



UNIVERSITÀ
DEGLI STUDI
DI PADOVA

Head Office: Università degli Studi di Padova

Department of Biology

Ph.D. COURSE IN: Biosciences

CURRICULUM: Genetics, Genomics and Bioinformatics

SERIES: 35th

The metagenomics revolution and its impact on virology: the case of anaerobic environments

Coordinator: Prof. Ildikó Szabó

Supervisor: Prof. Stefano Campanaro

Ph.D. student: Alessandro Rossi

Abstract

In the last twenty years, the advances in sequencing capabilities and data analysis have allowed microbiologists to transcend cultivation based techniques, giving rise to the field of metagenomics. Metagenomics nowadays includes a diverse array of techniques that make it possible to gather knowledge regarding many different aspects of microbial life, but there are many grey areas which deserve attention and improvement. This thesis takes a tour into some of these areas, with a main focus on viruses, key players in the global ecology and evolution of every lifeform.

Viruses are currently recognized as one of the most important hubs of genetic evolution, as well as active components of biogeochemical cycles. They also pose numerous challenges to scholars, due to their small genomes and their extreme genetic variability, making the knowledge about viruses severely lag behind compared to cellular organisms. Reconstructing their evolution, for instance, presents singular difficulties, as their genomes frequently follow a mosaic evolution model, in which different genes follow different evolutionary trajectories, and may get integrated in the genome or lost with ease. The first work hereby presented deals with this issue in the regards of a peculiar taxon of bacteriophages, the order *Crassvirales*, revealing the genes that are most representative of the evolution of the taxon as a whole.

An environment in which knowledge about viruses is particularly lacking is anaerobic digestion. Anaerobic digestion is a metabolic pathway that converts organic material into simple compounds, mainly carbon dioxide and methane. Its use by humans is and will be growing in importance in the near future as the climate change crisis pushes humanity to abandon fossil fuels and shift to a circular, green economy. This metabolic process is carried out by a complex community of microorganisms, which are well known thanks to whole shotgun sequencing and metagenomic techniques. The viral community which inhabits this environment, on the other hand, has received much less attention by scholars, in spite of the potential impact it has on the microbiome. The second work presented in this thesis deals with the characterization of the viral and prokaryotic community of anaerobic digestion under different conditions, and how the relation between temperate viruses and their hosts changes when a much wider database is taken into account. Improving the production of methane via anaerobic digestion can be attained by manipulating the microbial community, and viruses are a promising tool to accomplish this.

Another aspect of metagenomics that pushes researchers is the annotation of genes. The assignment of functionalities to novel gene or protein sequences is far from being trivial, and many new sequences remain uncharacterized. This problem is conceptually related to the lack of knowledge about viral sequences: as annotation methods mainly rely on databases of previously characterized sequences, genes of unknown function remain uncharacterized, perpetuating the outcome. At genome level, metabolic pathways may remain incomplete either because of either the missed gene annotation or the incompleteness of the reconstructed genome. This thesis includes the publication of KEMET, a software which assesses the completeness of metabolic pathways in genomes. Besides providing an understanding of the metabolism of the chosen genomes, KEMET allows for the improvement of the annotation of the genome, as the user is allowed to identify missing genes by means of a Hidden Markov Model search.

Abstract	1
Acronyms	2
Introduction	3
The era of metagenomics	3
Standard metagenomic analysis	3
Functional Annotation	5
Viruses	5
Metaviromics: novelties and challenges	6
Virus-oriented metagenomic techniques	8
Anaerobic Digestion	8
Viruses in anaerobic digestion	9
Anaerobic digestion database	10
Materials and Methods	10
Functional Annotation	11
Publications	13
References	14

Acronyms

AD - Anaerobic Digestion

BC - Baltimore Class

CRISPR - Clustered Regularly Interspaced Short Palindromic Repeats

dsDNA - double-stranded DNA

dsRNA - double-stranded RNA

HGT - Horizontal Gene Transfer

HMM - Hidden Markov Model

ICTV - International Committee on Taxonomy of Viruses

KEGG - Kyoto Encyclopedia of Genes and Genomes

MAG - Metagenome-Assembled Genome

MGE - Mobile Genetic Element

NGS - Next Generation Sequencing

OTU - Operational Taxonomic Unit

VFA - Volatile Fatty Acids

vOTU - viral Operational Taxonomic Unit

Introduction

The era of metagenomics

The advances of computing systems that began last century have irreversibly changed the ways of science. The development of tools with more and more computational power have allowed scientists to gather, manipulate, store and visualize more data than ever before.

This applies to microbiology as well. DNA sequencing became commonplace in the eighties, with Sanger sequencing, but in the early 2000s the development of Next Generation Sequencing (NGS) made it possible to produce orders of magnitude more data than the previous techniques. The computing power needed to manipulate the short reads also became available at that time. These advancements made it much easier to gather data from uncultured media, paving the way for the study of environmental sequences.

When the term “metagenomics” first appeared in 1998 [1], the only technique used to read environmental DNA was Sanger sequencing, but with the invention and more and more widespread usage of NGS sequencing the number of sequences produced grew exponentially. The International Nucleotide Sequence Database Collaboration (INSDC), which unifies the three sequence databases GenBank, ENA and DDBJ, doubles its content approximately every 18 months.

Traditionally, the description of new microbial species relied on the isolation of microorganisms, with classical microbiological techniques such as *in vitro* cultures. Since most bacterial or archaeal species are either extremely difficult or downright impossible to cultivate, the isolation approach excludes many taxa from description via traditional means. The introduction of NGS techniques thus brought to the discovery of many taxonomic groups among prokaryotes, such as DPANN and Asgard archaea and the CPR bacterial phyla [2–4].

A newer revolution in sequencing is ongoing: long read sequencing technologies have entered the scene in the last decade. These technologies allow the sequencing of reads thousands of base pairs long, reducing the uncertainty that originates from the fragmentation of DNA into short reads and the subsequent assembly [5]. They make it possible, in fact, to obtain complete sequences of entities with short genomes, such as viruses and plasmids [6].

Long read sequencing technologies, despite getting cheaper and more widespread every year, are still far from being a routine tool of microbiologists, and the vast majority of metagenomic analyses are carried out with short read sequencing.

Standard metagenomic analysis

There are two main types of metagenomic analysis, both carried out with short and long read sequencing: amplicon and shotgun. Amplicon sequencing is the amplification of a single DNA region via PCR and the usage of specific adapters for the generation of sequencing libraries; when choosing a region which is present across a broad range of species, it allows to retrieve a snapshot of the whole community. The gene coding for the 16S ribosomal RNA has proven to be an excellent candidate for this purpose. It is found in every prokaryotic organism, both bacteria and archaea, and it contains

both conserved and hypervariable regions. The alternation of conserved and variable regions make it possible to design universal primers, which sit on conserved regions, that amplify hypervariable regions, which provide enough variation to distinguish among low-level taxa. For these reasons, 16S ribosomal DNA (commonly abbreviated as 16S) sequencing is routinely used to characterize the species composition of environmental samples in an easy and relatively cheap fashion. Amplicon sequencing, though, is not a definitive tool, and leaves out many details. As it targets only one genomic region, an amplicon sequencing does not further the knowledge of the intra-genomic variety, and forces the researchers who employ it to rely on literature or previous data regarding, for instance, the functional composition of the community. Furthermore, the taxonomic resolution of 16S sequencing is limited: the lowest taxonomic rank that can be assessed is either species or genus, according to the region that has been chosen for amplification [7]. In this contest, the amplification of the whole 16S gene yields the most precise results [7], but it requires either low-throughput or long reads sequencing. The insight at strain-level resolution is increasingly revealed as crucial, since different strains of the same species may feature a great variety in accessory genomes, resulting in very different functional properties. Plus, 16S sequencing does not allow the detection of extra-genomic material, such as viral particles and plasmids. Integrated viral genomes too are invisible to amplicon sequencing, leaving the viral community neglected once more.

Shotgun sequencing is carried out with random primers and amplifies sequences from any source, indiscriminately. The presence in the dataset of reads coming from any part of any genome, though, opens the way to several problems, which relate with how to manage the data so far gathered. Short reads are incredibly numerous, in the order of magnitude of $10^6 - 10^7$, but only 150-200 bp long, while a single prokaryotic genome easily exceeds one million base pairs. This problem is compounded by the presence, in a single sample, of several species, whose reads are, on a first glance, indistinguishable from each other. The first step to reconstruct the original genomes is the assembly of single reads into longer sequences, called “contigs” or “scaffolds”. There are two main approaches to the assembly of reads, both of which formalize the process of assembly as a graph problem. The first assembly algorithm to be conceived was the Overlap Layout Consensus, in which reads are aligned pairwise; the alignments are then used to build a graph that is traversed in order to find a Hamiltonian path, *i.e.*, a path where the nodes are visited only once [8]. A final step, typically a multiple sequence alignment, is used to close the assembly [8]. A more recent approach, less costly in terms of memory and speed, is the De Bruijn graphs assembly, in which reads are fragmented into sequences of a given length k , named k -mers. k -mers are then interpreted as edges that link $k-1$ mers, generating the graph. The graph is then traversed in order to find the longest non-redundant path, similarly to the previous technique [8,9]. Depending on the assembling program, paths may be Hamiltonian or Eulerian, *i.e.*, paths in which edges, and not nodes, are visited only once [9].

The techniques discussed so now are referred to as *de-novo* assembly, as they do not require pre-existing data to be performed. It is also possible to use template genomes to assemble new genomes, a technique dubbed “reference-based assembly”. Reference-based assembly is a rarely used approach in metagenomics, as the collection of complete genomes does not reflect the actual variety of species present in environments and would introduce even more severe biases in the outcome.

The assembly step produces far from complete genomes; only a fraction of the contigs or scaffolds usually gets close to the length of actual genomes. Since original genomes are thus represented by fragments, the solution to obtain nearly-complete genomes is to group the contigs into collections, a process called “binning”. This is accomplished by gathering one or more statistics, the most commonly used of which is read coverage. Contigs that are part of the same genome are assumed to show similar coverage profiles across different samples; these coverage statistics are obtained by mapping reads from several samples on the assembled sequences, and the more samples are used the more accurate the binning process is. Other statistics that can be used are GC content and

tetranucleotide composition [10,11]. Neural network approaches have also been used to develop binning tools [12]. It is possible to estimate the quality of the recovered bins as genomes with tools like CheckM and CheckM2 [13,14], which estimate completeness and contamination of genomic bins. Bins that correspond to are commonly referred to as Metagenome-Assembled Genomes (MAGs). For more accurate results, it is possible to run multiple binning tools and then pick the highest quality ones for each species, a process called dereplication.

MAGs are thus typically used as the starting point for subsequent analyses. These may include taxonomic assignment, gene prediction and annotation.

Functional Annotation

The huge data-retrieval capabilities offered by next-generation sequencing techniques provide researchers with far more data than it can be analyzed with standard wet-lab techniques. Since *in vivo* and *in vitro* techniques are nowadays the bottleneck in microbiological knowledge, the fastest way to gather new data is via *in silico* analyses. Functional annotation, *i.e.*, assigning a more or less specific function to a sequence, is one of the problems to which this approach can be applied. All the knowledge about protein function ultimately originates from wet-lab studies, but for those sequences that have not been analyzed this way - as in the vast majority of metagenomic studies - functional annotation is carried out via homology searches with genes or proteins of known function. Every type of homology is applied to *in silico* functional annotation, from basic sequence alignment to Hidden Markov Model (HMM) searches, to structural homology and neural network-based annotation. Despite the best intent in annotating every predicted ORF, the utmost majority remains uncharacterized. One more reason contributing to the difficulty of *in silico* annotation is that genomes recovered via metagenomics approaches are often fragmented, lacking several genes useful to frame the species's metabolic properties. Furthermore, recognizing a genome as incomplete isn't straightforward, and genomes belonging to poorly-known taxa are the most difficult to characterize as such. Similar difficulties arise when trying to assess the level of contamination of a MAG, *i.e.*, the quantity of extraneous sequences included in the same metagenomic bin. Approaches based on neural networks seem to be improving in this regard, but any other useful approach is welcome. It is also possible to hypothesize the presence of certain genes in a certain organism based on the presence of genes in specific pathways, as is routinely done in Genome Scale Metabolic Modeling.

Viruses

Viruses are a type of Mobile Genetic Element (MGE) characterized by a high selfishness and mobility [15], able to reproduce only by taking advantage of the metabolic system of the host. They are universally reported as being the most abundant biological entities on Earth, and yet their status as living beings is uncertain, and hangs on the definition itself of life, which they stretch. Viruses, in fact, are not made up of cells, do not have an independent metabolism, nor are they capable of reproduction on their own, but for every biochemical activity they have to rely on the enzymatic machinery of their host. An attempt to solve this conundrum is the definition of virocell, in which the host cell in the altered state of infection is considered as a separate organism than the non-infected cell, which reproduces via the production of viral particles [16]. No matter the semantics, viruses are of massive importance for the evolution of life on earth, for a whole range of reasons. Virtually every branch of the tree of life is host to viral parasites, and play a fundamental role in shaping biogeochemical cycles by freeing the matter used by their hosts. They also shape the evolution of their hosts both by being a source of selective pressure and by enacting Horizontal Gene Transfer (HGT) across hosts.

Most viruses subscribe to one of two types of life cycles: virulent or temperate. Virulent viruses immediately hijack the replication machinery of the host cell after the infection, leading to the production of new viral particles and the eventual destruction of the host, in a process called lytic cycle. Temperate viruses have the capability to perform what's known as lysogenic cycle, in which the infection of a host is followed by the integration of the viral genetic material into the host's own genome. The lysogenic cycle allows the virus to hitchhike the reproductive success of the host, and temperate viruses feature different mechanism that lead the virus to enter the lytic cycle under certain conditions where it is more evolutionarily advantageous for the virus to sacrifice their host and reproduce via the production of viral particles instead. These two lifestyles are not the only ones present in viruses; some species, *e.g.* haloarchaeal viruses, keep the host cell alive while continuously producing new viral particles, at a rate that does not impede the life of the host [17]; other do not even produce viral particles, relying only on vertical transmission or cellular anastomosis to spread to other hosts [18,19].

The interaction between a virus and its host must not be mandatorily seen as a one-way parasitic relation, in which the virus is the only recipient. It is well known, especially in temperate species, that the virus also benefits the host, a phenomenon understood as an evolutionary advantage for the virus. Viruses may provide new functionalities for the host in the guise of genes; this is especially recognized in virulent bacterial strains that acquire their virulence via HGT from the bacteriophages that infect them, as attested in *Vibrio cholera*, *Escherichia coli* and other species [20,21]. Another remarkable case is the presence of mechanisms that prevent the same cell from being targeted by additional viruses, either from the same or different viral species. In this context, viral genomes may carry CRISPR-Cas loci targeting viruses, an adaptation that peaks in huge bacteriophages, some of which carry up to 95 spacers that target 32 viral species[22]. These mechanisms secure the host to a single virus while providing advantageous features to the host as well. It's a perfect example of the "guns for hire" model, where defense systems, which are costly for the host, acquire MGE-like characteristics to ensure their own survival and MGEs acquire defense genes in order to provide an advantage for themselves when they infect a host [15].

Metaviromics: novelties and challenges

Metagenomics is also applied to the study of viruses, thus dubbed metaviromics. Just as in the study of prokaryotes, metagenomics has brought invaluable knowledge to the field of virology too. The most famous example is the case of crAssphage, a tailed bacteriophage which lives in the human gut.

Unknown until 2014, crAssphage was recognized as the most abundant virus of the human gut, where it covers up to 90% of the viral reads [23]. It was discovered thanks to the cross-assembly program crAss – hence the name – which builds a depth profile for each contig generated in a co-assembly of different samples, a then novel technique [24]. Further studies recognized a whole clade of similar bacteriophages, now classified as the order *Crassvirales* in the *Caudoviricetes* class [25,26]. More broadly speaking, the vast majority of viral sequences recovered nowadays is metagenomic in origin, so much so that the International Committee on Taxonomy of Viruses (ICTV) has taken the decision to allow for viral taxa to be defined on the basis of metagenomic data alone [27]. Dedicated databases of viral sequences derived from metagenomic analyses are commonplace; the largest database of environmental viral sequences, IMG/VR, has had its third release in 2021 with 2,332,702 viral genomes, grouped into 935,362 viral Operational Taxonomic Units (vOTUs) [28].

A large part of the metagenomic data cannot be taxonomically assigned nor functionally annotated, and for this reason it bears the moniker “dark matter”. The fraction of unclassified viral sequences is much higher than prokaryotic sequences. This is due to some noteworthy challenges posed by the nature itself of viruses. Most viruses have relatively small genomes; whereas the largest viral genomes might reach hundreds of kilobases, plenty of viral families have genomes as short as a few kilobases. A short genome means that, even when correctly assembled, the viral genome in question could be discarded due to a threshold on the length of contigs. Lowering the threshold would increase the amount of noise present in the dataset, and incomplete genomes belonging to these taxa would be even more difficult to spot, as they would be even smaller and more difficult to detect than complete genomes. This is compounded by the bias of viral sequence databases towards tailed bacteriophages, which have larger genomes; such a bias makes it more difficult to detect marker genes for the completeness of virus belonging to underrepresented taxa.

Viruses have short generation times and large effective population size, and are thus characterized by very fast evolutionary rates in which selective pressure, not genetic drift, is the main driver of genomic evolution. Many viral polymerases also have no proofreading activity, thus enhancing the rate of mutation accumulation in viruses. Another factor that explains viruses’ evolutionary velocity is their ease of access to recombination, leading to a mosaic evolution of the genome. This is even more pronounced in segmented viruses, whose genome is divided into several nucleotide strands, and multipartite viruses, whose genome is contained in different viral particles. These viruses can additionally undergo a phenomenon named “reassortment”, in which entire strands of genetic material are exchanged inter- or intra- individuals [29–31].

Differently from Bacteria, Archaea, and Eukaryotes, “viruses” is not a monophyletic taxon, but a descriptive category. Their diversity even reaches their genetic material, which has been used since 1971 to divide viruses into six “Baltimore Classes” (BCs), named after their ideator, David Baltimore [32]. Baltimore Classes I and II comprise double strand DNA (dsDNA) and single-strand DNA (ssDNA) viruses respectively; BCs III to V represent double-strand RNA, positive-, and negative-sense RNA viruses (dsRNA, (+)RNA, and (-)RNA); finally, BCs VI and VII are dedicated to RNA and DNA retroviruses (RNA-RT and DNA-RT). Their origin is debated: two models for the origin of the viruses describe them as either primitive organisms or intracellular replicators that gradually developed until they acquired genes, such as capsid genes, that allowed them to produce viral particles. Any trace of monophyly, as improbable as it is, has been erased by evolution, as there is no universally conserved sequence among viruses. The most recent attempts to model the evolution of all viruses focus on the phylogenetic tree of “superviral hallmark genes” (“super-VHGs”), *i.e.*, extremely widespread gene families, whose distributions even cross the borders of different BCs. The phylogenies of super-VHGs are combined with gene-sharing networks, providing a scaffold for the organization of the viral diversity at a global level [33].

Traditional virology depended on culture growth of the host, and as such featured the same problems of the former - as most bacteria and archaea cannot be cultivated, neither can their viruses. Traditional virology thus focused on a few viral species of a few selected hosts, focusing on coliphages such as lambda and T4. This started the enduring bias in favor of tailed bacteriophages and archaeal viruses (class *Caudoviricetes*) which are still today, the most represented in databases. This bias has repercussions in every type of analysis dedicated to prokaryotic viruses, from the annotation of genes to the identification of the viral genomes themselves.

Virus-oriented metagenomic techniques

The peculiarities of viruses require the use of alternative techniques for their reconstruction and characterization. It is possible to recover viral sequences both from regular metagenomic datasets and from samples enriched in viral sequences. In order to enrich environmental samples in viral sequences, the most apt technique is filtration through 0.22 μm pores, typically with a tangential flow filtration system. Despite its effectiveness, this technique introduces a few biases in the final composition of the sample, mainly by excluding the larger viruses, such as the amoeba-infecting giant viruses (*Nucleocytoviricota*) and huge bacteriophages. Furthermore, samples obtained by filtration of viral particles do not take into account temperate viruses integrated in their host's genome. Another technique, far less popular, for the enrichment of viral particles, is iron chloride flocculation [34], which consists in the addition of iron chloride (FeCl_3) to the medium. Viral proteins bind to FeCl_3 , which makes them precipitate and easy to recover. Other common steps for enriching samples in viral particles are centrifugation and freeze-drying.

Whereas amplicon sequencing may be used to characterize specific taxa [35–38], is not fit to characterize an entire viral community, due to the lack of common marker genes, and the most common approach is shotgun sequencing. Reads are assembled with the same software used for samples which are not enriched in viruses, even though specialized assemblers for viral genomes have been published [39–41]. Binning of assembled contigs/scaffolds is typically not considered a necessary step for the reconstruction of viral genomes, but dedicated pieces of software have been developed for this purpose and have brought forward promising results [42–44].

It is desirable, when a viral genome has been recovered, to assess its quality, as it is the case for the genomes of living organisms. The main tool aimed at computing completeness and contamination of viral genomes is the CheckM-inspired CheckV[45]. CheckV, while generally reliable, suffers from the same biases of other tools dedicated to the analysis of viral sequences, as it is biased towards the best known taxa: tailed bacteriophages. Finally, viruses too can be dereplicated into viral Operational Taxonomic Units (vOTUs). Like in the MAG dereplication mentioned above, this process involves the clustering of genomes into groups meant to represent single species; the best quality genomes are then picked as representative of the entire species. The advised threshold is 95% average nucleotide identity over 85% alignment fraction, relative to the shorter sequence [46]. Software like RedRed and vConTACT 2 are dedicated to clustering viral genomes [47,48].

Anaerobic Digestion

Anaerobic Digestion (AD) is a complex metabolic process that leads to the breakdown of organic substrates into simple compounds, in absence of oxygen. It is carried out by a complex community of bacteria and archaea, which has been thoroughly characterized thanks to metagenomic techniques. The functional analysis of the community has brought to the reconstruction of the network of reactions, which is commonly depicted as a funnel, where the more complex compounds enter at the top and which lets the final electron acceptor - methane - exit at the bottom. The reactions are commonly grouped in four steps: hydrolysis, acetogenesis, acidogenesis, and methanogenesis. Each one of these steps is carried out by a specific guild of microorganisms, with a few generalist species being able to play multiple roles at once. In the step of hydrolysis the more complex molecules such as carbohydrates, proteins, and lipids are digested into simpler compounds. The microbial guild that carries out this step is the largest and most diverse. In the step of acidogenesis the single monomers produced via hydrolysis are metabolized into molecules such as Volatile Fatty Acids (VFA). These compounds are then used by acetogenic microbes, which produce acetate, carbon dioxide and molecular hydrogen. Finally, the guild of methanogens, composed exclusively of archaeal species, employs these molecules as substrates, producing methane. Methanogenesis is not a unique pathway, but up to four main types of methanogenesis can be differentiated, according to the substrate. The two main methanogenesis pathways are acetoclastic and hydrogenotrophic methanogenesis, which operate on acetate and CO₂ respectively. A pathway of lesser importance is methylotrophic methanogenesis, which uses substances such as methanol and methylamines as substrates. An additional pathway was only recently described, in which methane is produced from aromatic compounds. Up to now, only a species has been described as having this metabolic capability [49].

Many natural environments harbor microbial communities that carry out AD, whose main prerequisites are an abundance of organic matter and the absence of oxygen. For this reason, AD is common in bogs, swamps, and the gut of herbivore animals, from ruminants to termites. The main research done on AD, though, relates to its usage by humans in industrial-scale settings. AD reactors are mainly used to dispose of organic waste, *e.g.* of urban or agricultural origin, while producing biogas and a solid or liquid component known as digestate. The digestate has many industrial uses, such as fertilizer, bedding for livestock, and even components of bioplastics. The production of biogas, *i.e.*, a gaseous mixture primarily composed of methane and CO₂, is particularly relevant as it allows to reduce the dependency of humanity on fossil fuels and the creation of a circular economy, reducing the production of greenhouse gases. Given the importance of AD in the fight against climate change, understanding and optimizing the process is worthy of attention. The optimization of the AD process is, nowadays, mainly based on the manipulation of physicochemical parameters. Some ways to improve the process are the balancing of temperature and a limited exposure to atmospheric oxygen [50–53]. The addition of conductive materials, such as magnetite or biochar, to reactors is a common way to improve the production of methane, as it improves the electron transfer among microorganisms [54,55]. Finally, a series of techniques can convert residual CO₂ into CH₄, a process called “biogas upgrading” [56]. Biogas upgrading can be achieved through physico-chemical processes, as well as biologically, by injecting H₂ into the reactor and thus boosting methanogenesis [57,58].

Viruses in anaerobic digestion

Whereas the microbial community, as mentioned earlier, is well known today thanks to modern sequencing techniques and metagenomic data analysis approaches, the viral community that accompanies this microbiome is severely understudied. As in every environment, viruses have a potentially great importance on the dynamics of AD. For these reasons, they could be used as tools for manipulating the community, by targeting unwanted microbial species or by using them as vectors to confer new metabolic characteristics to key microorganisms in AD.

Until now, the number of published studies regarding the virome of AD is low [59–61]. There are a few reasons for this: first, AD is mainly studied from an engineering point of view, by engineering research groups; second, the complexity of the medium makes it difficult to apply proper techniques to enrich the samples in viral particles, such as filtration; third, because of the aforementioned problems that are always found when studying viruses. Nevertheless, a few metaviromic studies have been published on AD over the last decade. These unanimously paint the viral community as being composed for the most part by tailed bacteriophages (class *Caudoviricetes*, formerly order *Caudovirales*), accompanied by a small number of viral genomes belonging to less known families, such as *Inoviridae*[59] and *Tectiviridae* [59].

One study did not limit to the analysis of DNA viruses, but studied RNA viruses as well [60], recording an overwhelming presence of plant viruses. These plant viruses have, presumably, no impact on the AD process, as they are associated with the medium. The same study also reported the presence of amoeba-infecting giant viruses (*Mimiviridae*), which are also thought to be associated with the medium, *i.e.*, human feces. Similar results were obtained with a metaproteomics approach: tailed bacteriophages make up the largest part of the virome, and traces of plant- and animal-infecting viruses were retrieved [61].

Anaerobic digestion database

The AD database is a collection of metagenomics AD studies, developed in the Genomics and Bioinformatics unit of the University of Padova. In its latest iteration, it includes samples from 18 studies. The assembly and binning of these studies, followed by a dereplication step, allowed the recovery of 1635 non-redundant MAGs. A project during my PhD career was the expansion of the database with new studies and extending the analysis to the viral community. The results are too preliminary to be gathered in a preprint article; for this reason they are presented here as part of the body of the thesis.

Materials and Methods

I have gathered shotgun sequencing data from 12 more AD studies, downloaded with fastq-dump v2.10.8[62].

Following the procedures of the previous iteration of the database, raw reads were trimmed with Trimmomatic v0.39 and adapters were removed with BBDuk v38.86[63]. Assembly was performed independently for all the samples of each single experiment with MEGAHIT v1.2.9[64], and the statistics of each assembly were calculated with QUAST v5.0.2[65]. Every assembly was binned with MetaBAT2 v2.12.1[66] and MaxBin v2.2.7[67]. In the previous version of the database only MetaBAT2 was used for binning; those assemblies were rebinned with MaxBin.

Completeness and contamination of each MAG was calculated with CheckM v1.1.2[13]

The search for viral sequences was performed by scanning every assembled contig with three software for the detection of viral sequences: CheckV v0.7.0, VIBRANT v1.2.0, and PPR-Meta v1.1 [45,68,69]. A contig was assigned as viral if the prediction was independently assigned by at least two

of these programs. A length threshold of 5 kbp was also applied to viral sequences, in order to reduce the noise.

Results

The binning process allowed the recovery of more than 11,000 MAGs in total, which were reduced to 4568 after clustering those belonging to the same species according to ANI calculation. 2,217 MAGs (48.5%) were of high quality (completeness $\geq 90\%$; contamination $\leq 10\%$), while 2,351 (51.5%) were of medium quality (50% \leq completeness $\geq 90\%$; contamination $\leq 10\%$). Those having lower completeness or higher contamination were discarded.

Taxonomic investigation revealed that the prokaryotic community is, as described in previous works, dominated by members of Firmicutes, which comprise 1750 MAGs, *i.e.*, 38% of the database, followed by *Proteobacteria* (554 MAGs, 12%) and *Bacteroidetes* (466 MAGs, 10%).

Archaea are represented by 198 species, mainly belonging to the Euryarchaeota phylum (151 species, 76.3%). While Archaeal MAGs are dominated by *Euryarchaeota*, *Candidatus* Bathyarchaeota and *Candidatus* Diapherotrites are represented by 13 and 12 members respectively. One MAG taxonomically assigned to the *Candidatus* Lokiarchaeota was also recovered, and could present an acetoclastic or methylotrophic metabolism.

Analysis of the MAGs relative abundance in all the samples under examination allowed us to determine the distribution of microbial species in the database. The calculation of the fraction of samples in which each MAG had abundance greater than 1% showed that some phyla, including *Euryarchaeota*, *Synergistetes* and *Candidatus* Cloacimonetes species, tend to be more widespread, while others, such as *Fibrobacteres*, *Ignavibacteriae*, *Planctomycetes* and many *Candidatus* phyla have a more scattered distribution.

The detection of viral sequences allowed the recovery of 26,513 viral sequences. Around half of these sequences (13,437, 50.7%) belong to the *Caudoviricetes* class of tailed bacteriophages, but a vast variety of other taxa, such as *Inoviridae* and *Rudiviridae*, was also found.

