_cells\_e} \sim$ Niche) and from 51.35 to 17.81% for mobile elements ($D_{5\_mobile} \sim$ Niche and $D_{5\_mobile\_e} \sim$ Niche). The strong influence of the ecological niche on the 5-mer profiles is thus significantly but not exclusively explained by codon frequencies.

Joint analysis of plasmid, viral and cellular genomes from *Archaea* highlights the influence of coevolution and of the extrachromosomal element families on 5-mer profiles

To visualize a dendrogram encompassing both archaeal cells and their extrachromosomal elements, we created a smaller subset by randomly selecting approximately half of the sequences in each category (cell, virus and plasmid) and we jointly analyzed the corresponding 5-mer profiles. This subset comprised a total of 296 genome sequences, of which 119 were from cells, 106 were from plasmids and 71 were from viruses.

Based on hierarchical clustering (Fig. 3) and at the global scale, viruses and plasmids did not form a separate cluster. Rather, they tended to group with archaea sharing the same taxonomy as their hosts. This was best evidenced by the class *Halobacteria*, for which most members and their associated extrachromosomal elements were grouped in a single specific cluster (Fig. 3, letter a). This trend was also visible for the orders *Sulfolobales*, *Thermococcales*, and *Methanococcales* (Fig. 3, clusters b, c, d, respectively). It was less clear for the orders *Methanobacteriales*, *Thermoproteales* and *Desulfurococcales*, as well as *Marine Group II*, which were more dispersed at various locations of the dendrogram.

Bize *et al. BMC Genomics*   (2021) 22:186

Page 11 of 22

However, several host-virus or host-plasmid associations were still visible in some of these smaller isolated clusters (e.g., for *Methanobacteriales* and *Desulfurococcales*, Fig. 3, letters e and f, respectively). While this trend of 5-mer profile similarity between extrachromosomal elements and hosts has its exceptions, it still highlights the influence of the coevolution between hosts and their mobile elements on their short k-mer composition.

Within each of the 4 abovementioned groups for which the association was the strongest (the class *Halobacteria* and orders *Sulfolobales*, *Thermococcales*, and *Methanococcales*), the cell and extrachromosomal element branches were not fully intertwined. Rather, they tended to form several aggregates rich in either cells or extrachromosomal elements. This is particularly well illustrated by the case of the *Sulfolobales* order (Fig. 3, letter b).

Importantly, although the 5-mer profiles of archaeal extrachromosomal elements are strongly influenced by the coevolution with the hosts, they also retain a specific component, likely due to their different nature. To better understand the nature of these interactions, we focused on *Halobacteria* and *Sulfolobales*, for which many families of extrachromosomal elements, either plasmids or viruses, have already been defined.

Megaplasmids and other mobile elements from *Halobacteria* have 5-mer profiles distinct from those of *Halobacteria* cells

The class *Halobacteria* comprises exclusively halophilic archaea that thrive in high-salt environments. We focused specifically on the sequenced mobile elements of *Halobacteria* members, which are numerous and diverse [25, 26, 48, 49]. Our dataset comprised 53 cellular *Halobacteria* genomes, as well as 118 plasmids and 36 viruses of hosts from the orders *Halobacteriales*, *Haloferacales*, and *Natrialbales* (Additional file 19). A particularity of *Halobacteria* is the abundance of megaplasmids, considered here as plasmids longer than 150 kb (51 represented in our dataset), and of large plasmids, with sizes ranging from 100 to 150 kb (23 represented in our dataset). The 44 other plasmids had sizes ranging from 1.1 kb to 96 kb. There is currently a scientific debate on the nature of megaplasmids. Indeed, some of them encode essential genes and could hypothetically be currently evolving into chromosomes [50]. In our dataset, 5 distinct elements were classified as second chromosomes according to public databases. Associated with the *Haloarcula* or *Halorubrum* genus, these elements had sizes ranging from 288 kb to 526 kb.

Using PERMANOVA, it appeared again that the genomic GC content and the taxonomic family together explained an important proportion of the 5-mer profile dissimilarity variance of extrachromosomal elements, namely, 55.52% (Additional file 20, $D_{5\_mobile\_halo}$ ~ GC%*Family). By contrast, the taxonomy of the host explained only a very limited proportion of the variance, 5.28%, consistent with the loss of phylogenetic signal from the hosts within the class *Halobacteria* (Additional file 20, $D_{5\_mobile\_halo}$ ~ Host order*Host genus).

The pattern obtained by hierarchical clustering was quite complex (Fig. 4a, Additional file 21). It still evidenced the presence of cell-rich clusters (Fig. 4a, clusters a1 to a4), while other clusters were rich in megaplasmids and large plasmids (Fig. 4a, clusters b1 to b3), in other plasmids (Fig. 4a, cluster c), in viruses (Fig. 4a, clusters d1 to d3), or in a mixture of other plasmids and viruses (Fig. 4a, clusters e1 and e2). Some clusters were enriched in plasmids or viruses belonging to well-defined families. In particular, we noticed clusters rich in *Caudovirales* (Fig. 4a, clusters d2), *Sphaerolipoviridae* (Fig. 4a, clusters d3), or RC-Rep SF I elements (Fig. 4a, one subcluster of e2). We also noticed that the *Halobacterium halobium* plasmid ehsp was identical to the *Halobacterium salinarum* plasmid pHSB, a small rolling-circle replication plasmid of 1.7 kb [25] (in cluster e2). For *Caudovirales*, we observed a certain consistency between the viral types and clustering patterns. Except for HHTV-1, HGTV-1 and the *Natrialba magadii* provirus (Nmag-Pro1), *Caudovirales* members were distributed among 3 main clusters (Fig. 4a, cluster d2, one subcluster of e1, one subcluster of e2). The first one exclusively comprised 9 *Caudovirales* members (Fig. 4a, cluster d2), with an average genome length of 83.3 kb. Within this cluster, the 3 HCTV-type *Siphoviridae* members grouped together (HCTV-1, HCTV-5 and HVTV-1); in the *Myoviridae* family, similar results were observed for the 4 HF2-type viruses (HF1, HF2, HRTV-8 and HRTV-5) and for both HRTV-7-type viruses (HRTV-7 and HSTV-2). Moreover, HF2-type and HRTV-7-type viruses that are evolutionarily related [49] also clustered together. In contrast, other *Caudovirales*-rich clusters also comprised plasmids of limited size as well as *Pleiolipoviridae* and *Sphaerolipoviridae* members. *Caudovirales* members in these mixed clusters had a smaller average genome size, of 43.5 kb. Finally, HHTV-1 (*Caudovirales* order) was one of the outermost elements in the haloarchaea dendrogram (Fig. 4a, in cluster d1), consistent with its description as the most divergent among sequenced haloarchaeal tailed viruses [49].

A gene-sharing network based on protein similarity was constructed (Fig. 4b) and supported the same observation when the weak edges were filtered out. This reinforces the conclusion since gene sharing networks address a different type of information, depending on the genome functional content.

The network (Fig. 4b) also showed that cells shared few strong edges with plasmids of limited size (< 100 kb), in contrast to large plasmids and megaplasmids. This

Fig. 4 (See legend on next page.)

was further confirmed by basic statistics on the number of edges among these different types of elements (Fig. 4c). For the smaller plasmid category (< 100 kb), the level of this indicator was actually similar to that of viruses (Fig. 4c). *Halobacteria* plasmids therefore seem to have heterogeneous properties with respect to genetic connections with their hosts. Plasmid size appears to act as a major influential factor, possibly by increasing the probability of gene exchange.

Good congruence between mobile element families and 5-mer composition in *Sulfolobales*

Viruses and plasmids present in *Sulfolobales* (genera *Sulfolobus*, *Metallosphaera* and *Acidianus*) are among the best characterized archaeal mobile elements. *Sulfolobales* members produce viruses with unique morphotypes (e.g., fusiform, bottle-shaped), which has aroused important scientific interest during the last two decades [51]. *Fuselloviridae*, *Lipothrixviridae*, and *Rudiviridae*, reviewed in [24]) and 2 distinct plasmid families (cryptic pRN-like, conjugative pNOB8-like [52]) have been studied extensively. A total of 119 *Sulfolobales* sequences of cells, plasmids and viruses were studied here (Additional File 22).

The cellular genomes were distributed between 2 distant clusters, one corresponding to *Metallosphaera* and the other to *Sulfolobus* and *Acidianus* (Fig. 5a, black color, codes starting with m, s and a respectively). The average genomic GC content in *Metallosphaera* was of 45.4% ± 1.6 SD, compared to 35.2% ± 1.6 SD in the other *Sulfolobales* genomes, which possibly influenced this partition. In the *Sulfolobus-Acidianus* cluster (Fig. 5a), the subclusters were consistent with the distinct species, namely, *Sulfolobus islandicus* (codes starting with si), *Sulfolobus solfataricus* (codes starting with sso or so), *Sulfolobus acidocaldarius* (codes starting with sac or sa) and *Acidianus* species (codes starting with a). The only exception was *Sulfolobus tokkodai* (code sto), which was located in the *Acidianus* subcluster.

The *Sulfolobales* extrachromosomal elements were grouped primarily according to their taxonomic family rather than to the taxonomy of their hosts (Fig. 5a). This general pattern appeared once more to be partly linked to the GC content of the sequences (Fig. 5a, Additional file 23). There were notable exceptions, such as the *Fuselloviridae* proviruses previously described in [53] (Fig. 5a, SSV-m1425, SSV-ls215 and SSV-yg5714): their sequences were less GC rich than those of the other

Fuselloviridae members (19.6% ± 0.7 SD compared to 39.2% ± 2.3 SD) but they were still located in the main *Fuselloviridae* cluster.

For viruses, 14 out of 16 *Rudiviridae* genomes, 12 out of 13 *Fuselloviridae* genomes and 7 out of 8 *Lipothrixviridae* genomes clustered together (Fig. 5a). A similar trend was observed for less represented families, with all *Ampullaviridae* and *Turriviridae* members grouping into consistent clusters. For the plasmids, all pRN-like cryptic plasmids and 2 related phage-plasmid hybrid entities (pSSVx and pSSVi) (Fig. 5a, magenta color) formed a single cluster that also included *Turriviridae*. Finally, 12 out of the 13 pNOB8-like conjugative plasmids clustered together (Fig. 5a, green color). Interestingly, the main pNOB8-like plasmid cluster (with sizes ranging from 20.4 to 42.2 kb) was located very close to the main cell cluster, whereas the pRN-like cryptic plasmid cluster (with sizes ranging from 5 to 13.6 kb) was much more distant (Fig. 5A). Similar to our observations for *Halobacteria*, this finding highlights that larger plasmids are more similar to cells than shorter plasmids and viruses in terms of 5-mer composition.

This could reflect the occurrence of frequent genetic exchange between *Sulfolobales* cells and pNOB8-like conjugative plasmids. Based on PERMANOVA, the viral and plasmid families together with the genomic GC content explained 77.68% of the 5-mer profile dissimilarity variance among *Sulfolobales* mobile elements (Additional file 23, $D_{5\_mobile\_sulfo}$ ~ Family*GC%).

A gene sharing network also showed that *Sulfolobales* mobile elements tended to group according to their family. The proximity of pNOB8-like conjugative plasmids and *Sulfolobales* cells was visible, whereas connections between cells and pRN-like plasmids or viruses were less striking (Fig. 5b, Fig. 5c). A noticeable difference between the dendrogram based on the 5-mer profiles and the gene sharing network regarded the links between the *Lipothrixviridae* and *Rudiviridae* families, which together form the *Ligamenvirales* order [54]. While this evolutionary connection was clear in the gene sharing network (Fig. 5b), it was not clear from the 5-mer-based analysis (Fig. 5a), confirming the idea that sequence composition changes more rapidly than gene content and that similarity in sequence composition can identify only close evolutionary relationships. The different 5-mer compositions between *Lipothrixviridae* and *Rudiviridae* may be explained by the low genomic GC contents

Fig. 5 (See legend on next page.)

of *Rudiviridae* (28.25% ± 6.17% SD on average). We also noticed that *Rudiviridae* members seem to have an unusual 5-mer composition since their main cluster had a long branch and they were isolated not only from *Lipothrixviridae* but also from all other mobile elements (Fig. 5a). In addition to their very low GC content, several factors could possibly explain the specific 5-mer composition of *Rudiviridae*, such as unusual DNA packaging constraints or their DNA replication mode (hypothetically complex mechanisms, not yet fully identified [55], reviewed in [24]).

Outliers and host transfers

Genomes with unexpected 5-mer composition (outliers) could presumably reveal singular evolutionary trajectories. We identified a total of 51 outlier plasmids and viruses (Additional File 2) by combining a systematic approach (see Materials and Methods) and visual examination of the dendrograms. These elements had unexpected 5-mer compositions compared to the average in their taxonomic group or the 5-mer composition of their hosts.

For 4 of them, their very short length (< 4 kb) likely explains their atypical composition. The presence of tRNA genes in viral genomes has previously been identified as a possible factor explaining the divergence between host and viral genome k-mer compositions, acting by reducing the selective pressure on the viral genome for adaptation to host codon usage [14, 56]. Such a phenomenon was not prevalent here, since only 3 out of 51 outliers encoded tRNAs in their genomes (Additional File 2).

Assuming that recent host transfer could also explain atypical 5-mer compositions, we specifically examined *Thermococcales* and *Methanococcales*, which are evolutionarily closely related and known to share evolutionarily-related plasmids. One of the previously described interorder host transfer events was indeed visible by PCA (Fig. 6a) or hierarchical clustering (Additional File 24), suggesting that the *Methanocaldococcus* plasmid pMETVU01 originated from a *Thermococcales* host [25]. More ancient evolutionary connections detected previously between some *Methanococcales* plasmids, such as pMEFER01, and the pT26–2 *Thermococcales* plasmid family [25] were not visible based on the 5-mer profiles. This suggests that the 5-mer composition of newly transferred mobile elements must evolve rapidly, so only recent transfers can be detected by this approach.

We then considered more closely the 13 pNOB8-like *Sulfolobales* conjugative plasmids because in a previous version of the dataset, two pNOB8-like plasmids, namely, pMGB1 and pTC, were located close to *Metallosphaera* genomes, far from the main pNOB8-like cluster (Additional File 25). This suggested that pTC and pMGB1 could replicate in *Metallosphaera* archaea, in addition to *Sulfolobus*. Interestingly, we identified a remnant plasmid very similar to pMGB1 in the genome of *Metallosphaera sedula* (Fig. 6b), consistent with this hypothesis. We named this new integrated conjugative plasmid ICEmse, for "Integrative Conjugative Element of *M. sedula*", and we included it in the dataset. ICEmse was consistently located in the same cluster as the pTC, pMGB1 and *Metallosphaera* genomes in the previous dataset version (Additional File 26). In our latest dataset version, the trends were less clear, since *Metallosphaera* formed a fully separate cluster. Moreover, only pTC grouped with ICEmse (Fig. 5a) and was detected as an outlier. By contrast, pMGB1 was located in the main pNOB8-like plasmid cluster, but was the outermost element. The PCA result was in good agreement with the host transfer scenario, since pTC, pMGB1 and ICEmse were located roughly at mid-distance between *Sulfolobus* and *Metallosphaera* cells (Fig. 6c). Finally, consistent with the high GC content of *Metallosphaera* genomes, the pMGB1, pTC and ICEmse genomic GC contents were 39.6, 41.4 and 41.5%, respectively, compared to only 36.7% ± 0.6 SD for the other pNOB8-like elements, again supporting the host transfer hypothesis.

Discussion

The influence of their phylogenetic position on the 5-mer composition of archaeal cell genomes is clearly visible in our dataset, consistent with the genome-wide importance of short k-mers, which could play a role in speciation and be critical to recombination (reviewed and defended in [2]). However, the global topology that we obtained by hierarchical clustering was not fully consistent with the phylogeny of archaea, as detailed in the results section. It could be interesting to evaluate whether more sophisticated methods [16–18] and the use of various k-mer sizes would enable us to obtain a global topology more consistent with the phylogeny of archaea. Whether it could be achieved is, however, uncertain. The fact that we could detect recent HGTs but that several ancient evolutionary connections [54, 57]

Fig. 6 (See legend on next page.)

were not detected in our analysis suggests that the genome composition in short k-mers must evolve rapidly. The acquisition or loss of adaptation to extreme conditions played a strong role in the evolution of archaea (e.g. [29, 34]). It was proposed that the last archaeal common ancestor was a hyperthermophile [29], and the subsequent adaptation to other niche constraints may likely have blurred the phylogenetic signal of k-mer profiles in *Archaea*. This must have resulted in certain cases in convergent evolution of sequence composition, which could also blur the phylogenetic signal.

Our results were mostly consistent with previous studies, but they provide a different view since most of the latter focused on amino acid composition [35, 41, 42] and codon usage (e.g., [35]), rather than k-mers and absolute codon frequencies. Our analysis shows that the ecological niche also has a strong link with the 5-mer composition of archaeal extrachromosomal elements. For virions in particular, it would be interesting to determine whether the composition results exclusively from the coevolution with the hosts or whether other selective pressures are exerted, for instance on the packaging structure properties during the extracellular stage, corresponding to a more direct effect of the extracellular environment.

