Mechanisms of Genetic Transfer Between Archaea and Bacteria via Membrane Vesicles, Plasmid Vesicles and Viruses

Dissertation zur Erlangung des Grades
eines Doktors der Naturwissenschaften
- Dr. rer. nat. -

Dem Fachbereich II (Biologie / Chemie)
vorgelegt von

Dominik Lücking

Bremen Dezember 2023

Die Forschungsarbeit, die dieser Thesis zugrunde liegt, wurde in der Zeit von November 2019 bis Dezember 2023 in der Arbeitsgruppe Archaea Virology am Max-Planck-Institut für Marine Mikrobiologie unter der Leitung von Dr. Susanne Erdmann durchgeführt. Die Thesis wurde im Rahmen der International Max Planck Research School of Marine Microbiology (IMPRS MarMic) angefertigt.

The scientific work that is the foundation of this thesis was conducted from November 2019 to December 2023 under the supervision of Dr. Susanne Erdmann, group leader of Archaeal Virology at the Max Planck Institute for Marine Microbiology. The thesis was prepared under the guidelines of the International Max Planck Research School for Marine Microbiology (IMPRS MarMic).

Examination Committee

Gutachterinnen / Reviewers:

Dr. Susanne Erdmann	MPI Bremen
Prof. Dr. Alexander Probst	University of Duisburg-Essen

Prüfende / Doctoral committee:

Prof. Dr. Rudolf Amann	MPI Bremen
Prof. Dr. Alexander Probst	University of Duisburg-Essen
Dr. Luis 'Coto' Orellana	MPI Bremen

Kolloquiumsvorsitz / Chairwoman/man:

Prof. Dr. Jan-Hendrik Hehemann	MPI / Marum Bremen

Datum des Kolloquiums / Date of the Defense:

30.01.2024

Acknowledgements

First, I want to thank my supervisor Susanne Erdmann. Thank you, for an always open door, the opportunities and the freedom you gave me. Four years is a long time to work together and I felt encouraged to pursue my interests and ideas throughout this time. Thanks for telling me when to stop and thanks for telling me when to go for it.

Thanks to my thesis advisory committee, Alexander Probst, Ahmed Zayed, Luis 'Coto' Orellana and Bernhard Fuchs, for guidance, open discussions and support. I'm thankful for the opportunity to interact with such incredibly smart and knowledgeable, and at the same time kind and respectful scientists.

Thank you to my colleagues, Coco, Lauren, Johanna, Tomás and Jay. For uncomplicated friendships, advice and laughs. Thanks Lauren, for dealing with me on a daily basis. I mean, I've spent more time with you than with most people over the last couple of years and yet you still have an open ear for my nonsense.

I'm incredibly thankful to my friends. There is this Tolkien (for Lennart: that's the Lord of the Ring guy. No, not Gandalf) quote - "I don't know half of you half as well as I should like; and I like less than half of you half as well as you deserve." - and I swear to god I don't know what it means and I don't think it applies here. Anyway, I like all of you so much and all of you deserve it, for hours spent in our beloved Mensa, for discussion, for support, for laughter, for latex, for taking my mind off of science. Special thanks to the person closest to me: Thank you Greta, without you I would not have managed.

To my Mom and my Dad, not only for the obvious things but for a seemingly impossible perspective on things like this thesis: Yep, it's very important, go pursue what you want, be curious and invest in it. And no, don't worry, it's just a piece of paper (many in this case). Go, touch some grass and look into the distance. Thanks to my siblings, Debora, Svenja, Julia, Kathrin. You cannot grasp the importance you have to me, not only for this thesis, but for everything.

- - | - -

Summary

Microbial life on earth is shaped by horizontal gene transfer (HGT) events, which themselves often are driven by genetic mobile elements (MGE). These events drive the evolution of bacteria and archaea and have paramount ecological implications on all biomes of Earth. By inducing heritable changes to the DNA content of bacteria and archaea, HGT blurs phylogenetic boundaries and challenges existing classification schemes and nomenclature. Even more, HGT events are not restricted to microbial organisms but influence MGEs as well. In the age of next-generation sequencing, the analysis of interactions of MGEs with their microbial counterparts and other forms of DNA exchange relies on highly sophisticated bioinformatic tools and pipelines. In this dissertation, two exciting cases of HGT are explored. Both cases challenge existing nomenclature and classification schemes and highlight the need for dedicated bioinformatic pipelines. The archaeal plasmid pR1SE is disseminated in self-encoded plasmid vesicles (PVs), which enables a virus-like lifestyle. In Chapter 2, we describe the identification of 40 novel pR1SE-like plasmids, subsequently termed archaeal plasmids of *Haloarchaea* potentially transferred in plasmid vesicles (apHPVs). We show their distribution and host range, explore their genetic organization, and expand on the implications of their life cycles. Lastly, we establish that apHPVs genetically link viruses and plasmids of archaea by facilitating the exchange of genes. In Chapter 3, the current state of the art of bioinformatic analysis of HGT and MGEs is reviewed. Additionally, the challenges with identifying gene transfer in extracellular vesicles (EVs) are elaborated, and the approach presented in Chapter 4 circumvents these challenges. By applying the developed pipeline to metagenomic datasets (Chapter 4), we shed light on the composition of HGT mechanisms in the ocean and analyze the functional potential of the genetic material transferred in EVs. Finally, we establish a novel nomenclature in order to better describe the datasets analyzed.