Future perspectives

The AD database will be further developed by delving into the associations among its components. One of the current aims of the project is determining the correlations between physicochemical parameters and the composition of both the microbial and viral communities, by mapping the sample reads onto the genomes and determining the relative abundance. The interactions between viruses and their hosts will be made clearer by identifying the CRISPR spacers in microbial genomes and using them in a strict similarity sequence search against viral genomes. These results, in conjunction with functional prediction, will help in determining the presence of virus-mediated HGT in AD communities. In conclusion, this database hopes to shape a perspective of how viruses impact the AD microbial community and, consequently, the production of biogas.

Overview on the selected manuscripts and conclusions

These studies were selected thanks to their relevance in my academic path, as well as their insights into relevant metagenomics topics: viral metagenomics and functional annotation.

Paper I is the published version of my Master's Degree thesis. It is an investigation on the evolution of single genes in *Crassvirales* viruses (then named crass-like viruses), which includes the reconstruction of the phylogenetic tree of each gene, followed by a pairwise comparison of the phylogenetic trees with the mirrortree method. The results show a mosaic evolution of crAss-like viruses' genes, with a core genome formed by genes coevolving with each other, and accessory genes which do not follow the same evolutionary path as the rest of the genome. The "core" genome is fundamentally formed by capsid, replication and structural genes; more specifically, the capsid genes are revealed as the most widespread across crAss-like genomes, as well as the most diverse in terms of sequence identity. These characteristics consolidate the capsid proteins as the most appropriate to represent the evolution of the clade as a whole - as shown in the ICTV proposal for the establishment of the *Crassvirales* order, in which the taxonomy of the order is based on the concatenated alignment of the major capsid protein, large terminase subunit, and DNA primase. This article was included in the paper collection as new analyses were implemented in my first year of Ph.D., which then provided a foundation for the following studies.

Paper II presents a software that faces the problems related to functional annotation, based on the reconstruction of functional pathways present in a single genome. KEMET uses the functional annotation of genes provided by eggNOG-mapper, KofamKOALA, and KAAS. It attests the completeness of functional pathways formalized as KEGG modules, allowing to hypothesize the functional capabilities of the organism. Furthermore, when a pathway is characterized as incomplete, KEMET allows the user to attempt to recover a homologous sequence in the genome by downloading the reference sequences from the KEGG GENES database and running an HMM search. I contributed to the coding part, especially regarding the function that attests the completeness of a module, as well as polishing the code of the entire program.

Paper III is a study about the viral and prokaryotic community of AD and is, to my knowledge, the first study to examine the variation of these communities in relation to different environmental conditions. A set of AD batches was created starting from the same inoculum, and were then inflicted with different conditions which were deemed both stressful for the prokaryotic species living in it and likely to cause induction in temperate viruses. The results show that the change in abundance of different viral genomes can be neatly divided into groups which feature group-exclusive genes. It's also possible to appreciate a widespread presence of tail-associated sialidases in viruses, hypothesized to be used during the infection. Sixty-four out of 120 prokaryotic genomes were found to be hosting integrated proviruses; the reads from the AD database were mapped onto these virus-host couples, in order to gather insight about their behavior in a much wider range of physical and chemical conditions. It can be attested that temperature is the main driver of the abundance ratio between viruses and their hosts - but, even more strikingly, that simplified medium cultures show an unusually high amount of viruses 10 times or more abundant than their hosts, which we hypothesize is due to the stress that a simplified medium, with little variety of nutrients, brings to the bacteria.

Publications

I **Alessandro Rossi**, Laura Treu, Stefano Toppo, Henrike Zschach, Stefano Campanaro, Bas E. Dutilh (2020). Evolutionary study of the crAssphage virus at gene level. *Viruses*, 12, 1035.

II Matteo Palù, Arianna Basile, Guido Zampieri, Laura Treu, **Alessandro Rossi**, Maria Silvia Morlino, Stefano Campanaro (2022). KEMET – A python tool for KEGG Module evaluation and microbial genome annotation expansion. *Computational and Structural Biotechnology Journal*, 20, 1481–1486.

III **Alessandro Rossi**, Maria Silvia Morlino, Maria Gaspari, Arianna Basile, Panagiotis Kougias, Laura Treu, Stefano Campanaro (2022). Analysis of the anaerobic digestion metagenome under environmental stresses stimulating prophage induction. *Microbiome*, 10, 125.

References

1. Handelsman J, Rondon MR, Brady SF, Clardy J, Goodman RM. Molecular biological access to the chemistry of unknown soil microbes: a new frontier for natural products. *Chem Biol.* 1998;5:R245–9.
2. Dombrowski N, Lee J-H, Williams TA, Offre P, Spang A. Genomic diversity, lifestyles and evolutionary origins of DPANN archaea. *FEMS Microbiol Lett* [Internet]. 2019 [cited 2022 Sep 26];366. Available from: <https://academic.oup.com/femsle/article/doi/10.1093/femsle/fnz008/5281434>
3. Liu Y, Makarova KS, Huang W-C, Wolf YI, Nikolskaya AN, Zhang X, et al. Expanded diversity of Asgard archaea and their relationships with eukaryotes. *Nature.* 2021;593:553–7.
4. Bokhari RH, Amirjan N, Jeong H, Kim KM, Caetano-Anollés G, Nasir A. Bacterial Origin and Reductive Evolution of the CPR Group. Baptiste E, editor. *Genome Biol Evol.* 2020;12:103–21.
5. Smith SE, Huang W, Tiamani K, Unterer M, Khan Mirzaei M, Deng L. Emerging technologies in the study of the virome. *Curr Opin Virol.* 2022;54:101231.
6. Beaulaurier J, Luo E, Eppley JM, Uyl PD, Dai X, Burger A, et al. Assembly-free single-molecule sequencing recovers complete virus genomes from natural microbial communities. *Genome Res.* 2020;30:437–46.
7. Johnson JS, Spakowicz DJ, Hong B-Y, Petersen LM, Demkowicz P, Chen L, et al. Evaluation of 16S rRNA gene sequencing for species and strain-level microbiome analysis. *Nat Commun.* 2019;10:5029.
8. Dida F, Yi G. Empirical evaluation of methods for *de novo* genome assembly. *PeerJ Comput Sci.* 2021;7:e636.
9. Liao X, Li M, Zou Y, Wu F-X, Yi-Pan, Wang J. Current challenges and solutions of *de novo* assembly. *Quant Biol.* 2019;7:90–109.
10. Herath D, Tang S-L, Tandon K, Ackland D, Halgamuge SK. CoMet: a workflow using contig coverage and composition for binning a metagenomic sample with high precision. *BMC Bioinformatics.* 2017;18:571.
11. Graham ED, Heidelberg JF, Tully BJ. BinSanity: unsupervised clustering of environmental microbial assemblies using coverage and affinity propagation. *PeerJ.* 2017;5:e3035.
12. Mao G, Wu Y, Zhang Y, Wang X, Zhu Y, Liu B, et al. DRBin: metagenomic binning based on deep representation learning. *J Genet Genomics.* 2022;49:681–4.
13. Parks DH, Imelfort M, Skennerton CT, Hugenholtz P, Tyson GW. CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome Res.* 2015;25:1043–55.
14. Chklovski A, Parks DH, Woodcroft BJ, Tyson GW. CheckM2: a rapid, scalable and accurate tool for assessing microbial genome quality using machine learning [Internet]. *Bioinformatics*; 2022 Jul. Available from: <http://biorxiv.org/lookup/doi/10.1101/2022.07.11.499243>
15. Koonin EV, Makarova KS, Wolf YI, Krupovic M. Evolutionary entanglement of mobile genetic elements and host defence systems: guns for hire. *Nat Rev Genet.* 2020;21:119–31.
16. Forterre P. To be or not to be alive: How recent discoveries challenge the traditional definitions of viruses and life. *Stud Hist Philos Sci Part C Stud Hist Philos Biol Biomed Sci.* 2016;59:100–8.
17. Luk A, Williams T, Erdmann S, Papke R, Cavicchioli R. Viruses of Haloarchaea. *Life.* 2014;4:681–715.
18. Ghabrial SA, Castón JR, Jiang D, Nibert ML, Suzuki N. 50-plus years of fungal viruses. *Virology.* 2015;479–480:356–68.
19. Nibert ML, Tang J, Xie J, Collier AM, Ghabrial SA, Baker TS, et al. 3D Structures of Fungal Partitiviruses. *Adv Virus Res* [Internet]. Elsevier; 2013 [cited 2022 Sep 28]. p. 59–85. Available from: <https://linkinghub.elsevier.com/retrieve/pii/B9780123943156000039>
20. Schroven K, Aertsen A, Lavigne R. Bacteriophages as drivers of bacterial virulence and

- their potential for biotechnological exploitation. *FEMS Microbiol Rev.* 2021;45:fuaa041.
21. Wagner PL, Waldor MK. Bacteriophage Control of Bacterial Virulence. *Infect Immun.* 2002;70:3985–93.
 22. Al-Shayeb B, Sachdeva R, Chen L-X, Ward F, Munk P, Devoto A, et al. Clades of huge phages from across Earth's ecosystems. *Nature.* 2020;578:425–31.
 23. Dutilh BE, Cassman N, McNair K, Sanchez SE, Silva GGZ, Boling L, et al. A highly abundant bacteriophage discovered in the unknown sequences of human faecal metagenomes. *Nat Commun.* 2014;5:4498.
 24. Dutilh BE, Schmieder R, Nulton J, Felts B, Salamon P, Edwards RA, et al. Reference-independent comparative metagenomics using cross-assembly: crAss. *Bioinformatics.* 2012;28:3225–31.
 25. Yutin N, Makarova KS, Gussow AB, Krupovic M, Segall A, Edwards RA, et al. Discovery of an expansive bacteriophage family that includes the most abundant viruses from the human gut. *Nat Microbiol.* 2018;3:38–46.
 26. Current ICTV Taxonomy Release | ICTV [Internet]. [cited 2022 Jul 21]. Available from: <https://ictv.global/taxonomy>
 27. Simmonds P, Adams MJ, Benkő M, Breitbart M, Brister JR, Carstens EB, et al. Virus taxonomy in the age of metagenomics. *Nat Rev Microbiol.* 2017;15:161–8.
 28. Roux S, Páez-Espino D, Chen I-MA, Palaniappan K, Ratner A, Chu K, et al. IMG/VR v3: an integrated ecological and evolutionary framework for interrogating genomes of uncultivated viruses. *Nucleic Acids Res.* 2021;49:D764–75.
 29. Chare ER, Holmes EC. A phylogenetic survey of recombination frequency in plant RNA viruses. *Arch Virol.* 2006;151:933–46.
 30. Lefeuvre P, Lett J-M, Varsani A, Martin DP. Widely Conserved Recombination Patterns among Single-Stranded DNA Viruses. *J Virol.* 2009;83:2697–707.
 31. McDonald SM, Nelson MI, Turner PE, Patton JT. Reassortment in segmented RNA viruses: mechanisms and outcomes. *Nat Rev Microbiol.* 2016;14:448–60.
 32. Baltimore D. Expression of animal virus genomes. *Bacteriol Rev.* 1971;35:235–41.
 33. Koonin EV, Dolja VV, Krupovic M, Varsani A, Wolf YI, Yutin N, et al. Global Organization and Proposed Megataxonomy of the Virus World. *Microbiol Mol Biol Rev.* 2020;84:e00061-19.
 34. John SG, Mendez CB, Deng L, Poulos B, Kauffman AKM, Kern S, et al. A simple and efficient method for concentration of ocean viruses by chemical flocculation: Virus concentration by flocculation with iron. *Environ Microbiol Rep.* 2011;3:195–202.
 35. Frantzen CA, Holo H. Unprecedented Diversity of Lactococcal Group 936 Bacteriophages Revealed by Amplicon Sequencing of the Portal Protein Gene. *Viruses.* 2019;11:443.
 36. Sullivan MB, Coleman ML, Quinlivan V, Rosenkrantz JE, DeFrancesco AS, Tan G, et al. Portal protein diversity and phage ecology: Portal protein diversity and phage ecology. *Environ Microbiol.* 2008;10:2810–23.
 37. Sabar MA, Honda R, Haramoto E. CrAssphage as an indicator of human-fecal contamination in water environment and virus reduction in wastewater treatment. *Water Res.* 2022;221:118827.
 38. Maurier F, Beury D, Fléchon L, Varré J-S, Touzet H, Goffard A, et al. A complete protocol for whole-genome sequencing of virus from clinical samples: Application to coronavirus OC43. *Virology.* 2019;531:141–8.
 39. Oluniyi PE, Ajogbasile F, Oguzie J, Uwanibe J, Kayode A, Happi A, et al. VGEA: an RNA viral assembly toolkit. *PeerJ.* 2021;9:e12129.
 40. Dovrolis N, Kassela K, Konstantinidis K, Kouvela A, Veletza S, Karakasiliotis I. ZWA: Viral genome assembly and characterization hindrances from virus-host chimeric reads; a refining approach. Ouzounis CA, editor. *PLOS Comput Biol.* 2021;17:e1009304.
 41. Al Qaffas A, Nichols J, Davison AJ, Ourahmane A, Hertel L, McVoy MA, et al. LoReTTA, a user-friendly tool for assembling viral genomes from PacBio sequence data. *Virus Evol.* 2021;7:veab042.
 42. Johansen J, Plichta DR, Nissen JN, Jespersen ML, Shah SA, Deng L, et al. Genome

- binning of viral entities from bulk metagenomics data. *Nat Commun.* 2022;13:965.
43. Kieft K, Anantharaman K. Virus genomics: what is being overlooked? *Curr Opin Virol.* 2022;53:101200.
44. Kieft K, Adams A, Salamzade R, Kalan L, Anantharaman K. vRhymer enables binning of viral genomes from metagenomes. *Nucleic Acids Res.* 2022;50:e83–e83.
45. Nayfach S, Camargo AP, Schulz F, Eloë-Fadrosch E, Roux S, Kyrpides NC. CheckV assesses the quality and completeness of metagenome-assembled viral genomes. *Nat Biotechnol.* 2021;39:578–85.
46. Roux S, Adriaenssens EM, Dutilh BE, Koonin EV, Kropinski AM, Krupovic M, et al. Minimum Information about an Uncultivated Virus Genome (MIUViG). *Nat Biotechnol.* 2019;37:29–37.
47. Bin Jang H, Bolduc B, Zablocki O, Kuhn JH, Roux S, Adriaenssens EM, et al. Taxonomic assignment of uncultivated prokaryotic virus genomes is enabled by gene-sharing networks. *Nat Biotechnol.* 2019;37:632–9.
48. RedRed/RedRed at master · kseniaarkhipova/RedRed [Internet]. GitHub. [cited 2022 Sep 21]. Available from: <https://github.com/kseniaarkhipova/RedRed>
49. Kurth JM, Nobu MK, Tamaki H, de Jonge N, Berger S, Jetten MSM, et al. Methanogenic archaea use a bacteria-like methyltransferase system to demethoxylate aromatic compounds. *ISME J.* 2021;15:3549–65.
50. Nguyen D, Khanal SK. A little breath of fresh air into an anaerobic system: How microaeration facilitates anaerobic digestion process. *Biotechnol Adv.* 2018;36:1971–83.
51. Tsapekos P, Kougias PG, Vasileiou SA, Lyberatos G, Angelidaki I. Effect of micro-aeration and inoculum type on the biodegradation of lignocellulosic substrate. *Bioresour Technol.* 2017;225:246–53.
52. Kim JK, Oh BR, Chun YN, Kim SW. Effects of temperature and hydraulic retention time on anaerobic digestion of food waste. *J Biosci Bioeng.* 2006;102:328–32.
53. Wilson CA, Murthy SM, Fang Y, Novak JT. The effect of temperature on the performance and stability of thermophilic anaerobic digestion. *Water Sci Technol.* 2008;57:297–304.
54. Wu Y, Wang S, Liang D, Li N. Conductive materials in anaerobic digestion: From mechanism to application. *Bioresour Technol.* 2020;298:122403.
55. Jing Y, Wan J, Angelidaki I, Zhang S, Luo G. iTRAQ quantitative proteomic analysis reveals the pathways for methanation of propionate facilitated by magnetite. *Water Res.* 2017;108:212–21.
56. Angelidaki I, Treu L, Tsapekos P, Luo G, Campanaro S, Wenzel H, et al. Biogas upgrading and utilization: Current status and perspectives. *Biotechnol Adv.* 2018;36:452–66.
57. Fu S, Angelidaki I, Zhang Y. In situ Biogas Upgrading by CO₂-to-CH₄ Bioconversion. *Trends Biotechnol.* 2021;39:336–47.
58. Kapoor R, Ghosh P, Kumar M, Vijay VK. Evaluation of biogas upgrading technologies and future perspectives: a review. *Environ Sci Pollut Res.* 2019;26:11631–61.
59. Willenbücher K, Wibberg D, Huang L, Conrady M, Ramm P, Gätcke J, et al. Phage Genome Diversity in a Biogas-Producing Microbiome Analyzed by Illumina and Nanopore GridION Sequencing. *Microorganisms.* 2022;10:368.
60. Calusinska M, Marynowska M, Goux X, Lentzen E, Delfosse P. Analysis of ds DNA and RNA viromes in methanogenic digesters reveals novel viral genetic diversity. *Environ Microbiol.* 2016;18:1162–75.
61. Heyer R, Schallert K, Siewert C, Kohrs F, Greve J, Maus I, et al. Metaproteome analysis reveals that syntrophy, competition, and phage-host interaction shape microbial communities in biogas plants. *Microbiome.* 2019;7:69.
62. Fastq-dump [Internet]. Bioinforma. Noteb. [cited 2022 Sep 22]. Available from: <https://rnh.github.io/bioinfo-notebook/docs/fastq-dump.html>
63. Bushnell, B. BMAP [Internet]. Available from: <http://sourceforge.net/projects/bbmap/>
64. Li D, Liu C-M, Luo R, Sadakane K, Lam T-W. MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. *Bioinformatics.* 2015;31:1674–6.
65. Gurevich A, Saveliev V, Vyahhi N, Tesler G. QUAST: quality assessment tool for genome

assemblies. *Bioinformatics*. 2013;29:1072–5.

66. Kang DD, Li F, Kirton E, Thomas A, Egan R, An H, et al. MetaBAT 2: an adaptive binning algorithm for robust and efficient genome reconstruction from metagenome assemblies. *PeerJ*. 2019;7:e7359.






67. Wu Y-W, Simmons BA, Singer SW. MaxBin 2.0: an automated binning algorithm to recover genomes from multiple metagenomic datasets. *Bioinformatics*. 2016;32:605–7.

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

69. Fang Z, Tan J, Wu S, Li M, Xu C, Xie Z, et al. PPR-Meta: a tool for identifying phages and plasmids from metagenomic fragments using deep learning. *GigaScience*. 2019;8:giz066.

Article

Evolutionary Study of the CrAssphage Virus at Gene Level

Alessandro Rossi ¹, Laura Treu ^{1,*}, Stefano Toppo ², Henrike Zschach ³,
Stefano Campanaro ^{1,4} and Bas E. Dutilh ⁵

¹ Department of Biology, University of Padova, 35131 Padova, Italy; alessandro.rossi.23@phd.unipd.it (A.R.); stefano.campanaro@unipd.it (S.C.)

² Department of Molecular Medicine, University of Padova, 35131 Padova, Italy; stefano.toppo@unipd.it

³ Department of Biology, University of Copenhagen, 1017 Copenhagen, Denmark; henrike.zschach@bio.ku.dk

⁴ CRIBI Biotechnology Center, University of Padua, 35131 Padova, Italy

⁵ Institute of Biodynamics and Biocomplexity, University of Utrecht, 3508 Utrecht, The Netherlands; bedutilh@gmail.com

* Correspondence: laura.treu@unipd.it; Tel.: +39-049-827-6165

Received: 7 August 2020; Accepted: 14 September 2020; Published: 17 September 2020



Abstract: crAss-like viruses are a putative family of bacteriophages recently discovered. The eponym of the clade, crAssphage, is an enteric bacteriophage estimated to be present in at least half of the human population and it constitutes up to 90% of the sequences in some human fecal viral metagenomic datasets. We focused on the evolutionary dynamics of the genes encoded on the crAssphage genome. By investigating the conservation of the genes, a consistent variation in the evolutionary rates across the different functional groups was found. Gene duplications in crAss-like genomes were detected. By exploring the differences among the functional categories of the genes, we confirmed that the genes encoding capsid proteins were the most ubiquitous, despite their overall low sequence conservation. It was possible to identify a core of proteins whose evolutionary trees strongly correlate with each other, suggesting their genetic interaction. This group includes the capsid proteins, which are thus established as extremely suitable for rebuilding the phylogenetic tree of this viral clade. A negative correlation between the ubiquity and the conservation of viral protein sequences was shown. Together, this study provides an in-depth picture of the evolution of different genes in crAss-like viruses.

Keywords: metaviromics; gene evolution; crAssphage; mirrortree; human gut

1. Introduction

Metagenomics, i.e., the discipline focused on studying genomic sequences from environmental samples, is a relatively new field. The first occurrence of the word itself in the literature dates to 1998 [1]; over the last two decades metagenomics, fueled by the rise in computing power and the development of high-throughput sequencing techniques, has become a major force behind the development of microbiology and environmental biology [2,3]. This revolution is happening for viruses too, whereas the discovery and isolation of new viruses was traditionally linked to the infection of cultivated bacteria, nowadays novel viral sequences are routinely discovered via analysis of environmental samples [4,5]. Thanks to these approaches, the understanding of phage biodiversity has been widely enlarged, with new clades of human- and animal-associated jumbophages and megaphages and novel viruses (without known isolates) being recently described [6,7]. The bulk of the newly discovered viral species is such that the International Committee for the Taxonomy of Viruses (ICTV) has proposed to allow for the classification of new species on the basis of sequence data alone [8]. In summary, metagenomics

has given the possibility to look at the microbiological world from a perspective which transcends the bias associated with the methods previously used. These findings led also to a more-comprehensive analysis of the microbial communities and to a more detailed evaluation of the diversity, prevalence, and ecosystem distribution of phages.

One of the most exciting discoveries from metaviromics is the existence of the crAssphage bacteriophage, found in about half of the human population and being the reference for up to 90% of the viral reads in human gut metagenomes [9,10]. It has subsequently been found, in association with human feces, in every part of the world [11]. This viral species, despite its ubiquity, has been discovered only recently, this is due to the divergence of the crAssphage genomic sequence from other viral genomes and the difficulty of growing its host, a species of the *Bacteroides* genus, *in vitro*. A method based on single assembly from more than one metagenome, named cross-Assembly (hence the phage's name) was used for this purpose [12]. An additional study proved that crAssphage is a member of an entirely new viral clade [13]. The putative crAss-like family has been divided into four subfamilies and ten candidate genera [14]; it includes viruses from diverse ecological niches such as termite gut and marine sediments, as well as another known member of the human gut, the immunodeficiency-associated stool virus (IAS). Given the importance of the microbial communities with regards to human health, it makes sense to investigate every aspect of the most abundant virus of one of the most abundant bacteria residing in the human gut. Furthermore, crAssphage's association with humans is not recent: members of the same viral family, even from the same putative subfamily, are present in non-human primate guts [11]. This suggests that crAss-like viruses are long-time companions of the human lineage. Among the many reasons to study phages, the fact that they can kill specific microbes and can transfer antibiotic resistance or pathogenicity and, consequently, alter host metabolism is one of the most relevant [15,16]. In particular, the impact of variations in the gut microbiome, and potentially of the associated virome as well, on human health is becoming evident [17–19]. Although microbial correlations with human pathologies have been observed, no significant association of crAssphage genomic features with health or disease was found [11]. However, as major components of the human gut, crAss-like viruses deserve more attention, and the dynamics among dominant and rare populations recently described shall be further investigated [11]. In this respect, understanding the overall crAssphage genome evolution and the evolutionary relationships between its individual genes may provide new information.

The crAssphage genome is composed of a dsDNA 97 kbp long circular sequence, divided in two sections in which genes have different functions: replication-related genes are found in the forward strand, while all the other genes are encoded in the reverse strand [9,13]. Similarly, transcription and capsid-related proteins are found in single-purpose blocks of the genome. This organization is common among viruses, and particularly bacteriophages; in fact, this tendency is used in order to identify prophages amidst bacterial scaffolds [20]. Not all genes fit in this model, though. Despite all the efforts, the gene functions in a vast part of the genome remain unknown. Thus, evaluating the coevolution between interacting protein families could provide additional insights on their putative biological function. Specifically, host–parasite interactions are known to be deeply influenced by coevolution, as well as it is known to influence conspecific populations that may diverge or co-adapt under different circumstances [21,22]. Evaluation of the similarity between proteins evolutionary histories has been successfully achieved by correlating mutations using multiple sequence alignments [23]. In this respect, while the previous studies mainly focused on the crAss-like genome as a whole, in the current work we performed a gene-centric analysis of crAssphage evolution. In particular, we examined the conservation degree of protein sequences and the degree of coevolution among them, as well as the relation to their functional roles. The overall aim of our study is to measure the frequency and ubiquity of crAssphage proteins divided by function in order to define a global vector of conservation.

2. Materials and Methods

2.1. Sequences Selection, Retrieval, and Data Preparation

A database of 805 crAss-like assembled sequences was created (Supplementary Materials Table S1), retrieved from previous works [11,13]. The reference crAssphage genome was included in the database. Redundant copies of the same sequences were removed with custom software developed using the Python programming language version 3.6.3 [24] and the Biopython package [25]. PRODIGAL v.2.6.3 [26] was used to predict the genes for all the crAss-like contigs in the database, using the meta-procedure appropriate for viruses and small plasmids. Afterwards, the proteins predicted from the reference crAssphage genome were used to recover the homologous proteins from all the crAss-like contigs (Table S2). For this step, each protein was used as a PSI-BLAST [27] query against the whole crAss-like contig database, with maximum e-value of 0.001 and minimum query coverage of 80% to minimize potential domain walking artifacts. In order to validate the robustness of the PSI-BLAST homology detection, the search was repeated with the query coverage percentages of 50% and 95%. The PSI-BLAST searches were run until convergence. The annotation and function of crAssphage reference genes were retrieved from a previous study [13]. The ORFs of the previous and current studies were compared by BLASTp [28] search considering 90% of query coverage as lower threshold.

The protein functions were loosely grouped into six functional modules, following the spatial division along the reference crAssphage genome, as previously proposed [13]. The functional groups are “Uncharacterized”, in which all the proteins with unknown function are grouped, “Replication”, “Transcription”, “Tail and structural”, “Capsid”, and “Other”, which comprises four proteins whose function has been identified but do not fall into any of the previous categories. The amino-acid sequences were grouped into clusters for each reference protein they matched in the PSI-BLAST search, resulting in 92 clusters, one for each reference crAssphage ORF as predicted by PRODIGAL. Only contigs which included two or more ORFs were included in the database, and this filtering lowered the number used in this study from 805 to 370. Protein sequences that were duplicated or had introns in the reference genome were treated separately. Three different clusters were created for the two RepL genes: one cluster comprises all the genes that hit only the Reference_crAssphage.1_45 reference ORF, another comprises all the genes that hit the Reference_crAssphage.1_91 ORF and the last one includes both groups. The clusters were aligned using MAFFT v7.271 [29] with standard parameters and the L-INS-I algorithm.

2.2. Phylogenetic Tree Reconstruction and Comparison

For every homologous group of proteins a ML tree was built with IQ-TREE v1.6.1 [30] with standard settings and 1000 replicates ultrafast bootstrap [31]. For this analysis, the dUTPase homologous group was manually split into two halves, using AliView v1.19 [32]. Despite the difference in length among the sequences in the RepL homologous group, the alignment did not need manual trimming, as it does not present badly aligned regions which would impair the phylogenetic reconstruction. Furthermore, the phylogenetic tree reconstruction software treats gaps as not containing information, and as such the presence of a well-aligned subset of longer sequences does not have an impact on the overall quality of the reconstruction. The ML matrices generated by IQ-TREE were used for the comparison of phylogenetic trees. Matrices were compared pairwise with the Mirrortree method [33]. In each comparison, rows and columns were trimmed in order to keep only the nodes coming from the same contigs; the matrices were then converted into vectors and compared using Pearson correlation. Pairs of homologous groups which shared less than 5 sequences coming from the same genome were discarded, in order to avoid biases due to low sample size. These operations were performed via custom scripts in Python, using the Scipy package [34].

2.3. Quantifying Sequence Conservation

The conservation of the amino acid sequences was measured as the average Shannon information content across the entire alignment of sequences, as defined by Shannon [35], via custom Python scripts. As a comparison, all Prokaryotic Viral Orthologous Groups (pVOGs), which include proteins coming from viruses belonging to the Podoviridae family, were retrieved in May 2020 [36]. The annotation was retrieved and used to classify the proteins according to their functional categories, using a custom Perl script and based on crAssphage categorization [13]. They were aligned as described above and the Shannon information content was calculated using the same scripts. Linear correlation (R and *p*-values) between Shannon information content and log₁₀ value of the number of genes in the group was calculated independently per each functional class and plotted using ggpubr R package.

2.4. Function Prediction and Distant Similarity Searches

To further characterize the proteins assigned to the “Uncharacterized” group, recent and best performing algorithms for protein function prediction were applied; these include the sequence-based Argot2.5 server [37] and 3D structure-based I-TASSER [38]. Both methods have been chosen according to their recent best performance in community-wide challenges of prediction methods CAFA3 [39] and CASP13 [40].

2.5. Data Visualization and Availability

All the data visualization was performed using the Matplotlib package for Python [41], embedded in various in-house developed scripts. Phylogenetic trees were displayed via the iTOL v4.42 website [42]. All the in-house developed scripts, as well as the complete dataset including all proteins, annotations, and homologous groups are publicly available on a GitHub repository <https://github.com/Ale-Rossi/crAssphage-gene-evolution>.