Halobacteria members and their extrachromosomal elements showed a very strong signature at all studied levels: GC content, 5-mer and 3-mer compositions of the whole genome sequences and codon composition. *Halobacteria* was clearly separated from the other clades of archaea, most likely as a consequence of their evolution in high-salt environments. Halophiles have an exceptionally high GC content among archaea ($\sim 60\%$) (Additional file 4), possibly to prevent the formation of thymidine dimers following extensive exposure of these archaea to UV at the surface of solar salterns [58]. *H. walsbyi* genomes are notable exceptions, and their low GC-content (48%) may be partly compensated by the presence of 4 encoded photolyases in their genomes [59]. In addition, proteins of halophiles have specific features that enable them to be functional under the high salt concentration in the cytoplasm (up to 4 M KCl) [35]. Their surface is typically enriched in acidic [42] and negatively charged residues [43], while their core has a moderate hydrophobicity [43].

Regarding the signature for hyperthermophily, many differences in the methods and datasets could explain the imperfect agreement with previous studies [44, 45].

Primarily, our information on amino acids is indirect, based on absolute codon frequency analysis, while most cited studies directly focused on amino acid composition. An additional explanation could be that several previous analyses included both archaea and bacteria, whereas we focused exclusively on archaea, mainly on *Desulfurococcales*, *Thermoproteales* and *Thermococcales*. In addition, our dataset includes more sequences, and finally, the statistical methods employed are slightly different. In particular, Lambros et al. [60] considered the optimal growth temperature as a quantitative variable, pointing out that most changes in response to growth temperature occur below 60 °C. We therefore may have missed some of the compositional changes that start to occur at lower temperatures. It is, however, interesting that discriminant 5-mers could be identified from our diverse dataset and when considering a high temperature threshold to partition the dataset into two categories.

We observed that mobile elements of archaea harbor some specificity in their 5-mer composition compared to their hosts, with two major types of situations. The first corresponds to major compositional differences between the mobile elements and their hosts. Such mobile elements are outliers and do not represent the most frequent cases. According to the literature, such differences could be explained by the presence of tRNA genes in the mobile element genome, enabling the uncoupling of codon usage constraints of the hosts from those of the mobile element [14, 48]; by a large genome size of the mobile element, which is indicative of a more autonomous replication cycle [14]; or by a recent acquisition by the host, such that the composition of the mobile element has not yet undergone host adaptation [31]. In the present study, we found a very limited presence of annotated tRNA genes in mobile elements (Additional file 2). We identified two recent host transfers, one previously described (pMETVU01) [25] and a newly described one (iCEmse). We hypothesize that the fact that the *Halobacteria* viruses His1 and His2 encode their own family B DNA polymerase [24] could possibly contribute to their atypical 5-mer composition. Apart from these few cases, no obvious factors could be identified at first glance for most outliers.

A second type of case, the most frequent, corresponds to a small 5-mer composition difference between the mobile elements and their hosts. In the literature, the influence of the host range and mode of transmission have been proposed, such as frequent changes of hosts [31] or

a wide host range [19]. For horizontally transferred mobile elements, occasional exposure to the extracellular environment could also create particular selective pressures [31]. Competition for metabolic resources has also been suggested to explain differences in GC content [61]. Beyond these general factors, we suggest that the specific composition of mobile elements could primarily result from the intrinsic properties of mobile element families. This idea is best illustrated by *Sulfolobales* plasmids and viruses that cluster mainly according to their own taxonomic family, rather than those of their host strains. This suggests that each mobile element family has its own specificity in terms of 5-mer composition and indicates that their 5-mer composition does not simply reflect their adaptation to their hosts or to the extracellular environment. This notion is echoed by [15], the authors of which could classify viruses based on their tetramer composition. One could imagine other selective forces shaping the k-mer composition of mobile elements. There could hypothetically be constraints related to the replication mode or the functional content. For plasmidions [62, 63] and viruses, additional constraints linked to packaging or structure can be imagined, in relation to but not limited to the properties of the extracellular environment.

Interestingly, we observed a lower difference in sequence composition between hosts and large plasmids or megaplasmids, than between hosts and smaller plasmids and viruses. A similar trend was previously observed by several authors who suggested that the low difference in the case of large plasmids could be explained by a stronger adaptation to the host for large plasmids [32] whereas the larger difference in the case of small plasmids could result either from the limited compositional representativeness of short sequences [32] or by their greater host range [19]. We hypothesize that the lower difference in the case of large plasmids could also be due to the fact that they exchange more genes with their hosts and also lack the selective pressures related to packaging or stability in the extracellular environment. Paul et al. [35] mentioned that the difference in codon usage between chromosomes I and II of *Haloarcula marismortui* must be linked to the more recent acquisition of the second chromosome. Our study shows that second chromosomes in the class *Halobacteria* have a 5-mer signature similar to that of large or megaplasmids, and distinct from that of first chromosomes. Therefore, the distinct nucleotide composition of chromosome II of *H. marismortui* could also result from its different origin from that of chromosome I, supporting the idea that chromosome II belongs to the plasmid realm.

Our simple gene sharing network analyses yielded consistent trends, again highlighting a stronger link between larger plasmids and cells than between short mobile elements (plasmids or viruses) and cells. Similar analyses have previously highlighted the important role of mobile elements in gene dissemination, enabling the identification of those more specifically involved in this process [64, 65]. Halary et al. [65] in particular contrasted viruses and plasmids, the latter being, according to their study, the major key players of HGT. Even if our study covers a single domain of life, our observations suggest that the size of the mobile elements (plasmid or viruses) might be in fact the most important factor determining its importance in the evolutionary relationships with hosts. Moreover, the delineation between plasmids, viruses and other types of mobile elements, such as plasmidions, is becoming increasingly blurred [62].

Conclusions

Our study provides a useful framework for the interpretation of k-mer approaches applied to cell or extrachromosomal elements of the domain *Archaea*. For cells, the global topologies based either on 5-mer profiles or on phylogeny are inconsistent. At a finer level, the results, however, show the strong influence of phylogenetic relationships and of adaptation to environmental constraints on 5-mer compositions. These two factors are interdependent to a significant extent, and the respective weight of their contribution varies according to the clade. Our analysis highlighted the possibility of differential adaptation to the environmental niche between chromosomal DNA and extrachromosomal element DNA. In addition, we clearly observed different patterns depending on the mobile element type and size. For mobile elements, coevolution with the host has a clear influence on their 5-mer composition. However, strikingly, viral and plasmid families also retain a specific imprint in their 5-mer profile. Our analysis also enabled us to detect two host transfer events, but exclusively recent ones, which suggests the fast adaptation of short k-mer profiles in a fluctuating environment. The genome composition difference observed here between mobile genetic elements and their hosts suggests that using k-mer based methods to analyze mobile elements in metagenomic data may lead to spurious results. Incorrect host prediction could occur [66], as well as missed detection of integrated elements during MAG reconstruction [67].

Our results thus call for caution when using k-mers for the identification of mobile elements in metagenomics data, for host prediction of mobile elements, and for phylogenetic reconstruction, especially for ancestral events.

Methods

Presentation of the dataset and of the approach

Basic information about the genomes included in the dataset is available in Additional file 2, such as the taxonomy, length and GC content of each element.

Additional file 4 provides a synthetic view of GC% values across the dataset, according to the taxonomic order of the host and to the type of element; Additional file 27 shows the GC% values according to the Niche and type of element; finally, an analysis of variance (ANOVA) of these GC% values is presented in Additional file 28.

We selected 11 taxonomic groups (at the order level) of the domain *Archaea* (Additional files 19 and 29) for which a significant number of extrachromosomal element sequences were available (plasmids or viruses). For these 11 taxonomic orders, we gathered a total of 589 whole genome sequences of cells, plasmids, viruses and proviruses. The dataset covered 3 and 8 orders of the phyla *Crenarchaeota* and *Euryarchaeota*, respectively. It comprised exclusively halophiles, acidothermophiles, hyperthermophiles and methanogens.

For each genome, we established a profile consisting of its 5-mer absolute frequencies. To select the k-mer length, a compromise needed to be established: longer k-mers are more informative; however, excessively long k-mers result in data scarcity due to low average counts, leading to artifacts during subsequent statistical analyses. For plasmids and viruses, k-mer length of 5 was selected as a good compromise. Indeed, their average genome length in the dataset was 89,814 bases; since there are 4^k distinct possible k-mers, the average counts were 88 per 5-mer (89,814 divided by 4^5), which we considered sufficiently representative, and slightly more specific than tetramers. For cells, although they have a much higher average genome length, we also used 5-mers to compare their profiles with those of extrachromosomal elements.

The obtained 5-mer frequency profiles included 1024 proportions (4^5) and constituted a highly multidimensional dataset. To gain insight into these complex data, the landscape of these profiles across the dataset was explored with four methods: hierarchical clustering, PCA, PERMANOVA and PLS-DA. PCA aims to project highly multidimensional data on a set of orthogonal axes to visualise them easily while preserving their variance as best possible. PERMANOVA is a generalized form of ANOVA used to analyze the variance of multidimensional values, here the 5-mer profile distance matrix, and relate them to potential structuring factors. Finally, PLS-DA was used to identify the most discriminant k-mers between several categories of genomes, such as genomes from halophiles, versus nonhalophiles.

Genome sequences
We collected 534 publicly available whole genome sequences of cells, plasmids, viruses and proviruses (Additional file 2) from the NCBI genome database. We performed a final update on the 7th of August 2018. In addition, we retrieved 28 provirus sequences directly from cellular genome sequences based on literature

information [53, 68, 69]. Finally, we included 26 magrovirus sequences [70] available on a specific website (https://github.com/BejaLab/Magrovirus/tree/master/Supp_files) and the assembly of a Marine Group II archaeon (GCA_003324605). When the mobile elements were not classified into well-defined families, we categorized them according to the taxonomy of their host (e.g. *Halobacteriales* megaplasmid).

Establishment of profiles based on the sequence 5-mer composition
Two types of profiles were established for each sequence based on its 5-mer composition, as described in more detail below. The profiles of the different genomes were then combined across the dataset to obtain two distinct matrices, one for each type of profile.

The first type of profile was based on the 5-mer frequencies of the whole genome sequences. The 5-mer counts were calculated with Jellyfish 2.2.6 on the INRAE-MIGALE cluster (URL https://migale.inrae.fr/). The obtained count data were imported into R [71] (version 3.4.2) and transformed into a frequency matrix to obtain normalized data: for each genome, the sum of the 5-mer frequencies was equal to 1.

The second type of profile relied exclusively on the coding regions; it reflected the exceptionality of the different 5-mers in the coding regions after correcting for differences in codon composition in the studied genome. The exceptionality scores were calculated with R'MES software [72], with the following options: Gaussian model, k-mer length of 5, second-order Markov chain model, and 3 phases. Briefly, R'MES fits a Markov chain on each genome's concatenated coding regions to compute the expected frequencies of 5-mers based on observed codon frequencies. Exceptionality scores are then computed as standardized deviations between observed and expected 5-mer frequencies. The exceptionality score values obtained for each 5-mer were directly used to generate the second type of 5-mer profile of each genome. R'MES was run on the INRAE-MIGALE cluster.

Statistical analyses of the profiles based on 5-mer composition
All statistical analyses were performed using R (version 3.4.2). PCA were performed with the dudi.pca function of the ade4 package [73], on scaled and centered data. We performed PLS-DA analyses with the caret package [74], using a 10-times repeated 10-fold cross-validation and the "accuracy" metrics to select the number of components, again on centered and scaled data. Hierarchical clustering was realized with the hclust function from R applied to Euclidian distance matrices with the Ward.D2 method. PERMANOVA of Euclidian distance matrices were conducted with the adonis function of the vegan

Bize *et al. BMC Genomics* (2021) 22:186

Page 20 of 22

package [75], with *p*-values computed on 9999 permutations. PERMANOVA assumes that 5-mer profiles respond linearly to changes in the covariates and that the variance of profiles is comparable across conditions of the data. The *p*-values were computed by permutations: this nonparametric approach is robust to model misspecification. The wilcox.test function from R CRAN was employed to test the equality of means through Mann-Whitney-Wilcoxon statistical tests.

Most plots were prepared with ggplot2 package [76]. Dendrograms constructed with hclust were exported in newick format and used in the online tool Interactive Tree Of Life (iTOL) [77] to construct the tree figures.

Network analyses
Gene sharing network data were generated with EGN 1.0 software [78]. For this purpose, whole proteomes were downloaded from the NCBI website; the resulting multifasta file was formatted according to the EGN manual's instructions. Blastp [79] searches were computed within EGN software, which acts as a wrapper. The EGN parameters were set as follows: e-value threshold of 1e-05, hit identity threshold of 30%, hit coverage of the shortest sequence of 60%, hit coverage of both sequences of at least 30%, minimal hit length of 20 amino acids, best reciprocity threshold of 10%. The EGN results consisted in the number of similar genes shared between each pair of genomes. These values were subsequently normalized by dividing them by the smallest genome length of the concerned pair.

The obtained networks were visualized with Cytoscape 3.7.1 [80] by using the edge-weighted spring embedded layout and by filtering out the weaker interactions (edge values), as specifically indicated in each case.

Genome comparison
BLAST comparisons between selected genomes were visualized with Easyfig 2.2.2 [81].

Outlier identification
For each viral or plasmid family, the distance of each element's 5-mer profile to the profile barycenter of the considered family was calculated. A gamma distribution was fitted to the histogram of all distances. A 0.95 confidence threshold was selected to define outliers, corresponding to a distance value of 1.654. With this approach, implemented by a homemade R script, 18 outliers were identified, of which 3 were removed after visual examination of the 5-mer frequency-based dendrograms. In addition to this systematic method, 36 other outliers were identified by visual examination of these dendrograms (e.g. genomes not clustering with other genomes from the same family), resulting in a total of 51 outlier elements.

Abbreviations
ANOVA: Analysis of variance; HGT: Horizontal gene tranfer; MAG: Metagenome-assembled genome; PCA: Principal component analysis; PERMANOVA: Permutational multivariate analysis of variance; PLS-DA: Partial least squares discriminant analysis; A: Alanine; D: Aspartic acid; E: Glutamic acid; F: Phenylalanine; H: Histidine; I: Isoleucine; L: Leucine; N: Asparagine; P: Proline; Q: Glutamine; R: Arginine; S: Serine; T: Threonine; V: Valine; W: Tryptophane; Y: Tyrosine

Supplementary Information
The online version contains supplementary material available at https://doi.org/10.1186/s12864-021-07471-y.

> **Additional files 1 and 3 to 29.** Additional tables, figures and text.
>
> **Additional file 2.** Excel file with genome list and genomic features.

Availability of data and materials
"The dataset supporting the conclusions of this article is included within the article and its additional files."

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare that they have no competing interests.

Author details
[1]Université Paris-Saclay, INRAE, PROSE, F-92761 Antony, France. [2]Université Paris-Saclay, INRAE, MaIAGE, F-78350 Jouy-en-Josas, France. [3]Université Paris-Saclay, INRAE, BioinfOmics, MIGALE bioinformatics facility, F-78350 Jouy-en-Josas, France. [4]Institut Pasteur, Unité de Virologie des Archées, Département de Microbiologie, 25 Rue du Docteur Roux, 75015 Paris, France. [5]Université Paris-Saclay, CEA, CNRS, Institute for Integrative Biology of the Cell (I2BC), 91198 Gif-sur-Yvette, France.

References
1. Zielezinski A, Vinga S, Almeida J, Karlowski WM. Alignment-free sequence comparison: benefits, applications, and tools. Genome Biol. 2017;18(1):186.

2.   Forsdyke DR. Success of alignment-free oligonucleotide (k-mer) analysis confirms relative importance of genomes not genes in speciation and phylogeny. Biol J Linn Soc. 2019;128(2):239–50.

3.   Kariin S, Burge C. Dinucleotide relative abundance extremes: a genomic signature. Trends Genet. 1995;11(7):283–90.

4.   Karlin S, Mrázek J, Campbell AM. Compositional biases of bacterial genomes and evolutionary implications. J Bacteriol. 1997;179(12):3899.

5.   Kang DD, Froula J, Egan R, Wang Z. MetaBAT, an efficient tool for accurately reconstructing single genomes from complex microbial communities. PeerJ. 2015;3:e1165.

6.   Alneberg J, Bjarnason BS, de Bruijn I, Schirmer M, Quick J, Ijaz UZ, Lahti L, Loman NJ, Andersson AF, Quince C. Binning metagenomic contigs by coverage and composition. Nat Methods. 2014;11:1144.

7.   Wood DE, Salzberg SL. Kraken: ultrafast metagenomic sequence classification using exact alignments. Genome Biol. 2014;15(3):R46.

8.   Richter M, Rosselló-Móra R. Shifting the genomic gold standard for the prokaryotic species definition. Proc Natl Acad Sci. 2009;106(45):19126.

9.   Teeling H, Meyerdierks A, Bauer M, Amann R, Glöckner FO. Application of tetranucleotide frequencies for the assignment of genomic fragments. Environ Microbiol. 2004;6(9):938–47.

10.   Benoit G, Peterlongo P, Mariadassou M, Drezen E, Schbath S, Lavenier D, Lemaitre C. Multiple comparative metagenomics using multiset k-mer counting. PeerJ Comput Sci. 2016;2:e94.

11.   Dubinkina VB, Ischenko DS, Ulyantsev VI, Tyakht AV, Alexeev DG. Assessment of k-mer spectrum applicability for metagenomic dissimilarity analysis. BMC Bioinformatics. 2016;17(1):38.