Zusammenfassung

Das mikrobielle Leben auf der Erde wird durch Ereignisse des horizontalen Gentransfers (HGT) geformt, welche ihrerseits oft durch genetisch mobile Elemente (MGE) angetrieben werden. Diese Ereignisse treiben die Evolution von Bakterien und Archaeen voran und haben weitreichende ökologische Auswirkungen auf alle Biome der Erde. Da HGT vererbbare Veränderungen in der DNA von Bakterien und Archaeen verursacht, verschwimmen phylogenetische Grenzen und bestehende Klassifikationsschemata und Nomenklaturen werden in Frage gestellt. Darüber hinaus sind HGT-Ereignisse nicht auf mikrobielle Organismen beschränkt, sondern beeinflussen auch MGEs. Im Zeitalter des Next-Generation Sequencing, ist die Analyse der Interaktionen von MGEs mit ihren mikrobiellen Gegenspielern und anderen Formen des DNA-Austauschs auf hochentwickelte bioinformatische Werkzeuge und Pipelines angewiesen. In dieser Dissertation werden zwei interessante Fälle von HGT untersucht. Beide Fälle stellen bestehende Nomenklaturen und Klassifikationsschemata in Frage und verdeutlichen den Bedarf an speziellen bioinformatischen Pipelines. Das archaeische Plasmid pR1SE wird in selbstkodierten Plasmidvesikeln (PVs) verbreitet, die einen virusähnlichen Lebensstil ermöglichen. In Kapitel 2 beschreiben wir die Identifizierung von 40 neuartigen pR1SE ähnlichen Plasmiden, die nachfolgend als archaeal plasmids of haloarchaea potentially transferred in plasmid vesicles (apHPVs) bezeichnet werden. Wir zeigen ihre Verbreitung und ihr Wirtsspektrum, untersuchen ihre genetische Organisation und erläutern die Auswirkungen ihrer Lebenszyklen. Schließlich stellen wir fest, dass apHPVs Viren und Plasmide von Archaeen genetisch miteinander verbinden. In Kapitel 3 wird der aktuelle Stand der bioinformatischen Analyse von HGT und MGEs besprochen. Darüber hinaus werden die Herausforderungen bei der Identifizierung von Gentransfer in extrazellulären Vesikeln (EVs) erläutert, welche durch den in Kapitel 4 entwickelten Ansatz umgangen werden. Durch die Anwendung der entwickelten Pipeline auf metagenomische Datensätze (Kapitel 4) beleuchten wir die Zusammensetzung von HGT-Mechanismen im Ozean und analysieren das funktionelle Potenzial des in EVs übertragenen genetischen Materials. Schließlich führen wir eine neue Nomenklatur ein, um die analysierten Datensätze besser beschreiben zu können.

List of Publications

Lücking, D., Alarcón-Schumacher, T., & Erdmann, S. (2023). Distribution and implications of haloarchaeal plasmids disseminated in self-encoded plasmid vesicles.
Submitted to *microorganisms*.
(Chapter 2)

Lücking, D. & Erdmann, S. (2023). Bioinformatic Analysis of Drivers of Horizontal Gene Transfer.
In preparation.
(Chapter 3)

Lücking, D., Mercier, C., Alarcón-Schumacher, T., & Erdmann, S. (2023). Extracellular vesicles are the main contributor to the non-viral protected extracellular sequence space. *ISME communications*, 3(1), 112.
Published in *ISME communications*.
(Chapter 4)

Contents

List of Figures

Chapter 1

General Introduction

When the vast complexity of life is to be studied by researchers, the need for practical classification, helpful terminology and abstract concepts becomes evident. Generalizations are crucial for a fundamental understanding of too-big-to-grasp systems, e.g., the biological world. However, if the developed generalization does not reflect the details of the system, the conventional theories, classifications or terminology must be addressed, not the least because they often determine our hypotheses, our thinking and our future approaches to the system. Carl Woese highlighted this need when he wrote to Ernst Mayr "*A biological classification is...an overarching evolutionary theory that guides our thinking and experimentation, and it must be structured...to reflect evolutionary reality*" (Woese 1998). Mayr had argued that the morphological differences between eukaryotes and prokaryotes (archaea and bacteria) were more fundamental than differences in ribosomal subunits (van der Gulik et al. 2017), based on which Woese had proposed a three-domain model consisting of eukaryotes, bacteria and archaea (Woese et al. 1990). By now, Woese's classification model has been widely accepted as the model that better reflects the evolutionary history of organisms. With the discovery of complex archaea of the superphylum (Spang et al. 2015), the gap between archaea and eukaryotes seemed to be closed slowly. Subsequently, proposals of a modern tow-domain system, with archaea and bacteria as the fundamental domains, gained traction (Koonin 2015), (Raymann et al. 2015). While the present work will not dare to deal with this specific fundamental question, the overarching theme is shared: The discovery of novel mechanisms and entities challenges conventional terminology (see Chapter 4) or classification schemes (see Chapter 2).