3. Results

3.1. Protein Identification and Clustering

Whereas the length of the crAssphage reference genome is 97 kbp, only 5% of the contigs (41) lie in the 80–100 kbp range; thus, most of them do not represent complete genome sequences. Eighty-one percent of the contigs used in this study (302) are shorter than 40 kbp and the average length is 26 kbp (Figure S1). While the high number of shorter contigs is probably due to incompletely assembled genomes, we were still able to use these sequences to quantify the co-evolutionary signal between protein families that are encoded on the same unit. Moreover, some long contigs share relatively few ORFs with the crAssphage reference genome, representing distantly related crAss-like bacteriophages [13,14].

A total of 92 ORFs were identified in the reference crAssphage genome and, using similarity search against ORFs identified in all other crAss-like genome sequences, 92 protein clusters were produced. Most proteins were assigned to a single cluster, with two notable exceptions as described in the Methods section and further analyzed below. Although the Prodigal software is not optimized for viral genomes, it identified the same genes as previous studies that used Glimmer and MetaGeneMark [9,13]. More specifically, the first study identified 80 protein-encoding genes; the latter identified 90 protein-encoding genes. We further classified the proteins into functional groups: the replication and structural modules are the largest, featuring 24 and 17 genes, respectively; the capsid, replication and “other” groups are very small, with only 3, 2, and 4 genes respectively. No annotation was available for 42 genes representing the “Uncharacterized” group. The predictions of this group and a tentative consensus between Argot2.5 and I-TASSER showed an improvement in annotation and partial agreement between the two tools for 4 out of 42 proteins (Table S3).

Each protein cluster comprises from 8 to 174 sequences, not equally distributed across the functional categories. In fact, some clusters have a less variable number of sequences, partly due to the

linkage of genes closely located on the genome. Nonetheless, all the functional groups have a median of more than 50 sequences per cluster. Additionally, most of the proteins with a low cluster size are uncharacterized (Figure 1). The capsid proteins stand out as the most ubiquitous ones, with the major capsid protein (MCP) being the most widespread gene, and present in 174 contigs.

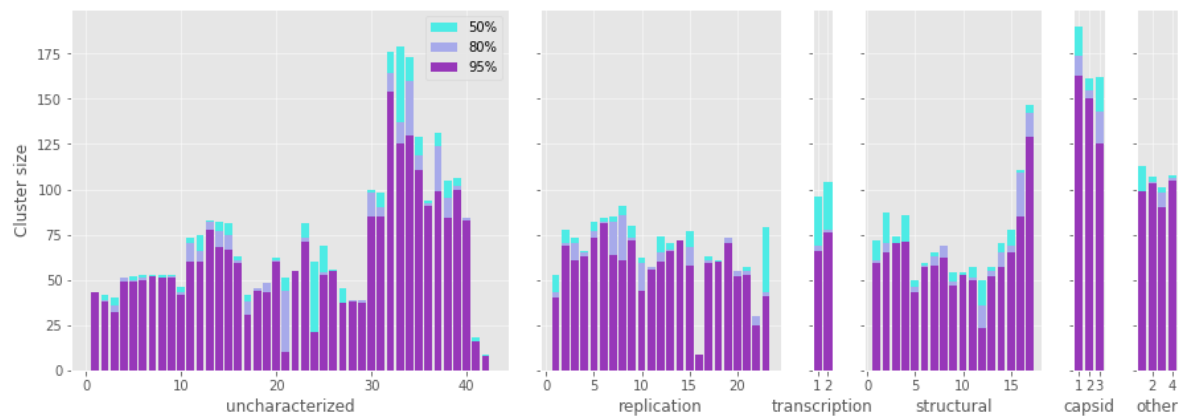


Figure 1. Number of sequences in each group of protein homologs (cluster size) according to functional categories. The three color bars refer to the different similarity thresholds applied to filter alignments (50%, 80%, and 95% respectively). The capsid proteins are the most frequently present in the crAss-like contigs. See Table S2 for the names of all homologous groups of proteins and values for all statistics.

Several proteins presented exceptions to the standard clustering applied in this work. The gene encoding RepL is found in two copies in the crAssphage reference genome. The paralogs arose through an ancient gene duplication, as supported by the RepL phylogenetic tree, which shows a distinct separation between the two ORFs (Figure S2a). In addition, a total of three gene duplications have been newly detected in the collection of crAss-like phage contigs, including duplications of Reference_crAssphage.1_44, Reference_crAssphage.1_73, and Reference_crAssphage.1_74. In particular, Reference_crAssphage.1_73 encodes for a tail sheath protein that, following the reference genome structure, is included in the short module composed of tail and structural proteins. Six contigs contain duplications of the Reference_crAssphage.1_74 sequence (i.e., Gut.14, Gut.03, Gut.07, Gut.05, Activated_sludge.3, and Gut.06). In the phylogenetic tree of the protein, these sequences are very closely related (Figure S2b). No annotation is available for this ORF, but the genomic context and the presence in close proximity of genes encoding capsid proteins are suggestive of a structural role. Additional evidence comes from the iVIREONS structural protein score (0.62) [9], and putative structural homology to a kinase (max TM-score: 0.62, max C-score: -3.71) according to I-TASSER [38] (Table S3). RepL proteins, which were first found in *Staphylococcus aureus* plasmids, are known to increase the number of copies of the plasmids they are found in [43]; this prompts us to speculate that the duplication could have given the virus an evolutionary advantage relative to a heightened reproductive ability. In-depth studies are needed to confirm this. There is not enough data to allow for speculation regarding the other instances of gene duplication, as most events are found in single genomes, and it is then not possible to predict how the duplications play a role in the phage life cycle.

Additionally, another protein cluster contains two ORFs encoded on the crAssphage reference genome recognized as two distinct regions of a dUTPase gene that are split by the insertion of an intron-encoded endonuclease belonging to the HNH protein family, Reference_crAssphage.1_35. The intron insertion is relatively recent and was found in nine contigs, all of which were assembled from shotgun reads collected from gut samples of the twin sisters of a single family [44]. This can also be seen in the phylogenetic tree where the sequences having the insertion form a single clade (Figure S2c).

3.2. Protein Conservation

In order to improve interpretation of the findings in the crAss-like family, all the prokaryotic viral orthologous groups (pVOGs) were analyzed and trends in the distribution of the conservation degree were identified. In pVOGs, there is a negative correlation between the number of sequences in the alignment and the Shannon information content (Figure 2). The relation between the two variables can be described with a logarithmic regression with the correlation coefficient ranging from -0.6 to -0.81 according to the functional category ($p < 0.05$ for all categories). The pVOGs as a whole show a correlation coefficient of -0.67 ($p < 2.2 \times 10^{-16}$). A similar trend, albeit not as pronounced, can be seen in crAss-like protein homologous groups too. Both the structural, the uncharacterized and the replication proteins have a negative correlation coefficient, but the p -value calculated for the latter category was not significant ($R = -0.41$, $p = 0.0073$; $R = -0.65$, $p = 0.0049$; $R = -0.29$, $p = 0.19$). The remaining functional categories, having a very small size, were ignored. crAss-like proteins, as a whole, show this correlation as well ($R = -0.39$, $p = 1 \times 10^{-4}$). The most likely explanation for this negative correlation is a sampling effect where many, widespread sequences may contain more diversity than few sequences with a narrow occurrence distribution. The stronger correlation coefficients of pVOGs relative to the homologous groups of proteins in crAss-like viruses that were built in this study is likely due to other factors. Firstly, pVOGs are built from proteins found across all the prokaryotic viruses and include several homologous groups with a wide distribution that strongly contribute to the low correlation coefficient (note the difference in scale of the X-axes in Figure 2). Conversely, the homologous groups of proteins built in this study only include sequences from the relatively closely related group of crAss-like viruses, which have been proposed to represent a single viral family. Thus, no homologous groups are observed with more than a few hundred sequences and the correlation coefficient is less extreme than for the pVOGs. Secondly, the high threshold of 80% query coverage used in the BLASTp has raised the degree of conservation between the detected homologous proteins that formed the crAss-like clusters. For this same reason, the average Shannon information content of crAss-like proteins is rather high; both the mean and the median lie between 3.0 and 3.5 bits, compared to a maximum of about 4.32 bits. The variation in information content in proteins is consistent across almost all the different protein functions (Figure S3). In all the six groups of proteins we found both highly conserved and divergent sequences, the exception being the capsid proteins. The capsid proteins are unusually low in Shannon information content, as their average value is just 1.95 bits. As with other widely spread genes, it is possible that their ubiquity has led to a great divergence, and that there is less negative evolutionary pressure on them.

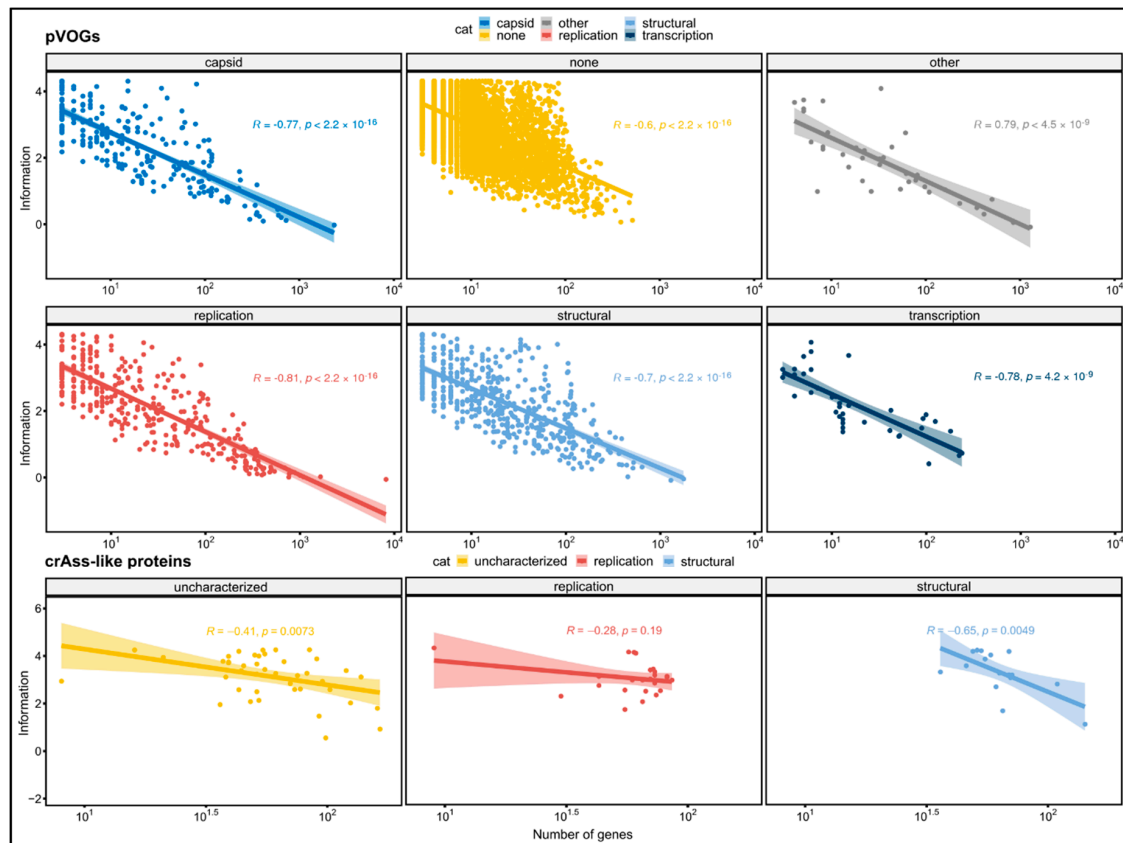


Figure 2. Correlation between the logarithm of the number of sequences in a protein family and the Shannon information content of the positions in the protein sequence alignment. Inverse correlations were obtained both for crAss-like viruses and pVOGs. For proteins in crAss-like viruses, only functional classes having more than four genes are reported.

3.3. Tree Comparison: Consistency of Evolutionary Signal

In order to measure to what extent the encoded proteins followed a similar evolutionary history, we ran a Mirrortree algorithm among pairs of homologous groups of proteins, which consists of the correlation between the distance matrices derived from each pair of phylogenetic trees. The distribution of the 4186 pairwise Pearson coefficients follows a two-peaks distribution: the higher peak ranges from around 0.7 to 1 and consists of 357 pairwise correlations, while the lower peak is from -0.1 to 0.4 , including 1409 pairs (Figure 3). It can be seen that there are regionally defined groups of highly correlating genes distributed in the genome (Figure 3). The first of these regions (gene position 12–28) encompasses part of the replication module, including the DNA polymerase, a helicase, a primase, and a DNA ligase, i.e., the core of the DNA replication machinery, as well as enzymes involved in the protection of DNA, such as an uracil-DNA glycosylase and a thymidylate synthase. The second large region (gene position 72–88) including high-correlation genes ranges from the end of the structural to the “other” functional module, encompassing all the capsid encoding genes. Other, narrower regions include transcription and structural proteins (gene positions 31, 33, 47–48, 51–54, and 58–59). These results suggest the presence of a subset of genes evolving coherently with each other while others undergo a different evolutionary history, a possible cause of which is genomic mosaicism, which is common in bacteriophages. Overall, the distributions of average correlation coefficients across the functional groups is relatively even (Figure S4); all the functional categories feature highly correlating genes, and the three most abundant groups (uncharacterized, replication, and structural) all feature low-correlating genes (Figure S4). The capsid proteins, though, stand out as the group with the highest

correlations, since none of their average mirrortree coefficients are lower than 0.6 ($p = 0.0039$, one-tailed Mann–Whitney U test).

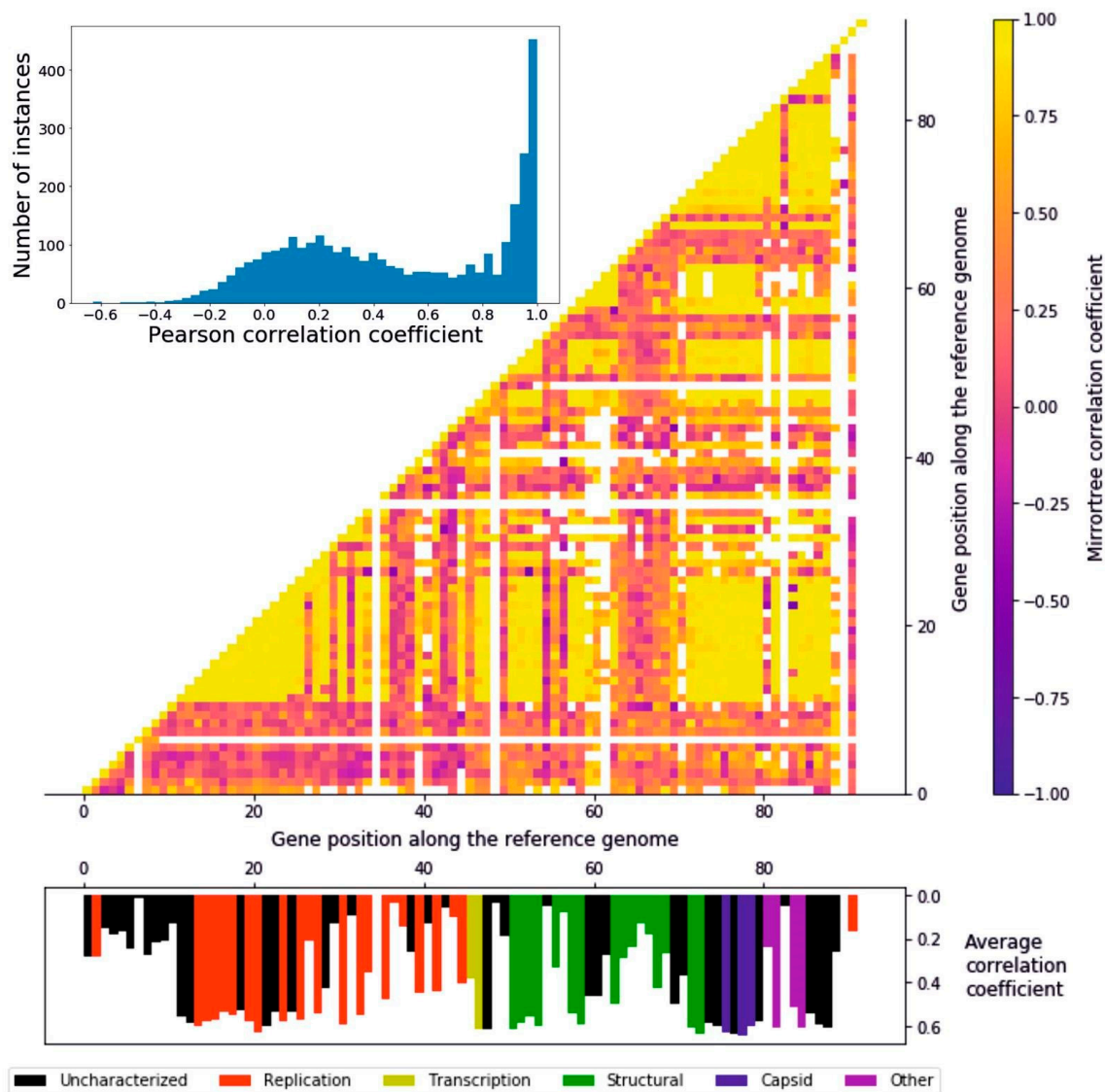


Figure 3. Mirrortree algorithm applied to the homologous groups of proteins. Histogram representing the distribution of the pairwise Pearson's r coefficients. A great number of genes appear to be coevolving. Heatmap of the Pearson correlation coefficient of each protein with any other. Along the x and y axis the 92 ORFs identified in the reference crAssphage genome are represented. The interactions between clusters sharing less than five sequences were colored in white, in order to avoid confusion due to a low size. Histogram representing the average correlation coefficient of every protein represented in their position on the reference genome. The different colors represent the six functional groups.

4. Discussion

Genomes exhibiting mosaicism are composed of parts with different evolutionary origins and history. This phenomenon is particularly evident in viral genomes, due to evolutionary processes such as horizontal gene transfer and recombination both with the host and other viral species. In fact, their evolutionary model has been described as the accretion and exchange of various genetic modules [45]. We propose that mosaicism is a prominent feature of crAss-like viruses as well. The heatmap representing the pairwise correlation coefficients (Figure 3) bears witness to this observation: the presence in the genome of regions with a high degree of coevolution, detected with the mirrortree

coefficients that were calculated from pairs of phylogenies of homologous groups of proteins, can be interpreted as a result of these evolutionary mechanisms, with the low-correlating genes being accessories to a core genome.

Furthermore, the protein prediction and clustering steps reveal the existence of crAss-like viruses sharing only a small fraction of genes with the reference genome, these genes belonging mostly to the capsid and structural functional modules.

Our approach and the cutoffs used for detecting homologs were rather strict, as nearly full-length homology was required (50%, 80%, and 95% of the query length). Based on these cutoffs and the dataset used, the capsid proteins were found to be the most widespread proteins in the crAss-like viruses. They have been used as phylogenetic markers by Yutin and colleagues because of this very reason [13]. In fact, they are among the genes typically used as signature genes for the identification of viral sequences in metagenomic samples and phylogenetic trees reconstruction. Other genes frequently used as such are portal proteins, tail sheath proteins and polymerases, and specific metabolic genes [46]. Their sequence conservation is below average; although surprising, this could reflect the variety of crAss-like phage species they are found in. After all, an inverse correlation between sequence conservation and spread seems to be the norm, as seen in crAss-like genomes and pVOGs alike. Whereas all the categories feature proteins with high average correlation coefficients, the capsid proteins are by far the group with the highest statistic; this observation highlights how crucial these genes are in phage evolution. From both observations it emerges that capsid proteins in crAss-like phages are confirmed to be one of the most important protein families when it comes to reconstructing their evolutionary history [47]. In summary, whereas other proteins could be regarded as either less plastic/adaptable than capsid proteins, or having too much variability, capsid proteins show themselves as among the most malleable of the crAss-like phage proteins. Indeed, these proteins may be able to reflect the evolution of crAssphage and to develop variation without losing functionality.

It would be easy to speculate that the genes which make up the most phylogenetically consistent part of a viral genome are well conserved, but this is not the case. For each protein cluster the mirrortree statistics done in comparison with every other cluster was averaged. In none of the functional groups was it possible to find a correlation between the average mirrortree statistics and the Shannon information content of a homologous group of proteins. Nonetheless, there can be a discrepancy between the species tree and the gene trees: overall the capsid proteins are probably the best way to retrieve the species tree [47]. Nevertheless, they have been confirmed to be the most ubiquitous among all genes, being present in 174, 155, and 143 crAss-like contigs each. The only homologous groups featuring a similar number of sequences, i.e., 160 and 164, belong to uncharacterized proteins whose genes are close to the capsid-encoding genes in the reference genome. Despite being not characterized, they had already been theorized as capsid proteins [13]. This finding invigorates this hypothesis. The capsid proteins are actually an excellent example of how genes can have a very consistent evolution with other genes while keeping a low level of sequence conservation: this could be due to a similarity in tree topology, with branch lengths varying proportionately. Moreover, it is possible that the higher variability and high average correlation coefficient of these proteins are due to the larger number of homologs; the high correlation statistic would emerge from the subtrees being consistent with the trees of other proteins. On the one hand, it would be assumed that genes which lie on the same genome show similar evolutionary patterns; on the other hand, the mirrortree method was developed in order to identify interacting proteins, not necessarily from the same organism. Nevertheless, many of the proteins within an organism do interact reciprocally: structural proteins interlock with each other in order to form the virion's structure, and proteins that build complexes similarly do. However, different evolutionary pressures acting on different genes (and different regions on the same gene), and phenomena such as recombination, transposition, horizontal gene transfer, and gene duplications often lead to different genes of the same genome having trees with different topologies [48–50]. In light of this, such a high degree of coevolution is definitely remarkable in the evolution of crAss-like viruses.

No functional prediction is currently available for almost half of the genes in the reference genome. There is good reason to believe that many of these genes are of great interest, when studying the biology of crAss-like viruses: some uncharacterized genes are ubiquitous in the viral contigs. Another important finding to point out is that, among the proteins which have a high mirrortree coefficient with other proteins, all the functional groups are represented. These protein categories only loosely share a function, so, with some hindsight, it is not surprising to attest how much they diverge. It is worth noticing, though, that even among the top-scoring proteins the uncharacterized category is represented, but including also highly conserved proteins. This is the same for every group, with the exception of the capsid proteins. This means that many genes are highly conserved and strictly co-evolving with other genes, thus potentially interacting with other genes and important in the virus biology, have an unknown function. The uncharacterized functional group in the crAssphage genome is by far the most numerous among the six categories, boasting 42 genes, whereas the second most abundant one only has 24.

The difficulty of annotating genomes is a huge problem when trying to decipher viruses' biology. In this study, we attempted to annotate these proteins with available top-performing function prediction tools. Unfortunately, all the proteins revealed to be difficult targets and only for a small portion of them it was possible to extract functional predictions from Argot2.5 and I-TASSER. In spite of the scores not being very high, results provided potential function for four proteins, i.e., Reference_crAssphage.1_12, Reference_crAssphage.1_13, Reference_crAssphage.1_50, and Reference_crAssphage.1_62 (Table S3). This may be, indeed, due to many factors such as the short length of some proteins submitted for the prediction that could be artifacts of the gene prediction step. One brute-force approach to deal with the lack of annotation could just be ignoring all the uncharacterized proteins and focus our attention only on the already annotated ones. In fact, the 89 ORFs previously reported [13] are only slightly different from the 92 genes identified in the present study. The additional sequences are very short and are not matched by other viral genes in the BLAST search. In fact, many of the homologous groups of proteins with the lowest sequence count belong to the "Uncharacterized" functional group (Figure 1).

One of our findings is that crAss-like family phages follow the trend of prokaryotic viruses in which more widespread genes appear to be less conserved, with the capsid proteins being the utmost representatives of this situation. crAssphage's genes are clearly consistent in their evolution: there is a bulk of genes showing a high coefficient of coevolution. The capsid proteins are confirmed to be a good choice when building a viral phylogenetic tree. They seem to coevolve consistently with many other genes and are thus fit to represent the evolution of the genome as a whole.

Gene duplications are common across all biological entities, and crAss-like viruses are no exception. While some are recent and found in a small number of genomes, a few are ancient and widespread, suggesting relevance in the evolutionary success of the virus lineage. More specifically, it can be easily speculated that the duplication event involving RepL possibly lends to a reproductive advantage, since such gene is known to increase the number of copies of the plasmid it is found in [43]. Still, in-depth studies are needed to confirm this. Another duplication event found in more than one genome is the one concerning the Reference_crAssphage.1_74 ORF, found in six closely related genomes. This protein has not been characterized and our functional prediction has not been able to unequivocally predict a putative function, though a weak hypothesis has emerged that this could be a structural protein with a trans-membrane region, and potentially kinase-like activity. Transposing elements too are ubiquitous, and one, a HNH endonuclease, has made its way into a crAssphage lineage. It is likely too early to have an idea about its impact on the bacteriophage biology, as this insertion appears to be very recent and specifically localized into a few individuals.

Metaviromics has allowed to investigate more in detail the ecosystem distribution of phages transcending the biases of classical isolation-based studies. The fast technological development of the high-throughput sequencing will allow to overcome the limitations encountered in the study of crAss-like viruses and other metagenomic data: new sequencing techniques can reduce—or eliminate—the need of computationally intensive assembly programs. For example, the application of

long read-producing sequencing methods, such as Oxford nanopore [51], would greatly improve the quality of the assembly. While annotation issues are more difficult to address, more investigations, such as 3D structure analysis, are becoming possible. Furthermore, the newly found possibility to grow crAssphage in vitro might open the door to experimental annotation of its proteins [51].

Supplementary Materials: The following are available online at <http://www.mdpi.com/1999-4915/12/9/1035/s1>, Figure S1: Contig length distribution, Figure S2: Phylogenetic trees of relevant evolutionary events, Figure S3: Distribution of average Shannon information, Figure S4: Distribution of average Pearson coefficients, Table S1: Starting dataset of the project with 805 crAss-like contigs, Table S2: Homologous groups of proteins and corresponding statistics, and Table S3: Results of the 3D function prediction.

Author Contributions: Conceptualization A.R. and B.E.D.; methodology A.R., S.T., and B.E.D.; formal analysis A.R.; investigation, A.R. and L.T.; resources H.Z., S.C., and B.E.D.; writing—original draft preparation A.R.; writing—review and editing A.R., L.T., S.T., S.C., and B.E.D.; visualization A.R., L.T., and S.C.; and supervision L.T., S.C., and B.E.D. All authors have read and agreed to the published version of the manuscript.

Funding: B.E.D. was supported by NWO Vidi grant 864.14.004 and ERC Consolidator grant 865694: DiversiPHI.