12.   Ren J, Ahlgren NA, Lu YY, Fuhrman JA, Sun F. VirFinder: a novel k-mer based tool for identifying viral sequences from assembled metagenomic data. Microbiome. 2017;5(1):69.

13.   Krawczyk PS, Lipinski L, Dziembowski A. PlasFlow: predicting plasmid sequences in metagenomic data using genome signatures. Nucleic Acids Res. 2018;46(6):e35.

14.   Galiez C, Siebert M, Enault F, Vincent J, Söding J. WIsH: who is the host? Predicting prokaryotic hosts from metagenomic phage contigs. Bioinformatics. 2017;33(19):3113–4.

15.   Wang T, Herbster M, Mian IS: Virus genome sequence classification using features based on nucleotides, words and compression. arXiv preprint arXiv: 180903950 2018.

16.   Wen J, Chan RHF, Yau S-C, He RL, Yau SST. K-mer natural vector and its application to the phylogenetic analysis of genetic sequences. Gene. 2014; 546(1):25–34.

17.   Bernard G, Chan CX, Ragan MA. Alignment-free microbial phylogenomics under scenarios of sequence divergence, genome rearrangement and lateral genetic transfer. Sci Rep. 2016;6(1):28970.

18.   Déraspe M, Raymond F, Boisvert S, Culley A, Roy PH, Laviolette F, Corbeil J. Phenetic comparison of prokaryotic genomes using k-mers. Mol Biol Evol. 2017;34(10):2716–29.

19.   Pride DT, Meinersmann RJ, Wassenaar TM, Blaser MJ. Evolutionary implications of microbial genome Tetranucleotide frequency biases. Genome Res. 2003;13(2):145–58.

20.   Bernard G, Ragan MA, Chan CX. Recapitulating phylogenies using k-mers: from trees to networks. F1000Research. 2016;5:2789.

21.   Bernard G, Greenfield P, Ragan MA, Chan CX. K-mer similarity, networks of microbial genomes, and taxonomic rank. mSystems. 2018;3(6):e00257–18.

22.   Dufraigne C, Fertil B, Lespinats S, Giron A, Deschavanne P. Detection and characterization of horizontal transfers in prokaryotes using genomic signature. Nucleic Acids Res. 2005;33(1):e6.

23.   Huang G-D, Liu X-M, Huang T-L, Xia L-C. The statistical power of k-mer based aggregative statistics for alignment-free detection of horizontal gene transfer. Synthetic Syst Biotechnol. 2019;4(3):150–6.

24.   Krupovic M, Cvirkaite-Krupovic V, Iranzo J, Prangishvili D, Koonin EV. Viruses of archaea: structural, functional, environmental and evolutionary genomics. Virus Res. 2018;244:181–93.

25.   Forterre P, Krupovic M, Raymann K, Soler N. Plasmids from Euryarchaeota. In Plasmids (eds M.E. Tolmasky and J.C. Alonso). 2015. https://doi.org/10.1128/9781555818982.ch20.

26.   Wang H, Peng N, Shah SA, Huang L, She Q. Archaeal Extrachromosomal genetic elements. Microbiol Mol Biol Rev. 2015;79(1):117–52.

27.   Roux S, Enault F, Ravet V, Colombet J, Bettarel Y, Auguet J-C, Bouvier T, Lucas-Staat S, Vellet A, Prangishvili D, et al. Analysis of metagenomic data

reveals common features of halophilic viral communities across continents. Environ Microbiol. 2016;18(3):889–903.

28.   Ackermann HW. Frequency of morphological phage descriptions in the year 2000. Arch Virol. 2001;146(5):843–57.

29.   Groussin M, Gouy M. Adaptation to environmental temperature is a major determinant of molecular evolutionary rates in Archaea. Mol Biol Evol. 2011; 28(9):2661–74.

30.   Campbell A, Mrázek J, Karlin S. Genome signature comparisons among prokaryote, plasmid, and mitochondrial DNA. Proc Natl Acad Sci. 1999; 96(16):9184.

31.   van Passel MWJ, Bart A, Luyf ACM, van Kampen AHC, van der Ende A. Compositional discordance between prokaryotic plasmids and host chromosomes. BMC Genomics. 2006;7(1):26.

32.   Bohlin J, Skjerve E, Ussery DW. Reliability and applications of statistical methods based on oligonucleotide frequencies in bacterial and archaeal genomes. BMC Genomics. 2008;9(1):104.

33.   Bohlin J, Skjerve E, Ussery DW. Investigations of oligonucleotide usage variance within and between prokaryotes. Plos Comput Biol. 2008;4: e1000057.

34.   Boussau B, Blanquart S, Necsulea A, Lartillot N, Gouy M. Parallel adaptations to high temperatures in the Archaean eon. Nature. 2008;456:942.

35.   Paul S, Bag SK, Das S, Harvill ET, Dutta C. Molecular signature of hypersaline adaptation: insights from genome and proteome composition of halophilic prokaryotes. Genome Biol. 2008;9(4):R70.

36.   Reimer LC, Vetcininova A, Carbasse JS, Söhngen C, Gleim D, Ebeling C, Overmann J. BacDive in 2019: bacterial phenotypic data for high-throughput biodiversity analysis. Nucleic Acids Res. 2018;47(D1):D631–6.

37.   Botzman M, Margalit H. Variation in global codon usage bias among prokaryotic organisms is associated with their lifestyles. Genome Biol. 2011; 12(10):R109.

38.   Slonczewski JL, Fujisawa M, Dopson M, Krulwich TA. Cytoplasmic pH Measurement and Homeostasis in Bacteria and Archaea. In: Poole RK, editor. Advances in Microbial Physiology, vol. 55. Academic Press; 2009. p. 1–317.

39.   Lin F-H, Forsdyke DR. Prokaryotes that grow optimally in acid have purine-poor codons in long open reading frames. Extremophiles. 2007;11(1):9–18.

40.   Roy Chowdhury A, Dutta C. A pursuit of lineage-specific and niche-specific proteome features in the world of archaea. BMC Genomics. 2012;13(1):236.

41.   Nath A. Insights into the sequence parameters for halophilic adaptation. Amino Acids. 2016;48(3):751–62.

42.   Fukuchi S, Yoshimune K, Wakayama M, Moriguchi M, Nishikawa K. Unique amino acid composition of proteins in Halophilic Bacteria. J Mol Biol. 2003; 327(2):347–57.

43.   Kastritis PL, Papandreou NC, Hamodrakas SJ. Haloadaptation: insights from comparative modeling studies of halophilic archaeal DHFRs. Int J Biol Macromol. 2007;41(4):447–53.

44.   Singer GAC, Hickey DA. Thermophilic prokaryotes have characteristic patterns of codon usage, amino acid composition and nucleotide content. Gene. 2003;317:39–47.

45.   Zeldovich KB, Berezovsky IN, Shakhnovich EI. Protein and DNA sequence determinants of Thermophilic adaptation. PLoS Comput Biol. 2007;3(1):e5.

46.   Kreil DP, Ouzounis CA. Identification of thermophilic species by the amino acid compositions deduced from their genomes. Nucleic Acids Res. 2001; 29(7):1608–15.

47.   Tekaia F, Yeramian E, Dujon B. Amino acid composition of genomes, lifestyles of organisms, and evolutionary trends: a global picture with correspondence analysis. Gene. 2002;297(1):51–60.

48.   Luk A, Williams T, Erdmann S, Papke R, Cavicchioli R. Viruses of Haloarchaea. Life. 2014;4(4):681.

49.   Sencilo A, Roine E. A Glimpse of the genomic diversity of haloarchaeal tailed viruses. Front Microbiol. 2014;5(84):1–6. https://www.frontiersin.org/articles/10.3389/fmicb.2014.00084/full.

50.   Ng WV, Ciufo SA, Smith TM, Bumgarner RE, Baskin D, Faust J, Hall B, Loretz C, Seto J, Slagel J, et al. Snapshot of a large dynamic replicon in a Halophilic Archaeon: Megaplasmid or Minichromosome? Genome Res. 1998;8(11): 1131–41.

51.   Leigh JA, Albers S-V, Atomi H, Allers T. Model organisms for genetics in the domain Archaea: methanogens, halophiles, Thermococcales and Sulfolobales. FEMS Microbiol Rev. 2011;35(4):577–608.

52.   Greve B, Jensen S, Brügger K, Zillig W, Garrett RA. Genomic comparison of archaeal conjugative plasmids from Sulfolobus. Archaea. 2004;1(4):231–9.

53. Held NL, Whitaker RJ. Viral biogeography revealed by signatures in Sulfolobus islandicus genomes. Environ Microbiol. 2009;11(2):457–66.

54. Iranzo J, Koonin EV, Prangishvili D, Krupovic M. Bipartite network analysis of the Archaeal Virosphere: evolutionary connections between viruses and Capsidless Mobile elements. J Virol. 2016;90(24):11043–55.

55. Martínez-Alvarez L, Bell SD, Peng X. Multiple consecutive initiation of replication producing novel brush-like intermediates at the termini of linear viral dsDNA genomes with hairpin ends. Nucleic Acids Res. 2016;44(18): 8799–809.

56. Edwards RA, McNair K, Faust K, Raes J, Dutilh BE. Computational approaches to predict bacteriophage–host relationships. FEMS Microbiol Rev. 2016;40(2): 258–72.

57. Badel C, Erauso G, Gomez AL, Catchpole R, Gonnet M, Oberto J, Forterre P, Da Cunha V. The global distribution and evolutionary history of the pT26-2 archaeal plasmid family. Environ Microbiol. 2019;21(12):4685–705.

58. Kennedy SP, Ng WV, Salzberg SL, Hood L, DasSarma S. Understanding the adaptation of Halobacterium species NRC-1 to its extreme environment through computational analysis of its genome sequence. Genome Res. 2001;11(10):1641–50.

59. Bolhuis H, Palm P, Wende A, Falb M, Rampp M, Rodriguez-Valera F, Pfeiffer F, Oesterhelt D. The genome of the square archaeon Haloquadratum walsbyi : life at the limits of water activity. BMC Genomics. 2006;7:169.

60. Lambros RJ, Mortimer JR, Forsdyke DR. Optimum growth temperature and the base composition of open reading frames in prokaryotes. Extremophiles. 2003;7(6):443–50.

61. Rocha EPC, Danchin A. Base composition bias might result from competition for metabolic resources. Trends Genet. 2002;18(6):291–4.

62. Forterre P, Da Cunha V, Catchpole R. Plasmid vesicles mimicking virions. Nat Microbiol. 2017;2(10):1340–1.

63. Erdmann S, Tschitschko B, Zhong L, Raftery MJ, Cavicchioli R. A plasmid from an Antarctic haloarchaeon uses specialized membrane vesicles to disseminate and infect plasmid-free cells. Nat Microbiol. 2017;2(10):1446–55.

64. Tamminen M, Virta M, Fani R, Fondi M. Large-scale analysis of plasmid relationships through gene-sharing networks. Mol Biol Evol. 2011;29(4): 1225–40.

65. Halary S, Leigh JW, Cheaib B, Lopez P, Bapteste E. Network analyses structure genetic diversity in independent genetic worlds. Proc Natl Acad Sci. 2010;107(1):127–32.

66. Badel C, Da Cunha V, Catchpole R, Forterre P, Oberto J. WASPS: web-assisted symbolic plasmid synteny server. Bioinformatics. 2020;36(5):1629–31.

67. Maguire F, Jia B, Gray KL, Lau WYV, Beiko RG, Brinkman FSL Metagenome-assembled genome binning methods with short reads disproportionately fail for plasmids and genomic islands. Microb Genom. 2020;6(10): mgen000436.

68. Krupovič M, Forterre P, Bamford DH. Comparative analysis of the mosaic genomes of tailed Archaeal viruses and proviruses suggests common themes for Virion architecture and assembly with tailed viruses of Bacteria. J Mol Biol. 2010;397(1):144–60.

69. Krupovič M, Bamford DH. Archaeal proviruses TKV4 and MVV extend the PRD1-adenovirus lineage to the phylum Euryarchaeota. Virology. 2008; 375(1):292–300.

70. Philosof A, Yutin N, Flores-Uribe J, Sharon I, Koonin EV, Béjà O. Novel abundant oceanic viruses of uncultured marine group II Euryarchaeota. Curr Biol. 2017;27(9):1362–8.

71. Team RC: R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria**.** In.; 2016.

72. Schbath S, Hoebeke M. R'MES: a tool to find motifs with a significantly unexpected frequency in biological sequences. In: Advances in Genomic Sequence Analysis and Pattern Discovery. Volume 7. World Scientific; 2011. p. 25–64.

73. Chessel D, Dufour AB, Thioulouse J. The ade4 package-l-one-table methods. R news. 2004;4(1):5–10.

74. Kuhn M. Building predictive models in R using the caret package. J Stat Softw. 2008;28(5):1–26.

75. Oksanen J, Guillaume Blanchet F, Kindt R, Legendre P: vegan: Community ecology package. R package version 2.3–5. In.; 2016.

76. Wickham H. ggplot2: elegant graphics for data analysis. New York: Springer; 2016. ISBN 978-3-319-24277-4. https://ggplot2.tidyverse.org.

77. Letunic I, Bork P. Interactive tree of life (iTOL): an online tool for phylogenetic tree display and annotation. Bioinformatics. 2007;23(1):127–8.

78. Halary S, McInerney JO, Lopez P, Bapteste E. EGN: a wizard for construction of gene and genome similarity networks. BMC Evol Biol. 2013;13(1):146.

79. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J Mol Biol. 1990;215(3):403–10.

80. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, Amin N, Schwikowski B, Ideker T. Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res. 2003;13(11): 2498–504.

81. Sullivan MJ, Petty NK, Beatson SA. Easyfig: a genome comparison visualizer. Bioinformatics. 2011;27(7):1009–10.

Publisher's Note

RESEARCH ARTICLE

Diversity of novel archaeal viruses infecting methanogens discovered through coupling of stable isotope probing and metagenomics

Vuong Quoc Hoang Ngo[1] 🔘 | François Enault[2] | Cédric Midoux[1,3,4] |
Mahendra Mariadassou[3,4] | Olivier Chapleur[1] | Laurent Mazéas[1] |
Valentin Loux[3,4] | Théodore Bouchez[1] | Mart Krupovic[5] | Ariane Bize[1] 🔘

[1]Université Paris-Saclay, INRAE, PRocédés biOtechnologiques au Service de l'Environnement, Antony, France

[2]Université Clermont Auvergne, CNRS, LMGE, Clermont-Ferrand, France

[3]Université Paris-Saclay, INRAE, MaIAGE, Jouy-en-Josas, France

[4]Université Paris-Saclay, INRAE, BioinfOmics, MIGALE Bioinformatics Facility, Jouy-en-Josas, France

[5]Institut Pasteur, Université de Paris, CNRS UMR6047, Archaeal Virology Unit, Paris, France

Correspondence
Ariane Bize, Université Paris-Saclay, INRAE, PRocédés biOtechnologiques au Service de l'Environnement, 92761 Antony, France.
Email: ariane.bize@inrae.fr

Abstract

Diversity of viruses infecting non-extremophilic archaea has been grossly understudied. This is particularly the case for viruses infecting methanogenic archaea, key players in the global carbon biogeochemical cycle. Only a dozen of methanogenic archaeal viruses have been isolated so far. In the present study, we implemented an original coupling between stable isotope probing and complementary shotgun metagenomic analyses to identify viruses of methanogens involved in the bioconversion of formate, which was used as the sole carbon source in batch anaerobic digestion microcosms. Under our experimental conditions, the microcosms were dominated by methanogens belonging to the order Methanobacteriales (*Methanobacterium* and *Methanobrevibacter* genera). Metagenomic analyses yielded several previously uncharacterized viral genomes, including a complete genome of a head-tailed virus (class *Caudoviricetes*, proposed family *Speroviridae*, *Methanobacterium* host) and several near-complete genomes of spindle-shaped viruses. The two groups of viruses are predicted to infect methanogens of the *Methanobacterium* and *Methanosarcina* genera and represent two new virus families. The metagenomics results are in good agreement with the electron microscopy observations, which revealed the dominance of head-tailed virus-like particles and the presence of spindle-shaped particles. The present study significantly expands the knowledge on the viral diversity of viruses of methanogens.

INTRODUCTION

Viruses infecting archaea represent one of the most unique parts of the global virosphere (Krupovic et al., 2018). Despite the limited number of archaeal viruses described so far compared to bacterial viruses, archaeal viruses show a great diversity of gene content and morphological properties. In particular, several morphotypes are specific to archaeal viruses, showing no

similarity to viruses infecting bacteria and eukaryotes, such as bottle-shaped (*Ampullaviridae*), coil-shaped (*Spiraviridae*), or spindle-shaped (*Bicaudaviridae*, *Fuselloviridae*, *Halspiviridae*, *Thaspiviridae*) ones. Archaeal viruses are currently classified into 33 families (Baquero et al., 2020; Liu et al., 2021), including cosmopolitan icosahedral viruses and archaea-specific viruses.

Methanogenic archaea play a major role in carbon cycling at the global scale, through methanogenesis.