Nucleic acids are the genetic backbone of all living organisms. The vertical transmission of DNA, i.e., the production of offspring with nearly identical genetic material, allows the classification of organisms into operational taxonomic units (OTUs). Mutations, recombination events, and the transfer of genetic material between organisms, so-called horizontal gene transfer, blur the lines between discrete OTUs (see Figure 1.1). One can surmise the ecological and evolutionary implications of genomic flexibility: On one hand, changes to the genetic potential of a species can increase the adaptability and, therefore, fitness of a species to a changing environment. These changes in gene content seldom affect all organisms within a population at once; for example, only a single cell might have acquired a defense system against a specific parasite. The sum of all genes within a given population,

including genes only found in a fraction of all organisms of the species, are considered the pangenome of a species. In the described case, the single cell, well-equipped to deal with the attack of a parasite, ensures the survival of the entire species. On the other hand, the same changes are the driving force behind evolution (Acar Kirit et al. 2022). Species move from one ecological niche to another, split into distinct populations (speciation), and, over time, entire lineages develop. Therefore, the analysis of the genetic material is, next to phenological analysis, the key to entangling the evolutionary history of an organism and its role within a microbial community.

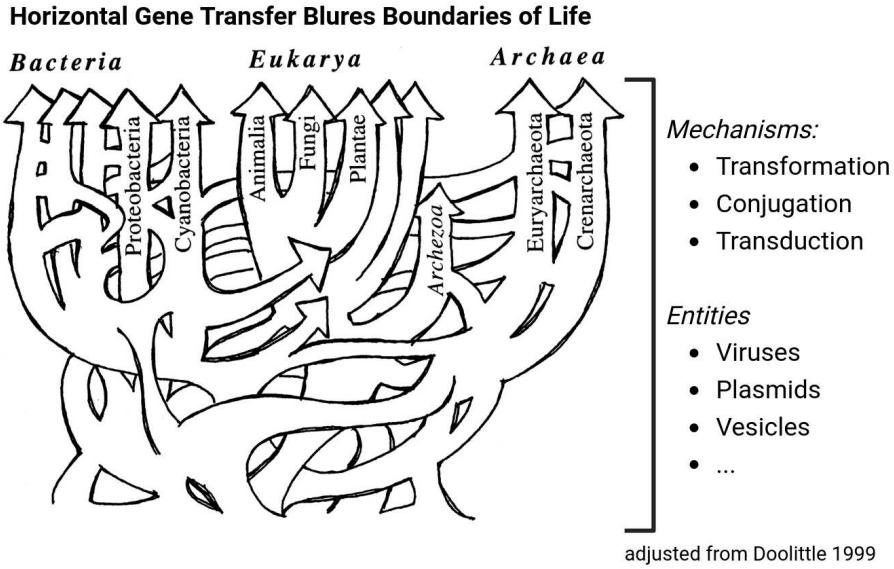


Figure 1.1: Horizontal gene transfer is a major mechanism that blurs the boundaries of species (Doolittle 1999).

1.1 Horizontal Gene Transfer in Archaea and Bacteria

Horizontal or lateral gene transfer (HGT or LGT) is the movement of genetic material between organisms other than through the vertical transmission of DNA. Evidence for HGT events is present in nearly all prokaryotic genomes, demonstrating HGT's fundamental role in shaping the 'web of life' itself (Soucy et al. 2015). Not all gene-sharing events provide necessarily an adaptive benefit or are necessarily detrimental to the organism or its population. Genes often are moved from a donor to a recipient by chance, thus it becomes a stochastic selective process that determines which changes to the genome are stable and might propagate through the population. Sequences that provide a positive effect will be positively selected for, while negative and neutral effects will be outcompeted (Arnold et al. 2022). The three main mechanisms of horizontal gene transfer are transformation (the uptake of free environmental DNA by a cell), transduction (the transport of host DNA in virus capsids) and conjugation (the direct transfer of DNA, e.g. plasmids, through direct cell-to-cell contact or via pili) (Burmeister 2015). Additionally, there are an ever growing number of non-canonical ways to transport DNA from one organism to another. HGT is therefore driven - but not exclusively! - by a set of mobile genetic elements (MGEs). MGEs

consist of a set of core genes and additionally carry auxiliary genes, potentially providing their host with a selective advantage. One famous example of this is two dsDNA viruses[1] of the class *Caudoviricetes* that provide genes for two crucial proteins to their photosynthetic host, *Porchlorococcus sp.*: the photosystem II core reaction center protein D1 (psbA) and a high-light-inducible protein (hli) (Lindell et al. 2004). The lines between different MGEs are intriguingly blurry since MGEs are defined by their set of core genes, which, as a result of interactions with other MGEs or independent recombination events, change over time (Koonin et al. 2021). Additionally, not all HGT is driven by MGEs: DNA received via gene transfer agents or extracellular vesicles and the uptake of free environmental DNA (transformation) are HGT-driving mechanisms outside of influence of MGEs but essential pieces in understanding the flux of genes within a microbial community (Hall et al. 2020; Soucy et al. 2015).