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

References

1. Handelsman, J.; Rondon, M.R.; Brady, S.F.; Clardy, J.; Goodman, R.M. Molecular biological access to the chemistry of unknown soil microbes: A new frontier for natural products. *Chem. Biol.* **1998**, *5*, R245–R249. [[CrossRef](#)]
2. Koonin, E.V. Environmental microbiology and metagenomics: The Brave New World is here, what's next? *Environ. Microbiol.* **2018**, *20*, 4210–4212. [[CrossRef](#)]
3. Koonin, E.V.; Dolja, V.V. Metaviromics: A tectonic shift in understanding virus evolution. *Virus Res.* **2018**, *246*, A1–A3. [[CrossRef](#)]
4. Al-Shayeb, B.; Sachdeva, R.; Chen, L.-X.; Ward, F.; Munk, P.; Devoto, A.; Castelle, C.J.; Olm, M.R.; Bouma-Gregson, K.; Amano, Y.; et al. Clades of huge phages from across Earth's ecosystems. *Nature* **2020**, *578*, 425–431. [[CrossRef](#)] [[PubMed](#)]
5. Schulz, F.; Roux, S.; Paez-Espino, D.; Jungbluth, S.; Walsh, D.A.; Denev, V.J.; McMahon, K.D.; Konstantinidis, K.T.; Eloë-Fadrosh, E.A.; Kyrpides, N.C.; et al. Giant virus diversity and host interactions through global metagenomics. *Nature* **2020**, *578*, 432–436. [[CrossRef](#)] [[PubMed](#)]
6. Yuan, Y.; Gao, M. Jumbo Bacteriophages: An Overview. *Front. Microbiol.* **2017**, *8*. [[CrossRef](#)] [[PubMed](#)]
7. Devoto, A.E.; Santini, J.M.; Olm, M.R.; Anantharaman, K.; Munk, P.; Tung, J.; Archie, E.A.; Turnbaugh, P.J.; Seed, K.D.; Blekhan, R.; et al. Megaphages infect *Prevotella* and variants are widespread in gut microbiomes. *Nat. Microbiol.* **2019**, *4*, 693–700. [[CrossRef](#)]
8. Simmonds, P.; Adams, M.J.; Benkő, M.; Breitbart, M.; Brister, J.R.; Carstens, E.B.; Davison, A.J.; Delwart, E.; Gorbalenya, A.E.; Harrach, B.; et al. Virus taxonomy in the age of metagenomics. *Nat. Rev. Microbiol.* **2017**, *15*, 161–168. [[CrossRef](#)]
9. Dutilh, B.E.; Cassman, N.; McNair, K.; Sanchez, S.E.; Silva, G.G.Z.; Boling, L.; Barr, J.J.; Speth, D.R.; Seguritan, V.; Aziz, R.K.; et al. A highly abundant bacteriophage discovered in the unknown sequences of human faecal metagenomes. *Nat. Commun.* **2014**, *5*, 4498. [[CrossRef](#)]
10. Koonin, E.V.; Yutin, N. The crAss-like Phage Group: How Metagenomics Reshaped the Human Virome. *Trends Microbiol.* **2020**, *28*, 349–359. [[CrossRef](#)]
11. Edwards, R.A.; Vega, A.A.; Norman, H.M.; Ohaeri, M.; Levi, K.; Dinsdale, E.A.; Cinek, O.; Aziz, R.K.; McNair, K.; Barr, J.J.; et al. Global phylogeography and ancient evolution of the widespread human gut virus crAssphage. *Nat. Microbiol.* **2019**, *4*, 1727–1736. [[CrossRef](#)] [[PubMed](#)]
12. Dutilh, B.E.; Schmieder, R.; Nulton, J.; Felts, B.; Salamon, P.; Edwards, R.A.; Mokili, J.L. Reference-independent comparative metagenomics using cross-assembly: crAss. *Bioinformatics* **2012**, *28*, 3225–3231. [[CrossRef](#)] [[PubMed](#)]
13. Yutin, N.; Makarova, K.S.; Gussow, A.B.; Krupovic, M.; Segall, A.; Edwards, R.A.; Koonin, E.V. Discovery of an expansive bacteriophage family that includes the most abundant viruses from the human gut. *Nat. Microbiol.* **2018**, *3*, 38–46. [[CrossRef](#)]

14. Guerin, E.; Shkoporov, A.; Stockdale, S.R.; Clooney, A.G.; Ryan, F.J.; Sutton, T.D.S.; Draper, L.A.; Gonzalez-Tortuero, E.; Ross, R.P.; Hill, C. Biology and Taxonomy of crAss-like Bacteriophages, the Most Abundant Virus in the Human Gut. *Cell Host Microbe* **2018**, *24*, 653–664.e6. [CrossRef] [PubMed]
15. Balcazar, J.L. Bacteriophages as Vehicles for Antibiotic Resistance Genes in the Environment. *PLoS Pathog.* **2014**, *10*, e1004219. [CrossRef]
16. Penadés, J.R.; Chen, J.; Quiles-Puchalt, N.; Carpena, N.; Novick, R.P. Bacteriophage-mediated spread of bacterial virulence genes. *Curr. Opin. Microbiol.* **2015**, *23*, 171–178. [CrossRef] [PubMed]
17. Carding, S.R.; Davis, N.; Hoyles, L. Review article: The human intestinal virome in health and disease. *Aliment. Pharmacol. Ther.* **2017**, *46*, 800–815. [CrossRef]
18. de la Cuesta-Zuluaga, J.; Corrales-Agudelo, V.; Velásquez-Mejía, E.P.; Carmona, J.A.; Abad, J.M.; Escobar, J.S. Gut microbiota is associated with obesity and cardiometabolic disease in a population in the midst of Westernization. *Sci. Rep.* **2018**, *8*, 11356. [CrossRef]
19. Kashyap, P.C.; Quigley, E.M. Therapeutic implications of the gastrointestinal microbiome. *Curr. Opin. Pharmacol.* **2018**, *38*, 90–96. [CrossRef]
20. Akhter, S.; Aziz, R.K.; Edwards, R.A. PhiSpy: A novel algorithm for finding prophages in bacterial genomes that combines similarity- and composition-based strategies. *Nucleic Acids Res.* **2012**, *40*, e126. [CrossRef]
21. Juan, D.; Pazos, F.; Valencia, A. High-confidence prediction of global interactomes based on genome-wide coevolutionary networks. *Proc. Natl. Acad. Sci. USA* **2008**, *105*, 934–939. [CrossRef] [PubMed]
22. Burton, R.S.; Rawson, P.D.; Edmands, S. Genetic architecture of physiological phenotypes: Empirical evidence for coadapted gene complexes. *Am. Zool.* **1999**, *39*, 451–462. [CrossRef]
23. Pazos, F.; Valencia, A. Protein co-evolution, co-adaptation and interactions. *EMBO J.* **2008**, *27*, 2648–2655. [CrossRef] [PubMed]
24. Van Rossum, G.; Drake, F.L., Jr. *Python Reference Manual*; Centrum voor Wiskunde en Informatica: Amsterdam, The Netherlands, 1995.
25. Cock, P.J.A.; Antao, T.; Chang, J.T.; Chapman, B.A.; Cox, C.J.; Dalke, A.; Friedberg, I.; Hamelryck, T.; Kauff, F.; Wilczynski, B.; et al. Biopython: Freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics* **2009**, *25*, 1422–1423. [CrossRef]
26. Hyatt, D.; Chen, G.-L.; LoCascio, P.F.; Land, M.L.; Larimer, F.W.; Hauser, L.J. Prodigal: Prokaryotic gene recognition and translation initiation site identification. *BMC Bioinform.* **2010**, *11*, 119. [CrossRef] [PubMed]
27. Altschul, S. Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. *Nucleic Acids Res.* **1997**, *25*, 3389–3402. [CrossRef]
28. Altschul, S.F.; Gish, W.; Miller, W.; Myers, E.W.; Lipman, D.J. Basic local alignment search tool. *J. Mol. Biol.* **1990**, *215*, 403–410. [CrossRef]
29. Katoh, K.; Standley, D.M. MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability. *Mol. Biol. Evol.* **2013**, *30*, 772–780. [CrossRef]
30. Nguyen, L.-T.; Schmidt, H.A.; von Haeseler, A.; Minh, B.Q. IQ-TREE: A Fast and Effective Stochastic Algorithm for Estimating Maximum-Likelihood Phylogenies. *Mol. Biol. Evol.* **2015**, *32*, 268–274. [CrossRef]
31. Hoang, D.T.; Chernomor, O.; von Haeseler, A.; Minh, B.Q.; Vinh, L.S. UFBoot2: Improving the Ultrafast Bootstrap Approximation. *Mol. Biol. Evol.* **2018**, *35*, 518–522. [CrossRef] [PubMed]
32. Larsson, A. AliView: A fast and lightweight alignment viewer and editor for large datasets. *Bioinformatics* **2014**, *30*, 3276–3278. [CrossRef] [PubMed]
33. Pazos, F.; Valencia, A. Similarity of phylogenetic trees as indicator of protein–protein interaction. *Protein Eng. Des. Sel.* **2001**, *14*, 609–614. [CrossRef] [PubMed]
34. Jones, E.; Oliphant, T.; Peterson, P. SciPy: Open Source Scientific Tools for Python. 2001. Available online: <https://www.scienceopen.com/document?vid=ab12905a-8a5b-43d8-a2bb-defc771410b9> (accessed on 14 September 2020).
35. Shannon, C.E. A Mathematical Theory of Communication. *Bell Syst. Tech. J.* **1948**, *27*, 623–656. [CrossRef]
36. Graziotin, A.L.; Koonin, E.V.; Kristensen, D.M. Prokaryotic Virus Orthologous Groups (pVOGs): A resource for comparative genomics and protein family annotation. *Nucleic Acids Res.* **2017**, *45*, D491–D498. [CrossRef]
37. Lavezzo, E.; Falda, M.; Fontana, P.; Bianco, L.; Toppo, S. Enhancing protein function prediction with taxonomic constraints—The Argot2.5 web server. *Methods* **2016**, *93*, 15–23. [CrossRef]
38. Yang, J.; Zhang, Y. I-TASSER server: New development for protein structure and function predictions. *Nucleic Acids Res.* **2015**, *43*, W174–W181. [CrossRef]

39. Zhou, N.; Jiang, Y.; Bergquist, T.R.; Lee, A.J.; Kacsoh, B.Z.; Crocker, A.W.; Lewis, K.A.; Georghiou, G.; Nguyen, H.N.; Hamid, M.N.; et al. The CAFA challenge reports improved protein function prediction and new functional annotations for hundreds of genes through experimental screens. *Genome Biol.* **2019**, *20*, 244. [[CrossRef](#)] [[PubMed](#)]
40. Kryshchak, A.; Schwede, T.; Topf, M.; Fidelis, K.; Mout, J. Critical assessment of methods of protein structure prediction (CASP)—Round XIII. *Proteins Struct. Funct. Bioinform.* **2019**, *87*, 1011–1020. [[CrossRef](#)]
41. Hunter, J.D. Matplotlib: A 2D Graphics Environment. *Comput. Sci. Eng.* **2007**, *9*, 90–95. [[CrossRef](#)]
42. Letunic, I.; Bork, P. Interactive Tree Of Life (iTOL): An online tool for phylogenetic tree display and annotation. *Bioinformatics* **2007**, *23*, 127–128. [[CrossRef](#)]
43. Dwidar, M.; Yokobayashi, Y. Riboswitch Signal Amplification by Controlling Plasmid Copy Number. *ACS Synth. Biol.* **2019**, *8*, 245–250. [[CrossRef](#)] [[PubMed](#)]
44. Reyes, A.; Haynes, M.; Hanson, N.; Angly, F.E.; Heath, A.C.; Rohwer, F.; Gordon, J.I. Viruses in the faecal microbiota of monozygotic twins and their mothers. *Nature* **2010**, *466*, 334–338. [[CrossRef](#)] [[PubMed](#)]
45. Hendrix, R.W.; Lawrence, J.G.; Hatfull, G.F.; Casjens, S. The origins and ongoing evolution of viruses. *Trends Microbiol.* **2000**, *8*, 504–508. [[CrossRef](#)]
46. Adriaenssens, E.M.; Cowan, D.A. Using Signature Genes as Tools To Assess Environmental Viral Ecology and Diversity. *Appl. Environ. Microbiol.* **2014**, *80*, 4470–4480. [[CrossRef](#)]
47. Krupovic, M.; Koonin, E.V. Multiple origins of viral capsid proteins from cellular ancestors. *Proc. Natl. Acad. Sci. USA* **2017**, *114*, E2401–E2410. [[CrossRef](#)]
48. Szöllősi, G.J.; Tannier, E.; Daubin, V.; Boussau, B. The Inference of Gene Trees with Species Trees. *Syst. Biol.* **2015**, *64*, e42–e62. [[CrossRef](#)]
49. Nichols, R. Gene trees and species trees are not the same. *Trends Ecol. Evol.* **2001**, *16*, 358–364. [[CrossRef](#)]
50. Pamilo, P.; Nei, M. Relationships between gene trees and species trees. *Mol. Biol. Evol.* **1988**. [[CrossRef](#)]
51. Shkoporov, A.N.; Khokhlova, E.V.; Fitzgerald, C.B.; Stockdale, S.R.; Draper, L.A.; Ross, R.P.; Hill, C. ΦCrAss001 represents the most abundant bacteriophage family in the human gut and infects *Bacteroides intestinalis*. *Nat. Commun.* **2018**, *9*, 4781. [[CrossRef](#)]



© 2020 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<http://creativecommons.org/licenses/by/4.0/>).



KEMET – A python tool for KEGG Module evaluation and microbial genome annotation expansion

Matteo Palù^{a,1}, Arianna Basile^{a,1}, Guido Zampieri^{a,*}, Laura Treu^{a,**}, Alessandro Rossi^a, Maria Silvia Morlino^a, Stefano Campanaro^{a,b}

^a Department of Biology, University of Padova, Via U. Bassi 58/b, 35121 Padova, Italy

^b CRIBI Biotechnology Center, University of Padova, 35131 Padova, Italy

ARTICLE INFO

Article history:

Received 8 November 2021

Received in revised form 17 March 2022

Accepted 18 March 2022

Available online 26 March 2022

Keywords:

Gene annotation

Microbial genome

Metabolic pathway

Hidden Markov model

Genome-scale metabolic model

ABSTRACT

Background: The rapid accumulation of sequencing data from metagenomic studies is enabling the generation of huge collections of microbial genomes, with new challenges for mapping their functional potential. In particular, metagenome-assembled genomes are typically incomplete and harbor partial gene sequences that can limit their annotation from traditional tools. New scalable solutions are thus needed to facilitate the evaluation of functional potential in microbial genomes.

Methods: To resolve annotation gaps in microbial genomes, we developed KEMET, an open-source Python library devised for the analysis of Kyoto Encyclopedia of Genes and Genomes (KEGG) functional units. KEMET focuses on the in-depth analysis of metabolic reaction networks to identify missing orthologs through hidden Markov model profiles.

Results: We evaluate the potential of KEMET for expanding functional annotations by simulating the effect of assembly issues on real gene sequences and showing that our approach can identify missing KEGG orthologs. Additionally, we show that recovered gene annotations can sensibly increase the quality of draft genome-scale metabolic models obtained from metagenome-assembled genomes, in some cases reaching the accuracy of models generated from complete genomes.

Conclusions: KEMET therefore allows expanding genome annotations by targeted searches for orthologous sequences, enabling a better qualitative and quantitative assessment of metabolic capabilities in novel microbial organisms.

© 2022 The Authors. Published by Elsevier B.V. on behalf of Research Network of Computational and Structural Biotechnology. This is an open access article under the CC BY-NC-ND license (<http://creativecommons.org/licenses/by-nc-nd/4.0/>).

1. Introduction

Metagenomics investigates environmental, engineered, and host-associated microbiomes, stimulating new fast-growing applications in biomedicine and biotechnology [1,2]. The shift towards a holistic approach in microbiome studies can uncover biological activities emerging from synergistic cooperation of microorganisms [3]. Many environments are now being inspected to decipher inhabiting microbial communities, with the aim of predicting their functions and interactions. Thanks to recent improvements in genome-resolved metagenomics, the recovery of metagenome-assembled genomes (MAGs) of high quality is becoming accessible

and fast [4]. Functional analysis of genomes derived from metagenomic approaches allows estimating the metabolic potential of species present in a given microbiota. Several dedicated databases, such as the Kyoto Encyclopedia of Genes and Genomes (KEGG), are used as knowledgebases for metabolic pathway inference and reconstruction [5], while tools such as KEGG Mapper [6] and eggNOG-Mapper [7] can assign open reading frames to their function and predict metabolic capabilities at the genome level. However, newly generated metagenomes contain a large number of poorly characterized species, which can be hardly annotated exhaustively with traditional tools [8].

Moreover, genome-scale metabolic models (GSMMs) are now starting to be applied on a metagenome scale [3,9]. GSMM are directly informed by annotation databases and can be automatically reconstructed using tools like CarveMe [10] or gapseq [11]. Such models are useful to infer interactions among microbial species, but the application to uncultured and non-model species can be challenging. In fact, MAG-based GSMMs are especially prone to

* Corresponding author.

** Co-corresponding author.

E-mail addresses: guido.zampieri@unipd.it (G. Zampieri), laura.treu@unipd.it (L. Treu).

¹ Equally contributing authors.

reconstruction errors due to the gapped nature of metagenomic assemblies. Starting from GSMM reconstructions, several algorithms for network gap-fill enable *in silico* growth simulation and phenotype data fitting. Nevertheless, reactions added this way are not always supported by genomic evidence [12], possibly resulting in erroneous predictions.

To obtain a more exhaustive functional annotation of microbial genomes and improve associated GSMMs, we present KEMET. KEMET - KEgg Module Evaluation Tool - is a command-line, open-source Python toolbox aiming at summarizing and expanding KEGG annotation by comparing microbial sequences to orthologs with curated annotations. With KEMET, annotation recovery from trusted knowledgebases can strengthen the biological fidelity and phenotype prediction in GSMMs and lower the manual refinement effort.

2. Methods and implementation

Starting from genome sequences and associated KEGG annotations, KEMET serves three main goals: functional annotation evaluation, HMM-driven ortholog search in the original sequence, and integration of the corresponding metabolic reactions into GSMMs (Fig. 1). KEMET is a system-independent tool and every dependency is available to UNIX-based and Windows systems. KEMET is freely available and can be downloaded from the GitHub page <https://github.com/Matteopaluh/KEMET>, where all the procedures to reproduce the results presented in this manuscript are available.

2.1. Module completeness evaluation

The evaluation of metabolic functions present in microbial genomes of interest is performed according to KEGG Modules [5], which consist of manually curated logical expressions of ortholog genes defining the biochemical steps (blocks) of a given function. Functional annotations deriving from different software can directly serve as input data for the Module completeness evaluation, allowing for a flexible downstream implementation of KEMET on pre-existing pipelines. Examples of the supported input files are available in the “toy” folder of the dedicated GitHub repository. At the present time, eggNOG-mapper [7], KofamKOALA [13], and KAAS [14] annotations are supported, and they can be selected through the *-a* parameter. Blocks having KEGG Ortholog (KO) annotations can be identified in target genomes by running KEMET, which allows scaling up the analysis to hundreds of MAGs. Present or missing ortholog blocks in the original annotation can be identified by querying files with KO Module structures. This analysis brings a considerable advantage with respect to the use of KEGG tools alone, allowing to point out single missing orthologs, thus aiding in targeted queries regarding metabolic capabilities of input genomes. The output includes a human-readable tabular file and a flat file indicating the sequential position of missing KOs.

To implement this feature, KEGG Module files are downloaded via the KEGG application-programming interface (API) and parsed to generate intermediate files (<module_id>.kk files in the GitHub repository) that are used as Module block structure templates and queried during script usage. The logic behind the block structure in KEMET is devised so as to better identify missing orthologs connected to a single biochemical step. Specifically, the number of blocks in a Module is given by the highest number of individual KOs involved in any alternative reaction path. For example, in Module M00308 the terminal glyceraldehyde-3-phosphate conversion can either be performed via a mechanism involving two KO genes, or via a single dehydrogenase ortholog. While in KEGG Mapper these KOs belong to a single block, KEMET decomposes the longer path into two blocks. These alternative algorithms lead to

the same results in terms of numbers of missing orthologs but can give slightly different results when the Module completeness is inspected, as shown in the Results and discussion Section.

2.2. Identification of missing KEGG orthologs

KOs missing from functional annotation can result in incomplete KEGG Modules. This phenomenon can be due to real biological gaps in the species metabolic potential, gene truncation resulting from gaps in the assembly, or limitations of the functional annotation procedure. Missing genes can be sought more in-depth in the genomic sequences, using nucleotidic hidden Markov models (HMM) automatically generated by KEMET, when the *--hmm_mode* parameter is indicated. KEMET has different options for HMM profile generation and for missing KO search. The set of input sequences for the HMM profiles can derive from KOs in an input user-defined list (*--hmm_mode kos*) or from KOs in Modules of interest, e.g. those pointing to specific metabolic functions in the input genomes (*--hmm_mode modules*). Alternatively, HMMs can be built from the KOs of all Modules with one incomplete ortholog block (*--hmm_mode onebm*).

When this analysis is performed, the following workflow is employed with every KO of interest:

1. A taxonomically relevant subset of the KEGG GENES database is downloaded via the KEGG API. This subset includes sequences for every species included in a clade, defined by a C-level KEGG BRITE taxonomical hierarchy (br08601). Such taxonomy is generally almost coincident to that on the phylum level, or to that on the class level for a few specific taxa (e.g. Euryarchaeota).
2. A filtering step is performed to obtain a non-redundant set of sequences. A multiple sequence alignment is built up from these sequences using MAFFT v7.475 [15]. The *--auto* parameter is used here, to choose the appropriate strategy among the possible algorithms according to the size of the alignment dataset.
3. A HMM is generated from the aligned sequences using the *hmmbuild* command from the HMMER suite v3.1b2 [16]. Only the subset of KOs indicated in the *--hmm_mode* argument is utilized.
4. The obtained profiles are searched in the genome of interest with the *nhmmer* program from HMMER version 3.1b2 [16].

The default threshold value depends on the *nhmmer* score divided by the length of the profile HMM. Preliminary tests were performed to fine-tune this value, comparing translated BLASTp hits against the NCBI nr dataset (performed in March 2021), which were manually checked for two different MAG datasets. The threshold identified the highest number of hits with sequence names matching the correct KEGG ortholog gene descriptors, while pointing to the lowest number of false positives. Values obtained from the aforementioned tests resulted in 4.6–7.5% of the hits, depending on the input dataset. Stringency of the scoring for significant hits can be modulated with the *--threshold_value* parameter.

2.3. Integration of recovered biochemical reactions into genome-scale metabolic models

In automated draft GSMM reconstruction, metabolic reactions are collected based on genome or protein sequence alignment scores. Using KEMET, the HMM best scoring hits can be selected, providing new insights into the metabolic network obtained from the initial gene calling process. One option is the generation of a novel GSMM with newly identified orthologs. Alternatively, the HMM prioritization process determines a different set of reactions to be included in an existing GSMM. KEMET implements the --

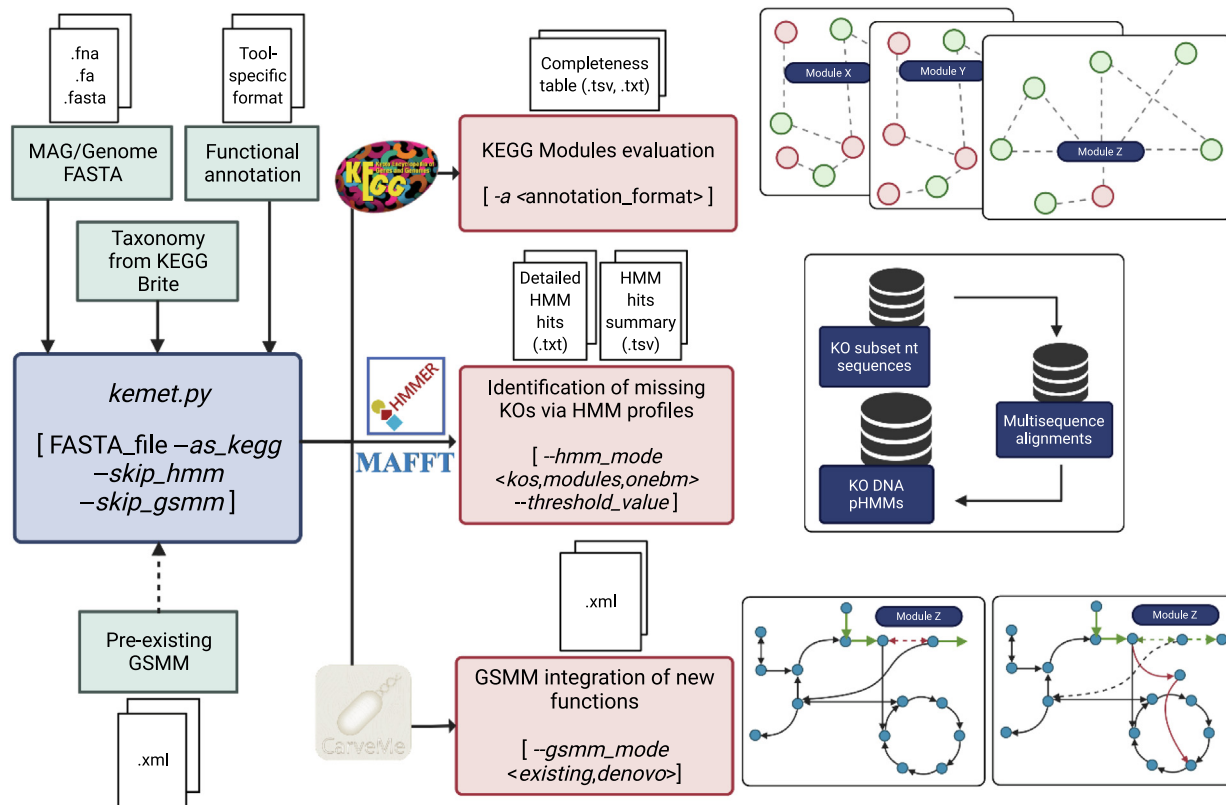


Fig. 1. Workflow of KEMET reporting the input files, outputs, and main parameters for all the tasks that can be executed: KEGG Module evaluation, identification of missing KOs, and integration of identified KOs in GSMMs. On the right side, the rationale of each task is visually outlined.

gsmm_mode parameter to include the newly predicted biochemical functions obtained from genomic evidence into the GSMMs.

KEMET links HMM-identified KOs to their corresponding biochemical reactions present in reference databases for GSMM, namely BiGG [17] and ModelSEED [18]. Their namespaces are adopted by popular GSMM reconstruction tools, such as CarveMe [10] and gapseq [11]. The retrieved reactions can then be incorporated in input GSMMs. As a second option, the translated HMM KO hits can be directly added to the input sequences used for *de novo* genome-scale model generation.

3. Results and discussion

To validate KEMET, we first compared its KEGG Module partitioning with those performed by KEGG Mapper and METABOLIC v4.0 [19] across all the KEGG Modules present at the time of the tests. As shown by Fig. 2A, the three tools interpret the Module block structure in a largely consistent way. However, KEMET is able to capture more Modules in the evaluation and has a block structure that more closely resembles that of KEGG as compared to METABOLIC.

Next, we validated KEMET annotation expansion by two different approaches: (a) an annotation removal strategy to test its ability to identify known KO annotations, and (b) a draft GSMM reconstruction strategy to verify that newly identified annotations produce more sound quantitative models of microbial metabolism, and thus reflect correctly identified functions.

Strategy (a) was used to test *kemet.py --hmm_mode* capability to retrieve the proper annotated sequences when either the original annotation was removed or the sequence was truncated. The rationale was to simulate misassembly-derived gene disruptions and

other problems impairing functional prediction in MAGs. KEMET was tested on 12 MAGs derived from a contig-level assembly resulting from a previous work [20] as well as 5 complete genomes downloaded from NCBI (details in Supplementary Data). In terms of taxonomic “novelty”, the MAGs were highly different and included species assigned at different levels (spanning from class to species) using GTDB-tk v1.5.0 [21]. The gene calling was performed using Prodigal v2.6.3 [22] with default options. Functional annotations of predicted genes were performed using eggNOG-mapper v2 [7] with default parameters. While in principle alternative gene predictions can impact the subsequent functional annotation, previous empirical investigations found negligible performance variation among different tools [23,24]. For this reason, our tests focused on benchmarking functional annotation prediction by using a single state-of-the-art gene prediction tool.

The test consisted in the removal of three KO annotations from the input set of each genome (i.e. from eggNOG results) before running KEMET with the *--hmm_mode onebm* option. The selected KOs were annotated once per genome, only on a single gene. Moreover, removed KOs were chosen from different Modules marked “Complete” by KEMET, among the mandatory orthologs for a given biochemical step. In this way, removing them would result in the change of Module completeness to “1 block missing”. Altogether, 20/36 and 8/12 KO mock removals (55% and 67% true positive rate) resulted in the correct gene and annotation recovery for MAGs and complete genomes, respectively (Fig. 2B and Supplementary Data).

To model MAG construction issues more closely, the removal strategy was repeated two more times by simulating the deletion of tested KO-annotated gene sequences, either by 30% or 70% of their original length. This was done to mimic the typical scenario of a highly fragmented assembly where gene sequences can be

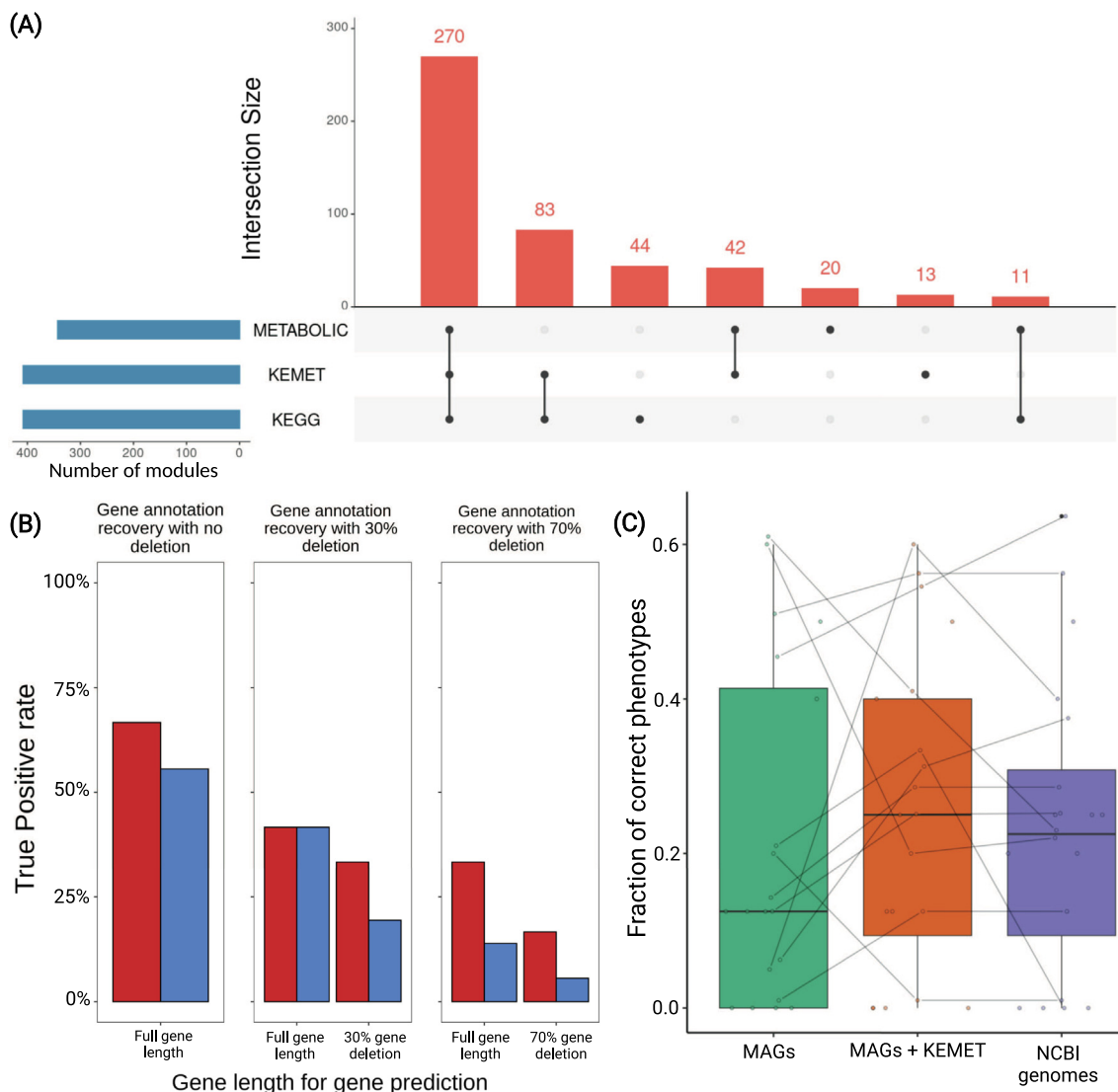


Fig. 2. Results of KEMET quality tests. (A) Comparison between KEMET and METABOLIC in terms of KEGG Module block structure with respect to the original KEGG Modules obtained through KEGG Mapper. The plot shows the intersections among the Module datasets for the three tools, together with the total number of Modules evaluated by each of them. (B) True positive rate for gene sequence identification by HMMs. Results for both isolated genomes (red) and MAGs (blue) are reported. Gene deletions of different extents were performed prior to running KEMET. When deletions were performed, gene annotation recovery was evaluated both with the gene prediction resulting from the original sequences and from those truncated, in order to account for the impact of deletions on gene prediction. (C) Fraction of correct metabolic phenotypes predicted by GSMMs reconstructed from microbial MAGs (green), the same MAGs with an expanded annotation through KEMET (orange), and the corresponding genomes from isolates (purple), based on the literature. The lines track the performance of individual GSMMs corresponding to the same strain. For readability purposes, only lines between points having performance differences across the datasets were drawn. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

split between two different contigs, resulting in a missed gene prediction or improper functional annotation. These additional tests resulted in a decreased performance using both the complete genomes and the MAGs dataset, as expected, but nonetheless gave a significant annotation recovery rate for gene truncations shorter than 50%. Specifically, an annotation recovery between 20% and 33% was achieved when accounting for the impact of sequence truncation on the gene prediction step, whereas a recovery rate of 42% was obtained assuming an unbiased gene prediction. This interval therefore captures KEMET performance in the presence of minor gene deletions. Similarly, for 70% gene truncations the annotation recovery rate further decreases, more clearly for the MAG dataset, as it is sensible with most of the gene sequence lacking. Hence, these results provide a proof-of-principle of KEMET annotation recovery in the occurrence of gene sequence disruption.