Known methanogenic archaea are currently grouped into eight different orders in the phyla Euryarchaeota and Halobacteriota, and a few candidatus taxa in the Euryarchaeota, Halobacteriota, and in the TACK group (Evans et al., 2019; Lyu et al., 2018) These were isolated from very diverse natural ecosystems such as wetlands, termite, human and livestock digestive tracts, rice fields, and deep-sea hydrothermal vents, and also from anaerobic digesters (ADs) (Lyu et al., 2018). Indeed, their unique metabolic features are exploited in AD processes (Ahring, 2003) for valorization of organic waste and effluents into methane-rich biogas, a renewable energy source. Among archaeal viruses, 10 have been reported to infect methanogenic archaea (methanogens) (Krupovič et al., 2010a; Meile et al., 1989; Molnár et al., 2020; Nölling et al., 1993; Pfister et al., 1998; Thiroux et al., 2021; Weidenbach et al., 2017; Weidenbach et al., 2021; Wolf et al., 2019; Wood et al., 1989). In addition, several proviruses integrated in the genomes of diverse methanogens have been described (Krupovič et al., 2010; Krupovič & Bamford, 2008). Almost all of the viruses of methanogens described so far have been isolated from AD samples (Table S1). A total of five head-tailed viruses or proviruses originate from thermophilic ADs, all infecting Methanobacteriales hosts: ΨM1 and ΨM2 (family *Leisingerviridae*; siphovirus morphology) (Liu et al., 2021; Meile et al., 1989; Pfister et al., 1998) and related defective provirus ΨM100 (Luo et al., 2001) infect *Methanothermobacter* strains, whereas ΦF1 (unclassified) and ΦF3 (unclassified; siphovirus morphology) (Nölling et al., 1993) infect *Methanobacterium* species. Moreover, four viruses or virus-like particles (VLPs) have been isolated from mesophilic ADs: *Methanobacterium*-infecting virus Drs3 (family *Anaerodiviridae*; siphovirus morphology) (Liu et al., 2021; Wolf et al., 2019), Blf4 (unclassified; siphovirus morphology), infecting *Methanoculleus* strains (Methanomicrobiales) (Weidenbach et al., 2021), MetSV (unclassified) (Weidenbach et al., 2017), a spherical virus infecting *Methanosarcina* strains (Methanosarcinales), and finally the oblate or spindle-shaped *Methanococcus* (Methanococcales)-infecting A3 VLPs (Wood et al., 1989). MFTV1, an unclassified temperate head-tailed virus (siphovirus morphology), has been induced from a *Methanocaldococcus fervens* AG86 strain (Methanococcales) isolated from deep-sea hydrothermal vents (Thiroux et al., 2021); it is the first characterized virus infecting hyperthermophilic methanogens. In addition, MetMV (unclassified) (Molnár et al., 2020) has been suggested to infect mesophilic *Methanosarcina* strains, but this still needs to be confirmed. For the other four known orders of methanogens (Evans et al., 2019), no viruses have been isolated so far. Thus, the diversity of viruses infecting methanogenic archaea remains largely unexplored.

Metagenomics can provide a less biased view on the diversity of viruses infecting methanogens, by circumventing the challenge of cultivating some of the methanogenic archaeal strains, as well as biases associated with virus isolation. In such context, AD ecosystems prove to be particularly well suited as they are relatively easy to access to, and to establish in laboratory reactors. Moreover, they encompass methanogenic archaea from several orders, such as Methanobacteriales, Methanomicrobiales and Methanosarcinales (Evans et al., 2019; Lin et al., 2016). Yet, identifying viruses infecting methanogens is challenging in AD metagenomes due to the complexity of the catalytic microbial communities, dominated by diverse bacteria, and due to usually low proportions of methanogenic archaea in AD processes.

In the present study, an original experimental approach was applied with the aim of favouring the enrichment of AD microbial communities in methanogens, to help the discovery of their viruses in metagenomes. To this end, AD microcosms were fed with ^{13}C-labelled formate, one of the known substrates for methanogenesis through the following equation (Sun et al., 2021):

$$4HCOO^- + H^+ + H_2O \rightarrow CH_4 + 3HCO_3^- \left(\Delta G^0 : -170.2 kJ/mol\right)$$
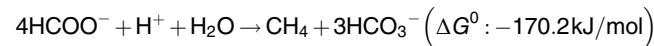
In addition to favouring methanogens, such an experimental approach has the advantage of preserving the AD process and a certain level of microbial diversity. In particular, it can enable to reach higher proportions of methanogens, and it is also compatible with the presence of several methanogenic species or genera, offering the possibility of a relatively broad view on the diversity of viruses of methanogens. Another benefit of this method is the possibility to identify DNA viruses targeting microorganisms that actively assimilated the substrate. Indeed, within complex microbial communities, not all of the microorganisms are active or involved in the degradation of a specific substrate. Thus, identifying active microorganisms, and their associated viruses, are key issues in linking viral diversity with functional aspects. Hence, we coupled stable isotope probing (SIP) (Radajewski et al., 2000), applied to the total cellular DNA, to *in silico* host prediction of the viral contigs. Host prediction was applied by using the cellular metagenomes obtained from the heavy (^{13}C-enriched) and light (not enriched in ^{13}C) DNA sequences, which were used to build host databases. Whenever the predicted host was identified in the heavy cellular DNA fraction, we considered it as evidence that the virus infected active microorganisms, assimilating the ^{13}C-formate.

This original coupling (SIP and bioinformatic analyses for host prediction) led to the discovery of several previously uncharacterized genomes of DNA viruses of methanogens. In particular, they included two contigs likely representing new families of spindle-shaped

viruses, thereby expanding the knowledge on the diversity of viruses infecting methanogens. One of these families was associated with the active hydrogenotrophic methanogenic archaea, while the other one seemed to target methanogens that were not assimilating the formate in the studied microcosms, which could only be evidenced thanks to the SIP analysis.

EXPERIMENTAL PROCEDURES

AD microcosm experiments and monitoring procedure

AD batch microcosms consisted in glass plasma bottles with a total volume of 500 ml. Either ^{13}C-labelled-formate or unlabelled formate was employed as the sole carbon source. A mineral solution suitable for AD (standard NF EN ISO 11734) was also added, as well as an inoculum prepared from a lab-scale digester fed with biowaste. To obtain the inoculum, the sampled anaerobic sludge was incubated in anaerobic conditions at 35°C for 20 days, for the majority of fermentable compounds to decompose. It was then centrifuged (10,000g, 20 min, 15°C), aliquoted and stored at −80°C until further use. The microcosms were hermetically sealed with an aluminium screw cap and a rubber septum. The headspaces were flushed with nitrogen gas (Linde). For each treatment (^{13}C-labelled or unlabelled formate), microcosms were prepared in triplicates with a working volume of 300 ml, using the mineral solution, 3 g of inoculum, and either 0.2 M ^{13}C-sodium formate (Cortecnet) or 0.2 M unlabelled sodium formate (Sigma Aldrich) (equivalent to 4.08 g).

During incubation, the following physicochemical parameters were monitored according to the protocols described in Puig-Castellví et al. (2022): biogas production, biogas composition, pH, isotopic composition of the biogas, volatile fatty acids (VFAs), total inorganic carbon (TIC), total organic carbon (TOC) and chemical oxygen demand.

To characterize the dynamics of microbial communities (archaea, bacteria and in particular their viruses), a pair of microcosms (one fed with ^{13}C-formate, one with unlabelled formate) was sacrificed at each of three different time points (days 8, 13 and 17), for cellular and viral DNA extraction. For each sacrificed microcosm, the liquid phase was centrifuged (8000g, 20 min, 15°C). The pellet was stored at −80°C for subsequent cellular DNA extraction, and the supernatant (containing VLPs) was collected and stored at 4°C for virion preparation.

Preparation and observation of VLPs

The supernatant resulting from the previous step was centrifuged once more (6000g, 20 min, 4°C) to further remove cells. The newly obtained supernatant was subjected to a three-step filtration with 1.2 and 0.8 μm polyethersulfone filters (Sartorius), and finally with 0.2 μm acetate cellulose filters. The filtrations were realized in 500 ml filtration units (Thermo Scientific) under vacuum. The final filtrate was centrifuged (40,000g, 3 h, 4°C) to pellet VLPs. The final pellet was suspended in 2 ml SM buffer (0.1 M NaCl, 0.1 M MgSO$_4$, 0.05 M Tris–HCl, pH 7.5) and stored at 4°C until subsequent analysis.

VLPs were observed by transmission electron microscopy (TEM) at MIMA2 MET - GABI, INRAE, AgroParisTech (78352 Jouy-en-Josas, France). Virion-containing solutions were adsorbed onto a carbon film membrane on a 300-mesh copper grid, stained with 1% uranyl acetate, dissolved in distilled water, and dried at room temperature. Grids were examined with a Hitachi HT7700 electron microscope operated at 80 kV (Elexience – France), and images were acquired with a charge-coupled device camera (AMT).

DNA extraction and quantification

Total cellular DNA was extracted with Qiagen DNeasy PowerSoil Kit, according to the manufacturer's instructions.

For VLPs, 360 μl of virion solution were treated with 2 U of Turbo DNase (2 U μl^{-1}, Ambion), and 22 μl of RNaseA (1 mg/ml, Ambion) for 1 h at 37°C, to remove contaminant non-encapsidated DNA. Inactivation buffer from the Ambion kit was added to stop the DNase activity. DNA extraction was then based on a classical phenol-chloroform method (Pickard, 2009), using 1.5 ml Phase lock gel light (VWR). The detailed procedure for viral DNA extraction is provided in Supplementary information (Section 1.1).

Concentrations of viral DNA were determined with Qubit dsDNA HS Assay Kit (Invitrogen), according to the manufacturer's instructions.

DNA ultracentrifugation

Ultracentrifugation on a CsCl gradient was applied to the total cellular DNAs to separate them according to their density, as described in Chapleur et al. (2016). Around 1 μg of DNA was employed for each sample. For each collected fraction, the DNA quantification was performed with Qubit dsDNA HS Assay Kit (Invitrogen). Based on DNA density profiles [Figure 1(D)], the fractions were pooled into three main groups (referred to as 'fraction pools' thereafter) and DNA was purified by glycogen-polyethylene glycol (PEG) 6000 precipitation, according to Neufeld et al. (2007).

The DNA of each fraction pool was amplified with RTG GenomiPhi™ V3 DNA Amplification Kit (illustra),
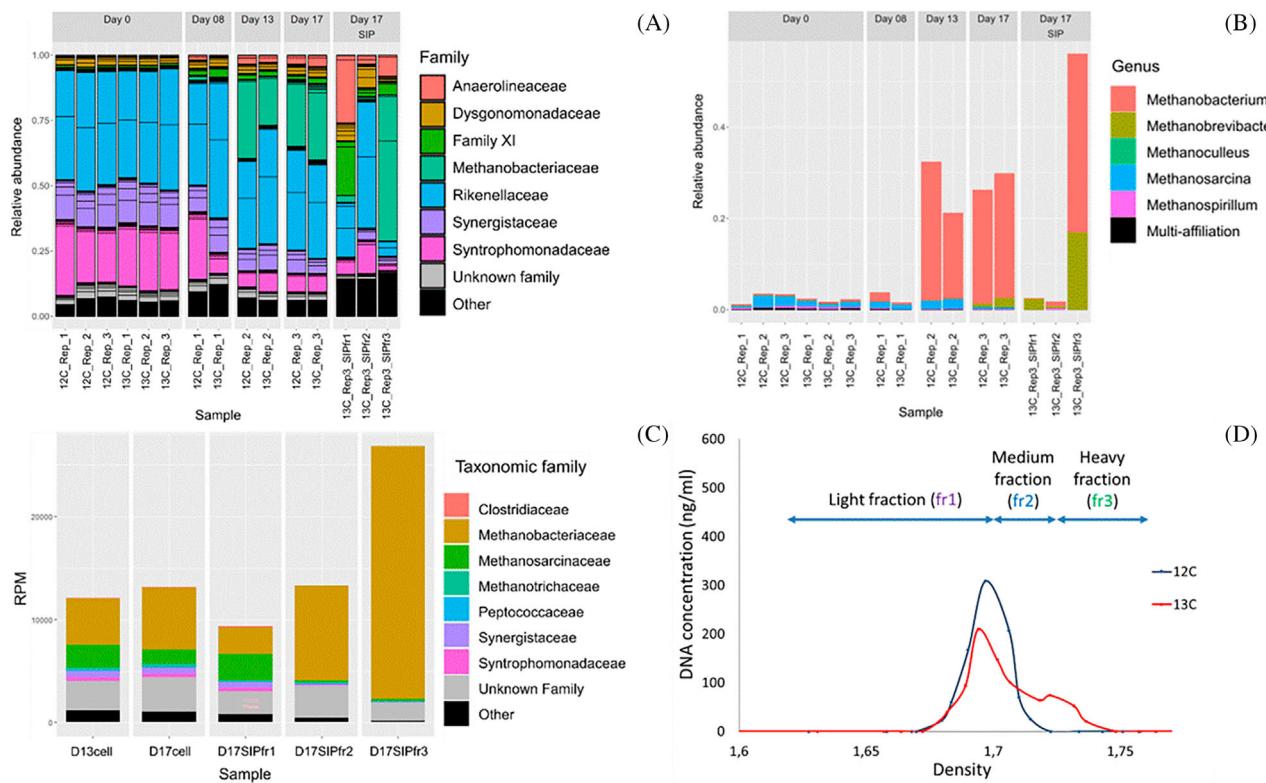
FIGURE 1 Microbial dynamics during AD incubation. (A) Microbial community composition of archaea and bacteria at the family level, based on 16S rRNA gene metabarcoding. (B) Composition of archaea only, at the genus level, based on 16S rRNA gene metabarcoding. (C) Taxonomic assignation (order level) of the microorganisms involved in methane metabolism, based on the functional analysis of the cellular shotgun metagenomes. (D) Density profiles of the cellular DNAs obtained from microcosms fed with unlabelled formate (blue) or with ^{13}C-formate (red) at day 17. The arrows labelled as fr1, fr2 and fr3 indicate how the DNA fractions were pooled in the case of labelled formate. fr1 contained only unlabelled DNA, whereas fr3 contained only ^{13}C-labelled DNA. fr2 likely contained a mix of ^{13}C-labelled and unlabelled DNA

according to the manufacturer's instructions. It was then purified by glycogen-PEG 6000 precipitation, and the concentration of the collected DNA was determined with Qubit dsDNA HS Assay Kit (Invitrogen).

16S rRNA gene metabarcoding

Archaeal and bacterial hypervariable regions V4–V5 of the 16S rRNA gene were amplified and then sequenced according to the protocol described in Poirier et al. (2016) and Madigou et al. (2019), with some modifications. The targeted region was amplified by PCR with fusion primers 515F (5′-Ion A adapter–Barcode–GTGYCAGCMGCCGCGGTA-3′) (Wang et al., 2007) and 928R (5′-Ion trP1 adapter–CCCCGYCAATTCMTTTRAGT-3′) (Wang & Qian, 2009). The detailed protocol for library preparation is provided in the Supplementary Information (Section 1.2). Sequencing was performed on an Ion Torrent Personal Genome Machine using Ion 316 Chip V2 (Life Technologies) and Ion PGM Hi-Q View Sequencing Kit (Life Technologies) according to the manufacturer's instructions. Sequencing data were processed with the Torrent Suite

Software. All diversity analyses were performed with the R phyloseq package.

Shotgun metagenomic sequencing

Illumina sequencing was performed at the I2BC sequencing platform (University of Paris-Saclay, Gif-sur-Yvette, France). 250–500 ng of genomic DNA was fragmented (400 bp mean size) on a Covaris S220 sonicator. DNA fragments were end-repaired and dA-tailed (NEB#E7595), Illumina TruSeq adapters were ligated (NEB#E6040), and the PCR-free library fragments were purified using AMPure XP beads (Beckman Coulter). Final library quality was assessed on an Agilent Bioanalyzer 2100, using an Agilent High Sensitivity DNA Kit. Libraries were pooled in equimolar proportions and sequenced using paired-end 2×150 pb runs, on an Illumina NextSeq500 instrument, using NextSeq 500 High Output 300 cycles kit.

Demultiplexing was performed with bcl2fastq2 v2.18.12. Adapters were trimmed with Cutadapt v1.15, and only reads longer than 10 b were kept for further analysis.

Metagenomic pipe-line

The most generic steps of our pipeline were scripted as a snakemake (Köster & Rahmann, 2012) workflow (https://forgemia.inra.fr/cedric.midoux/workflow_metagenomics/-/tree/v21.04), and applied to both cellular and viral metagenomes. After a pre-processing step (adapter removal, and trimming according to quality scores and length with fastp), reads were assembled with metaSPADES (Nurk et al., 2016) (individual assembly by sample for cellular metagenomes and coassembly for metaviromes). Coding regions were predicted with Prodigal (Hyatt et al., 2010). Taxonomic affiliations of contigs and their predicted genes were respectively obtained with CAT (von Meijenfeldt et al., 2019) and kaiju (Menzel et al., 2016), against the NCBI nr database. Genes were annotated using comparison to NCBI nr database with Diamond (Buchfink et al., 2015). For each dataset, the cleaned reads were mapped to the assembled contigs using Bowtie2 (Langmead & Salzberg, 2012) and counted with samtools (Li et al., 2009). Details of versions and parameters for the snakemake workflow are available in the GitLab repository.