1.1.1 Transformation

Transformation describes the uptake of free DNA from the environment into the cell. The presence of this DNA is the result of disintegrated cells, e.g. through lysis. The chromosome and all secondary replicons get released and quickly break down into smaller fragments due to the instability of longer DNA segments, which, depending on the organism, can be multiple megabases per cell. The smaller, fragmented DNA molecules are relatively stable in the environment and can be taken up by transformable organisms, an attribute also called 'competence'. The competence of cells is determined by the presence of two crucial proteins: a DNA-binding protein on the outside of the cell, so-called autolysins, and nucleases. Within the population of a species, not all cells may take up DNA (Johnsborg et al. 2007; Johnston et al. 2014). For example, cells of *Acinetobacter* species are naturally competent and close to 100 % may do so, while *Bacillus* are also competent, but only around 20 % of a population will be transformed (Brock et al. 2003). In Chapter 4, this plays a crucial role since the goal in this chapter is to demonstrate the uptake of DNA from extracellular vesicles. Therefore, the careful removal of free, environmental DNA in preparation of the dataset is essential and could otherwise introduce biases in our findings. In order to have a heritable, transformative effect, the absorbed DNA needs to be incorporated into a replicon. This happens either via homologous recombination, using the protein RecA, or site-specific recombinases and transposases, which require a site-specific, short recognition site of a few bases (Arnold et al. 2022 and the references therein). The recombination enables the receiving cell to express the newly gained genetic material. A beneficial transformation might be selected positively for and spread in the population. An alternative purpose of DNA uptake, outside of incorporating new genetic information, is using transformed DNA as a carbon source (Finkel and Kolter 2001). The two purposes are most likely not mutually exclusive, and the usage might depend on environmental circumstances. Indeed, not all transformation events lead to a genetic or even heritable change in the uptaking cell, yet transformation plays a significant role in microbial evolution and

[1]Note: Traditionally, viruses infecting bacteria are called bacteriophages or phages in short. However, I will avoid this terminology, since it excludes viruses infecting archaea, and instead refer to them as viruses of archaea and bacteria, respectively.

HGT in general.

1.1.2 Transduction

The exchange of genetic material via viruses, coined transduction, was first described for Salmonella phage P22 (Zinder and Lederberg 1952). Since then, a plethora of transduction events have been observed, involving hosts from all domains of life and a similarly diverse set of viruses. Historically, two main transduction mechanisms were described: specialized and generalized transduction. A third mechanism was recently established: lateral transduction (Chiang et al. 2019) (see Figure 1.2). Regardless of the virus's life cycle (lytic, lysogenic or chronic, see subsection 1.2.1), the transduction mode depends on the assembly of the viral particle, precisely the type of terminase responsible for DNA packaging into the virus particle. Both types of terminases recognize a phage-specific packaging site (*pac*- or *cos*-site), where it cleaves the DNA and begins the packaging process. *Pac*-type terminases will detect when the carrying capacity of the virus particle is reached and cut the DNA at a non-specific site (Rao and Feiss 2015), while *cos*-type terminases need to recognize a second motif (*cos*-site) for the second cut (Feiss and Siegele 1979). *Pac*-type packaging is typically associated with generalized transduction (Thierauf et al. 2009). Here, a homolog to a *pac*-site, termed 'pseudo-*pac*-site', is detected by the phage on the chromosome or other replicons and thus mistakenly packaged into the viral particle, resulting in a particle packaged completely with non-viral, microbial DNA. Such pseudo-*pac*-sites could, in theory, be located all over the genome. Thus all parts of the host chromosome or its plasmids could be transduced, resulting in generalized transduction. For *cos*-type terminases to mistakenly package microbial DNA, two *cos*-site homologs would have to occur at just the right distance from each other. Therefore, *cos*-type packaging is less associated with generalized transduction. Contrastingly, specialized transduction is more often associated with *cos*-type packaging, although not exclusively. In specialized transduction, microbial DNA is co-packaged with the viral genome into the virus particle. This only occurs with genes adjacent to the integrated viral genome or when the excised viral genome recombines with short microbial DNA fragments. Since the required erroneous excision or exosomal recombination events are rare, specialized transduction is thought to be less frequent than generalized transduction (Chiang et al. 2019). In contrast to both mechanisms, lateral transduction does not result from an error in packaging but instead appears as a regular part of the virus life cycle (Chen et al. 2018; Davidson 2018). Traditionally, the excision of a virus happens shortly after induction, and the expression of the responsible *Xis* excisionase occurs very early in the virus life cycle. However, as shown by Chen *et al.*, in temperate viruses of *Staphylococcus aureus*, *Xis*-activity is delayed, and therefore, the excision of the virus occurs later. At this point, the integrated viral genome has been replicated in situ next to microbial DNA. Packaging is initiated at a bona fide *pac*-site, and multiple segments of DNA, first the viral, then host DNA is packaged into multiple capsids (Chen et al. 2018). The prevalence of this mechanism in other organisms needs to be clarified, especially for archaeal hosts and their integrated viruses. However, its efficiency and frequency of transduction make it an essential case of horizontal gene transfer.