Detailed results are included in the GitHub page at <https://github.com/Matteopaluh/KEMET/blob/main/tests/README.md>.

Strategy (b) was implemented to assess the impact of recovering missing KO annotation on downstream metabolic analyses, i.e. via GSMM reconstruction. Specifically, we compared microbial phenotypes recovered from the literature (indicated in [Supplementary Data](#)) in terms of metabolite production or consumption capabilities, to their corresponding *in silico* model predictions. This analysis was performed starting from MAGs and their corresponding complete genomes recovered from the NCBI or from the PATRIC database (as pointed by <https://github.com/snayfach/IGGdb>), by selecting species collected from the anaerobic digestion microbiome [20]. MAG quality metadata were recovered and included genome completeness and contamination. If more than one MAG per species was present in the database, those with $\geq 90\%$ com-

pleteness and $\leq 5\%$ contamination were considered for the subsequent analysis. Both MAGs and the complete genomes of isolates were used to check the Module completeness. MAGs were also used to search for missing KOs by using *kemet.py --hmm_mode onebm*. GSMMs were reconstructed from complete genomes and MAGs using CarveMe v1.4.1 [10] with the options *--fbc2 -u*, using as input both the MAG original gene calling and this same data added with the translated nucleotide sequences identified with the HMM via KEMET using the *--gsmm_mode denovo* parameter. Moreover, KEMET performance times were monitored and are included in [Supplementary Data](#).

To benchmark how the addition of newly identified sequences affects GSMM ability to describe *in silico* microbial physiology, metabolic capabilities retrieved from the literature were compared with predictions obtained starting from three types of input for GSMM reconstruction: MAG annotation, MAG annotation expanded with KEMET, and complete genome annotation. Flux variability analysis (FVA) was performed on the obtained GSMMs for assessing such metabolic capabilities, as follows. For each metabolite export reaction, it was determined whether the range of possible fluxes was directed towards metabolite consumption or production (respectively, having flux ranges consisting only of negative or positive values), while maintaining a fixed maximal growth rate. FVA results showing blocked reactions or flux ranging both positive and negative values were considered as incorrect predictions. The results show a nearly 10% improvement in the ability of MAG-derived GSMMs to produce and consume metabolites predicted from wet lab experiments, with an acquired accuracy comparable with the accuracy of GSMMs reconstructed from the genomes of isolates (both around 33%, [Fig. 2B](#) and [Supplementary Data](#)). On the annotation level, HMMs used on MAGs resulted in 84.76% hits in common with the respective reference isolate genome selected; 7.62% hits were present solely in the MAG dataset (false positives), and 7.62% hits were present in the complete genome dataset alone (false negatives). According to the selected dataset, KEMET HMM predictions therefore display a 91.75% precision and 91.75% sensitivity ([Supplementary Data](#)). Despite the addition of a limited number of protein sequences, the resulting models can thus be sensibly more accurate, leading to more precise inferences based on metabolic capabilities. For example, *Selenomonas ruminantium* MAG-derived GSMM (PATRIC genome id: 971.16) phenotype predictions were improved after KEMET usage. The original GSMM could not predict any known metabolic capability of *S. ruminantium*, while the modified GSMM could correctly reproduce metabolic exchanges involving cellobiose, salicin, mannitol, xylose, arabinose, fructose, maltose, lactose, and sucrose. In contrast, the GSMM based on the full genome annotation captured the correct exchanges for glycerol, cellobiose, salicin, mannitol, xylose, and arabinose.

These results demonstrate that KEMET efficiently tackles the summarization of (meta)genomic potential in a user-friendly and scalable way. Other bioinformatics tools allow the evaluation of microbial genome annotation completeness (e.g. METABOLIC [19]). However, to date and up to our knowledge, this is the only tool able to selectively fill the gaps in the annotation, and seamlessly add newly gathered information into GSMMs. At the moment, KEMET relies on KEGG given its structure allowing a systematic pathway completeness evaluation. Further development could include support towards other knowledgebases, such as MetaCyc [25], to further expand the tool compatibility and predictive power. While other published programs, such as DRAM and Anvi'o [26,27] rely on specific KEGG releases, KEGG databases are constantly updated due to newly added sequences, or newly defined KO classifications. In contrast, KEMET allows users to update the downloaded KEGG GENES database through the KEGG API, in order to use the most up-to-date version of KEGG database

without relying on fixed versions. The download of such a database represents the only limiting computational factor in KEMET ([Supplementary Data](#)), being a mandatory step to comply with the KEGG license. More efficient communication with KEGG servers could be obtained via license, while better solutions will be explored and implemented in future versions of KEMET. Nevertheless, this step is required only once at each database update, which can be decided by the user. Further, KEMET is based on HMMs given their broad applicability in the genomics and metagenomics fields. Other probabilistic graphical models, such as conditional random fields or Bayesian networks could be implemented in future versions of the software.

Altogether, our experiments show that focusing on Module completeness down to single orthologs can aid in identifying missing annotations and enable their correction, not only supporting qualitative evaluation of microbial functions but also improving quantitative models of microbial metabolism. This enables a better mechanistic investigation of microbial ecological roles, allowing us to gather insights without relying necessarily on cultivation or in-depth characterization, which is impractical for most metagenomic studies.

Author Statement

The authors declare that all of them have seen and approved the final version of the manuscript. The manuscript is the authors' original work, has not received prior publication and is not under consideration for publication elsewhere.

Funding

This work was financially supported by the “Budget Integrato della Ricerca Dipartimentale” (BIRD198423) PRID 2019 of the Department of Biology of the University of Padua, entitled “SyM-MoBio: inspection of Syntrophies with Metabolic Modelling to optimize Biogas Production”, by the project “Sviluppo Catalisi dell’Innovazione nelle Biotecnologie” (MIUR ex D.M.738 dd 08/08/19) of the Consorzio Interuniversitario per le Biotecnologie” (CIB), and by the project LIFE20 CCM/GR/001642 – LIFE CO2toCH4 of the European Union LIFE+ program.

CRediT authorship contribution statement

Matteo Palù: Conceptualization, Methodology, Software, Validation, Investigation, Data curation, Writing – original draft, Writing – review & editing, Visualization. **Arianna Basile:** Conceptualization, Methodology, Software, Writing – original draft, Writing – review & editing, Visualization. **Guido Zampieri:** Methodology, Validation, Supervision, Writing – review & editing, Visualization. **Laura Treu:** Conceptualization, Supervision, Writing – review & editing, Funding acquisition. **Alessandro Rossi:** Methodology, Software. **Maria Silvia Morlino:** Methodology, Writing – review & editing. **Stefano Campanaro:** Conceptualization, Supervision, Project administration, Funding acquisition.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix A. Supplementary data

Supplementary data to this article can be found online at <https://doi.org/10.1016/j.csbj.2022.03.015>.

References

- [1] Nayfach S, Roux S, Seshadri R, Udway D, Varghese N, Schulz F, et al. A genomic catalog of Earth's microbiomes. *Nat Biotechnol* 2021;39:499–509. <https://doi.org/10.1038/s41587-020-0718-6>.
- [2] D'Hondt K, Kostic T, McDowell R, Eudes F, Singh BK, Sarkar S, et al. Microbiome innovations for a sustainable future. *Nat Microbiol* 2021;6:138–42. <https://doi.org/10.1038/s41564-020-00857-w>.
- [3] Basile A, Campanaro S, Kovalovszki A, Zampieri G, Rossi A, Angelidaki I, et al. Revealing metabolic mechanisms of interaction in the anaerobic digestion microbiome by flux balance analysis. *Metab Eng* 2020;62:138–49. <https://doi.org/10.1016/j.ymben.2020.08.013>.
- [4] Bowers RM, Kyrpidis NC, Stepanauskas R, Harmon-Smith M, Doud D, Reddy TBK, et al. Minimum information about a single amplified genome (MISAG) and a metagenome-assembled genome (MIMAG) of bacteria and archaea. *Nat Biotechnol* 2017;35:725–31. <https://doi.org/10.1038/nbt.3893>.
- [5] Kanehisa M, Goto S, Sato Y, Kawashima M, Furumichi M, Tanabe M. Data, information, knowledge and principle: back to metabolism in KEGG. *Nucleic Acids Res* 2014;42:D199–205. <https://doi.org/10.1093/nar/gkt1076>.
- [6] Kanehisa M, Sato Y, Kawashima M. KEGG mapping tools for uncovering hidden features in biological data. *Protein Sci* n.d.;n/a. <https://doi.org/10.1002/pro.4172>.
- [7] Cantalapiedra CP, Hernández-Plaza A, Letunic I, Bork P, Huerta-Cepas J. eggNOG-mapper v2: Functional Annotation, Orthology Assignments, and Domain Prediction at the Metagenomic Scale. *Mol Biol Evol* 2021. <https://doi.org/10.1093/molbev/msab293>.
- [8] Frioux C, Dittami SM, Siegel A. Using automated reasoning to explore the metabolism of unconventional organisms: a first step to explore host-microbial interactions. *Biochem Soc Trans* 2020;48:901–13. <https://doi.org/10.1042/BST20190667>.
- [9] Zorrilla F, Buric F, Patil KR, Zelezniak A. metaGEM: reconstruction of genome scale metabolic models directly from metagenomes. *Nucleic Acids Res* 2021. <https://doi.org/10.1093/nar/gkab815>.
- [10] Machado D, Andrejev S, Tramontano M, Patil KR. Fast automated reconstruction of genome-scale metabolic models for microbial species and communities. *Nucleic Acids Res* 2018;46:7542–53. <https://doi.org/10.1093/nar/gky537>.
- [11] Zimmermann J, Kaleta C, Waschina S. gapseq: informed prediction of bacterial metabolic pathways and reconstruction of accurate metabolic models. *Genome Biol* 2021;22:81. <https://doi.org/10.1186/s13059-021-02295-1>.
- [12] Bernstein DB, Sulheim S, Almaas E, Segrè D. Addressing uncertainty in genome-scale metabolic model reconstruction and analysis. *Genome Biol* 2021;22:64. <https://doi.org/10.1186/s13059-021-02289-z>.
- [13] Aramaki T, Blanc-Mathieu R, Endo H, Ohkubo K, Kanehisa M, Goto S, et al. KofamKOALA: KEGG Ortholog assignment based on profile HMM and adaptive score threshold. *Bioinformatics* 2020;36:2251–2. <https://doi.org/10.1093/bioinformatics/btz859>.
- [14] Moriya Y, Itoh M, Okuda S, Yoshizawa AC, Kanehisa M. KAAS: an automatic genome annotation and pathway reconstruction server. *Nucleic Acids Res* 2007;35:W182–5. <https://doi.org/10.1093/nar/gkm321>.
- [15] Nakamura T, Yamada KD, Tomii K, Katoh K. Parallelization of MAFFT for large-scale multiple sequence alignments. *Bioinformatics* 2018;34:2490–2. <https://doi.org/10.1093/bioinformatics/bty121>.
- [16] Wheeler TJ, Eddy SR. nhmmer: DNA homology search with profile HMMs. *Bioinformatics* 2013;29:2487–9. <https://doi.org/10.1093/bioinformatics/btt403>.
- [17] Norsigian CJ, Pularia N, McConn JL, Yurkovich JT, Dräger A, Pálsson BO, et al. BIGG Models 2020: multi-strain genome-scale models and expansion across the phylogenetic tree. *Nucleic Acids Res* 2020;48:D402–6. <https://doi.org/10.1093/nar/gkz1054>.
- [18] Seaver SMD, Liu F, Zhang Q, Jeffries J, Faria JP, Edirisinghe JN, et al. The ModelSEED Biochemistry Database for the integration of metabolic annotations and the reconstruction, comparison and analysis of metabolic models for plants, fungi and microbes. *Nucleic Acids Res* 2021;49:D575–88. <https://doi.org/10.1093/nar/gkaa746>.
- [19] Zhou Z, Tran PQ, Breister AM, Liu Y, Kieft K, Cowley ES, et al. METABOLIC: high-throughput profiling of microbial genomes for functional traits, metabolism, biogeochemistry, and community-scale functional networks. *Microbiome* 2022;10:33. <https://doi.org/10.1186/s40168-021-01213-8>.
- [20] Campanaro S, Treu L, Rodriguez-R LM, Kovalovszki A, Ziels RM, Maus I, et al. New insights from the biogas microbiome by comprehensive genome-resolved metagenomics of nearly 1600 species originating from multiple anaerobic digesters. *Biotechnol Biofuels* 2020;13:25. <https://doi.org/10.1186/s13068-020-01679-y>.
- [21] Chaumeil P-A, Mussig AJ, Hugenholtz P, Parks DH. GTDB-Tk: a toolkit to classify genomes with the Genome Taxonomy Database. *Bioinformatics* 2020;36:1925–7. <https://doi.org/10.1093/bioinformatics/btz848>.
- [22] Hyatt D, Chen G-L, LoCascio PF, Land ML, Larimer FW, Hauser LJ. Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinf* 2010;11:119. <https://doi.org/10.1186/1471-2105-11-119>.
- [23] Korandla DR, Wozniak JM, Campeau A, Gonzalez DJ, Wright ES. AssessORF: combining evolutionary conservation and proteomics to assess prokaryotic gene predictions. *Bioinformatics* 2020;36:1022–9. <https://doi.org/10.1093/bioinformatics/btz714>.
- [24] Dimonaco NJ, Aubrey W, Kenobi K, Clare A, Creevey CJ. No one tool to rule them all: prokaryotic gene prediction tool annotations are highly dependent on the organism of study. *Bioinformatics* 2022;38:1198–207. <https://doi.org/10.1093/bioinformatics/btab827>.
- [25] Caspi R, Billington R, Keseler IM, Kothari A, Krummenacker M, Midford PE, et al. The MetaCyc database of metabolic pathways and enzymes - a 2019 update. *Nucleic Acids Res* 2020;48:D445–53. <https://doi.org/10.1093/nar/gkz862>.
- [26] Eren AM, Kiefl E, Shaiber A, Veseli I, Miller SE, Schechter MS, et al. Community-led, integrated, reproducible multi-omics with anvio. *Nat Microbiol* 2021;6:3–6. <https://doi.org/10.1038/s41564-020-00834-3>.
- [27] Shaffer M, Borton MA, McGivern BB, Zayed AA, La Rosa SL, Solden LM, et al. DRAM for distilling microbial metabolism to automate the curation of microbiome function. *Nucleic Acids Res* 2020;48:8883–900. <https://doi.org/10.1093/nar/gkaa621>.

RESEARCH

Open Access



Analysis of the anaerobic digestion metagenome under environmental stresses stimulating prophage induction

Alessandro Rossi^{1†}, Maria Silvia Morlino^{1†}, Maria Gaspari², Arianna Basile¹, Panagiotis Kougiyas³, Laura Treu^{1*} and Stefano Campanaro^{1,4}

Abstract

Background: The viral community has the potential to influence the structure of the microbiome and thus the yield of the anaerobic digestion process. However, the virome composition in anaerobic digestion is still under-investigated. A viral induction experiment was conducted on separate batches undergoing a series of DNA-damaging stresses, in order to coerce temperate viruses to enter the lytic cycle.

Results: The sequencing of the metagenome revealed a viral community almost entirely composed of tailed bacteriophages of the order *Caudovirales*. Following a binning procedure 1,092 viral and 120 prokaryotic genomes were reconstructed, 64 of which included an integrated prophage in their sequence.

Clustering of coverage profiles revealed the presence of species, both viral and microbial, sharing similar reactions to shocks. A group of viral genomes, which increase under organic overload and decrease under basic pH, uniquely encode the *yopX* gene, which is involved in the induction of temperate prophages. Moreover, the in-silico functional analysis revealed an enrichment of sialidases in viral genomes. These genes are associated with tail proteins and, as such, are hypothesised to be involved in the interaction with the host. *Archaea* registered the most pronounced changes in relation to shocks and featured behaviours not shared with other species. Subsequently, data from 123 different samples of the global anaerobic digestion database was used to determine coverage profiles of host and viral genomes on a broader scale.

Conclusions: Viruses are key components in anaerobic digestion environments, shaping the microbial guilds which drive the methanogenesis process. In turn, environmental conditions are pivotal in shaping the viral community and the rate of induction of temperate viruses. This study provides an initial insight into the complexity of the anaerobic digestion virome and its relation with the microbial community and the diverse environmental parameters.

Background

Anaerobic digestion (AD) is a functional process carried out by microbial communities composed of *Bacteria* and *Archaea* which degrade organic matter in anoxic conditions. AD occurs in natural environments such as aquatic sediments, wetlands, and animal gut, but it is also widely employed in industrial processes. It is particularly valuable as a way to produce methane while disposing of organic waste, playing an important role in the reduction

[†]Alessandro Rossi and Maria Silvia Morlino contributed equally to this work.

*Correspondence: laura.treu@unipd.it

¹ Department of Biology, University of Padua, via U. Bassi 58/b, 35131 Padova, Italy

Full list of author information is available at the end of the article



© The Author(s) 2022. **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>. The Creative Commons Public Domain Dedication waiver (<http://creativecommons.org/publicdomain/zero/1.0/>) applies to the data made available in this article, unless otherwise stated in a credit line to the data.

of the dependence from fossil fuels and the development of a circular economy approach [1].

The composition of AD microbiomes is extremely variable, and it reflects the wide variety of substrates and physicochemical conditions under which this degradation process occurs, both in natural and technical environments [2]. In AD, polymers are first broken down into simple molecules, which are then converted into Volatile Fatty Acids (VFA), then into acetate and finally into methane in the four steps of hydrolysis, acidogenesis, acetogenesis and methanogenesis. The first three steps are conducted by the bacterial community which, despite the great variation across different conditions, is dominated by the phylum *Firmicutes*, usually followed by *Bacteroidetes* and *Proteobacteria*. Archaeal species, mostly belonging to the phylum *Euryarchaeota*, are involved in the conversion of simple molecules to methane and usually account for a much smaller part of the community [3]. The microbial species present and their balance are crucial for optimisation of biogas production, and they have been extensively studied in the last two decades [4]. Among the numerous factors concurring to shape microbial communities, the importance of viruses, in particular bacteriophages, is increasingly recognised [5]. Viral concentration in samples from wastewater treatment plants (WWTPs) has been estimated to be greater in comparison to aquatic environments by one to three orders of magnitude [6, 7]. Furthermore, it has been observed that bacteriophages have a strong correlation with prokaryotic species across time in wastewater-treating bioreactors [8]. Despite the importance of this, most of the existing articles regarding the AD virome are limited to characterisations of the community and do not assess the impact of viruses on the microbial community [9–11]. Zhang and colleagues showed that there is a correlation between the viral community composition and the production of methane in anaerobic digesters of WWTPs and argued that the viral shunt has a positive impact on the production of methane [12], but such conclusions are drawn on a broad scale analysis, leaving many of the actual dynamics unaddressed.

The transition of a temperate phage from lysogenic to lytic cycle is known as induction. Temperate viruses spontaneously undergo induction at a low rate, but in several species of bacteriophages and archaeal viruses, this phenomenon is known to increase with DNA-damaging stresses [13]. For example, in a study targeting the response to different types of anaerobic stresses in *Nitrosospira multiformis* 25196, it was observed how *N. multiformis* cells reacted to a wide range of environmental stresses through prophage induction [14].

A prime example of the importance of bacteriophages in engineered systems is the dairy industry, which is

threatened by bacteriophages attacking *Streptococcus thermophilus* strains [15]. Moreover, viruses are known to be players in the regulation of global carbon and nitrogen cycles in natural ecosystems [12], e.g. aquifer sediments [16], and in phytoplankton dynamics and diversity [17].

As parasites, viruses apply strong selective pressures on their hosts. It has been estimated that in marine ecosystems, viruses kill about 20% of the microbial biomass daily [18]. The recycling of organic matter from lysed microbes, called viral shunt, plays a relevant role in the regulation of global carbon and nitrogen cycles. In biogas plants, phage-induced bacterial cell lysis can decrease biogas production when the key species associated with biogas production are affected. At the same time, auxotrophic microorganisms are benefitted as lysis serves as a source of cofactors, vitamins and amino acids [11]. Furthermore, as mobile genetic elements, viruses enact horizontal gene transfer (HGT) across microbes at different taxonomic ranks, from species to phyla. This potentially endows hosts with beneficial functions, increases the genetic diversity of the population, and plays a role in the complex co-evolutionary dynamics between viruses and hosts [19]. However, both HGT and viral lysis rates in engineered systems are still overlooked.

Despite the elucidation of virus-mediated mechanisms, most of the viral diversity remains unknown [20]. However, the advent of metagenomics has brought large advances in the description of microbial environmental communities. The introduction of binning methods in standard metagenomic data analysis pipelines has allowed for the recovery of many uncultivable AD microbial species [2] and the detailed description of key organisms of the microbiome. The exponential increase of available sequences from both bulk metagenomes and metaviromes has led to the creation of numerous databases of viral sequences [21, 22], paralleled by the development of predictors able to effectively find new phages [23–26]. All viral prediction algorithms depend to some extent on the previous knowledge associated with taxonomically assigned genomes reported in public databases. This applies to all the software, whether they are based on homology search, like CheckV and PHASTER, or leverage k-mer usage like VirFinder, or analyse sequence features within machine learning frameworks, like VIBRANT, VirSorter2 and PPR-Meta [23–28]. However, their application has proven effective in discovering novel viral clades, the most emblematic case being the crAss-like phage family [29–31]. In the light of the relevant results obtained from metaviromics the International Committee on Taxonomy of Viruses proposed the establishment of new classification methods based solely on genomic features [32].

Unravelling the “dark matter” of novel viral diversity is a daunting task, and the aforementioned exploratory studies conducted on the AD virome showed the potential that phages have in shaping the prokaryotic community [9, 11, 12]. The DNA virome of AD has been described as dominated by tailed bacteriophages of the *Siphoviridae*, *Podoviridae* and *Myoviridae* families, with a minor presence of *Tectiviridae*, *Inoviridae* and other families. The AD microbiome is extremely complex and composed of species involved in different functional tasks, including the hydrolysis of organic matter and the conversion of the derived by-products in simple organic molecules (e.g. volatile fatty acids and methane). However, little is known on which prokaryotic species can be potentially affected by phages and, therefore, which are the functional processes potentially influenced by lytic cycles. Heyer and colleagues [11] reported that species belonging to *Bacillaceae*, *Enterobacteriaceae*, and *Clostridiaceae* are among the favourite targets of bacteriophages, but these findings are not conclusive to determine whether specific parts of the AD funnel are more impacted by viruses. Identifying and characterising the viruses and their hosts in this system can lead not only to a better comprehension of AD microbial dynamics but also to applications such as phage-mediated treatment of the reactors in order to increase process performance. Bacteriophages are already used as tools for manipulating microbial communities in different fields, such as phage therapy and pathogen control in food and water, and have been used as control for biomass bulking in wastewater treatment [33–36]. An increased attention towards the AD viral community could lead to the development of similar techniques for the improvement of the AD process as well. This could be achieved by removing species like the sulphate reducers which compete with key players in pivotal steps of methanogenesis, or leveraging bacteriophage-mediated HGT in order to confer desirable metabolic characteristics to microbial species of interest. In this experiment, we attempted to use induction to explore the effect of diverse conditions potentially affecting the AD process on both the microbial and viral community. We then assessed the presence of the retrieved genomes in other AD metagenomes from the Biogas Microbiome collection [2, 37] (microbial-genomes.org).

Materials and methods

Inoculum and feedstock

Active inoculum was obtained from a lab-scale Continuous Stirred Tank Reactor (CSTR) (Waste Management and Bioprocessing Lab, Thessaloniki, Greece), treating cattle manure at mesophilic conditions ($37 \pm 1^\circ\text{C}$). Cattle manure was collected from a full-scale biogas plant located in northern Greece (Biogas Lagada S.A.,

Thessaloniki, Greece). The raw substrate was sieved using a separating net with a 2 mm opening to remove large particles and stored until usage at -20°C to prevent alterations in its composition.

Batch assays experimental setup

In order to test for perturbations of the AD process, 21 anaerobic batch experiments were performed aiming to define the microbial and viral composition in reactors under different conditions. These included addition of mitomycin, temperature shifts, high salt concentration, oxidative stress, pH shifts, and organic overload (details regarding the application of each condition are listed in Additional file 1). Four of these assays involved a combination of temperature change with salt or oxidation stresses. Finally, a control assay was conducted by incubating the inoculum without imposing any stressing condition. All the experiments were performed in triplicate using 300 mL serum glass bottles with a working volume of 50 mL and an organic load of 2 g VS/L, for a total of 66 batches. In addition to the batch experiments, an aliquot of inoculum was saved by storing it at -20°C immediately after sampling. Prior to incubation, bottles were flushed with nitrogen to achieve anaerobic conditions. Thereafter, the bottles were hermetically closed with butyl rubber stoppers and screw caps. The batch reactors were maintained at 37°C in a temperature-controlled incubator (BINDER BD260, Tuttlingen, Germany) for 24 h.

Analytical methods

At the end of each treatment, after 24 h of operation, biogas composition and VFA concentration were measured on all 21 assays plus the control bottles, with the aim of evaluating the effect of the different conditions in the digesters. To determine biogas composition, a gas chromatograph (Shimadzu GC-2014, Kyoto, Japan) equipped with a thermal conductivity detector (TCD) and a packed column (Molecular Sieve 5A, $1.8\text{ m} \times 2\text{ mm ID}$) was used. The VFA concentrations were defined with a gas chromatograph (Shimadzu GC-2010 Pro, Kyoto, Japan) provided with a flame ionisation detector (FID) and equipped with fused silica capillary column ($30\text{ m} \times 0.53\text{ mm ID}$, $1\text{ }\mu\text{m}$ film thickness). The oven temperature was initially set at 50°C for 3.5 min, subsequently increased at a rate of $25^\circ\text{C}/\text{min}$ to 130°C and, finally, increased at a rate of $10^\circ\text{C}/\text{min}$ until reaching the final temperature of 210°C , which was maintained stable for 2 min. The temperature in the injection port was 150°C and in the detector 230°C . Helium was used as carrier gas for the gas chromatograph.

DNA extraction and sequencing

The most promising conditions according to the literature and methane yield variation were selected for DNA extraction, along with the control bottles and the inoculum. Specifically, conditions with a decrease in methane yield between 0 and 30% compared to the control were selected, under the assumption that they were affected enough to potentially observe phage induction, but not to the point of killing the majority of cells (Fig. 1). For each bottle and the inoculum, pellet and supernatant samples were collected as described below for analysing the microbial and viral community, respectively. Centrifugation at high speed was used to separate the viral and bacterial fractions [38, 39]. For the microbial-enriched community (pellet) samples, before starting the extraction, 3 mL of the inoculum were centrifuged at 15000 rpm for 10 min in order to obtain the acquired quantity of 0.2–0.8 g pellet, while the supernatant was discarded. Hereupon, the genomic DNA was extracted. For the viral-enriched community (supernatant) samples, 45 mL of each bottle's content (or of inoculum from the reactor) were centrifuged (Thermo Scientific SL 16R, New York, USA) at 15,000×g for 10 min at 4°C. In order to further enrich supernatant samples in viral content, an attempt was performed to filter the supernatant with 0.22 µm syringe filters [40]. However, due to the high content of suspended and dissolved solids of the substrate, filtering was only possible with 1 µm syringe filters (Millex-GP, Merck Millipore Ltd). With the intention of reducing the final volume while using the whole viral content, the filtered flowthroughs were frozen overnight and lyophilised using a freeze dryer (Christ Alpha 1–2, Martin Christ Gefriertrocknungsanlagen GmbH, Germany) coupled with a vacuum pump (rotary vane vacuum pump, Vacuumbrand RZ 2.5, Vacuumbrand GmbH + CO KG, Germany) for 48 h at 0.4 mbar. Before the extraction of the obtained enriched viral community, the lyophilised samples were resuspended in 3 mL PCR water. Subsequently, the samples underwent DNA extraction with DNeasy® PowerSoil® Kit (QIAGEN, Hilden, Germany) following the manufacturer's protocol. Recovery of DNA from pellet and supernatant samples was ensured by qualitative and quantitative analyses on the samples, using NanoDrop Microvolume UV-Vis spectrophotometer (Thermo Fisher Scientific, USA) and Qubit Fluorometer (Thermo Fisher Scientific, USA). Importantly, DNA yield was limiting for the supernatant samples. Indeed, among all samples, only four tested conditions and the inoculum yielded enough DNA for library preparation, but only upon pooling the replicates for the supernatant. To ensure a coherent comparison, replicates for

pellet samples were also pooled before sequencing, and the four conditions and inoculum were further processed.