Several steps specifically dedicated to viral contig analysis and to host prediction were performed using homemade bash and python scripts. Metagenome-assembled genomes (MAGs) were constructed from assembled cellular metagenomic contigs with Metabat2 (Kang et al., 2019). MAG quality was improved with RefineM (Parks et al., 2017) and controlled with CheckM (Parks et al., 2015). Only MAGs with more than 60% completeness and less than 5% contamination were selected. Functional annotations of the cellular predicted genes were obtained with ghostKoala against KEGG database (Kanehisa et al., 2016). Concerning the viral metagenomes, their quality was assessed with ViromeQC (Zolfo et al., 2019) based on the trimmed reads. Viral genome detection was performed with VIBRANT (Kieft et al., 2020) and VirSorter2 (Guo et al., 2021), and the quality of viral contigs was assessed with CheckV (Nayfach et al., 2021). The predicted genes were further analysed by using HH-suite (Steinegger et al., 2019) against PHROGs (Terzian et al., 2021), a database dedicated to prokaryotic viruses.

To predict potential host for each viral contigs, two different methods were used. First, CRISPR spacers were detected in cellular metagenome contigs with CRISPRdetect (Biswas et al., 2016) and CRISPRCasFinder (Couvin et al., 2018). A non-redundant spacer database was built from the obtained spacer sequences. The viral contigs were subsequently aligned with BLASTn and Space-PHARER (Zhang et al., 2021) against this homemade database and a public spacer database (CRISPRCasdb spacer: https://crisprcas.i2bc.paris-saclay.fr/Home/DownloadFile?filename=spacer_34.zip), enabling host prediction based on an alignment-dependent method. Hosts were also predicted using the WIsH software (Galiez et al., 2017) to compare the genomic signatures of viral contigs to those of cellular MAGs.

After selecting viral contigs of interest as described in the result section, gene annotation was refined on HHpred server (Zimmermann et al., 2018) with default parameters, against four structural/domain databases: Pfam-A_v35, PDB_mmCIF70_12_Ot_2021, NCBI_Conserved_Domains(CD)_v3.18 and UniProt-SwissProt-viral70_3_Nov_2021.

The detailed parameters and versions for the python and bash scripts are presented in Supplementary Tables S2 and S3.

Bipartite network of viral contigs and protein orthologous groups

A bipartite network was built, where the two node classes were viral contigs or genomes on the one hand, and protein orthologous groups (OGs) on the other. Regarding 'viral nodes', most archaeal (pro)virus genomes available in public databases (RefSeqVirus and nr, 7th of October, 2021) or described in previous studies (Krupovič et al., 2010a; Philosof et al., 2017) were included, in addition to the viral contigs of interest. The final dataset consisted of 172 viral sequences, in total. All the proteins encoded in these sequences were categorized into OGs by a two-step procedure as described in Olo Ndela et al. (2021) and in the Supplementary Information (Section 1.2.2).

These OGs were functionally annotated by comparing their HMM profiles to those of the PHROGs database (Terzian et al., 2021) (annotation release v2), which contains well-annotated protein clusters of prokaryotic viruses. In addition, protein sequences were compared to RefSeqVirus using MMseqs (bit score ≥50), retaining only the best hit for each protein of an OG. These OGs were used as 'protein nodes' to construct the bipartite networks. The shared OGs and their functional prediction are listed in the Supplementary Table S4. All networks were computed using the igraph package of R (Csardi & Nepusz, 2005).

The data for this study have been deposited in the European Nucleotide Archive (ENA) at EMBL-EBI under accession number PRJEB46489 (https://www.ebi.ac.uk/ena/browser/view/PRJEB46489) (16S metabarcoding raw reads, shotgun metagenomics raw reads, annotated sequences of contigs C158, C889, C1359 and C1697).

RESULTS AND DISCUSSION

A significant proportion of formate was converted to methane through hydrogenotrophic methanogenesis

A total of six batch AD microcosms were established and monitored over time for a maximum of 17 days. They contained, as sole carbon source, either

unlabelled formate (three replicates) or ^{13}C-formate (three replicates). At each of days 8, 13 and 17, one pair of microcosms was sacrificed for metavirome analysis, one with unlabelled formate and one with ^{13}C-formate.

Chemical analyses (VFA, TIC, TOC) showed that both unlabelled and ^{13}C-labelled formate were consumed, and more than 85% of the initial quantity was metabolized after 17 days of incubation (Supplementary Figure S1). The bioconversion resulted in a pH increase from ∼7.50 to ∼9.20 on average, consistent with the methanogenesis equation mentioned above. Biogas was produced after a lag phase of about 5 days (Supplementary Figure S1) to reach final cumulated productions of ∼100 normo-ml. Furthermore, reproducible dynamics were observed in microcosms, irrespective of the formate type. Methane (CH_4) was by far the dominant biogas component, followed by carbon dioxide (CO_2) as well as traces of hydrogen (H_2) and hydrogen sulfide (H_2S). Collectively, these results support the successful establishment of the AD process in the microcosms. More precisely, the isotopic composition of the biogas demonstrated the dominance of ^{13}C in the methane (>95%) produced by the methanogens in the microcosms fed with ^{13}C-formate and indicated the dominance of the hydrogenotrophic methanogenesis pathway (Whiticar et al., 1986) in the microcosms fed with ^{12}C-formate. For unlabelled substrates, this method relies on the abundance of the stable isotope ^{13}C in nature (1.1%) and on the difference in reaction rate between ^{12}C- and ^{13}C-containing substrate molecules. The results and detailed calculation of isotopic signatures are provided in the Supplementary Information (Section 2.1).

The proportion of methanogenic archaea increased over time and they were dominated by *Methanobacterium* species

To identify the methanogenic archaea and more broadly the microbial community composition, we applied 16S rRNA gene metabarcoding [Figure 1(A)]. The relative abundance of the dominant archaea (Methanobacteriaceae, in green) increased from 1%–3% at day 0 to 26%–30% at day 17. Consistently, a functional analysis based on the cellular shotgun metagenomic data (Supplementary Figure S2) showed the dominance of methane metabolism, in relative abundance, at day 17. The KEGG category 'methane metabolism' includes methanogenesis, methane oxidation and metabolisms related to intermediate molecules of both these pathways. Among the microbial groups involved in methane metabolism [Figure 1(C)], archaea were the main actors (Methanobacteriaceae in brown, Methanosarcinacea in green and Methanotrichaceae in cyan). Similar proportions of archaea were observed in the metagenomic dataset (Supplementary Figure S3).

These results confirmed that the microbial communities were enriched in methanogenic archaea during the incubation.

The detected archaea were mostly methanogens [-Figure 1(B)]. At day 0, *Methanosarcina* (Methanosarcinales order) was dominant. However, at the end of the incubation, at day 17, the genus *Methanobacterium* became dominant, followed by *Methanobrevibacter* (both from the order Methanobacteriales). It can be assumed that Methanobacteriales members were selected during the incubation since they are able to grow on formate through the hydrogenotrophic pathway, and since they likely outcompeted Methanosarcinales methanogens, due to their faster growth rate: the doubling time is generally lower than 1 h in Methanobacteriales, compared to more than 10 h in Methanosarcinales (Thauer et al., 2008).

Concerning bacterial families present in these systems, a notable decrease in relative abundance was observed over time for *Synergistaceae* (phylum Synergistetes, from ∼15% down to ∼7%) and *Syntrophomonadaceae* (phylum Firmicutes, from ∼24% down to ∼7%) [Figure 1(A)]. *Synergistaceae* members generally consume amino acids to generate short-chain fatty acid (He et al., 2018), whereas *Syntrophomonadaceae* members are acetogens (Si et al., 2016). For both families, the decrease is understandable due to the lack of adequate substrate (proteins and short-chain fatty acids). A notable increase in relative abundance was observed at day 17 for members of *Anaerolineaceae* (phylum Chloroflexi, from >1% up to ∼4%), which are generally described as fermentatives or acetogens (Liang et al., 2015; Si et al., 2016), and some of them were previously shown to form syntrophic associations with methanogens (Lei et al., 2018). In the present experiment, *Anaerolineaceae* bacteria may degrade formate and/or play a role in electron transfer (Wang et al., 2021), in partnership with hydrogenotrophic methanogens.

For the six microcosms, the profiles of cellular DNA concentrations in function of their mass density were established (Supplementary Figure S4). For microcosms fed with ^{13}C-formate at day 17 [Figure 1(D), red line], a second peak, of denser DNA, was visible in the profile, indicating that the labelled substrate was assimilated by some of the microorganisms. The separation of DNA in CsCl gradient depends not only on the mass of the isotope but also on the GC content of the DNA (Eason & Campbell, 1978). Hence, we pooled the DNA into three distinct fractions, according to their density. The first fraction (fr1, density <1.705) corresponded exclusively to non-labelled DNA. The second fraction (fr2, 1.705 < density <1.738) possibly contained a mix of unlabelled and ^{13}C-labelled DNA. Finally, the third fraction (fr3, density >1.738) contained exclusively ^{13}C-labelled DNA. The 16S rRNA gene metabarcoding applied to these fractions showed that archaea reached
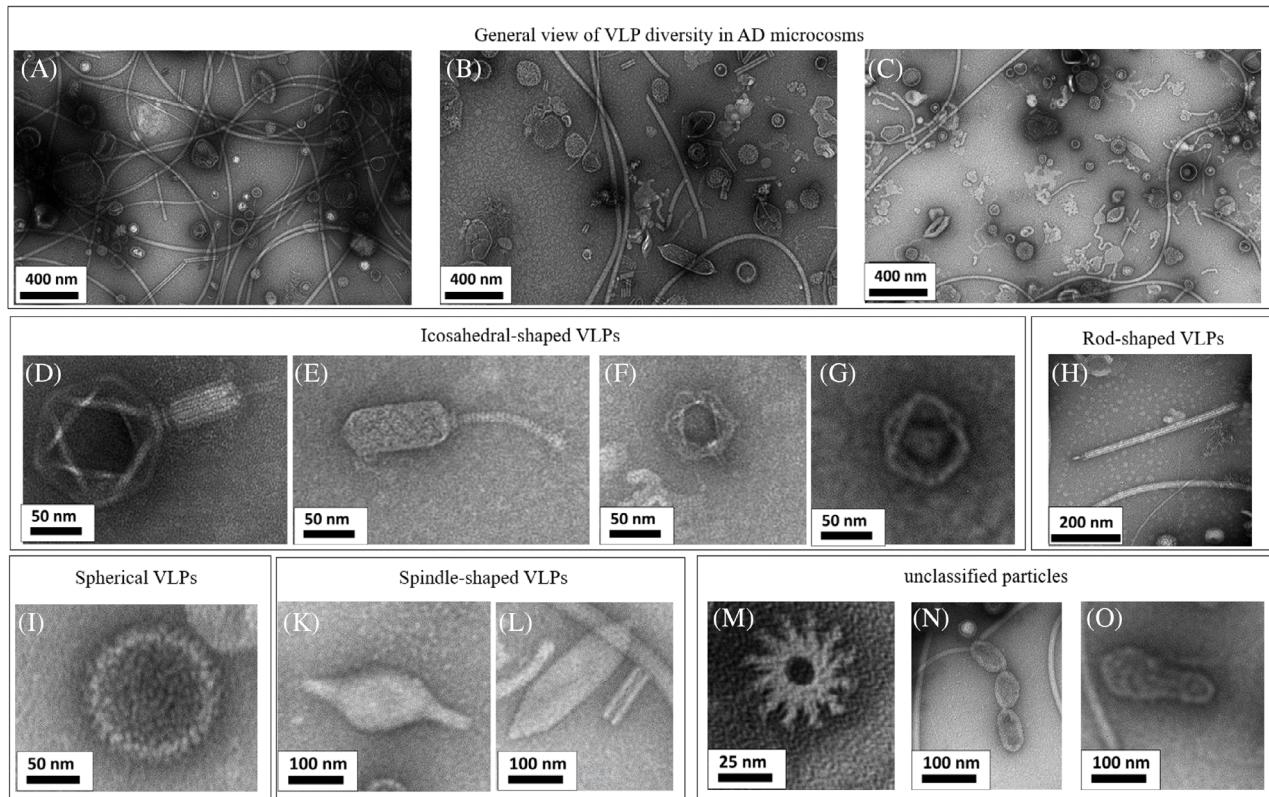
FIGURE 2 Morphotypic diversity of VLPs in different microcosms, observed by TEM. One representative sample is shown for each incubation time point (A: day 8, B: day 13, C: day 17)

their highest proportions (56%) in fraction fr3 at day 17 [Figure 1(B)], whereas their proportions in fractions fr1 and fr2 were lower than 3%. Consistently, the highest abundance of genes involved in methane metabolism was in fraction fr3 at day 17, based on metabolic pathway analysis from the shotgun metagenomics data [Figure 1(C)]. These observations confirmed that the ^{13}C-formate incorporated into the microbial biomass was consumed mostly by methanogenic archaea of the genera *Methanobacterium* and *Methanobrevibacter*.

TEM evidenced the presence of icosahedral and spindle-shaped VLPs

TEM observations revealed a great diversity of VLPs in the microcosms (Figure 2), especially after 13 and 17 days of incubation [Figure 2(A,C)]. Besides cosmopolitan morphotypes common to the domains Bacteria and Archaea, uncommon and especially archaea-specific morphotypes were also observed.

Cosmopolitan morphotypes included VLPs with a head-tailed morphology, typical of the class *Caudoviricetes*, as well as icosahedral tailless particles [Figure 2 (G)]. The latter could originate from tailless icosahedral viruses or from head-tailed viruses with a broken tail. In both samples, these VLPs were the most abundant and

presented a large diversity: myovirus-like [i.e. icosahedral capsids with contractile tails; Figure 2(D)], siphovirus-like [i.e. icosahedral capsids with long non-contractile tails; Figure 2(E)] and also podovirus-like [i.e. icosahedral capsids with short tails; Figure 2(F)] morphotypes were observed, with capsid diameters ranging from 50 to 200 nm.

Less common viral morphotypes were detected in the microcosms at days 13 and 17, such as rod-shaped [Figure 2(H)], spherical [Figure 2(I)] and spindle-shaped [Figure 2(K–L)]. Spindle-shaped morphotypes, which are specific to archaeal viruses, have been commonly observed in extreme geothermal and hypersaline environments (Krupovic et al., 2014), but have also been reported in moderate ones, such as freshwater and marine habitats (Borrel et al., 2012; Kim et al., 2019). Interestingly, the presence of spindle-shaped VLPs has also been reported in AD plants (Calusinska et al., 2016), suggesting that the hosts of these viruses could possibly be involved in the AD process.

Particles with unique morphotypes were also identified [Figure 2(M–O)]. Some of them had a chain structure with multiple of two or three monomers and their size ranged from 100 to 400 nm. Moreover, other particles appeared as round-shaped and less than 50 nm, with a hollow star-like structure [Figure 2(M)]. The nature of these particles is unclear.

TABLE 1 Overview of the shotgun metagenome datasets

	Cellular metagenomes (individual assemblies)					Viral metagenomes (co-assembly)	
	D13cell	D17cell	D17SIPfr1	D17SIPfr2	D17SIPfr3	D13vir	D17vir
Number of raw reads (millions)	85.68	63.76	75.41	78.78	65.54	38.92	46.48
Reads obtained after trimming (%)	98.39	98.08	97.80	98.35	98.40	98.41	98.61
Number of contigs (≥1 kb)	77,032	68,382	58,308	33,123	21,159	23,016	
Number of contigs (≥3 kb)	18,093	15,244	14,728	9432	5108	5571	
Max contig length (b)	622,910	622,910	449,342	388,850	389,221	372,273	
Contigs with taxonomic affiliation (%)	96.62	96.67	96.93	97.32	97.22	87.59	
Reads mapped to contigs (%)	89.35	87.74	89.03	93.17	93.73	87.20	79.85
Number of MAGs	81	67	75	55	28	–	
Number of archaeal MAGs	8	10	8	7	5	–	
Number of selected MAGs (≥60% completeness and ≤5% contamination)	81 (including 17 from archaea)						

Overview of the shotgun metagenomics datasets and of the host prediction approaches

Shotgun metagenomic sequencing was applied to seven selected DNA extracts. Four of the extracts were from the total cellular DNA and total viral DNA from samples collected at days 13 and 17 (D13cell, D13vir, D17cell, D17vir), selected due to their enrichment in methanogens observed in the metabarcoding analysis (Figure 1) and the presence of interesting VLP morphotypes (Figure 2). The three cellular DNA fractions obtained after density gradient centrifugation at day 17 were also sequenced (D17SIPfr1, D17SIPfr2, D17SIPfr3) [Figure 1(D)]. The purpose was primarily to analyse the metaviromes, and to use the cellular metagenomes for building specific databases for host prediction.

Based on classical metrics (Table 1), the sequencing data and the assemblies were of satisfactory quality. In the case of metaviromes, ViromeQC analysis indicated moderate contamination rates of 7.6% and 8.8% for D13vir and D17vir, respectively. Among 23,016 assembled contigs longer than 1 kb, 1927 were exclusively detected in D17vir, suggesting that most of the corresponding viruses were selected during the late stages of the incubation. Among the 87.59% of contigs that obtained a taxonomic annotation, 98.43% were affiliated to prokaryotes, which can be explained primarily by database biases: prokaryotes were more represented than their viruses in the NCBI nr database used for the taxonomic assignment, and many viral genes could therefore have their best match in prokaryotic genomes, in particular in proviruses. Such bias is a strong limitation for the identification of viruses and their taxonomic classification. Yet, it can bring information on the possible hosts of the viral contigs.