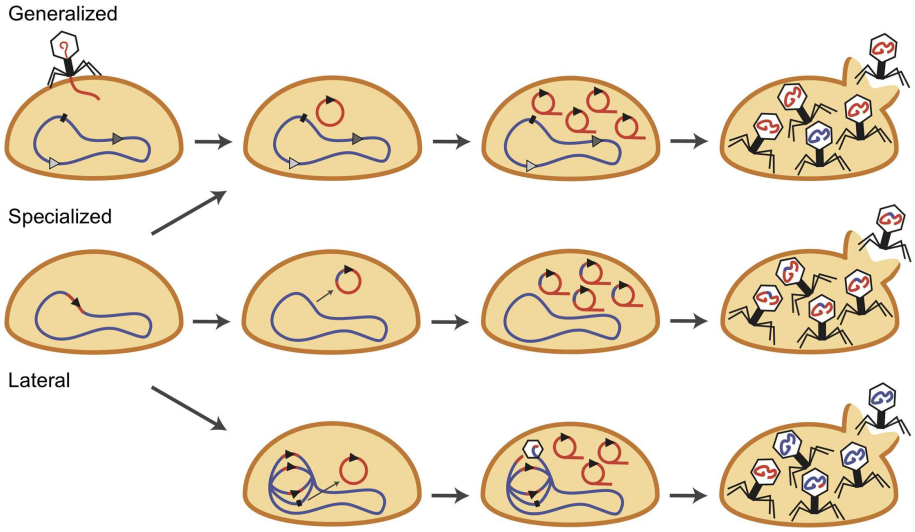


Figure 1.2: Types of transduction. Generalized, specialized and lateral transduction are shown. Blue - microbial chromosome, red - viral genome, triangles - *cos/pac*-packaging sites. Adjusted from (Chiang et al. 2019).

1.1.3 Conjugation

The third classic way of horizontal gene transfer is conjugation. Generally, conjugation describes the transmission of a genetic element, often a conjugative plasmid, from a donor to a recipient cell via direct cell-to-cell contact. The proteins that make up the conjugation apparatus are most often found directly on the plasmid being conjugated; e.g., the F plasmid of *E. coli* possesses a *tra*-region that encodes for the necessary proteins. In other cases, the host domesticated the system, and the necessary genes are located on the main chromosome (Brock et al. 2003). The cassette of proteins required consists of four components: i) a pilus that establishes first contact between the donor and the receiving cell. ii) a type V coupling protein iii) a T4SS protein complex that stabilizes the cell-to-cell contact and enables the transfer of DNA across cell membranes, and lastly, iv) a relaxosome that nicks and unwinds the DNA double-strand for cross-membrane transport (Beltran et al. 2023). After receiving the plasmid, the cells' surface receptors are altered so that they cannot establish future cell-to-cell contact, which ensures that only one copy of the plasmid or genetically similar plasmids are present in a cell. In archaea, conjugative plasmids have been described for various *Sulfolobus* strains, most prominently in *S. acidocaldarius* and *S. islandicus* (Grogan 1996; Prangishvili et al. 1998) and more recently for hypothermophilic Euryarchaeota (Catchpole et al. 2023). Furthermore, metagenomic analysis revealed the prevalence of homologs of the bacterial conjugation system in archaea (Krupovic et al. 2019). The group of archaeal plasmids described in Chapter 2 are non-conjugative, as are many other plasmids of archaea. Nevertheless, conjugation remains a driving force of HGT, and conjugative plasmids are a widespread and highly successful class of MGEs.

1.2 Drivers of Horizontal Gene Transfer

1.2.1 Viruses

Viruses are the most prominent drivers of horizontal gene transfer. They are abundant and highly diverse; newest studies estimate the number of virus particles worldwide at 10^{31} with over hundreds viral families and thousands of species (Gorbalenya et al. 2020). Finding a precise and unambiguous definition for a class of highly diverse entities with many combinations of characteristics has been challenging for over 70 years. Lwoff attempted a first concise summary of what constitutes a virus in 1957. Key points of the definition were their strictly intracellular nature, that they contained one type of nucleic acids, potential pathogenicity and their inability to grow or undergo binary fission (Lwoff 1957). However, this definition was challenged by the subsequent discovery of other MGEs that were virus-like but also strikingly different. One example of this was the discovery of 'viroids', which were distinctly smaller and without a viral capsid or viral particle whatsoever, but were replicated similarly, existed without a metabolism and had a strictly intracellular lifestyle (Diener and Raymer 1967). Over the following decades, multiple iterations of a concise virus definition were attempted, always focussing on integrating the newest discoveries of MGEs that were virus-like but did not adhere to the previous definition. Additionally, the rate and methods of virus detection and identification changed with the advent of next-generation sequencing and the development of 'viromics' (see subsection 1.4.1 and (Simmonds et al. 2017)). The current definition of a virus, refined by the International Committee on Taxonomy of Viruses (ICTV), defines viruses as follows:

> *"Viruses sensu stricto are defined operationally by the ICTV as a type of MGE that encodes at least one protein that is a major component of the virion encasing the nucleic acid of the respective MGE and therefore the gene encoding the major virion protein itself or MGEs that are clearly demonstrable to be members of a line of evolutionary descent of such major virion protein-encoding entities. Any monophyletic group of MGEs that originates from a virion protein-encoding ancestor should be classified as a group of viruses."*
>
> - ICTV (Koonin et al. 2021; Kuhn et al. 2021)

In the same study, Koonin *et al.* introduce the compelling concept of the virosphere within the general space of replicators. It is divided into two sub categories: the orthovirusphere (true viruses) and the perivirosphere, a space surrounding the former. The authors place other selfishly replicating MGEs in the vicinity of these two spheres, thus creating a multidimensional concept that allows the placement of unique and exotic MGEs as well as traditional viruses (see Figure 1.3).