DNA samples underwent library preparation using the Nextera DNA Flex Library Prep Kit (Illumina Inc., San Diego CA) and were sequenced using the Illumina Novaseq platform at the CRIBI Biotechnology Center sequencing facility (University of Padova, Italy). The sequencing run yielded 6.3 million 150bp reads on average per sample. Raw data have been deposited at NCBI, BioProject PRJNA767833.

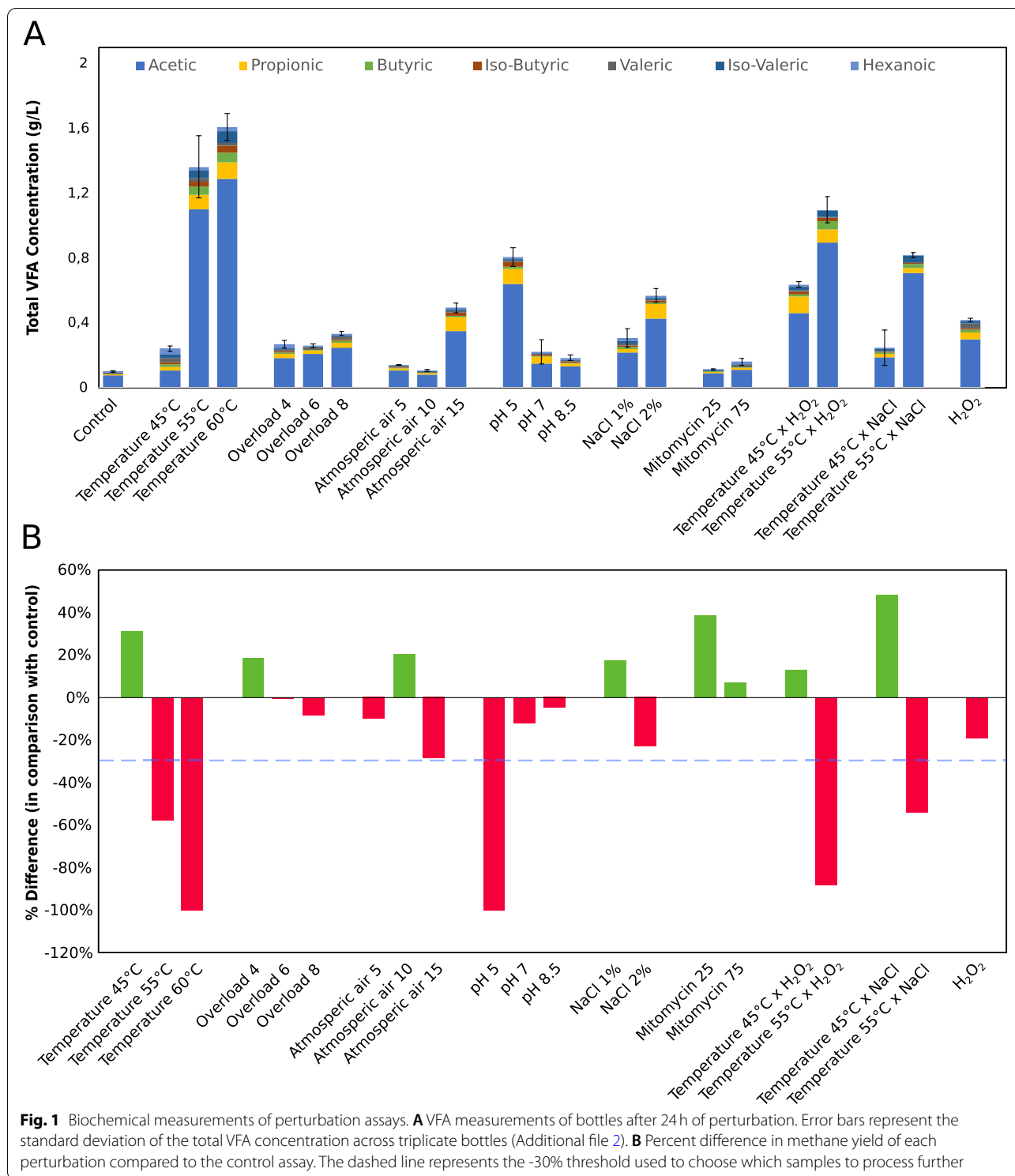
Assembly and binning

Reads were filtered with Trimmomatic v0.39 [41] and cleaned with BBDuk v38.86. The reads of all pellet and supernatant samples were co-assembled using MEGAHIT v1.2.9 [42]. The quality of the co-assembly was assessed with QUAST v5.0.2 [43]. The filtered and cleaned reads were then mapped back on the assembly with Bowtie 2 v2.3.5.1 [44]. Details and parameters used for these programs are reported in Additional file 1. The assembled contigs were analysed with PPR-Meta v1.1, CheckV v0.7.0, VIBRANT v1.2.0 and the PHASTER web server [23, 24, 27, 28]. CheckV results were filtered excluding predictions with “not determined” quality and no viral genes detected. PPR-Meta predictions were filtered for viral scores of 0.75 or higher. VIBRANT and PHASTER predictions were carried forward with no pre-filtering. A first list of contigs classified as viral was defined by considering predictions made by either PHASTER alone, or at least two of the other programs (Additional file 1: Figure 1).

A binning procedure was performed with MetaBAT2 v2.12.1 [45, 46] using a minimum bin size of 10,000 bp. Here, we refer to the output of the binning algorithm as “bins” and to bins which have passed quality control and are thus considered representative of prokaryotic genomes as “MAGs”. The bins yielded by MetaBAT2 were evaluated and divided into Metagenome-Assembled Genomes (MAGs), viral MAGs and unclassified contigs, according to a procedure described in Additional file 1. For prokaryotes, bin quality, completeness and contamination were measured with CheckM v1.1.2 [47]. Finally, CheckV was run again on the viral MAGs recovered from binning in order to calculate genome quality and completeness.

Coverage profiles

Relative abundance of prokaryotic and viral MAGs was calculated by performing genome count per million (CPM) normalisation, which takes into account genome length and sequencing depth, on read counts obtained from the reads mapped on the assembly. The values



obtained were highly similar with those obtained using CheckM software v1.1.2 (Additional file 1: Figure 2). For the 50 most abundant MAGs, the coefficient of variation, defined as the standard deviation divided by the mean,

was calculated. The mean was calculated across pellet samples.

The effect of the different shocks was evaluated by calculating the log ratio between the relative abundance of

genomes in each shock and the mean relative abundance across all the conditions considered. In this context, a positive log ratio refers to an abundance higher than average and vice versa. For an overall comparison of the conditions, Spearman correlation was calculated between log ratio values in different samples. The calculation of the Spearman correlation coefficients and the corresponding *P*-values was carried out with SciPy v1.3.1 [48]. The same analysis was performed by using the control sample as a reference.

The 50 most abundant MAGs and viral genomes were clustered by computing the Euclidean distance from the log ratios under different conditions and using an average linkage method. Correlation values between MAGs were computed by using all eight treated samples (four pellet, four supernatant). The software SparCC (commit 2ddc13f, February 2020) was used to calculate correlation coefficients while taking into account the compositional nature of the data [49]. The input was a matrix of mapped reads on each genome in the eight samples. Significance of the obtained correlation values was assessed by generating 1000 bootstraps and calculating two-sided pseudo *P*-values.

Taxonomic assignment and functional annotation

Prokaryotic MAGs were taxonomically assigned using GTDB-Tk v1.4.1 and converted to NCBI taxonomy with the script `gtdb_to_ncbi_majority_vote.py` [50]. Viral genomes were assigned via Hidden Markov Model against the Prokaryotic Virus Orthologous Groups (pVOGs) database using `hmmsearch` from the HMMER v3.3.2 suite [51, 52].

ORFs were predicted using Prodigal v2.6.3 [53]. Taxonomy was assigned on the basis of a consensus rule, as previously reported [21]. Taxonomy assignment is explained in detail in Additional file 1. Prodigal was also used to detect the presence of alternative stop codons in viral sequences, following the method used by Borges and colleagues [54].

Functional annotation was carried out on protein encoding genes predicted on prokaryotic and viral genomes using the `eggno-mapper` server [55]. The completeness of KEGG modules in each microbial genome was calculated with KEMET [56]. Furthermore, ORFs annotated with KEGG orthologs belonging to putative alternatives to the Wood-Ljungdahl pathway (WLP) were counted in each MAG to identify potential syntrophic acetate oxidising bacteria. Proteins involved in carbohydrate hydrolysis were searched against the `dbCAN` database [57] using `hmmsearch` and annotated. Proteins were also analysed with `gutSMASH` [58], in order to find gene clusters related to VFA production and metabolism. Fisher's exact test, implemented in SciPy v1.3.1, was

employed in order to assess whether the occurrence of the GH33 enzymatic family was significantly higher in viral genomes than in microbial genomes.

Detection of induction in integrated prophages

With the aim of evaluating the induction of putative integrated phages, an analysis was carried out on the eight samples analysed in this work and extended to 110 samples of the AD database [2]. The aim of the analysis was to check whether prophages detected in the induction experiment reported in this study were also present (and, possibly, induced) in other, unrelated AD communities. MAGs featuring integrated viruses were split into viral and nonviral sequences by extracting the viral sequence predicted within MAGs. This approach resulted in a dataset of 64 prokaryotic MAGs and 64 corresponding integrated viral sequences. Ten million reads were randomly extracted from fastq files and mapped on the database generated from the extracted prophages and MAGs using Bowtie 2. Genome coverage of MAGs and prophages in the different samples was calculated with CheckM coverage. One coverage value was obtained for each contig. Coverage values for each genome were obtained by averaging the values of individual contigs. The coverage threshold for a species to be considered was set to 0.01. The virus/MAG coverage ratios were calculated and the distributions of their values across samples and across genomes were inspected. Finally, log ratios were clustered with an average linkage algorithm based on Euclidean distance. Prophages were considered putatively induced in a sample when the log ratio was greater than 10. This threshold was chosen to exclude values resulting from noise, based on the exponential-like distribution which reaches a plateau around the value of 10 (Additional file 1).

Results

Anaerobic digestion perturbation assays

The current study investigated the effect of environmental stresses on viral and microbial composition of AD communities present in laboratory-scale reactors. The experimental plan included 21 different environmental perturbations known to affect the AD microbial community and to stimulate the induction of integrated prophages (Additional file 1) [14, 59–65]. Some of these conditions occur in biogas plants and negatively impact the reactor performance, but they do not completely disrupt the microbial community, and methane production can be recovered if the conditions are removed. Overall, the highest decreases in methane yield were observed with low pH and temperature shifts to 55 °C or 60 °C (Fig. 1). Temperature shifts had a similarly strong effect if combined with high salt concentration and a much

stronger one when paired with oxidation. A moderate temperature increase, on the other hand, registered positive effects on the production of methane, even when combined with other factors. Interestingly, the treatment with mitomycin did not result in a reduction of methanogenesis. In fact, for both concentrations of mitomycin assessed, the methane yield increased. Importantly, the aim was to identify the perturbations where viral induction was more likely to be detectable. A condition for this was that the microbial community had to be perturbed, but not completely disrupted. The most promising batches, either according to data found in literature, or according to the variation in methane yield, were carried over to DNA extraction and sequencing. DNA yields for the viral fraction were often exceptionally low, despite using all the available volume to perform the extraction. In several cases, including the control bottle, the insufficient yield of viral DNA made it impossible to generate sequencing libraries or provided raw reads of low quality that were discarded (data not shown). Ultimately, four conditions were successfully sequenced and analysed. All of them are characterised by a moderate decrease in the methane yield (up to -30%) with respect to the control bottle. These were organic overload of 8 g VS/L, exposure to atmospheric air at a concentration of 15 mL O₂/g VS, pH increase to the value of 8.5, and exposure to H₂O₂ at a concentration of 3 mM. The fourth condition, which obviously does not occur in biogas plants, was set up in order to mimic a strong oxidative shock, possibly happening during a massive oxygen influx in the system.

Viral and microbial community

Metagenomic analysis allowed the recovery of 1,092 viral genomes (virMAGs). A parallel binning approach recovered 120 microbial MAGs, 72 of them being of high quality according to MIMAG standards. It is reported in the literature that about half of known bacterial genomes feature integrated prophages in their sequence [66]; similarly, 64 of the 120 MAGs identified in this study harbour prophages (Additional file 3).

The normalised relative abundance of the viral component is very high, reaching almost 70% of the total community (viral + microbial) in some samples (Fig. 2). Viral MAGs and single-scaffolds viral genomes ranged in size from 1502 to 195,329 bp, covering the vast majority of the sequence lengths space occupied by prokaryotic viruses with the exclusion of jumbophages [67] (Additional file 3). This was done by design, as the three viral bins longer than 200 kpb were divided into single contigs (Additional file 1). Furthermore, CheckV showed that these bins featured a low completeness and high contamination, justifying the approach taken in their regard. The combined approach of viral prediction and binning

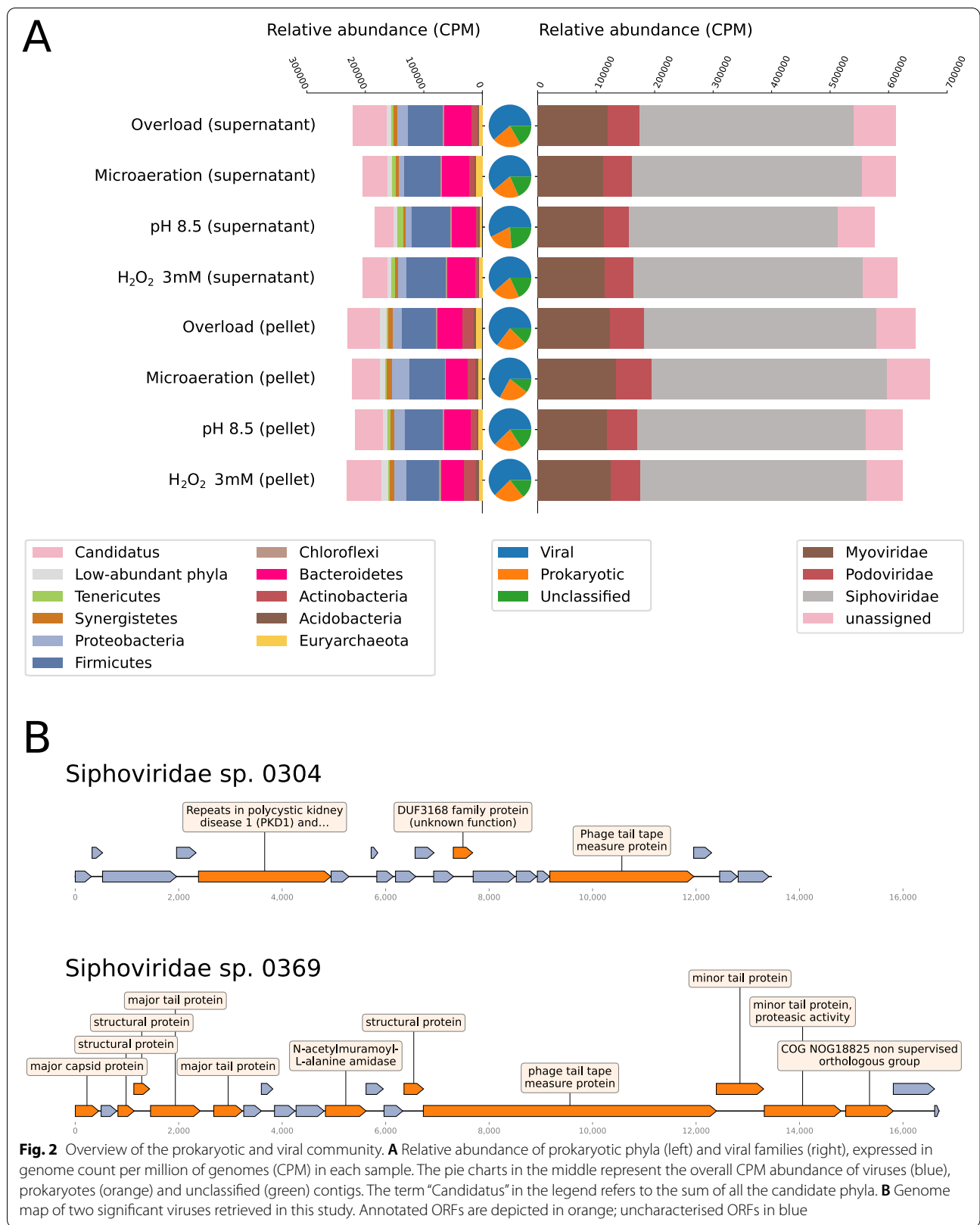
yielded a total of 16 high-quality and five complete viral MAGs according to MIUViG standards [68]. Other metaviromic studies report similar numbers of high quality viral genomes per dataset [9, 21, 69].

The DNA virome is largely dominated by *Caudovirales* phages belonging to the *Myoviridae*, *Podoviridae* and *Siphoviridae* families, confirming similar results observed in previous studies investigating AD viromes [9, 11]. A few small contigs (<5 kb) were classified as *Inoviridae* or *Microviridae*. This is consistent with the fact that species within these taxa tend to have very small genomes, in the order of kilobases. As they are short and underrepresented in databases, it is difficult both to detect them in metagenomes and to estimate completeness and contamination. Among the viral sequences investigated (considering all viral MAGs and prophages integrated in the prokaryotic MAGs), 4.5% were unequivocally assigned at species level. Furthermore, the vast majority (88%) of the viral genomes were assigned at family level.

The distribution of relative abundances among viral genomes is very skewed, with a few prominent viruses (including *Siphoviridae* sp. 0304, *Siphoviridae* sp. 0142, *Virus* sp. 0026, *Siphoviridae* sp. 0307; Additional file 1: Figure 3) and the majority having very low values. In particular, *Siphoviridae* sp. 0304 (Fig. 2B) is by far the most abundant, with a relative abundance of about 5.3% on average across samples.

Single-scaffold phage genome *Siphoviridae* sp. 0431, while not being as abundant as the virMAGs previously mentioned, is a prominent component of the viral community, with an average abundance of 5400 CPM (0.54%) across samples, which spikes at 13,009 CPM in the supernatant part of the sample subjected to alkaline condition.

The prokaryotic composition is consistent with the results of previous works describing anaerobic digestion communities [70, 71]. *Firmicutes* was the most abundant bacterial phylum (25–30% in relative abundance) followed by *Candidatus Cloacimonetes* (20%), *Bacteroidetes* (18%) and *Proteobacteria* (9.4%) (Additional file 3). Four of the 120 prokaryotic MAGs were classified as *Archaea*, accounting for 3 to 4% of the microbiome (Additional file 3). This set is represented by: *Methanoculleus* sp. 0064, *Methanotherix* sp. 0024, *Methanosarcina flavescens* 0114 and *Methanosarcina mazei* 0049. *Methanoculleus* species perform hydrogenotrophic methanogenesis [72, 73], archaea of *Methanotherix* genus perform acetoclastic methanogenesis, while the *Methanosarcina* are generalists [74]. The two *Methanosarcina* identified in this work harbour integrated viral sequences, belonging to the *Siphoviridae* family. The two genomes *Candidatus Cloacimonetes* spp. 057 and 073 are highly abundant, accounting for



a substantial percentage of the bacterial community, from 16% in the pH8.5 supernatant sample up to 23% in the organic overload supernatant sample. Other examples of AD communities dominated by members of *Candidatus* Cloacimonetes, not included in the AD database, have been described in literature [75–77]. Members of this phylum have been suggested as glycolytic in previous studies [78].

Effect of tested conditions on the metagenome

The effect of the different treatments on MAGs and viruses was evaluated by calculating the log ratio of the relative abundances with respect to the average value (see the “Materials and methods” section). The same analysis was performed by comparing the treated samples with the inoculum, the results of which are reported in Additional file 1: Figure 4 due to the marked differences in the microbial profiles existing between the control and the other conditions. A hierarchical clustering performed on the most abundant microorganisms and viruses highlighted groups of species with similar behaviours (Fig. 3). Correlation values between all genomes were calculated considering the compositional nature of the data and are reported in Additional file 3.

Groups of phages with very definite behaviours emerged from the hierarchical clustering (Fig. 3). These groups are heterogeneous in terms of viral taxonomies, yet certain genes are characteristic of each group. The 50 most abundant viral genomes match against a total of 371 HMM profiles of the pVOG database. Of these, 266 (72%) are cluster-specific. It is evident that basic pH and overload tend to have opposite effects on phages (Spearman's $\rho = -0.28$, $p < 0.05$), particularly in clusters “2” and “4”. The former is characterised by eight phages which decrease in relative abundance during overload treatment (log ratios between -0.5 and -1 , see Additional file 3). Three of these genomes have an increase in relative abundance only when exposed to basic pH, with *Siphoviridae* sp. 0431 and *Virus* sp. 0283 almost doubling. It should be noted that increased abundance could have different biological explanations including phage induction and increased abundance of the host. In the latter cluster, the opposite effect can be seen, and in particular two genomes (*Virus* sp. 0026 and *Virus* sp. 0212) are those most heavily affected by basic pH (log ratios -0.68 and -0.59 , respectively) and overload (log ratio 0.64 and 0.66 , respectively). Viruses of cluster “3”, similarly, increase in abundance during organic overload and decrease when exposed to basic pH, but less markedly (average log ratios 0.20 and -0.31 , respectively). This cluster includes *Siphoviridae* sp. 0307 and *Siphoviridae* sp. 0142, which are the second and third most abundant viruses present in the dataset. The *Yersinia* outer protein

gene (*yopX*), which is likely to be involved in life cycle regulation in temperate bacteriophages [79], is only present in genomes belonging to cluster “3”, chiefly *Siphoviridae* sp. 0142, contributing to the hypothesis that these viruses are temperate. The main dominant viral genome, *Siphoviridae* sp. 0304, is part of cluster “1”, entirely composed of *Siphoviridae* and the only cluster of genomes whose relative abundance increases during microaeration (log ratios from 0.11 to 0.31). Cluster “1” is the only one where matches against pVOGs VOG3653, VOG3654, and VOG9328 were found, all of which are annotated as tail proteins. The ORFs matching against these HMM profiles are always found in a duo: one matches with VOG3653, the other both with VOG3654 and VOG9328 (Fig. 2B). This result suggests that the proteins encoded by these two genes are complementary tail components. Overall, basic pH and overload affect phages the most. Basic pH is the condition with the most negative influence on the viruses considered, while overload is mostly associated with positive log ratio values. Unexpectedly, exposure to H_2O_2 is not responsible for great variations in relative abundance of phages. Considering the 50 most abundant phages, most log ratio values under hydrogen peroxide exposition fall between -0.1 and 0.1 , and only a small set of viruses shows a mild increase, including the aforementioned *Siphoviridae* sp. 0096 and *Siphoviridae* sp. 0307.

Microaeration and exposure to hydrogen peroxide have strong effects on MAGs and seem to have opposite effects on the whole microbial community, displaying a slight anticorrelation (Spearman's $\rho = -0.40$, $p < 0.005$). Basic pH is another condition that predominantly affects the microbial community. Genomes in cluster “A” (Fig. 3), sharply increase in relative abundance under basic pH and decrease under microaeration. This cluster shows a relative lack of integrated proviruses in the MAGs' genomes: only two MAGs out of 14 include viral sequences (14%), whereas in the entire dataset, 64 MAGs out of 120 do (53%). Contrariwise, in cluster “B” 6 out of 8 MAGs carry integrated prophages. However, the small number of genomes made it difficult to draw statistically sound conclusions regarding these differences.

The behaviour of the three archaeal MAGs included in the analysis is peculiar and, although they end up in distant places in the clustering (Fig. 3), they have common characteristics. First, they do not cluster with bacterial genomes, but rather exhibit a unique behaviour. Secondly, their relative abundance varies markedly (coefficients of variation between 0.34 and 0.46 , higher than 40 MAGs out of the 50 included in the heatmap), evidencing a marked response to the changes in the experimental conditions. For example, *M. flavescens* 114 and *Methanotherix*

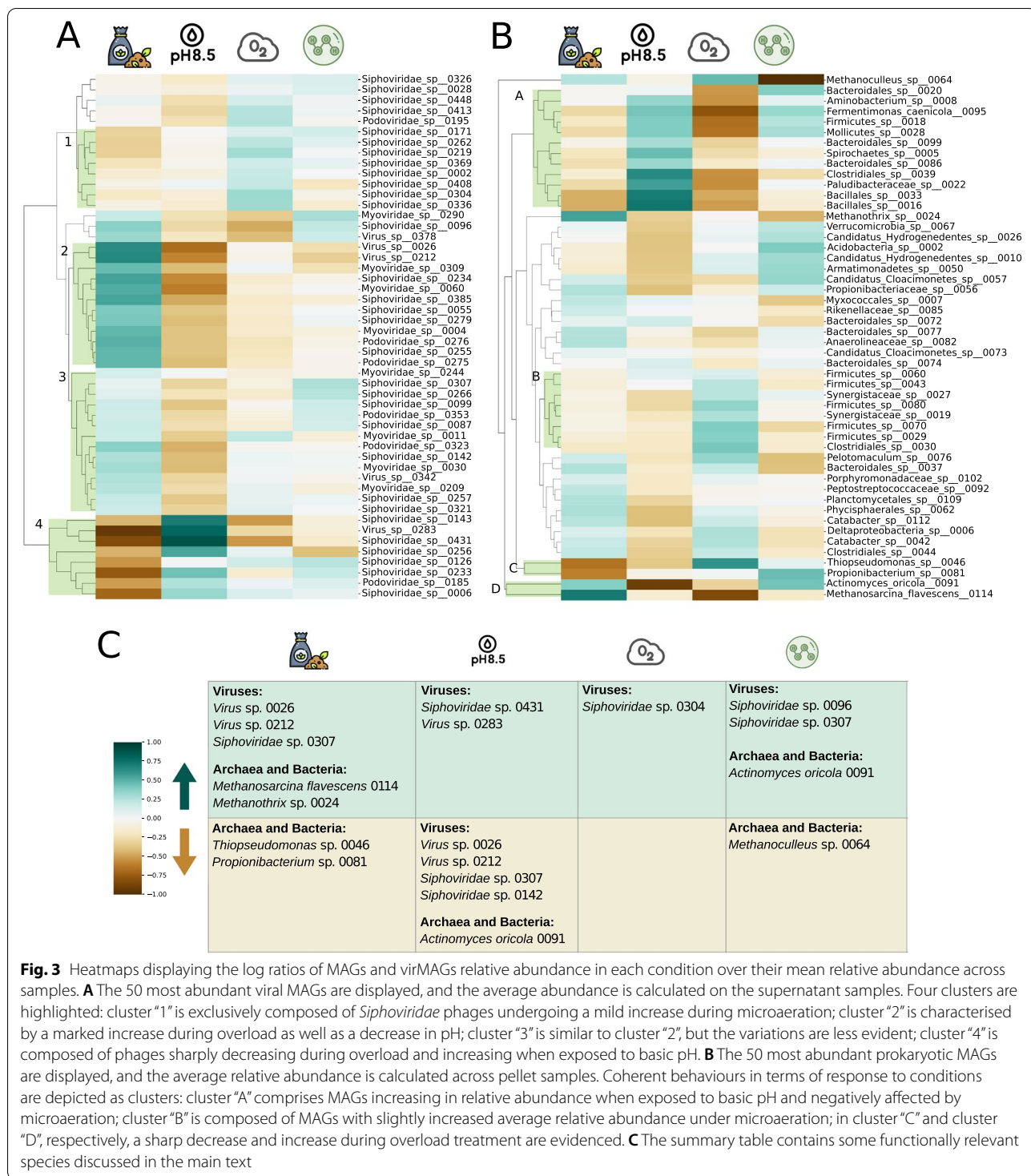


Fig. 3 Heatmaps displaying the log ratios of MAGs and virMAGs relative abundance in each condition over their mean relative abundance across samples. **A** The 50 most abundant viral MAGs are displayed, and the average abundance is calculated on the supernatant samples. Four clusters are highlighted: cluster “1” is exclusively composed of *Siphoviridae* phages undergoing a mild increase during microaeration; cluster “2” is characterised by a marked increase during overload as well as a decrease in pH; cluster “3” is similar to cluster “2”, but the variations are less evident; cluster “4” is composed of phages sharply decreasing during overload and increasing when exposed to basic pH. **B** The 50 most abundant prokaryotic MAGs are displayed, and the average relative abundance is calculated across pellet samples. Coherent behaviours in terms of response to conditions are depicted as clusters: cluster “A” comprises MAGs increasing in relative abundance when exposed to basic pH and negatively affected by microaeration; cluster “B” is composed of MAGs with slightly increased average relative abundance under microaeration; in cluster “C” and cluster “D”, respectively, a sharp decrease and increase during overload treatment are evidenced. **C** The summary table contains some functionally relevant species discussed in the main text

sp. 24 show a marked increase under organic overload, with log ratios of 0.71 and 0.56, respectively.