Viral genome detection was performed with CheckV, VIBRANT and VirSorter2. All these tools rely on protein similarity profiles and both latter detect protein profiles by machine learning. CheckV, VirSorter2 and VIBRANT identified 3898, 3968 and 3904 viral genomes (6484 distinct genomes in total), respectively, with most of the genomes being linear, double stranded and predicted to belong to virulent viruses. These numbers were low compared to the number of contigs longer than 1 kb, likely reflecting the fact that shortest contigs contain insufficient information for confident virus identification.

For cellular metagenomes, a taxonomic affiliation was obtained for more than 96% of the contigs, suggesting that they would constitute a good-quality database of host sequences. MAGs were reconstructed by binning of the cellular contigs: in total, 81 MAGs (completeness >60% and contamination <5%) were selected as reference hosts, including 17 assigned to the domain Archaea.

For host prediction, two complementary methods were used. The first one relied on matching of CRISPR spacers to the viral contigs. Spacers are short sequences (26–72 bp) (Makarova et al., 2011) of plasmid or viral origin, which are integrated into the host genomes by CRISPR-Cas (Nasko et al., 2019), adaptive immunity systems identified in 85% of Archaea and 40% of Bacteria (Makarova et al., 2020). The second approach relied on signatures, as most prokaryotic viruses have a genome k-mer composition similar to the one of their hosts, due to their co-evolution (Edwards et al., 2016). Similar genomic signatures were searched between viral contigs and the 81 microbial MAGs as potential hosts, using WIsH (Galiez et al., 2017). For contigs longer than 3 kb, the spacer-based method revealed a total of 375 contigs with a predicted host, whereas the signature-based method

23,016 contigs with length ≥1 kb

Filter 1: Selecting contigs with length ≥3 kb

Number of selected contigs: 5 571

Filter 2: Removing cellular contaminant contigs, fitting the three following criteria:
1. No hit in RefseqVir OR PHROGs
AND 2. Not detected as viral by CheckV NOR VIBRANT NOR Virsorter2
AND 3. All the genes have a hit in the home-made KEGG database impoverished in proviruses

Number of selected contigs: 3 435

Filter 3: Focus on archaeal OR viral contigs

Taxonomic affiliation

Contig level: Unannotated contigs | OR | Gene level: At least 1 gene affiliated to Archaea/virus

AND

Host prediction

Alignment-dependent methods (SpacePharer, Blastn): Archaea as predicted host | OR | k-mer profile comparison (WiSH): Archaea as predicted host & contig size ≥5 kb

Number of selected contigs: 267

Filter 4: Manual selection of putative archaeal virus contigs

39 contigs of interest

FIGURE 3 Strategy for the selection of contigs of interest, likely originating from viruses of methanogens

against our homemade MAG database predicted an acceptable host (*p*-value ≤0.05) for 3239 contigs. Coupling of these methods resulted in a total of 3466 contigs with a predicted host, showing their complementarity to improve the capability and accuracy of *in silico* host prediction.

The *Caudoviricetes* class dominated among the 39 contigs likely originating from viruses of methanogens

As our aim was to identify contigs of viruses infecting methanogenic archaea, we applied successive stringent filters and manual curation, relying on the integration of results from complementary bioinformatic analyses (Figure 3). A first filter based on contig length was applied to eliminate contigs which would contain only limited information. The second filter aimed at removing an important fraction of the cellular

contaminants. In the third step, contigs possibly originating from archaea or archaeal viruses were selected. Finally, the contigs obtained at this stage were analysed manually, to remove most remaining contaminant contigs (typically originating either from cells or from bacterial viruses).

A total of 39 contigs were thereby selected as contigs of interest for further analysis, as putatively originating from archaeal viruses. Due to the limited accuracy of some host prediction tools [e.g. of the order of 70%–80% at the phylum level (Galiez et al., 2017)], this selection of contigs still possibly contained a few contaminants. The selected contigs ranged in lengths from 4.2 to 53 kb (median: 9.4 kb) and the reads per kilobase per million mapped reads (RPKM) values from 0.07 to 468.27 for D13vir, and from 0.20 to 229.40 for D17vir. Nine of these contigs could not be assigned to any known taxon (either cellular or viral), suggesting them to be previously uncharacterized. Two contigs, C1661 and C1697, had best-predicted hosts as MAGs
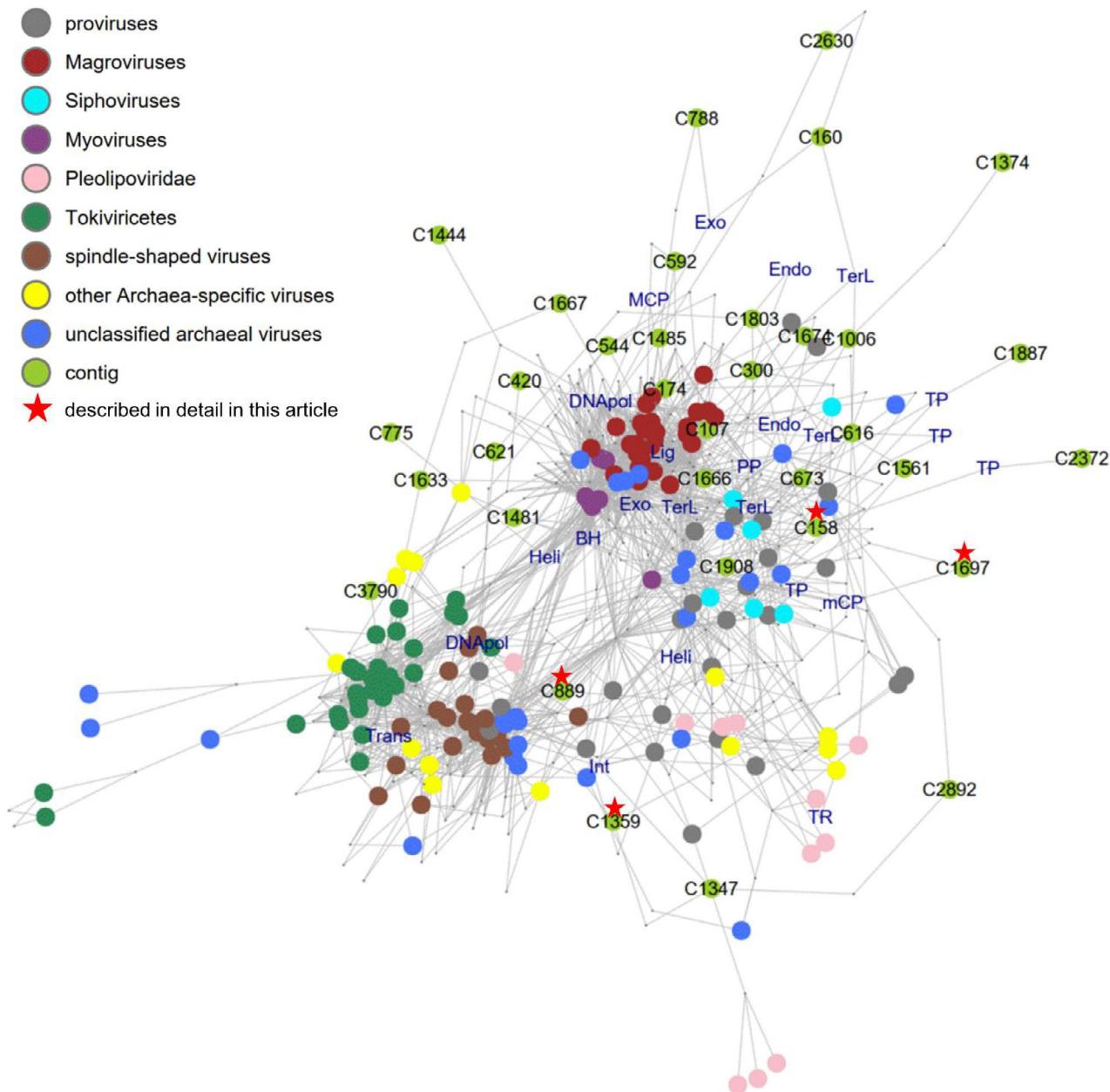
FIGURE 4 Bipartite network of known archaeal viruses and of the 39 contigs of interest, and of protein orthologous groups (OGs). Most of the contigs of interest originate from archaeal viruses. The genomes and contigs are represented as circles coloured according to the legend shown in the figure, and OGs are denoted by the intersections of edges. Some shared OGs with a functional annotation are represented. TerL: Terminase large subunit, MCP: major capsid protein, mCP: minor capsid protein, Exo: exonuclease, TP: tail protein, PP: portal protein, Lig: ligase, DNApol, DNA polymerase, BH: baseplate hub protein, Heli: helicase, Trans: transposase, TR transcription regulator

from the unlabelled metagenome D17SIPfr1, suggesting that they represent viruses infecting inactive and minor methanogens in our microcosms. Besides, 34 were predicted as virulent by VIBRANT and only one, C1485, was predicted as temperate (Supplementary Figure S5, Supplementary Table S5). Genome scaffold quality results were obtained through three different bioinformatic tools: VIBRANT, CheckV and VirSorter2: 35 contigs had low or medium quality. However, given that all of these tools were developed

based on the viral databases overwhelmingly dominated by bacterial *Caudoviricetes*, their accuracy on datasets including novel archaeal viruses, especially those with small or medium-sized genomes, is not to be expected. All the results from the bioinformatics tools are available for this set of 39 contigs, in Supplementary Table S5.

A bipartite network was built (Figure 4) to evaluate the similarity between the contigs of interest and the archaeal (pro)viruses described in the literature. Such

networks enable to represent complex systems comprising two distinct classes of components (nodes) (Iranzo et al., 2016b). Here, the two classes of node correspond to viral contigs or genomes on the one hand, and to protein OGs on the other. Two viral nodes can be connected only indirectly, through shared OG nodes. Previously published archaeal pro(viral) genomes were labelled according to their taxonomic affiliation, when established. The top half of the network contained viral genomes from (pro)viruses with a cosmopolitan head-tailed morphotype (siphovirus, myovirus, magrovirus, provirus). In the bottom half, the presence of various viral families specific to archaea was observed. Such a network topology is consistent with the spatial distribution in archaeal virus networks previously described in the literature (Iranzo et al., 2016a). Moreover, head–tail viruses (class *Caudoviricetes*) are known to be highly mosaic (Krupovič et al., 2010), hence generating tightly interconnected networks.

Out of 39 contigs of interest (green), 24 shared at least one OG with (pro)viruses belonging to the class *Caudoviricetes*, including virion morphogenesis proteins, such as terminase large subunit (TerL), various tail proteins, major/minor capsid proteins (MCP/mCP) and baseplate hub proteins (BH).

To further explore the organization of the network and relationships among the 39 contigs of interest and archaeal (pro)viruses, they were clustered according to the presence/absence of shared genes (see Experimental procedures). Consistent with the network analysis, the two main viral clusters (VC) were related to the class *Caudoviricetes* (Supplementary Figure S6). The first cluster, VC4, included four contigs, C1803, C174, C300 and C616, and was held together through OGs related to the capsid formation and packaging module, and/or DNA, RNA and nucleotide metabolism (Supplementary Figure S6). Moreover, in the longest contig of VC4, C174, a sheath protein was detected, also suggesting its affiliation to viruses with contractile tails (myovirus-like morphology) (Fokine & Rossmann, 2014). For all the contigs in this cluster, only WIsH predicted an archaeal host, whereas the other tools employed showed either a bacterial host or no identified one. To ascertain the host assignment, the taxonomic and functional annotations of each gene from the cluster were examined: no single gene was assigned to Archaea or annotated as an archaeal protein, while many were related to Bacteria, suggesting that bacterial hosts were more probable. It highlights the importance of relying on different complementary tools for more accurate host prediction.

The second cluster, VC10, included five contigs, C1666, C1908, C158, C673 and C1006. Several OGs shared among members of the cluster were related to various viral functional modules: virion morphogenesis (including capsid and tail formation, and genome packaging), DNA/RNA and nucleotide metabolism, as well as integration and excision. In particular, C158 seems to be closely related to C673 and C1908 as they share respectively six and two OGs. VC10 was located near the siphovirus nodes (cyan) in the network. Importantly, the taxonomic affiliation of these contigs and their host prediction by different tools were consistent with each other and confirmed that they originated from archaeal viruses, unlike contigs from VC4.

Contigs located on the periphery of the network shared very few OGs with other known viruses (only one or two OGs per contig). Moreover, those shared OGs were often uncharacterized or hypothetical proteins. Several contigs shared annotated OGs belonging to viral families specific to archaea, and they could have originated from previously uncharacterized viruses of methanogens. For example, C775 shared an OG annotated as helicase with eight genomes belonging to the family *Lipothrixviridae*. However, only the presence of signature genes, that is characteristic of particular virus groups, such as those encoding major structural proteins, can provide a reliable taxonomic affiliation of contigs, as it is the case for VC4 and VC10 contigs. For contigs outside of these two VC, such signature genes were identified only for two contigs (C1359 and C1697). Indeed, a structural protein typical of archaeal spindle-shaped viruses (Krupovic et al., 2014) was identified in these two contigs, suggesting that they represent new families of spindle-shaped viruses, as described in more detail in the next section.

Viruses infecting active methanogens during the incubation were identified

Among the 39 viral contigs of interest, several had predicted hosts corresponding to the dominant and active methanogens in the studied microcosms, namely, those from the order Methanobacteriales (*Methanobacterium* and *Methanobrevibacter* genera).

In particular, contig C158 (belonging to VC10) was predicted to infect *Methanobacterium* species, according to its taxonomic affiliation and the signature-based host prediction. *Methanobacterium* was the most abundant archaeal genus in the studied microcosms. C158 has a length of 42,490 bp and contains 49 predicted genes. Consistent with the abundance of the predicted host, C158 was the most abundant of the 39 contigs in metaviromes at days 13 and 17, with RPKM values of 468 and 229, respectively. Furthermore, C158 is a complete dsDNA circular genome, from a head-tailed virus. All functional modules typical for the class *Caudoviricetes* were identified in this contig (Figure 5), such as those required for virion morphogenesis (HK97-like MCP, mCP, tail proteins, terminase subunits, capsid maturation protease) and several

(A)

(B)
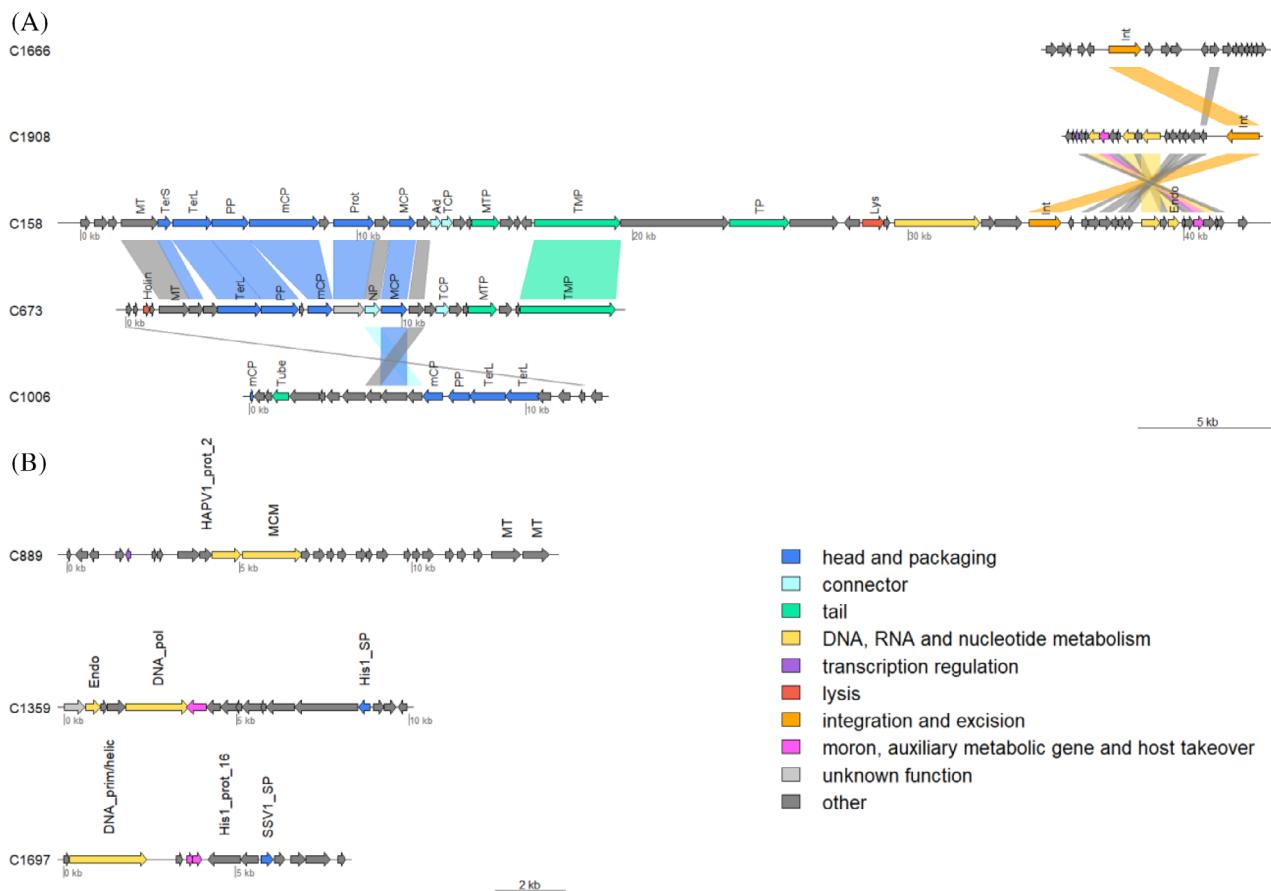
FIGURE 5 Genome map for selected contigs of interest. (A) Contigs from VC10 affiliated to class *Caudoviricetes*, including C158 (proposed new family *Speroviridae*). (B) Viral contigs not related to head-tailed viruses, including predicted spindle-shaped viruses. Genes were annotated with MMseq against PHROGs and with hhpred against four other databases (see Experimental procedures). Genes are represented as arrows. The main functional modules are indicated by colours. Abbreviations of core viral proteins: Ad, adaptor protein; BP, baseplate protein; BH, baseplate hub protein; BS, baseplate spike protein; BW, baseplate wedge protein; DNA_pol, DNA polymerase; DNA_prim, DNA primase; DNA_prim/helic, DNA primase/helicase; Endo, endonuclease; HAPV1_prot2, similar to protein 2 from *Halorubrum* pleomorphic virus 1; His1_prot_16, similar to protein 16 from *Haloarcula hispanica* virus 1; His1_SP, His1-like major capsid protein; Holin, holin; Int, integrase; Lys, endolysin; MCM, MCM helicase; mCP, minor coat protein; MCP, major capsid protein; MT, methyl transferase; MTP, major tail protein; NP, pre-neck appendage protein; NT, tRNA nucleotidyltransferase; PP, portal protein; Prot, capsid maturation protease; Sc, scaffolding; Sh, tail sheath; SSP1_SP, similar to the main structural protein of *Shigella* phage SSP1; SSV1_SP, SSV1-like major capsid protein; TCP, tail completion protein; TerL, terminase large subunit; TerS, terminase small subunit; TMP, tail tape measure protein; TP, tail protein; Tube, tail tube