There is great diversity among viruses that belong to the orthovirusphere, namely *Monodnaviria*, *Riboviria*, *Varidnaviria*, *Adnaviria*, and *Robzyviria* (Dion et al. 2020). Examples of the particles and genomic organization of a virus infecting archaea and a virus infecting bacteria is given in Figure 1.4. The composition of the nucleic acids of the viral genome

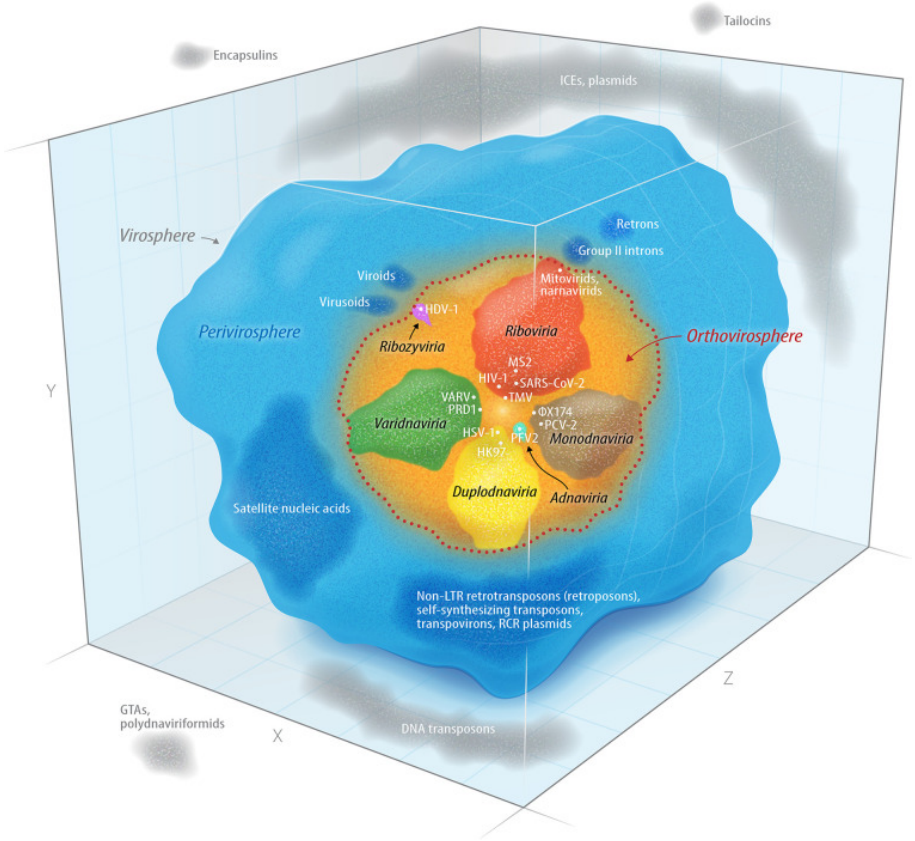


Figure 1.3: The replicator space, comprising the virosphere, which itself is divided into the perivirosphere and the orthovirosphere. In the center of the latter, true 'quintessential' viruses e.g. the first noncellular MGE identified Tobacco mosaic virus (TMV) are located. Further outside, MGEs that partially have lost virus-defining characteristics and for which the exact placements remains challenging (e.g. viriods). Outside of the virosphere, MGEs that are selfishly replicating, hence are within the replicator space, but are devoid of viral-like properties are located (e.g. plasmids, ICEs). Then, elements which are themselves not selfishly replicating are found outside the replicator space. Organelles and organisms, are not depicted for simplicity. Figure taken from (Koonin et al. 2021).

(single-stranded (ss) DNA, double-stranded (ds) DNA, dsRNA, positive (+) ssRNA, negativ (-) ssRNA), the presence or absence of a lipid envelope surrounding the viral capsid, the general morphology (helical, icosahedral, complex, pleomorphic), and lastly their lifestyle (lytic, lysogenic, chronic) can all vary drastically (Brock et al. 2003). Crucial for the description of viruses as drivers of HGT is their ability to transport the genetic material of the host in addition to their own genome within their capsid. This ability is closely connected to their lifestyle. Viruses following a lytic or productive life cycle will actively translate their genome, assemble virions and lyse their host in the process. Lysogenic viruses integrate into the host genome or exist extrachromosomally, replicating their genome alongside the host without producing new virions. In a chronic infection, the virus steadily produces virions without killing the host. Temperate phages can switch between a productive (chronic or lytic) and a dormant, non-productive (lysogenic) life cycle (Howard-Varona et al. 2017). Chronic infections were thought to be rare cases. However, recent studies reveal that such constant, non-lytic infections are more prevalent than thought (Alarcón-Schumacher et al. 2022), an oversight caused by the methodology often used for virus discovery: chronic and lysogenic infections are not always visible in traditional plaque assays, thus making lytic infections way more likely to be identified experimentally. All types of active infections pose the possibility of host genetic material being packaged into the virion. This could either be genes directly associated with an integrated virus genome or random fragments of mobilized DNA. The former is termed specialized and the latter generalized transduction, both of which are commonly viewed as missteps, either in excision or packaging of the virus genome (Chiang et al. 2019). Recently, a third form of transduction was described: lateral transduction. Here, larger chunks of DNA are packaged at a higher rate into the capsid, not as the result of a mistake but rather as a feature of the virus itself (Chen et al. 2018; Davidson 2018).