Another common trend of archaeal MAG abundance in relation to shocks is a decrease in response to H₂O₂ exposure. This is coherent with biochemical measurements,

as the sample is characterised by the highest concentration of acetate and VFA, and a substantially lower methane yield (Additional file 2, Fig. 1). A kinetic imbalance between acid producers and consumers, reflected by low methane production, is revealed by VFA accumulation

[62, 80]. This sample shows the highest concentration of VFA, including acetate. This is coherent with the idea that a halt in the activity of acetoclastic archaea leads to an accumulation of acetate and a decrease in methane production.

Two isolated small clusters dubbed “C” and “D”, comprising two MAGs each, comprise species which display marked variations in relative abundance among different conditions. Cluster “C” comprises *Thiopseudomonas* sp. 046 and *Propionibacterium* sp. 081, both decreasing with organic overload, but the former sharply increased with air. The *Thiopseudomonas* genus has been described as a facultative anaerobe, catalase- and oxidase-positive [81]. Furthermore, five ORFs in the two *Thiopseudomonas* MAGs are functionally related to oxidative stress (EC numbers 1.11.1.6, 1.11.1.15, and 1.8.1.9, Additional file 4), which could explain a possible air tolerance. Cluster “D” comprises *Actinomyces oricola* 091 and *M. flavescens* 114. The bacterium is negatively affected by basic pH, with its relative abundance almost halved with respect to the mean (mean abundance 6,631 CPM; abundance in basic pH 3541 CPM; log ratio -0.90). Its relative abundance also suggests an increase during H_2O_2 exposure.

Finally, *Pseudomonas formosensis* 084 and *Thiopseudomonas* sp. 083 display extreme variation between conditions. Their relative abundance is exceptionally low under overload (log ratios -4.8 and -2.2 respectively) which is their most striking characteristic. Furthermore, both revealed a higher relative abundance under microaeration, and a lower relative abundance with H_2O_2 in comparison to the average value across conditions.

Functional categories of proteins encoded in MAGs and viral MAGs

The tested conditions appear to have an important effect in shaping the structure of both the viral and microbial communities. Functional annotation was employed to investigate the link between the variation in community composition and the tested conditions. The AD process is carried out by a multitude of microorganisms, each one playing a number of roles in the degradation and conversion of organic matter [82]. A starting point for the metabolic characterisation is the functional annotation of genes. Analysis of the functional categories was performed on protein-coding genes identified in all predicted viruses and microbial genomes. Notably, 70% of the ORFs encoded by prokaryotic genomes registered a match in the KEGG Orthology database, while this percentage is as low as 30% in viral genomes. Viruses, although they do not perform metabolic activities in the community, can influence and modulate microbial functionality via infection, induction, and HGT. It was recently reported that, in the design of synthetic microbial communities, it is

of utmost importance to determine the absence of integrated, putative inducible prophages, to ensure the stability of the process [83, 84]. According to this approach, the presence of prophages was verified, in order to determine the putative level of vulnerability of specific steps of the AD process. AD is divided into four main steps: hydrolysis, acidogenesis, acetogenesis, and methanogenesis [85]. In the hydrolysis step, complex organic molecules are broken down to their monomers, which are then converted into VFAs by the guild of acidogenic bacteria. VFAs are then employed by the acetogenic species in the production of acetate, hydrogen and CO_2 , upon which the methanogenic archaea feed, producing methane. In an attempt to categorise the MAGs according to their role in the AD process, particular attention was paid to gene categories that are important in each of these steps. These gene categories were grouped as follows: (I) genes involved in binding and degradation of polysaccharides, especially cellulose-related protein families; (II) genes related to VFA production or metabolism; (III) genes pertaining to the Wood-Ljungdahl pathway (WLP) or pathways proposed as alternatives and potentially involved in syntrophic acetate oxidation; and (IV) genes involved in methanogenesis (Additional file 4).

Viruses have a double role in the funnel-shaped web of interaction of the AD. On one hand, they may carry genes conferring additional enzymatic functions to their hosts; on the other hand, they represent a threat to the host, as their lifestyle often involves the hijacking of the host metabolism and its death. For this reason, genes of interest were searched in the prophages integrated in the aforementioned MAGs, both with a positive and a negative impact on the host metabolism. Free viral genomes were also investigated as they could represent temperate viruses.

The MAGs encoding the largest number of ORFs annotated with carbohydrate-binding functions (guild I) belong to the candidate phyla Cloacimonetes and Hydrogenedentes. The two *Candidatus* Cloacimonetes genomes are characterised by the occurrence of ORFs annotated as CBM56 by dbCAN (Fig. 4A). This enzymatic family is associated with a beta-1-3-glucan binding function; hence, it is probably involved in favouring the binding of *Bacteria* to cellulose substrates. A glycolytic role has been suggested for *Candidatus* Cloacimonetes bacteria in a previous study [78]. Moreover, *Candidatus* Hydrogenedentes sp. 010 and *Candidatus* Hydrogenedentes sp. 026 feature, respectively, 13 and 27 genes annotated as dockerins, i.e., proteins that take part in the formation of cellulosomes.

Genes encoding enzymes related to carbohydrate synthesis, degradation and binding were found in 18 out of 64 integrated viral genomes. Three enzymatic families

Table 1 *Bacteria* related to VFA metabolism. The table reports the phylum of belonging, the family of integrated prophages, the completeness of the beta-oxidation KEGG module (M00087) and the presence of relevant metabolic gene clusters detected using the gutSMASH software

Genome ID	Phylum	Prophage	M00087	Metabolic gene clusters
<i>Deltaproteobacteria</i> sp. 006	<i>Proteobacteria</i>	<i>Myoviridae</i>	Complete	3 acetate to butyrate; 1 acetyl-CoA pathway
<i>Myxococcales</i> sp. 007	<i>Proteobacteria</i>	Absent	Complete	Absent
<i>Pseudomonas formosensis</i> 084	<i>Proteobacteria</i>	<i>Myoviridae</i>	Complete	Absent
<i>Paracandidimonas</i> sp. 097	<i>Proteobacteria</i>	<i>Siphoviridae</i>	Complete	3 Acetate to butyrate
<i>Bacillus ginsengihumi</i> 031	<i>Firmicutes</i>	<i>Siphoviridae</i>	Complete	3 Acetate to butyrate
<i>Ruminococcaceae</i> sp. 120	<i>Firmicutes</i>	Absent	Incomplete	1 Pyruvate to acetate-formate; 3 acetate to butyrate

mobile genetic elements can provide novel metabolic functionalities to their hosts [86].

GH33, instead, is a family of sialidases and neuraminidases, and is represented by 161 ORFs among prokaryotic sequences, which ranks it as the twelfth most represented. It is thus more frequent in viruses than in prokaryotic sequences ($p = 10^{-22}$, Fisher's exact test), hinting at their importance in the viral physiology (Fig. 4C).

The presence of genes related to VFA metabolism and acetate formation (guild II) was evaluated by checking the completeness of the beta-oxidation module and the presence of a selection of clusters of genes [58]. The beta-oxidation module (M00087) was complete in five genomes (Additional file 4). Three of these genomes, as well as *Ruminococcaceae* sp. 120, encode enzymes involved in acetate, pyruvate and butyrate metabolism (Table 1). In particular, the genome of *Deltaproteobacteria* sp. 006 features the complete M00087 beta-oxidation module, genes belonging to the "Acetate to butyrate" and "Acetyl-CoA pathway" metabolisms and an integrated prophage of the *Myoviridae* family (Table 1).

Bacteria potentially involved in syntrophic acetate oxidation (guild III) were identified by evaluating the presence of genes belonging to the Wood-Ljungdahl pathway (WLP) or its putative alternatives, the glycine synthase-reductase pathway (GSRP), and the reductive glycine pathway (RGP). In this dataset, these alternative WLP modules seem to be exclusive of *Firmicutes* (Additional file 4). Eleven MAGs comprise genes belonging to oxidative pathways, ten of which belong to *Firmicutes* (e.g. *Firmicutes* sp. 0060 and *Catabacter* sp. 0112) and one to *Chloroflexi* (*Anaerolineaceae* sp. 0082). *Firmicutes* bacteria have already been reported as capable of converting acetate to CO₂ through the reverse WL pathway [87, 88]. Furthermore, gene annotations in previous works [89] have already identified bacteria from the *Chloroflexi* phylum as potential

syntrophic acetate oxidising bacteria. In *Firmicutes* sp. 0060, both the glycine cleavage system and the GSRP were fully complete, while the RGP was 83% complete. A noteworthy viral genome recovered is *Siphoviridae* sp. 0243, which is an 84-kb phage with an estimated completeness between 80 and 100% and harbours a section of the glycine synthase-reductase pathway (GSRP). This suggests that this virus can confer additional enzymatic capabilities to its host, giving it an alternative to the Wood-Ljungdahl pathway.

Methanogenesis (guild IV) is exclusively carried out by methanogenic archaea. This guild is represented by a heterogeneous population of hydrogenotrophic (*Methanoculleus* sp. 0064), acetoclastic (*Methanotrix* sp. 0024) and generalist methanogens (*M. flavescens* 0114; *M. mazei* 49). The two MAGs assigned to *Methanosarcina* genus have integrated *Siphoviridae* proviruses.

Evaluation of selected prophages abundance in the AD database

Given the wide presence and importance of integrated proviruses across all metabolic guilds, the search was broadened by considering 123 additional shotgun sequencing experiments deposited in public AD databases [2]. Many of these experiments investigated AD reactors operating under a variety of parameters including temperatures ranging from 35 to 55 °C, different feedstocks and stressful conditions such as lipids overload [90] or high concentration of ammonia [91]. For each MAG featuring integrated proviruses, the viral/host ratio was defined as the ratio between the read coverage of the viral and the non-viral components. This proportion was calculated separately for each experiment, with the aim of showing whether a specific prophage increases in abundance with respect to the host under specific environmental conditions.

Microbial and viral diversity across the AD database

In the samples from the present study, the average prophage/host ratio was equal to 1.1 and the maximum value was 13.3. Contrariwise, considering all experiments from the AD database, the average ratio rises to 12.0 and the maximum is over 4200 (Additional file 5). These data reflect the diversity of environmental conditions across the AD database and underline their importance in shaping the microbial and viral community. The read coverage across the whole dataset shows that viruses and hosts are not always present in the same community. There are 64 MAGs featuring integrated proviruses, on which reads from the additional experiments were mapped, resulting in 7872 values of coverage ratio. In 2722 cases (34.6%) the integrated virus was not found, despite the presence of the host (Additional file 5). The opposite is much rarer: only in 22 cases (0.3%) reads map only onto the viral part of the MAG. These cases may be explained by the ability of the viral species itself or related strains to infect a different host [92]. In most cases (1221 occurrences, 15.5%) where the host microbe is not present, the bacteriophage is not present either.

Effects of temperature

Temperature is the strongest driver of clustering: most of the mesophilic samples end up in four sub-clusters of respectively 29, 6, 9 and 8 sequencing experiments (Fig. 5). One large cluster is entirely composed of thermophilic samples. Here, 73–98% of MAGs are still present, although their coverage is often lower than those registered in our samples. However, only 26–53% of the respective phages are present and the prophage/host ratio is 7.74 on average.

In order to evaluate the impact of the temperature in the composition of the viral community, the coverage of each provirus in mesophilic samples was compared with the coverage in thermophilic samples with a Mann-Whitney *U* test. Out of 64 proviruses, 54 (84%) are more abundant in the mesophilic samples, and this finding is expected because this study was performed on mesophilic reactors. However, four outliers were found to be more prevalent in thermophilic groups (*p*-value threshold = 0.05, Table 2), and six do not show significant differences.

Impact of simplified feedstocks

A 26-sample cluster named “simplified medium” is characterised by the predominance of samples in which the

feedstocks have a controlled or restricted composition: out of 26 reactors, 10 use as feedstock BA medium mixed with simple components as carbon sources, such as acetate, glucose, avicel, and VFA mixtures; 6 samples were fed with cheese whey 6 with acetate as the major substrate (Fig. 5). The peculiarity of this cluster is that they have a low number of MAGs and phages in common with our samples. On average, half of the MAGs identified in this study are present in this cluster, with two samples having as few as 6 and 9 MAGs. This is also due to the smaller number of taxa identified in the community present in these samples, which amplifies the difference with the samples investigated in this study. The integrated prophages are detected on average for 10% of MAGs, with only one sample having more than 20%.

The most ubiquitous MAG-prophage couples are *Clostridiales Family XIII Incertae Sedis* sp. 013, *Firmicutes* sp. 043 and *Bacteroidales* sp. 047, with their respective prophages. These MAG-prophage couples are found in association in over 90% of the samples.

The three least frequent MAG-prophage couples are *Bacillales* sp. 117, *Bacillales* sp. 090 and *Candidatus Cloacimonetes* sp. 057, all of them predominantly present at mesophilic conditions. Despite the prevalence of *Candidatus Cloacimonetes* sp. 057 in the data gathered in this study as mentioned earlier, this MAG is exclusively found in studies belonging to the group “meso 1”, characterised by mesophilic temperatures and cattle manure as feedstock (Fig. 5).

Behaviour of proviruses

Putatively induced phages in the different AD metagenomes have been estimated by using as a proxy the ratio between the coverage of the provirus compared with the rest of the MAG. A provirus has been considered induced in a certain condition when the ratio was larger than 10, based on the cumulative distribution of the viral/host abundance ratios (Additional file 1: Figure 5). However, the rise in abundance of a virus with respect to its host could be explained by factors different from induction, such as the presence of an alternative host or a low number of reads mapped on both genomes, and this bias should be taken into account when discussing this aspect.

In the “simplified medium” cluster, 13 out of the 24 MAG-virus couples have an average ratio indicating an induction. Most of the MAGs belong to the *Bacteroidales*

(See figure on next page.)

Fig. 5 Coverage of the MAGs with integrated phages in the 8 samples from the present experiment and in 110 samples from the AD database. The colour scale of coverage is logarithmic. The sample names are reported in Additional file 5 in the order in which they appear in the heatmap. MAGs are identified by the same sequential numbers that appear in their complete name (e.g., “*Firmicutes* sp. 043” is here “43 M”). “M” denotes the prokaryotic fraction of the MAGs, while “V” indicates the integrated viral part. Feedstock and temperature range are displayed as coloured labels at the top of the heatmap. Samples are grouped into clusters based on Euclidean distance. Clusters that are relevant to our analysis are highlighted at the top of the image with pastel colours

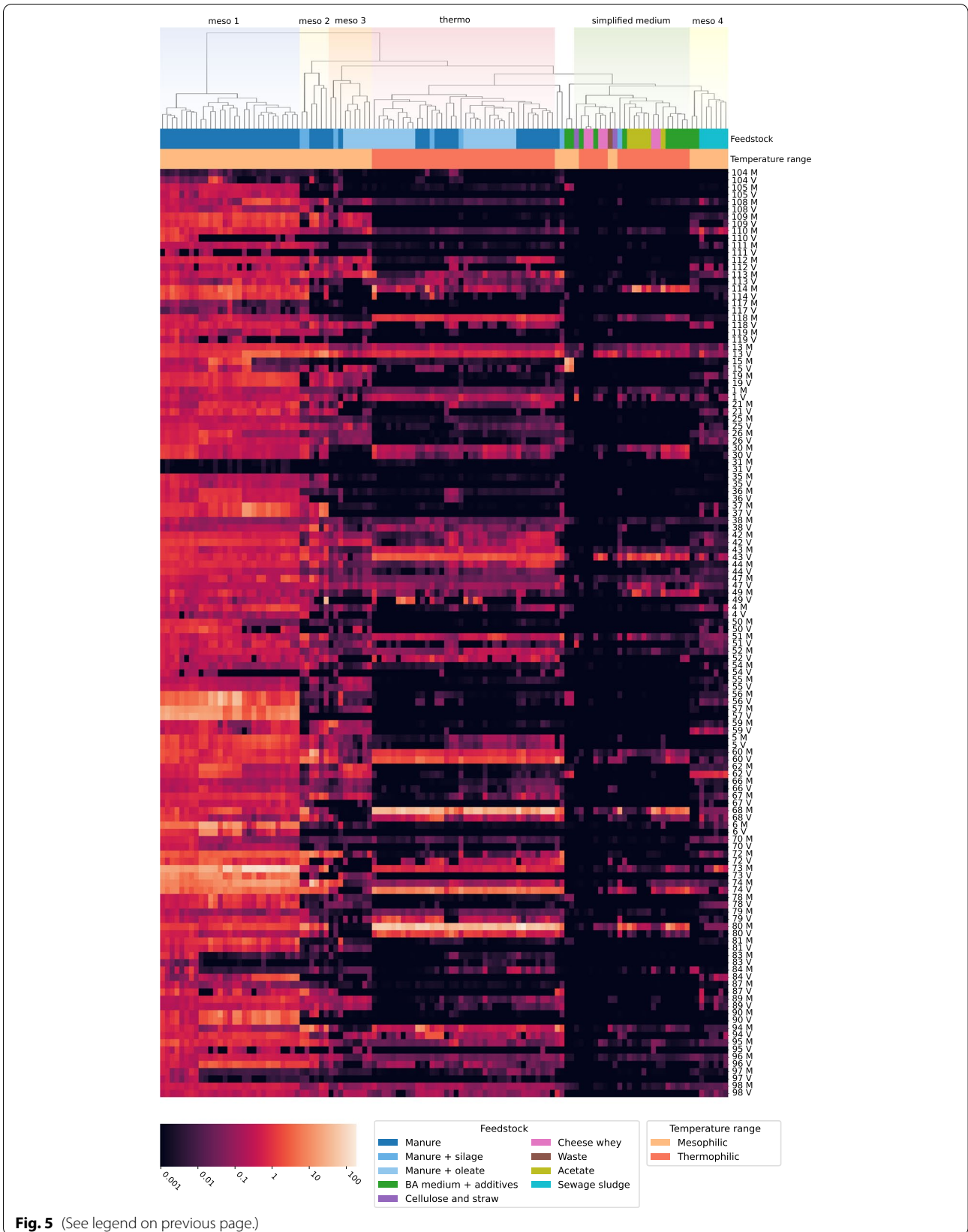


Fig. 5 (See legend on previous page.)

Table 2 Proviruses which score a higher read coverage in thermophilic samples, as confirmed by Mann-Whitney *U* tests

Provirus name	Taxonomy	<i>p</i> -value
provirus <i>Firmicutes</i> sp. 043	<i>Siphoviridae</i> sp.	1.5e−11
provirus <i>Firmicutes</i> sp. 060	<i>Siphoviridae</i> sp.	4.5e−2
provirus <i>Bacteroidales</i> sp. 074	<i>Myoviridae</i> sp.	2.9e−2
provirus <i>Firmicutes</i> sp. 080	<i>Siphoviridae</i> sp.	3.5e−05

and *Clostridiales* orders, with the exceptions of three members of *Firmicutes*, *Acholeplasmatales* sp. 079 and *Candidatus* Hydrogenedentes sp. 066. The latter is hypothesised to play a role in the hydrolysis of cellulose, and, as such, it is found as induced in a sample fed with cellulose and straw (virus/host coverage ratio = 94).

Provirus *Firmicutes* sp. 043, which infects one of the most ubiquitous MAGs, has an average coverage ratio of 493 in the “simplified medium” cluster and of 68 in the “meso 4” cluster. Provirus *Peptococcaceae* sp. 118 has an average ratio of 165 and 368 in the “meso 3” and “meso 4” clusters, respectively. Provirus *Pseudomonas formosensis* 084, whose host participates in the degradation of fatty acids via beta-oxidation, is detected as induced in the “meso 4” cluster, where it is present in five out of eight samples, and its coverage ratios range from 27 to 2,779. This cluster is characterised by mesophilic temperatures and most of the reactors in it are fed with sewage sludge.

Provirus *M. mazei* 049 scores a read coverage more than 10 times higher than its host in 9 samples belonging to the “thermo” cluster. In one sample, the virus has a coverage of 37, whereas the host only has 0.05. This is most likely a case in which the virus itself is present but infecting a different host.

Overall, the variety of environmental conditions under which AD occurs provides a range of opportunities to explore the interactions between viruses and their hosts, revealing large-scale trends which would otherwise be difficult to detect.

Discussion

In this study, the virome of AD communities undergoing several prophage-inducing stresses was investigated. As expected, much of the viral diversity hereby explored is novel, as shown by the challenges presented by functional annotation and taxonomic assignment. The DNA viral community is dominated by tailed bacteriophages belonging to *Siphoviridae*, *Podoviridae* and *Myoviridae* families. Members of single-strand DNA families such as *Microviridae* and *Inoviridae* were retrieved, as well as *Bicaudaviridae*. These families are characterised by small genomes and are underrepresented in sequence databases, which makes them more challenging to

detect [93]. This lack of representation means that the sequences retrieved in this study are going to contribute to the knowledge expansion about environmental viruses which has been going on for well over a decade, with no sign of decrease yet [22].

Viral particles have a different structure than organisms, and lack a metabolism [94], whereas living beings can rely on homeostatic mechanisms to face changes in external factors such as pH changes and oxidative stress [95, 96]. Temperate viruses, furthermore, are induced and enter the lytic cycle as a reaction to some DNA-damaging stresses. These factors are effective when looking at the reactions of both viruses and prokaryotes to the different conditions that were applied in the experimental setup. Viruses show a clear dichotomy between organic overload and basic pH (Spearman’s $\rho = -0.69$, $p < 0.001$), while prokaryotes show the greatest differences between basic pH and microaeration (Spearman’s $\rho = -0.40$, $p < 0.005$).

This said, there are characteristic responses to conditions, as shown by the log ratio clustering (Fig. 3). Viruses show these trends very clearly, and these can be linked to the presence of specific genes in the clusters. For instance, cluster “3” comprises viral genomes which include the *yopX* gene. This gene is known to be involved in the regulation of the life cycle of temperate bacteriophages [79], thus suggesting that members of this group are temperate, and *yopX* is involved in their induction. Among the members of cluster “3” is *Siphoviridae* sp. 0142, one of the most abundant viral genomes hereby retrieved, which encodes this gene.

The relative abundance of viruses is little affected by the action of H₂O₂. The literature regarding the effect of H₂O₂ on bacteriophages is scarce and focuses on the effects of H₂O₂ vapour, but it appears that non-enveloped viruses, as tailed bacteriophages, are more resistant to oxidation than enveloped viruses. The same studies show that the presence of a complex medium, in this case cattle manure, is able to shield the viral particles from the effect of the peroxide, either by acting as a physical barrier or by reacting with the oxidative agent, thereby diminishing its concentration [97].

Microorganisms, as well, respond differently to atmospheric air and H₂O₂. Both conditions are supposed to put microorganisms in a state of oxidative stress; however, H₂O₂ is a much stronger oxidising agent and, as consequence, several species of *Bacteria* and *Archaea* are less able to face and survive the damage. Since *Archaea* are anaerobes and at best oxytolerant, they are heavily affected by strong oxidative stress (Additional file 3).

It can be striking to observe that archaeal species, which are so reactive to oxidative stresses, increase during organic overload. Organic overload is associated with

an accumulation of VFAs and, consequently, the decrease of pH. This leads to inhibition of methanogenesis; however, this process takes several days to unfold, and the brief time span of this experiment could not allow it. Measurements of methane yield and VFA, particularly acetate concentration, are consistent with this explanation (Additional file 2, Fig. 1).

Some bacterial species decrease steadily in the presence of atmospheric air, the majority of which do not feature integrated proviruses. Accordingly, it can be speculated that bacteria more resistant to oxidative stress caused by microaeration are more likely to have integrated prophages. In fact, both *T. denitrificans* and the *Pseudomonas* genus are known as facultatively anaerobic [81, 98]: this clearly coincides with their ability to withstand a moderate oxidative stress as microaeration, but not a stronger one as the injection of H₂O₂. Both genomes have integrated phages, belonging to the *Myoviridae* and *Siphoviridae* families, respectively.

While 70% of the prokaryotic genes were annotated, this proportion drops to 30% in viral genomes, a proportion consistent with typical metaviromic studies [21, 69]. This striking discrepancy in terms of unclassified genes reiterates how vast the proportion of unknown genes in the viral world is. Yet, the annotated genes reveal the important role viruses play in shaping the microbial community of AD. Mobile Genetic Elements, viruses included, often contribute to the metabolic capability of their hosts by carrying genes conferring evolutionary advantage to the host [99], and the same can be observed in this community.

As mentioned before, viruses code for a considerable number of genes belonging to GT4 and GT2 families. These are enzymatic families of importance for biofilm synthesis, and it is known that in some bacterial species the lysogenic infection of a temperate phage increases the production of biofilm, benefitting both host and virus [100]. If such a phenomenon is confirmed to occur in AD environments, it will be reasonable to ponder the role of proviruses in the spreading of these genes.

The reductive acetyl-CoA pathway, also known as the Wood-Ljungdahl pathway, is a metabolic pathway characteristic of homoacetogenic bacteria and archaea. It allows the fixation of CO₂ and synthesis of acetate, which is then used by acetoclastic archaea to produce methane. Other species, known as syntrophic acetate oxidising bacteria, employ the reverse Wood-Ljungdahl pathway to digest acetate into H₂ and CO₂, both consumed by hydrogenotrophic archaea. Biochemical evidences regarding the activity of the WL pathway have been reported for a limited number of isolated species: the species dubbed as “acetate-oxidising, rod-shaped bacterium” (AOR) [97], *Clostridium ultunense* [101],

Thermoacetogenium phaeum [102], *Pseudothermotoga lettingae* [103], *Syntrophaceticus schinkii* [104] and *Tepidanaerobacter acetatoxydans* [105]. Genes belonging to these pathways were found in genomes retrieved in the present study, more specifically in 10 members of the *Firmicutes* phyla and one *Chloroflexi*. Members of both the *Firmicutes* and *Chloroflexi* phyla bacteria have been reported as either capable of converting acetate to CO₂ through the reverse WL pathway or as potential syntrophic acetate oxidising bacteria based on gene annotations [87–89]. However, alternative WL pathways mediated by the glycine cleavage system and tetrahydrofolate pathway have been proposed in recent studies [2, 106, 107]. Some genes belonging to these pathways are potentially involved in bacteriophage-mediated HGT: *Siphoviridae* sp. 0243 is particularly noteworthy, as its genome includes five genes of the GSRP pathway. The presence of these genes can derive from previous excision. The integration of *Siphoviridae* sp. 0243 in another bacterial genome can, in theory, confer to the host new metabolic capabilities. However, phage-mediated HGT of these genes has never been previously reported and can be targeted in future studies. The genome of *Siphoviridae* sp. 0163, similarly, includes an enzyme of the CBM56 family, involved in the degradation of polysaccharides, and thus could confer this metabolic function to its host via HGT. As a last example, five free viruses and six proviruses code for proteins annotated with the Gene Ontology term GO:0006979, which groups genes mediating oxidative stress response, and thus might increase the host's survivability to oxygen exposure. As a matter of fact, *Clostridiales* sp. 030 and *Synergistaceae* sp. 019, both of which include one of said proviruses, also show a positive log ratio under O₂ exposure.

The results also reveal some of the adaptations these parasites use against their hosts.

It is known that, in bacteriophages parasitizing *E. coli*, tail spikes present sialidases which degrade the host's coat of polysialic acid, allowing the interaction between phage and host [108]. Other depolymerases are known to enact similar processes in other bacteriophages [109]. However, our results indicate not only a presence of sialidases/neuraminidases, but an overrepresentation thereof, with respect to prokaryotic genomes. Hence, it is possible that these enzymes have an important, hitherto overlooked role in the mechanism of infection. In fact, 51% of the ORFs assigned to the GH33, GT2 and GT4 families (47 out of 91) are annotated by eggNOG as tail proteins or tape measure proteins, supporting the idea that these enzymatic activities are especially relevant in phage/host interactions.

Two of the archaeal genomes retrieved in this study, *M. mazei* sp. 049 and *M. flavescens* sp. 114, incorporate

in their sequence integrated proviruses of the Siphoviridae family. While most *Siphoviridae* are bacteriophages, there's evidence that some members of the family infect *Archaea*, including the methanogenic species *Methanoculleus bourgensis* and *Methanobacterium formicicum* [110, 111]. These members of *Siphoviridae* are lytic, i.e., do not integrate in the host of the genome; the newly recovered genomes show the existence of lysogenic archaeal *Siphoviridae* as well.

In order to better understand the dynamics between viruses and hosts, the reads from a large number of AD experiments were mapped on integrated proviruses and host genomes retrieved in this analysis. These experiments widened the exploration of prophage behaviour, allowing the identification of specific environmental conditions favouring prophage induction. Additionally, they provide insights on the presence/absence of each integrated prophage in the MAG across different conditions. The first observation is that, when host and virus are not both present, usually the virus is missing from the experiment. This is consistent with the idea that viruses tend to co-exist with their hosts, and that different communities may consist of different viral species even in the case where the same host is present.

Simplified medium communities are characterised by high viral abundance, as well as low number of reads mapped on the genomes retrieved in the current experiment. These characteristics are easily explained by the growth conditions: the specific nutrient source imposes a strong selection on the microbial species, e.g. the hydrolytic guild in the case of acetate-based media. It is feasible to think that such stresses lead to the induction of integrated proviruses. However, possible biases should be taken into account. A small number of reads mapping both on the host and the provirus may skew the ratio due to stochasticity; another factor which can lead to a high virus/host coverage ratio is the presence of alternative hosts.

Conclusions

The shift to a circular economy and the reduction of greenhouse gas emissions is pressing and requires a massive effort in terms of technology adoption. In this context, AD is a widely used technology; nevertheless, a key component of the process, the virome, is still relatively unknown. To our knowledge, this is the first time in which the viral community of the AD was inspected under a great variety of different conditions. This study reveals the pervasiveness of viruses in the AD microbiome. The data retrieved in this work and the analyses hereby carried out lay the bases towards the understanding of the complex role of the viral community in AD.