enzymes necessary for the life cycle, such as an intra-cellular proteinase inhibitor and an integrase, this latter suggesting a temperate lifestyle. According to its proximity to siphoviruses in the network (Figure 4) and VIR-FAM analysis (Lopes et al., 2014), C158 is likely to have a siphovirus-like morphology, that is long non-contractile tail. Considering the lack of very strong similarity with previously characterized viruses, it likely represents a new family of archaeal viruses within the *Caudoviricetes*. We suggest the name *Speroviridae* for this new viral family.

The second-most abundant contig of interest was C1359, one of the two contigs predicted to have a spindle-shaped morphology, with RPKM values at days 13 and 17 of 285 and 183, respectively. Surprisingly, this 9936 bp-long contig (17 detected genes) was predicted to infect *Methanoculleus* sp. (order

Methanomicrobiales), which was present at very low abundances over the incubation time (<0.5% of the archaea based on shotgun metagenomic data). This host prediction was based on spacer alignment (SpacePHARER) against a public spacer database. By contrast, WIsH and spacer alignment with blastn did not yield any high-confidence predicted host. Based on the archaeal community composition (see section 3.1.2), this is likely a false prediction, despite a highly significant p-value (2.03×10^{-7}). Nevertheless, C1359 gene content confirmed the probable archaeal nature of its host, and its high abundance suggests that it infected a dominant methanogen (order Methanobacteriales). Indeed, it encoded a protein showing a significant similarity with the major capsid proteins of two spindle-shaped archaeal viruses, the halophilic *Haloarcula hispanica* virus His1 (Bath & Dyall-Smith, 1998)

and the hyperthermophilic *Sulfolobus shibatae* virus SSV1 (Palm et al., 1991) (probability: 98.84% and 98.36% respectively). Moreover, spindle-shaped viruses are specific to archaea (Krupovic et al., 2018; Pina et al., 2011; Prangishvili et al., 2017; Snyder et al., 2015), excluding the possibility that the virus infects bacteria. This predicted morphotype is consistent with the presence of spindle-shaped VLPs observed by TEM [Figure 2(K–L)] and reported in AD ecosystems (Calusinska et al., 2016). Due to its limited similarity with previously characterized viruses, C1359 likely represents a new viral family, and corresponds to the first spindle-shaped virus reported in association with the order Methanobacteriales.

Viruses infecting minor methanogens in the formate incubation microcosms were identified

As mentioned above, methanogens of the orders Methanomicrobiales and Methanosarcinales were present in the studied microcosms at a low abundance (<5%). Nevertheless, previously uncharacterized viral genomes possibly infecting these methanogens were identified.

Contig C889 has a length of 14,120 bp and contains 26 predicted genes. Its RPKM values at days 13 and 17 were 2.3 and 3.3, respectively. Based on its taxonomic affiliation, this contig seemed to originate from a virus infecting a Methanomicrobiales host. In this contig, a protein similar to replicative minichromosome maintenance helicases was detected. This is not a structural protein but it has been identified in several families of archaeal viruses and archaeal plasmids (Krupovič et al., 2010a), strongly supporting an archaeal host. No structural protein could be identified in C889, suggesting either a partial genome or a new type of virus.

Contig C1697 has a length of 8207 bp, contains 11 predicted genes, and had RPKM values of 4.1 at day 13 and 8.4 at day 17. Based on its taxonomic affiliation and on host prediction with WIsH, this contig originates from a *Methanosarcina* virus. It encodes one protein showing significant profile similarity with the major capsid protein of SSV1 (*Fuselloviridae*) (probability: 94.92%), suggesting a spindle-shaped virion morphology. C1697 did not show pronounced similarity with C1359 (which also had a predicted spindle-shaped morphotype) or with known spindle-shaped viruses, suggesting that it could represent a second new family of spindle-shaped archaeal viruses. Notably, however, similar to spindle-shaped halspivirus His1 (Bath & Dyall-Smith, 1998) and thaspivirus *Nitrosopumilus spindle-shaped virus 1* NSV1 (Kim et al., 2019), C1697 encodes a protein-primed family B DNA polymerase, suggesting that the genome of this virus is linear.

CONCLUDING REMARKS

Predicting the hosts of viruses is a major bottleneck in microbial ecology. In recent years, besides the classic culture-dependent approaches, several promising methods have been developed to identify the hosts for viruses within complex microbial communities, such as PhageFISH (Allers et al., 2013; Barrero-Canosa & Moraru, 2019), Meta3C (Marbouty et al., 2014), viral tagging (Deng et al., 2014) or epicPCR (Sakowski et al., 2021; Spencer et al., 2016). These single cell level methods generated important new knowledge, but they also suffer from limitations, either due to the requirement of a pre-existing knowledge on the viral genome sequence, of the capability to cultivate the host, or due to their complexity. Several purely bioinformatic approaches of host prediction have been also developed (Coclet & Roux, 2021; Edwards et al., 2016). However, the accuracy of *in silico* methods is limited, in most cases, by the lack of microbial genomes closely related to that of the true host. Viral host prediction, therefore, remains challenging for metagenomics studies.

The original experimental approach developed in the present study is not strictly speaking a method for the identification of viral hosts but it still presents some advantages in this perspective. Relying on SIP enabled us to discriminate between the active microorganisms involved in formate metabolism and the other ones. Although this substrate was not consumed exclusively by methanogens, the use of formate strongly enriched the community in methanogens and their associated viruses, resulting in a simplified microbial community to study and, possibly, an improved MAG assembly. Coupling the results from different bioinformatic methods (taxonomic assignation of the contigs, alignment- and signature-based host prediction), we were thus able to detect several previously uncharacterized genomes of viruses infecting methanogens, with high accuracy, and to determine whether they targeted methanogens actively involved in the formate bioconversion. It was especially possible thanks to the establishment of specific host databases, through the shotgun sequencing of cellular metagenomes. Interestingly, one of them likely corresponds to a new archaeal virus family within the *Caudoviricetes* (proposed name *Speroviridae*), and two others to putative new families of spindle-shaped archaeal viruses. These results significantly expand the knowledge on the diversity of viruses of methanogens, since no spindle-shaped virus has been reported until now for the orders Methanobacteriales and Methanosarcinales. It illustrates the complementarity between SIP, metagenomics and specific bioinformatic tools for host-virus analysis in complex microbial communities. Our approach can be viewed as complementary to the available experimental methods for host identification, and could likely be combined with most of them. In particular, PhageFISH or epicPCR would be nicely

complementary as they require at least a partial knowledge of the viral sequence, and they would enable to fully confirm the host identity.

To sum up, we proposed an original coupling between SIP, an experimental method, and metagenomic analyses with complementary bioinformatic tools to identify viruses of methanogenic archaea within anaerobic digestion microcosms. It enabled successful enrichment of the microbial community in methanogenic archaea, and their labelling with ^{13}C. The *in silico* approach we developed led to the identification of several dozens of contigs predicted to originate from viruses infecting methanogenic archaea. Their analyses through gene-sharing network and comparative genomics highlighted the dominance of the *Caudoviricetes* class, with the discovery of a previously uncharacterized siphovirus infecting Methanobacteriales hosts and belonging to a new suggested viral family, *Speroviridae*. It also led to the discovery of two spindle-shaped viruses representing two new putative families, not previously reported for the orders Methanobacteriales and Methanosarcinales. Our results significantly expand the knowledge on the diversity of viruses of methanogens and reinforce the notion of the wide environmental and phylogenetic distribution of spindle-shaped viruses in Archaea (Krupovic et al., 2020). Our original experimental approach enables to identify viruses infecting key functional groups contributing to biogeochemical fluxes in communities of uncultured microbes. It can be invaluable for the study of viruses infecting metabolically active microorganisms in virtually any type of complex microbial community.

CONFLICT OF INTEREST
The authors declared to have no conflict of interest.

ORCID
Vuong Quoc Hoang Ngo https://orcid.org/0000-0001-6746-5521
Ariane Bize https://orcid.org/0000-0003-4023-8665

REFERENCES
Ahring, B.K. (2003) Perspectives for anaerobic digestion. In: Ahring, B.K., Angelidaki, I., de Macario, E.C., Gavala, H.N., Hofman-Bang, J., Macario, A.J.L. et al. (Eds.) *Biomethanation I. Advances in biochemical engineering/biotechnology*. Berlin, Heidelberg: Springer, pp. 1–30.

Allers, E., Moraru, C., Duhaime, M.B., Beneze, E., Solonenko, N., Barrero-Canosa, J. et al. (2013) Single-cell and population level viral infection dynamics revealed by phageFISH, a method to visualize intracellular and free viruses. *Environmental Microbiology*, 15, 2306–2318.

Baquero, D.P., Liu, Y., Wang, F., Egelman, E.H., Prangishvili, D. & Krupovic, M. (2020) Chapter four - structure and assembly of archaeal viruses. In: Kielian, M., Mettenleiter, T.C. & Roossinck, M.J. (Eds.) *Advances in virus research. Virus assembly and exit pathways*. Cambridge: MA: Academic Press, pp. 127–164.

Barrero-Canosa, J. & Moraru, C. (2019) PhageFISH for monitoring phage infections at single cell level. *Methods in Molecular Biology*, 1898, 1–26.

Bath, C. & Dyall-Smith, M.L. (1998) His1, an archaeal virus of the Fuselloviridae family that infects Haloarcula hispanica. *Journal of Virology*, 72, 9392–9395.

Biswas, A., Staals, R.H.J., Morales, S.E., Fineran, P.C. & Brown, C.M. (2016) CRISPRDetect: a flexible algorithm to define CRISPR arrays. *BMC Genomics*, 17, 356.

Borrel, G., Colombet, J., Robin, A., Lehours, A.-C., Prangishvili, D. & Sime-Ngando, T. (2012) Unexpected and novel putative viruses in the sediments of a deep-dark permanently anoxic freshwater habitat. *The ISME Journal*, 6, 2119–2127.

Buchfink, B., Xie, C. & Huson, D.H. (2015) Fast and sensitive protein alignment using DIAMOND. *Nature Methods*, 12, 59–60.

Calusinska, M., Marynowska, M., Goux, X., Lentzen, E. & Delfosse, P. (2016) Analysis of dsDNA and RNA viromes in methanogenic digesters reveals novel viral genetic diversity. *Environmental Microbiology*, 18, 1162–1175.

Chapleur, O., Mazeas, L., Godon, J.-J. & Bouchez, T. (2016) Asymmetrical response of anaerobic digestion microbiota to temperature changes. *Applied Microbiology and Biotechnology*, 100, 1445–1457.

Coclet, C. & Roux, S. (2021) Global overview and major challenges of host prediction methods for uncultivated phages. *Current Opinion in Virology*, 49, 117–126.

Couvin, D., Bernheim, A., Toffano-Nioche, C., Touchon, M., Michalik, J., Néron, B. et al. (2018) CRISPRCasFinder, an update of CRISRFinder, includes a portable version, enhanced performance and integrates search for Cas proteins. *Nucleic Acids Research*, 46, W246–W251.

Csardi, G. & Nepusz, T. (2005) The Igraph software package for complex network research. *InterJournal Complex Systems*, 695, 1–9.

Deng, L., Ignacio-Espinoza, J.C., Gregory, A.C., Poulos, B.T., Weitz, J.S., Hugenholtz, P. et al. (2014) Viral tagging reveals discrete populations in Synechococcus viral genome sequence space. *Nature*, 513, 242–245.

Eason, R. & Campbell, A.M. (1978) 8 - Analytical ultracentrifugation. In: Birnie, G.D. & Rickwood, D. (Eds.) *Centrifugal separations in molecular and cell biology*. London, UK: Butterworth-Heinemann, pp. 251–287.

Edwards, R.A., McNair, K., Faust, K., Raes, J. & Dutilh, B.E. (2016) Computational approaches to predict bacteriophage–host relationships. *FEMS Microbiology Reviews*, 40, 258–272.

Evans, P.N., Boyd, J.A., Leu, A.O., Woodcroft, B.J., Parks, D.H., Hugenholtz, P. et al. (2019) An evolving view of methane metabolism in the archaea. *Nature Reviews Microbiology*, 1, 219–232.

Fokine, A. & Rossmann, M.G. (2014) Molecular architecture of tailed double-stranded DNA phages. *Bacteriophage*, 4, e28281.

Galiez, C., Siebert, M., Enault, F., Vincent, J. & Söding, J. (2017) WIsH: who is the host? Predicting prokaryotic hosts from metagenomic phage contigs. *Bioinformatics*, 33, 3113–3114.

Guo, J., Bolduc, B., Zayed, A.A., Varsani, A., Dominguez-Huerta, G., Delmont, T.O. et al. (2021) VirSorter2: a multi-classifier, expert-guided approach to detect diverse DNA and RNA viruses. *Microbiome*, 9, 37.

He, J., Wang, X., Yin, X., Li, Q., Li, X., Zhang, Y. et al. (2018) Insights into biomethane production and microbial community succession during semi-continuous anaerobic digestion of waste cooking oil under different organic loading rates. *AMB Express*, 8, 92.

Hyatt, D., Chen, G.-L., LoCascio, P.F., Land, M.L., Larimer, F.W. & Hauser, L.J. (2010) Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics*, 11, 119.

Iranzo, J., Koonin, E.V., Prangishvili, D. & Krupovic, M. (2016a) Bipartite network analysis of the archaeal virosphere: evolutionary connections between viruses and capsidless mobile elements. *Journal of Virology*, 90, 11043–11055.

Iranzo, J., Krupovic, M. & Koonin, E.V. (2016b) The double-stranded DNA virosphere as a modular hierarchical network of gene sharing. *mBio*, 7, e00978-16.

Kanehisa, M., Sato, Y. & Morishima, K. (2016) BlastKOALA and GhostKOALA: KEGG tools for functional characterization of genome and metagenome sequences. *Journal of Molecular Biology*, 428, 726–731.

Kang, D.D., Li, F., Kirton, E., Thomas, A., Egan, R., An, H. et al. (2019) MetaBAT 2: an adaptive binning algorithm for robust and efficient genome reconstruction from metagenome assemblies. *PeerJ*, 7, e7359.

Kieft, K., Zhou, Z. & Anantharaman, K. (2020) VIBRANT: automated recovery, annotation and curation of microbial viruses, and evaluation of viral community function from genomic sequences. *Microbiome*, 8, 90.

Kim, J.-G., Kim, S.-J., Cvirkaite-Krupovic, V., Yu, W.-J., Gwak, J.-H., López-Pérez, M. et al. (2019) Spindle-shaped viruses infect marine ammonia-oxidizing thaumarchaea. *Proceedings of the National Academy of Sciences of the United States of America*, 116, 15645–15650.

Köster, J. & Rahmann, S. (2012) Snakemake—a scalable bioinformatics workflow engine. *Bioinformatics*, 28, 2520–2522.

Krupovič, M. & Bamford, D.H. (2008) Archaeal proviruses TKV4 and MVV extend the PRD1-adenovirus lineage to the phylum Euryarchaeota. *Virology*, 375, 292–300.