1.2.2 Gene Transfer Agents

Gene transfer agents (GTA) are morphologically similar to head-tail viruses. First isolated from *Rhodobacter capsulatus*, electron microscopy images showed a capsid head of ∼30 nm and a longer tail of ∼50 nm (Lang et al. 2012, 2017). To date, multiple types of GTAs have been identified: The mentioned RcGTA of *R. capsulatus* (see Figure 1.4 for the genomic organization and particle structure), Dd1 in *Desulfovibrio desulfuricans*, VTA in *Methanococcus voltae*, VSH-1 in *Serpulina hyodysenteriae* and BaGTA in the genus of *Bartonella* (Marrs 1974; Matson et al. 2005; Zhao et al. 2009; Lang et al. 2012; Québatte and Dehio 2019). All of them appear to be lytic, thus mimicking the life cycle of bacteriophages. However, contrary to viruses, the genetic material packaged inside the capsid is not specifically the capsid-, tail- or replication-encoding genes that produce the GTA particles, but rather random fragments of host DNA. The fragments packaged vary in size, but the small capsid size limits the fragments to ∼4.5 kbp in the case of RcGTA (Bárdy et al. 2020). The genes necessary for the production of GTA particles are located on the host's main chromosome, organized in clusters and subclusters, often spread out over the chromosome. Genes encoding for a terminase, major capsid protein, tail tube proteins and

proteinases are core proteins of GTAs and their closest homologs are their viral counterparts (Lang and Beatty 2000). Therefore, GTAs resemble a case of 'domesticated' viruses, which lost their selfish replicator characteristics but provide a mechanism of horizontal gene transfer for their host and, therefore, a fitness advantage for the population. For example, BcGTA is believed to have increased the genomic diversity of the population by inducing recombination events, thus increasing the adaptability to environmental changes (Québatte and Dehio 2019). GTAs therefore cannot be considered selfish elements within a replicator space, but rather a tool for HGT within a population, similar to the production of extracellular vesicles (see subsection 1.2.4).

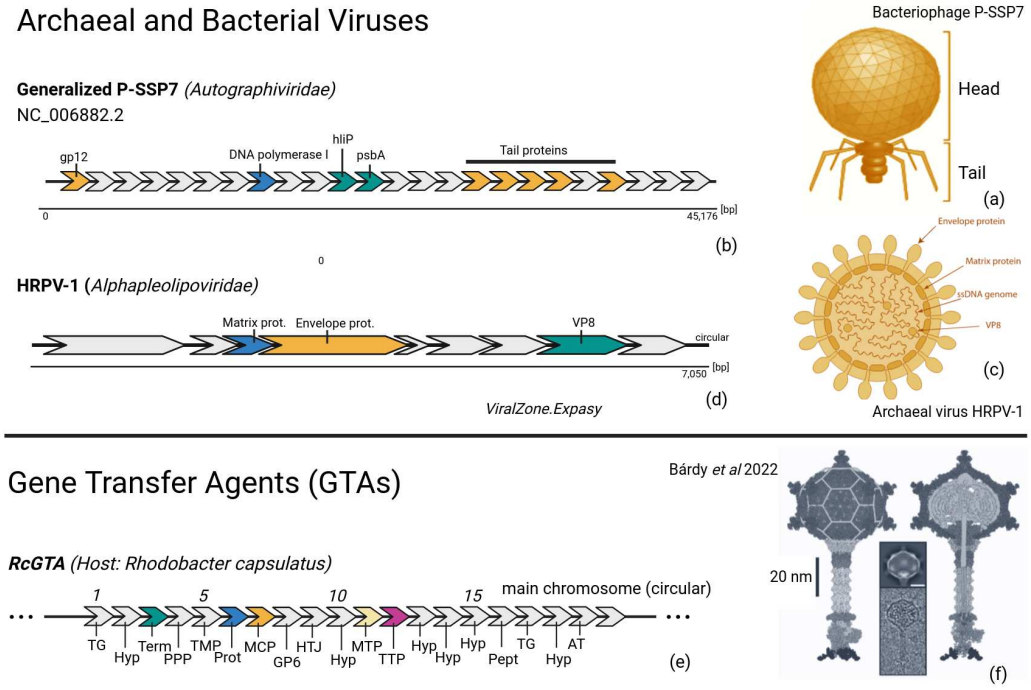


Figure 1.4: Overview of the drivers of horizontal gene transfer. (a) Head-tail structure (credit: `ViralZone.Expasy`) and (b) genome map of bacteriophage P-SSP7. (c) Structure and d) genome map of Archaeal virus HRPV1. (f) Structure (Bárdy et al. 2020) of *Rhodobacter capsulatus* gene transfer agent particles and (e) the corresponding genome map. AT acetyltransferase, MBP membrane bound protein, TG transglycosylase, Pept phage cell wall peptidase, TMP transmembrane protein, Hyp hypothetical, TAP tail assembly protein, TTP phage tail tube protein, MTP phage major tail protein, HTJ head-tail joining protein, GP6 gp6-like protein, MCP major capsid protein, Prot proteinase, PPP phage portal protein. Figure created with `biorender.com`.