Viral genomes featuring genes of relevance in the AD process were retrieved, opening up the possibility that HGT is carried out by viruses. Shocks impacted viruses and microbes in different ways, highlighting four taxonomically heterogeneous clusters of species.

Broadening the analysis to a wide array of AD studies enabled the consideration of the effect of more environmental parameters, such as temperature and medium composition, on the abundance of temperate viruses and their hosts. It also reveals that the viral community is more mutable than the microbial one, as viruses are often not found despite the presence of their hosts, while the opposite is much rarer. More in-depth studies on the microbiomes of samples of the AD database might elucidate if and how metabolic stresses and starvation placed on some microorganisms by the simplified feedstock affect phage induction. Although this study is limited to the analysis of DNA viruses, it can be expanded in the future to include the RNA viral community. In the next future, knowledge about the interactions between viruses and their host will have the potential to improve the efficiency of the AD process and the production of biogas, as it is already done in different environments such as wastewater treatment plants, food surfaces and even the human body. More studies with innovative approaches are needed to understand thoroughly the effects of conditions typical of AD on the lifestyle of the viruses that inhabit this engineered ecosystem. On a shorter timescale, the newly discovered viral genomes contribute to the ever-growing diversity of environmental viruses which is shifting our understanding of these entities.

Abbreviations

AD: Anaerobic digestion; WWTP: Wastewater treatment plants; HGT: Horizontal gene transfer; VFA: Volatile fatty acids; TCD: Thermal conductivity detector; FID: Flame ionisation detector; MAG: Metagenome-assembled genome; ORF: Open reading frame; CPM: Count per million; pVOG: Prokaryotic Virus Orthologous Group; KEGG: Kyoto Encyclopedia of Genes and Genomes; WLP: Wood-Ljungdahl pathway; GSRP: Glycine synthase-reductase pathway; RGP: Reductive glycine pathway.

Supplementary Information

The online version contains supplementary material available at <https://doi.org/10.1186/s40168-022-01316-w>.

Additional file 1: Supplementary methods regarding details of experimental setup, bin curation, calculation of relative abundance and taxonomy assignment. **Figure 1.** UpSet plot of viral sequences predicted by different tools. **Figure 2.** Scatter plots comparing CPM and CheckM relative abundance values. **Figure 3.** Relative abundance expressed in count per million (CPM) of the most abundant prokaryotic and viral genomes. **Figure 4.** Heatmaps displaying variations of phages and MAGs compared to supernatant and pellet control. **Figure 5.** Distribution of the provirus/host CPM abundance ratios across all datasets.

Additional file 2: Table 1. Measurements of VFA and methane yield.

Additional file 3: Table 2. General information about the prokaryotic and viral MAGs retrieved in this study: genome length, completeness, contamination, taxonomy, relative abundances, log ratios with respect to the average.

Additional file 4: Table 3. Functional annotation of MAGs with eggNOG mapper, KEMET, dbCAN, gutSMASH.

Additional file 5: Table 4. Coverage of prophage-host couples in samples from the AD database, results of hierarchical clustering of the samples, linear regressions between the relative abundances of proviruses and hosts.

Authors' contributions

Conceptualisation, A.R., and L.T.; methodology, A.R. and M.S.M.; software, A.R. and M.S.M.; formal analysis, A.R. and M.S.M.; investigation, M.G. and P.K.; data curation, A.R. and M.S.M.; writing—original draft, A.R. and M.S.M.; writing—review and editing, A.B., L.T., and S.C.; visualisation, A.R., M.S.M., and A.B.; supervision, L.T. and S.C.; the authors read and approved the final manuscript.

Funding

This work was financially supported by the "Budget Integrato della Ricerca Dipartimentale" (BIRD198423) PRID 2019 of the Department of Biology of the University of Padua, entitled "SyMMoBio: inspection of Syntrophies with Metabolic Modelling to optimise Biogas Production" and by the project "Sviluppo Catalisi dell'Innovazione nelle Biotecnologie" (MIUR ex D.M.738 dd 08/08/19) of the Consorzio Interuniversitario per le Biotecnologie" (CIB). This project has received funding from the Hellenic Foundation for Research and Innovation (HFRI) and the General Secretariat for Research and Innovation (GSRI), under grant agreement No 580.

Availability of data and materials

Raw reads data are available at NCBI SRA (<https://www.ncbi.nlm.nih.gov/sra>) under the BioProject ID PRJNA767833.

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

¹Department of Biology, University of Padua, via U. Bassi 58/b, 35131 Padova, Italy. ²Department of Hydraulics, Soil Science and Agricultural Engineering, Faculty of Agriculture, Aristotle University of Thessaloniki, GR-54124 Thessaloniki, Greece. ³Soil and Water Resources Institute, Hellenic Agricultural Organisation Demeter, Themi, 57001 Thessaloniki, Greece. ⁴CRIBI biotechnology center, University of Padua, via U. Bassi 58/b, 35131 Padova, Italy.

Received: 17 March 2022 Accepted: 25 June 2022

Published online: 15 August 2022

References

- Dutta S, He M, Xiong X, Tsang DCW. Sustainable management and recycling of food waste anaerobic digestate: a review. *Bioresour Technol.* 2021;341:125915.
- Campanaro S, Treu L, Rodriguez-R LM, Kovalovszki A, Ziels RM, Maus I, et al. New insights from the biogas microbiome by comprehensive genome-resolved metagenomics of nearly 1600 species originating from multiple anaerobic digesters. *Biotechnol Biofuels.* 2020;13:25.
- Ma S, Jiang F, Huang Y, Zhang Y, Wang S, Fan H, et al. A microbial gene catalog of anaerobic digestion from full-scale biogas plants. *GigaScience.* 2021;10:giaa164.
- Carabeo-Pérez A, Guerra-Rivera G, Ramos-Leal M, Jiménez-Hernández J. Metagenomic approaches: effective tools for monitoring the structure and functionality of microbiomes in anaerobic digestion systems. *Appl Microbiol Biotechnol.* 2019;103:9379–90.
- Paez-Espino D, Eloie-Fadrosch EA, Pavlopoulos GA, Thomas AD, Huntemann M, Mikhailova N, et al. Uncovering Earth's virome. *Nature.* 2016;536:425–30.
- Wommack KE, Colwell RR. Virioplankton: viruses in aquatic ecosystems. *Microbiol Mol Biol Rev.* 2000;64:69–114.
- Wu Q, Liu W-T. Determination of virus abundance, diversity and distribution in a municipal wastewater treatment plant. *Water Res.* 2009;43:1101–9.
- Shapiro OH, Kushmaro A, Brenner A. Bacteriophage predation regulates microbial abundance and diversity in a full-scale bioreactor treating industrial wastewater. *ISME J.* 2010;4:327–36.
- Calusinska M, Marynowska M, Goux X, Lentzen E, Delfosse P. Analysis of ds DNA and RNA viromes in methanogenic digesters reveals novel viral genetic diversity. *Environ Microbiol.* 2016;18:1162–75.
- Willenbücher K, Wibberg D, Huang L, Conrady M, Ramm P, Gätcke J, et al. Phage genome diversity in a biogas-producing microbiome analyzed by Illumina and Nanopore GridION sequencing. *Microorganisms.* 2022;10:368.
- Heyer R, Schallert K, Siewert C, Kohrs F, Greve J, Maus I, et al. Metaproteome analysis reveals that syntrophy, competition, and phage-host interaction shape microbial communities in biogas plants. *Microbiome.* 2019;7:69.
- Zhang J, Gao Q, Zhang Q, Wang T, Yue H, Wu L, et al. Bacteriophage-prokaryote dynamics and interaction within anaerobic digestion processes across time and space. *Microbiome.* 2017;5:57.
- Nanda AM, Thormann K, Frunzke J. Impact of spontaneous prophage induction on the fitness of bacterial populations and host-microbe interactions. Margolin W, editor. *J Bacteriol.* 2015;197:410–9.
- Choi J, Kotay SM, Goel R. Various physico-chemical stress factors cause prophage induction in *Nitrosospora multififormis* 25196- an ammonia oxidizing bacteria. *Water Res.* 2010;44:4550–8.
- Brüssow H, Bruttin A, Desiere F, Lucchini S, Foley S. Molecular ecology and evolution of *Streptococcus thermophilus* bacteriophages—a review. *Virus Genes.* 1998;16:95–109.
- Pan D, Watson R, Wang D, Tan ZH, Snow DD, Weber KA. Correlation between viral production and carbon mineralization under nitrate-reducing conditions in aquifer sediment. *ISME J.* 2014;8:1691–703.
- Brussaard CPD. Viral control of phytoplankton populations—a review. *J Eukaryot Microbiol.* 2004;51:125–38.
- Suttle CA. Marine viruses — major players in the global ecosystem. *Nat Rev Microbiol.* 2007;5:801–12.
- Harrison E, Brockhurst MA. Ecological and evolutionary benefits of temperate phage: what does or doesn't kill you makes you stronger. *BioEssays.* 2017;39:1700112.
- Krishnamurthy SR, Wang D. Origins and challenges of viral dark matter. *Virus Res.* 2017;239:136–42.
- Camarillo-Guerrero LF, Almeida A, Rangel-Pineros G, Finn RD, Lawley TD. Massive expansion of human gut bacteriophage diversity. *Cell.* 2021;184:1098–1109.e9.
- Roux S, Páez-Espino D, Chen I-MA, Palaniappan K, Ratner A, Chu K, et al. IMG/VR v3: an integrated ecological and evolutionary framework for interrogating genomes of uncultivated viruses. *Nucleic Acids Res.* 2021;49:D764–75.
- Kieft K, Zhou Z, Anantharaman K. VIBRANT: automated recovery, annotation and curation of microbial viruses, and evaluation of viral community function from genomic sequences. *Microbiome.* 2020;8:90.
- Nayfach S, Camargo AP, Schulz F, Eloie-Fadrosch E, Roux S, Kyrpides NC. CheckV assesses the quality and completeness of metagenome-assembled viral genomes. *Nat Biotechnol.* 2021;39:578–85.
- 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;5:69.
- Guo J, Bolduc B, Zayed AA, Varsani A, Dominguez-Huerta G, Delmont TO, et al. VirSorter2: a multi-classifier, expert-guided approach to detect diverse DNA and RNA viruses. *Microbiome.* 2021;9:37.

27. Arndt D, Grant JR, Marcu A, Sajed T, Pon A, Liang Y, et al. PHASTER: a better, faster version of the PHAST phage search tool. *Nucleic Acids Res.* 2016;44:W16–21.
28. Fang Z, Tan J, Wu S, Li M, Xu C, Xie Z, et al. PPR-Meta: a tool for identifying phages and plasmids from metagenomic fragments using deep learning. *GigaScience.* 2019;8:giz066.
29. Dutilh BE, Cassman N, McNair K, Sanchez SE, Silva GGZ, Boling L, et al. A highly abundant bacteriophage discovered in the unknown sequences of human faecal metagenomes. *Nat Commun.* 2014;5:4498.
30. Rossi A, Treu L, Toppo S, Zsach H, Campanaro S, Dutilh BE. Evolutionary study of the crassphage virus at gene level. *Viruses.* 2020;12:1035.
31. Yutin N, Benler S, Shmakov SA, Wolf YI, Tolstoy I, Rayko M, et al. Analysis of metagenome-assembled viral genomes from the human gut reveals diverse putative CrAss-like phages with unique genomic features. *Nat Commun.* 2021;12:1044.
32. Simmonds P, Adams MJ, Benkő M, Breitbart M, Brister JR, Carstens EB, et al. Virus taxonomy in the age of metagenomics. *Nat Rev Microbiol.* 2017;15:161–8.
33. Batinovic W, Knowler R, Stanton R, et al. Bacteriophages in natural and artificial environments. *Pathogens.* 2019;8:100.
34. Cristobal-Cueto P, García-Quintanilla A, Esteban J, García-Quintanilla M. Phages in food industry biocontrol and bioremediation. *Antibiotics.* 2021;10:786.
35. Jassim SAA, Limoges RG, El-Cheikh H. Bacteriophage biocontrol in wastewater treatment. *World J Microbiol Biotechnol.* 2016;32:70.
36. Kotay SM, Datta T, Choi J, Goel R. Biocontrol of biomass bulking caused by *Halscomenobacter hydrossis* using a newly isolated lytic bacteriophage. *Water Res.* 2011;45:694–704.
37. Rodriguez-R LM, Gunturu S, Harvey WT, Rosselló-Mora R, Tiedje JM, Cole JR, et al. The Microbial Genomes Atlas (MiGA) webserver: taxonomic and gene diversity analysis of Archaea and Bacteria at the whole genome level. *Nucleic Acids Res.* 2018;46:W282–8.
38. Ghosh D, Roy K, Williamson KE, White DC, Wommack KE, Sublette KL, et al. Prevalence of lysogeny among soil bacteria and presence of 16S rRNA and *trzN* genes in viral-community DNA. *Appl Environ Microbiol.* 2008;74:495–502.
39. Mahuku GS. A simple extraction method suitable for PCR-based analysis of plant, fungal, and bacterial DNA. *Plant Mol Biol Report.* 2004;22:71–81.
40. Santos-Medellin C, Zinke LA, ter Horst AM, Gelardi DL, Parikh SJ, Emerson JB. Viromes outperform total metagenomes in revealing the spatiotemporal patterns of agricultural soil viral communities. *ISME J.* 2021;15:1956–70.
41. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics.* 2014;30:2114–20.
42. Li D, Liu C-M, Luo R, Sadakane K, Lam T-W. MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. *Bioinformatics.* 2015;31:1674–6.
43. Gurevich A, Savelyev V, Vyahhi N, Tesler G. QUAST: quality assessment tool for genome assemblies. *Bioinformatics.* 2013;29:1072–5.
44. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. *Nat Methods.* 2012;9:357–9.
45. Kang DD, Li F, Kirton E, Thomas A, Egan R, An H, et al. MetaBAT 2: an adaptive binning algorithm for robust and efficient genome reconstruction from metagenome assemblies. *PeerJ.* 2019;7:e7359.
46. 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.
47. Parks DH, Imelfort M, Skennerton CT, Hugenholtz P, Tyson GW. CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome Res.* 2015;25:1043–55.
48. Virtanen P, Gommers R, Oliphant TE, Haberland M, Reddy T, Cournapeau D, et al. SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nat Methods.* 2020;17:261–72.
49. Friedman J, Alm EJ. Inferring correlation networks from genomic survey data. 2012. <https://doi.org/10.1371/journal.pcbi.1002687>.
50. Chaumeil PA, Mussig AJ, Hugenholtz P, Parks DH. GTDB-Tk: a toolkit to classify genomes with the Genome Taxonomy Database. 2020. <https://doi.org/10.1093/bioinformatics/btz848>.
51. Eddy SR. Profile hidden Markov models. *Bioinformatics.* 1998;14:755–63.
52. Graziotin AL, Koonin EV, Kristensen DM. Prokaryotic Virus Orthologous Groups (pVOGs): a resource for comparative genomics and protein family annotation. *Nucleic Acids Res.* 2017;45:D491–8.
53. Hyatt D, Chen G-L, LoCascio PF, Land ML, Larimer FW, Hauser LJ. Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics.* 2010;11:119.
54. Borges AL, Lou YC, Sachdeva R, Al-Shayeb B, Jaffe AL, Lei S, et al. Stop codon recoding is widespread in diverse phage lineages and has the potential to regulate translation of late stage and lytic genes. *bioRxiv.* 2021. <https://doi.org/10.1101/2021.08.26.457843>.
55. Huerta-Cepas J, Forslund K, Coelho LP, Szklarczyk D, Jensen LJ, von Mering C, et al. Fast genome-wide functional annotation through orthology assignment by eggNOG-mapper. *Mol Biol Evol.* 2017;34:2115–22.
56. Palù M, Basile A, Zampieri G, Treu L, Rossi A, Morlino MS, et al. KEMET—A python tool for KEGG Module evaluation and microbial genome annotation expansion. *Comput Struct Biotechnol J.* 2022;20:1481–6. <https://doi.org/10.1016/j.csbj.2022.03.015>.
57. Yin Y, Mao X, Yang J, Chen X, Mao F, Xu Y. dbCAN: a web resource for automated carbohydrate-active enzyme annotation. *Nucleic Acids Res.* 2012;40:W445–51.
58. Pascal Andreu V, Roel-Touris J, Dodd D, Fischbach MA, Medema MH. The gutSMASH web server: automated identification of primary metabolic gene clusters from the gut microbiota. *Nucleic Acids Res.* 2021;49:W263–70.
59. Binnenkade L, Teichmann L, Thormann KM. Iron triggers λ So prophage induction and release of extracellular DNA in *Shewanella oneidensis* MR-1 Biofilms. Spormann AM, editor. *Appl Environ Microbiol.* 2014;80:5304–16.
60. Long A, McDaniel LD, Mobberley J, Paul JH. Comparison of lysogeny (prophage induction) in heterotrophic bacterial and *Synechococcus* populations in the Gulf of Mexico and Mississippi river plume. *ISME J.* 2008;2:132–44.
61. Harris SM, Yue W-F, Olsen SA, Hu J, Means WJ, McCormick RJ, et al. Salt at concentrations relevant to meat processing enhances Shiga toxin 2 production in *Escherichia coli* O157:H7. *Int J Food Microbiol.* 2012;159:186–92.
62. Boe K, Batstone DJ, Steyer J-P, Angelidaki I. State indicators for monitoring the anaerobic digestion process. *Water Res.* 2010;44:5973–80.
63. Tsapekos P, Kougiyas PG, Vasileiou SA, Lyberatos G, Angelidaki I. Effect of micro-aeration and inoculum type on the biodegradation of lignocellulosic substrate. *Bioreour Technol.* 2017;225:246–53.
64. Angelidaki I, Treu L, Tsapekos P, Luo G, Campanaro S, Wenzel H, et al. Biogas upgrading and utilization: current status and perspectives. *Biotechnol Adv.* 2018;36:452–66.
65. Liu J, Jia R, Wang Y, Wei Y, Zhang J, Wang R, et al. Does residual H_2O_2 result in inhibitory effect on enhanced anaerobic digestion of sludge pretreated by microwave- H_2O_2 pretreatment process? *Environ Sci Pollut Res.* 2017;24:9016–25.
66. Touchon M, Bernheim A, Rocha EP. Genetic and life-history traits associated with the distribution of prophages in bacteria. *ISME J.* 2016;10:2744–54.
67. Yuan Y, Gao M. Jumbo Bacteriophages: An Overview. *Front Microbiol.* 2017;8:403.
68. Roux S, Adriaenssens EM, Dutilh BE, Koonin EV, Kropinski AM, Krupovic M, et al. Minimum information about an uncultivated virus genome (MIUViG). *Nat Biotechnol.* 2019;37:29–37.
69. Nayfach S, Páez-Espino D, Call L, Low SJ, Sberro H, Ivanova NN, et al. Metagenomic compendium of 189,680 DNA viruses from the human gut microbiome. *Nat Microbiol.* 2021;6:960–70.
70. Fontana A, Campanaro S, Treu L, Kougiyas PG, Cappa F, Morelli L, et al. Performance and genome-centric metagenomics of thermophilic single and two-stage anaerobic digesters treating cheese wastes. *Water Res.* 2018;134:181–91.
71. Kakuk B, Wirth R, Maróti G, Szuhaj M, Rakhely G, Laczi K, et al. Early response of methanogenic archaea to H_2 as evaluated by metagenomics and metatranscriptomics. *Microb Cell Factories.* 2021;20:127.
72. Tian H, Fotidis IA, Kissas K, Angelidaki I. Effect of different ammonia sources on acetitlastic and hydrogenotrophic methanogens. *Bioreour Technol.* 2018;250:390–7.
73. Maus I, Wibberg D, Stantscheff R, Eikmeyer F-G, Seffner A, Boelter J, et al. Complete genome sequence of the hydrogenotrophic,

- methanogenic archaeon *Methanoculleus bourgensis* strain MS2(T), Isolated from a sewage sludge digester. *J Bacteriol.* 2012;194:5487–8.
74. Evans PN, Boyd JA, Leu AO, Woodcroft BJ, Parks DH, Hugenholtz P, et al. An evolving view of methane metabolism in the Archaea. *Nat Rev Microbiol.* 2019;17:219–32.
 75. Ziels RM, Sousa DZ, Stensel HD, Beck DAC. DNA-SIP based genome-centric metagenomics identifies key long-chain fatty acid-degrading populations in anaerobic digesters with different feeding frequencies. *ISME J.* 2018;12:112–23.
 76. Calusinska M, Goux X, Fossépré M, Muller EEL, Wilmes P, Delfosse P. A year of monitoring 20 mesophilic full-scale bioreactors reveals the existence of stable but different core microbiomes in bio-waste and wastewater anaerobic digestion systems. *Biotechnol Biofuels.* 2018;11:196.
 77. Lucas R, Kuchenbuch A, Fetzer I, Harms H, Kleinstüber S. Long-term monitoring reveals stable and remarkably similar microbial communities in parallel full-scale biogas reactors digesting energy crops. *FEMS Microbiol Ecol.* 2015;91 Available from: <https://academic.oup.com/femsec/article-lookup/doi/10.1093/femsec/fiv004>. Cited 2021 Oct 15.
 78. Sun L, Liu T, Müller B, Schnürer A. The microbial community structure in industrial biogas plants influences the degradation rate of straw and cellulose in batch tests. *Biotechnol Biofuels.* 2016;9:128.
 79. Yasmin A, Kenny JG, Shankar J, Darby AC, Hall N, Edwards C, et al. Comparative genomics and transduction potential of *Enterococcus faecalis* temperate bacteriophages. *J Bacteriol.* 2010;192:1122–30.
 80. Ahring BK, Sandberg M, Angelidaki I. Volatile fatty acids as indicators of process imbalance in anaerobic digestors. *Appl Microbiol Biotechnol.* 1995;43:559–65.
 81. Tan W-B, Jiang Z, Chen C, Yuan Y, Gao L-F, Wang H-F, et al. *Thiopseudomonas denitrificans* gen. nov., sp. nov., isolated from anaerobic activated sludge. *Int J Syst Evol Microbiol.* 2015;65:225–9.
 82. Campanaro S, Treu L, Kougias PG, Luo G, Angelidaki I. Metagenomic binning reveals the functional roles of core abundant microorganisms in twelve full-scale biogas plants. *Water Res.* 2018;140:123–34.
 83. Cavaliere M, Feng S, Soyer OS, Jiménez JJ. Cooperation in microbial communities and their biotechnological applications. *Environ Microbiol.* 2017;19:2949–63.
 84. Rankin DJ, Rocha EPC, Brown SP. What traits are carried on mobile genetic elements, and why? *Heredity.* 2011;106:1–10.
 85. Sundberg C, Al-Soud WA, Larsson M, Alm E, Yekta SS, Svensson BH, et al. 454 pyrosequencing analyses of bacterial and archaeal richness in 21 full-scale biogas digesters. *FEMS Microbiol Ecol.* 2013;35:612–26.
 86. Hehemann J-H, Correc G, Barbeyron T, Helbert W, Czekaj M, Michel G. Transfer of carbohydrate-active enzymes from marine bacteria to Japanese gut microbiota. *Nature.* 2010;464:908–12.
 87. Buhlmann CH, Mickan BS, Jenkins SN, Tait S, Kahandawala TKA, Bahri PA. Ammonia stress on a resilient mesophilic anaerobic inoculum: methane production, microbial community, and putative metabolic pathways. *Bioresour Technol.* 2019;275:70–7.
 88. Mosbæk F, Kjeldal H, Mulat DG, Albertsen M, Ward AJ, Feilberg A, et al. Identification of syntrophic acetate-oxidizing bacteria in anaerobic digesters by combined protein-based stable isotope probing and metagenomics. *ISME J.* 2016;10:2405–18.
 89. Ruiz-Sánchez J, Campanaro S, Guivernau M, Fernández B, Prenafeta-Boldú FX. Effect of ammonia on the active microbiome and metagenome from stable full-scale digesters. *Bioresour Technol.* 2018;250:513–22.
 90. Chen S, Zamudio Cañas EM, Zhang Y, Zhu Z, He Q. Impact of substrate overloading on archaeal populations in anaerobic digestion of animal waste. *J Appl Microbiol.* 2012;113:1371–9.
 91. Kalamaras SD, Vasileiadis S, Karas P, Angelidaki I, Kotsopoulos TA. Microbial adaptation to high ammonia concentrations during anaerobic digestion of manure-based feedstock: biomethanation and 16S rRNA gene sequencing. *J Chem Technol Biotechnol.* 2020;95:1970–9.
 92. de Jonge PA, Nobrega FL, Brouns SJJ, Dutilh BE. Molecular and evolutionary determinants of bacteriophage host range. *Trends Microbiol.* 2019;27:51–63.
 93. Nasir A, Forterre P, Kim KM, Caetano-Anollés G. The distribution and impact of viral lineages in domains of life. *Front Microbiol.* 2014;5:194.
 94. Aljabali AA, Hassan SS, Pabari RM, Shahcheraghi SH, Mishra V, Charbe NB, et al. The viral capsid as novel nanomaterials for drug delivery. *Future Sci OA.* 2021;7:FSO744.
 95. Slonczewski JL, Fujisawa M, Dopson M, Krulwich TA. Cytoplasmic pH measurement and homeostasis in bacteria and archaea. *Adv Microb Physiol.* 2009;55(1–79):317.
 96. Khan MZ, Singha B, Ali MF, Taunk K, Rapole S, Gourinath S, et al. Redox homeostasis in *Mycobacterium tuberculosis* is modulated by a novel actinomycete-specific transcription factor. *EMBO J.* 2021;40:e106111.
 97. Wood JP, Richter W, Sunderman M, Calfee MW, Serre S, Mickelsen L. Evaluating the environmental persistence and inactivation of MS2 bacteriophage and the presumed Ebola virus surrogate phi6 using low concentration hydrogen peroxide vapor. *Environ Sci Technol.* 2020;54:3581–90.
 98. Yokoyama K, Yumura M, Honda T, Ajitomi E. Characterization of denitrification and net N₂O-reduction properties of novel aerobically N₂O-reducing bacteria. *Soil Sci Plant Nutr.* 2016;62:230–9.
 99. Johnson CN, Sheriff EK, Duerkop BA, Chatterjee A. Let Me Upgrade You: impact of mobile genetic elements on enterococcal adaptation and evolution. Margolin W, editor. *J Bacteriol.* 2021;203:e00177–21.
 100. Tan D, Hansen MF, de Carvalho LN, Røder HL, Burmølle M, Middelboe M, et al. High cell densities favor lysogeny: induction of an H₂O prophage is repressed by quorum sensing and enhances biofilm formation in *Vibrio anguillarum*. *ISME J.* 2020;14:1731–42.
 101. Schnürer A, Schink B, Svensson BH. *Clostridium ultunense* sp. nov., a mesophilic bacterium oxidizing acetate in syntrophic association with a hydrogenotrophic methanogenic bacterium. *Int J Syst Bacteriol.* 1996;46:1145–52.
 102. Hattori S, Kamagata Y, Hanada S, Shoun H. *Thermacetogenium phaeum* gen. nov., sp. nov., a strictly anaerobic, thermophilic, syntrophic acetate-oxidizing bacterium. *Int J Syst Evol Microbiol.* 2000;50:1601–9.
 103. Balk M, Weijma J, Stams AJM. *Thermotoga lettingae* sp. nov., a novel thermophilic, methanol-degrading bacterium isolated from a thermophilic anaerobic reactor. *Int J Syst Evol Microbiol.* 2002;52:1361–8.
 104. Westerholm M, Roos S, Schnürer A. *Syntrophaceticus schinkii* gen. nov., sp. nov., an anaerobic, syntrophic acetate-oxidizing bacterium isolated from a mesophilic anaerobic filter. *FEMS Microbiol Lett.* 2010;309(1):100–4. <https://doi.org/10.1111/j.1574-6968.2010.02023.x>.
 105. Westerholm M, Roos S, Schnürer A. *Tepidanaerobacter acetatodoxydans* sp. nov., an anaerobic, syntrophic acetate-oxidizing bacterium isolated from two ammonium-enriched mesophilic methanogenic processes. *Syst Appl Microbiol.* 2011;34:260–6.
 106. Nobu MK, Narihiro T, Rinke C, Kamagata Y, Tringe SG, Woyke T, et al. Microbial dark matter ecogenomics reveals complex synergistic networks in a methanogenic bioreactor. *ISME J.* 2015;9:1710–22.
 107. Zhu X, Campanaro S, Treu L, Seshadri R, Ivanova N, Kougias PG, et al. Metabolic dependencies govern microbial syntrophies during methanogenesis in an anaerobic digestion ecosystem. *Microbiome.* 2020;8:22.
 108. Bull JJ, Vimr ER, Molineux IJ. A tale of tails: Sialidase is key to success in a model of phage therapy against K1-capsulated *Escherichia coli*. *Virology.* 2010;398:79–86.
 109. Pires DP, Oliveira H, Melo LDR, Sillankorva S, Azeredo J. Bacteriophage-encoded depolymerases: their diversity and biotechnological applications. *Appl Microbiol Biotechnol.* 2016;100:2141–51.
 110. Wolf S, Fischer MA, Kupczok A, Reetz J, Kern T, Schmitz RA, et al. Characterization of the lytic archaeal virus Drs3 infecting *Methanobacterium formicicum*. *Arch Virol.* 2019;164:667–74.
 111. Weidenbach K, Wolf S, Kupczok A, Kern T, Fischer MA, Reetz J, et al. Characterization of Bf4, an archaeal lytic virus targeting a member of the methanomicrobiales. *Viruses.* 2021;13:1934.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.