Krupovič, M., Forterre, P. & Bamford, D.H. (2010) Comparative analysis of the mosaic genomes of tailed archaeal viruses and proviruses suggests common themes for virion architecture and assembly with tailed viruses of bacteria. *Journal of Molecular Biology*, 397, 144–160.

Krupovič, M., Gribaldo, S., Bamford, D.H. & Forterre, P. (2010a) The evolutionary history of archaeal MCM helicases: a case study of vertical evolution combined with hitchhiking of mobile genetic elements. *Molecular Biology and Evolution*, 27, 2716–2732.

Krupovic, M., Quemin, E.R.J., Bamford, D.H., Forterre, P. & Prangishvili, D. (2014) Unification of the globally distributed spindle-shaped viruses of the archaea. *Journal of Virology*, 88, 2354–2358.

Krupovic, M., Cvirkaite-Krupovic, V., Iranzo, J., Prangishvili, D. & Koonin, E.V. (2018) Viruses of archaea: structural, functional, environmental and evolutionary genomics. *Virus Research*, 244, 181–193.

Krupovic, M., Dolja, V.V. & Koonin, E.V. (2020) The LUCA and its complex virome. *Nature Reviews Microbiology*, 18, 661–670.

Langmead, B. & Salzberg, S.L. (2012) Fast gapped-read alignment with Bowtie 2. *Nature Methods*, 9, 357–359.

Lei, Y., Wei, L., Liu, T., Xiao, Y., Dang, Y., Sun, D. et al. (2018) Magnetite enhances anaerobic digestion and methanogenesis of fresh leachate from a municipal solid waste incineration plant. *Chemical Engineering Journal*, 348, 992–999.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N. et al. (2009) The sequence alignment/map format and SAMtools. *Bioinformatics*, 25, 2078–2079.

Liang, B., Wang, L.-Y., Mbadinga, S.M., Liu, J.-F., Yang, S.-Z., Gu, J.-D. et al. (2015) Anaerolineaceae and Methanosaeta turned to be the dominant microorganisms in alkanes-dependent methanogenic culture after long-term of incubation. *AMB Express*, 5, 37.

Lin, Q., De Vrieze, J., Li, J. & Li, X. (2016) Temperature affects microbial abundance, activity and interactions in anaerobic digestion. *Bioresource Technology*, 209, 228–236.

Liu, Y., Demina, T.A., Roux, S., Aiewsakun, P., Kazlauskas, D., Simmonds, P. et al. (2021) Diversity, taxonomy, and evolution of archaeal viruses of the class Caudoviricetes. *PLoS Biology*, 19, e3001442.

Lopes, A., Tavares, P., Petit, M.-A., Guérois, R. & Zinn-Justin, S. (2014) Automated classification of tailed bacteriophages according to their neck organization. *BMC Genomics*, 15, 1027.

Luo, Y., Pfister, P., Leisinger, T. & Wasserfallen, A. (2001) The genome of archaeal prophage ΨM100 encodes the lytic enzyme responsible for autolysis of Methanothermobacter wolfeii. *Journal of Bacteriology*, 183, 5788–5792.

Lyu, Z., Shao, N., Akinyemi, T. & Whitman, W.B. (2018) Methanogenesis. *Current Biology*, 28, R727–R732.

Madigou, C., Lê Cao, K.-A., Bureau, C., Mazéas, L., Déjean, S. & Chapleur, O. (2019) Ecological consequences of abrupt temperature changes in anaerobic digesters. *Chemical Engineering Journal*, 361, 266–277.

Makarova, K.S., Haft, D.H., Barrangou, R., Brouns, S.J.J., Charpentier, E., Horvath, P. et al. (2011) Evolution and classification of the CRISPR-Cas systems. *Nature Reviews Microbiology*, 9, 467–477.

Makarova, K.S., Wolf, Y.I., Iranzo, J., Shmakov, S.A., Alkhnbashi, O.S., Brouns, S.J.J. et al. (2020) Evolutionary classification of CRISPR-Cas systems: a burst of class 2 and derived variants. *Nature Reviews Microbiology*, 18, 67–83.

Marbouty, M., Cournac, A., Flot, J.-F., Marie-Nelly, H., Mozziconacci, J. & Koszul, R. (2014) Metagenomic chromosome conformation capture (meta3C) unveils the diversity of chromosome organization in microorganisms. *eLife*, 3, e03318.

von Meijenfeldt, F.A.B., Arkhipova, K., Cambuy, D.D., Coutinho, F.H. & Dutilh, B.E. (2019) Robust taxonomic classification of uncharted microbial sequences and bins with CAT and BAT. *Genome Biology*, 20, 217.

Meile, L., Jenal, U., Studer, D., Jordan, M. & Leisinger, T. (1989) Characterization of ψM1, a virulent phage of Methanobacterium thermoautotrophicum Marburg. *Archives of Microbiology*, 152, 105–110.

Menzel, P., Ng, K.L. & Krogh, A. (2016) Fast and sensitive taxonomic classification for metagenomics with Kaiju. *Nature Communications*, 7, 11257.

Molnár, J., Magyar, B., Schneider, G., Laczi, K., Valappil, S.K., Kovács, Á.L. et al. (2020) Identification of a novel archaea virus, detected in hydrocarbon polluted Hungarian and Canadian samples. *PLoS One*, 15, e0231864.

Nasko, D.J., Ferrell, B.D., Moore, R.M., Bhavsar, J.D., Polson, S.W. & Wommack, K.E. (2019) CRISPR spacers indicate preferential matching of specific virioplankton genes. *mBio*, 10, e02651-18.

Nayfach, S., Camargo, A.P., Eloe-Fadrosh, E., Roux, S. & Kyrpides, N. (2021) CheckV: assessing the quality of metagenome-assembled viral genomes. *Nature Biotechnology*, 39, 578–585.

Neufeld, J.D., Vohra, J., Dumont, M.G., Lueders, T., Manefield, M., Friedrich, M.W. et al. (2007) DNA stable-isotope probing. *Nature Protocols*, 2, 860–866.

Nölling, J., Groffen, A. & de Vos, W.M. (1993) φ F1 and φF3, two novel virulent, archaeal phages infecting different thermophilic strains of the genus Methanobacterium. *Microbiology*, 139, 2511–2516.

Nurk, S., Meleshko, D., Korobeynikov, A., & Pevzner, P. (2016) metaSPAdes: a new versatile de novo metagenomics assembler. arXiv:160403071 [q-bio].

Olo Ndela, E., Enault, F. & Toussaint, A. (2021) Transposable prophages in Leptospira: An ancient, now diverse, group predominant in causative agents of Weil's disease. *International Journal of Molecular Sciences*, 22, 13434.

Palm, P., Schleper, C., Grampp, B., Yeats, S., McWilliam, P., Reiter, W.-D. et al. (1991) Complete nucleotide sequence of the virus SSV1 of the archaebacterium Sulfolobus shibatae. *Virology*, 185, 242–250.

Parks, D.H., Imelfort, M., Skennerton, C.T., Hugenholtz, P. & Tyson, G.W. (2015) CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome Research*, 25, 1043–1055.

Parks, D.H., Rinke, C., Chuvochina, M., Chaumeil, P.-A., Woodcroft, B.J., Evans, P.N. et al. (2017) Recovery of nearly 8,000 metagenome-assembled genomes substantially expands the tree of life. *Nature Microbiology*, 2, 1533–1542.

Pfister, P., Wasserfallen, A., Stettler, R. & Leisinger, T. (1998) Molecular analysis of Methanobacterium phage ΨM2. *Molecular Microbiology*, 30, 233–244.

Philosof, A., Yutin, N., Flores-Uribe, J., Sharon, I., Koonin, E.V. & Béjà, O. (2017) Novel abundant oceanic viruses of uncultured marine group II Euryarchaeota. *Current Biology*, 27, 1362–1368.

Pickard, D.J.J. (2009) Preparation of bacteriophage lysates and pure DNA. *Methods in Molecular Biology*, 502, 3–9.

Pina, M., Bize, A., Forterre, P. & Prangishvili, D. (2011) The archeoviruses. *FEMS Microbiology Reviews*, 35, 1035–1054.

Poirier, S., Desmond-Le Quéméner, E., Madigou, C., Bouchez, T. & Chapleur, O. (2016) Anaerobic digestion of biowaste under extreme ammonia concentration: identification of key microbial phylotypes. *Bioresource Technology*, 207, 92–101.

Prangishvili, D., Bamford, D.H., Forterre, P., Iranzo, J., Koonin, E.V. & Krupovic, M. (2017) The enigmatic archaeal virosphere. *Nature Reviews Microbiology*, 15, 724–739.

Puig-Castellví, F., Midoux, C., Guenne, A., Conteau, D., Franchi, O., Bureau, C. et al. (2022) Metataxonomics, metagenomics and metabolomics analysis of the influence of temperature modification in full-scale anaerobic digesters. *Bioresource Technology*, 346, 126612.

Radajewski, S., Ineson, P., Parekh, N.R. & Murrell, J.C. (2000) Stable-isotope probing as a tool in microbial ecology. *Nature*, 403, 646–649.

Sakowski, E.G., Arora-Williams, K., Tian, F., Zayed, A.A., Zablocki, O., Sullivan, M.B. et al. (2021) Interaction dynamics and virus–host range for estuarine actinophages captured by epicPCR. *Nature Microbiology*, 6, 630–642.

Si, B., Liu, Z., Zhang, Y., Li, J., Shen, R., Zhu, Z. et al. (2016) Towards biohythane production from biomass: influence of operational stage on anaerobic fermentation and microbial community. *International Journal of Hydrogen Energy*, 41, 4429–4438.

Snyder, J.C., Bolduc, B. & Young, M.J. (2015) 40 Years of archaeal virology: expanding viral diversity. *Virology*, 479–480, 369–378.

Spencer, S.J., Tamminen, M.V., Preheim, S.P., Guo, M.T., Briggs, A.W., Brito, I.L. et al. (2016) Massively parallel sequencing of single cells by epicPCR links functional genes with phylogenetic markers. *The ISME Journal*, 10, 427–436.

Steinegger, M., Meier, M., Mirdita, M., Vöhringer, H., Haunsberger, S.J. & Söding, J. (2019) HH-suite3 for fast remote homology detection and deep protein annotation. *BMC Bioinformatics*, 20, 1–15.

Sun, H., Yang, Z., Shi, G., Arhin, S.G., Papadakis, V.G., Goula, M.A. et al. (2021) Methane production from acetate, formate and H_2/CO_2 under high ammonia level: modified ADM1 simulation and microbial characterization. *Science of the Total Environment*, 783, 147581.

Terzian, P., Olo Ndela, E., Galiez, C., Lossouarn, J., Pérez Bucio, R.E., Mom, R. et al. (2021) PHROG: families of prokaryotic virus proteins clustered using remote homology. *NAR Genomics and Bioinformatics*, 3, lqab067.

Thauer, R.K., Kaster, A.-K., Seedorf, H., Buckel, W. & Hedderich, R. (2008) Methanogenic archaea: ecologically relevant differences in energy conservation. *Nature Reviews Microbiology*, 6, 579–591.

Thiroux, S., Dupont, S., Nesbø, C.L., Bienvenu, N., Krupovic, M., L'Haridon, S. et al. (2021) The first head-tailed virus, MFTV1, infecting hyperthermophilic methanogenic deep-sea archaea. *Environmental Microbiology*, 23, 3614–3626.

Wang, Y. & Qian, P.-Y. (2009) Conservative fragments in bacterial 16S rRNA genes and primer design for 16S ribosomal DNA amplicons in metagenomic studies. *PLoS One*, 4, e7401.

Wang, Q., Garrity, G.M., Tiedje, J.M. & Cole, J.R. (2007) Naïve Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. *Applied and Environmental Microbiology*, 73, 5261–5267.

Wang, J., Zhao, Z. & Zhang, Y. (2021) Enhancing anaerobic digestion of kitchen wastes with biochar: link between different properties and critical mechanisms of promoting interspecies electron transfer. *Renewable Energy*, 167, 791–799.

Weidenbach, K., Nickel, L., Neve, H., Alkhnbashi, O.S., Künzel, S., Kupczok, A. et al. (2017) Methanosarcina spherical virus, a novel archaeal lytic virus targeting Methanosarcina strains. *Journal of Virology*, 91, e00955-17.

Weidenbach, K., Wolf, S., Kupczok, A., Kern, T., Fischer, M.A., Reetz, J. et al. (2021) Characterization of Blf4, an archaeal lytic virus targeting a member of the Methanomicrobiales. *Viruses*, 13, 1934.

Whiticar, M.J., Faber, E. & Schoell, M. (1986) Biogenic methane formation in marine and freshwater environments: CO_2 reduction vs. acetate fermentation—isotope evidence. *Geochimica et Cosmochimica Acta*, 50, 693–709.

Wolf, S., Fischer, M.A., Kupczok, A., Reetz, J., Kern, T., Schmitz, R.A. et al. (2019) Characterization of the lytic archaeal virus Drs3 infecting Methanobacterium formicicum. *Archives of Virology*, 164, 667–674.

Wood, A.G., Whitman, W.B. & Konisky, J. (1989) Isolation and characterization of an archaebacterial viruslike particle from Methanococcus voltae A3. *Journal of Bacteriology*, 171, 93–98.

Zhang, R., Mirdita, M., Levy Karin, E., Norroy, C., Galiez, C. & Söding, J. (2021) SpacePHARER: sensitive identification of phages from CRISPR spacers in prokaryotic hosts. *Bioinformatics*, 37, 3364–3366.

Zimmermann, L., Stephens, A., Nam, S.-Z., Rau, D., Kübler, J., Lozajic, M. et al. (2018) A completely reimplemented MPI bioinformatics toolkit with a new HHpred server at its core. *Journal of Molecular Biology*, 430, 2237–2243.

Zolfo, M., Pinto, F., Asnicar, F., Manghi, P., Tett, A., Bushman, F.D. et al. (2019) Detecting contamination in viromes using ViromeQC. *Nature Biotechnology*, 37, 1408–1412.

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

Titre : Déchiffrer le fonctionnement des communautés microbiennes de la méthanisation, et leurs virus, en combinant écologie moléculaire et métagénomique

Mots clés : écologie microbienne, biotechnologies environnementales, métaviromes, marquage isotopique, cellulose, archées

Résumé : La méthanisation est un procédé de valorisation des déchets organiques, qui permet d'obtenir une énergie renouvelable sous la forme de biogaz. Cette bioconversion anaérobie des déchets est catalysée par des communautés microbiennes complexes, qui sont sensibles à de nombreux facteurs abiotiques. Déchiffrer leur fonctionnement est un élément clé pour favoriser le développement de procédés de méthanisation stables et performants. En effet, cela permet de comprendre et d'exploiter les liens qui existent entre les conditions opératoires appliquées, la structuration et l'activité des communautés microbiennes, et enfin les performances globales du procédé.

Dans ce mémoire, je présente une synthèse de mes recherches en écologie microbienne. Une première phase de mes travaux a été dédiée à la caractérisation des commuanutés microbiennes impliquées dans la méthanisation de déchets cellulosiques, par des approches métagénomiques et métaprotéomiques.

Ces travaux ont notamment mis en évidence l'adaptation très fine de la composition et de l'activité des communautés microbiennes, aux propriétés des substrats cellulosiques. Par la suite, je me suis consacrée à l'étude des virus de microorganismes présents dans les méthaniseurs. Il est bien établi que les virus ont une influence sur les cycles biogéochimiques majeurs, mais cela a été encore très peu étudié dans le cas de la méthanisation. Nous avons combiné des approches isotopiques et métagénomiques pour étudier la diversité de virus d'archées méthanogènes hydrogénotrophes, ce qui a abouti à la découverte de nouveaux virus fusiformes. Enfin, souhaitant favoriser la valorisation et la réutilisation des données, je coordonne depuis plusieurs années le développement d'un système d'information, DeepOmics, pour les données métagénomiques issues de procédés de biotechnologies environnementales.

Title: Deciphering the functioning of anaerobic digestion microbial communities, and their viruses, by combining molecular ecology and metagenomics

Keywords: microbial ecology, environmental biotechnologies, metaviromes, stable isotope probing, cellulose, archaea

Abstract: Anaerobic digestion is a process for the valorization of organic waste, which produces biogas, a renewable energy. This anaerobic bioconversion of waste is catalyzed by complex microbial communities, which are sensitive to many abiotic factors. Deciphering their functioning is a key element to promote the development of stable and efficient anaerobic digestion processes. Indeed, it allows to understand and exploit the links between the applied operating conditions, the structuration and activity of the microbial communities, and finally the global performances of the process.

I present a synthesis of my research in microbial ecology. A first phase of my work was dedicated to the characterization of microbial communities involved in the anaerobic digestion of cellulosic waste, using metagenomics and metaproteomics.

This work has highlighted the very fine adaptation of the composition and activity of microbial communities to the properties of cellulosic substrates. Afterwards, I dedicated myself to the study of viruses of microorganisms in anaerobic digesters. It is well established that viruses have an influence on major biogeochemical cycles, but this has been little studied in the case of anaerobic digestion. We combined isotopic and metagenomic approaches to study the diversity of viruses from hydrogenotrophic methanogenic archaea, which led to the discovery of new spindle-shaped viruses. Finally, wishing to promote the valorization and reuse of data, I have been coordinating for several years the development of an information system, DeepOmics, for metagenomic data from environmental biotechnology processes.