1.2.3 Plasmids

Plasmids are extrachromosomal replicons. In archaea and bacteria, they occur in linear or circular form, and their size ranges from a few kilobases (Rozhon et al. 2010; Biet et al. 2002) to over two megabases (Hall et al. 2021; Almpanis et al. 2018), though typically less than 5 % of the size of a chromosome (Brock et al. 2003). Thousands of plasmids are described in the prokaryotic world, predominantly on the bacterial side. In 1978, the

first plasmid of *Halobacterium salinarium* was described, opening the door to the fascinating world of archaeal plasmids and their crucial role in HGT (Prangishvili et al. 1998; Wang et al. 2015). An example is given in Figure 1.5, where a genome map of the plasmid pKEF9 of *Sulfolobus islandicus* is shown. Plasmids are replicated in bacteria through theta-replication or rolling circle replication (RCR) (Del Solar et al. 1998), while for archaea additional replication mechanisms are known: replication initiation proteins, homologes to eukaryotic CDC6/Orc1-like proteins, are either recruited from the main chromosome, or directly found on the plasmid itself, and used for replication (Bell 2012; Erdmann et al. 2017). After replication, plasmids can occur in higher copy numbers than the corresponding chromosome of the host (Friehs 2004), which influences the plasmid's propagation. Plasmids with low copy numbers (<5), therefore rely on partitioning systems or toxin-antitoxin (TA) systems in order to ensure that each daughter cell contains a copy of the plasmid (Baxter and Funnell 2014; Díaz-Orejas et al. 2017). If the daughter cell does not receive the complete set of toxin-antitoxin genes, the daughter cell is killed, thus ensuring only plasmid-carrying cells are replicated. Additionally, conjugative plasmids are disseminated by direct cell-to-cell contact, making plasmids main drivers of HGT (see subsection 1.1.3). By definition, plasmids do not carry essential genes for the host to survive (Tazzyman and Bonhoeffer 2015), however, there is growing evidence that this clear-cut definition is inadequate (Harrison et al. 2010). Once again, nature does not follow strict definitions, and elements appear as distinct entities, but instead on a spectrum of characteristics.

1.2.4 Extracellular Vesicles

Organisms of domains of life produce small (10-300 µm), spherical, membrane-derived extracellular vesicles (EVs) (Deatherage and Cookson 2012). Bacteria and archaea employ several different mechanisms of EV formation. Since EVs are membrane-derived, EV formation mechanisms seemingly correlate with cell wall composition; however, the underlying biogenesis mechanisms, as do the associated genes, often remain unraveled. Like viral life cycles, EV formation mechanisms can generally be categorized into destructive cell-lysing and non-destructive blebbing mechanisms, with EV formation in archaea falling into the latter category (Gill et al. 2019; Toyofuku et al. 2023). Different archaeal lineages use two main machineries for EV formation. Most members of the DPANN *superphylum, Thaumarchaeota, Korarchaeota* and *Euryarchaeota* employ homologs of the eukaryotic ESCRT cell division system, and *Thermoproteota* and *Asgardarchaeota* employ homologs to bacterial FtSZ systems (van Wolferen et al. 2022). While the genes involved in EV formation have been identified for some species, an extensive set of potential 'marker genes' has yet to be established. The functions and interactions of extracellular vesicles are manifold. In cell-to-cell communication, removal of toxic substances from within the cell, stimulation of immune responses in eukaryotic cells in proximity, the transport of RNA or DNA and as a supply of nutrients for the population, EVs are used whenever the host needs to disseminate molecules, in a concentrated, protected way (Deatherage and Cookson 2012 and the references therein). Particularly, EVs enable the transport of molecules in biologically relevant concentrations, a distinct advantage over other secretion and transport systems

(Toyofuku et al. 2023). The function of EVs is, therefore, closely coupled with their cargo. Over the last decade, the transport of nucleic acids became the focus of environmental studies. Biller *et al.* showed that two Prochlorococcus strains, one of the most abundant bacteria in the oceans and likely the world, produce vast amounts of vesicles, which often contain fragments of host DNA (Biller et al. 2014, 2017; Linney et al. 2021, 2022). Additionally, it was shown that these DNA fragments are taken up by other cells, providing an inheritable genetic change to the recipient (Bitto et al. 2017; Grüll et al. 2018; Gill et al. 2019; Biller et al. 2022). Most recently, Hackl *et al.* demonstrated that EVs harbor novel integrative MGEs termed 'tycheposons', once again manifesting EVs as major drivers in HGT in the oceans (Hackl et al. 2023). Therefore, EVs are not mobile genetic elements or other selfishly replicating entities, such as e.g. viruses, but rather a vehicle that mediates the transfer of genetic material from one cell to another. Chapter 4 describes the first analysis of public, previously virus-focused datasets from an EV perspective. In contrast, section 3.3 describes the challenges of detecting EV-mediated horizontal gene transfer.

1.2.5 A special case: Plasmid Vesicles and 'apHPVs'

In 2017, Erdmann *et al.* reported the discovery and successful isolation of a plasmid with a curious set of characteristics. The ∼50 kbp circular plasmid pR1SE of the halophilic archaeon *Halorubrum lacusprofundi* R1S1 (see Figure 1.5), was shown to be transferred from one cell to another in membrane-derived vesicles (Erdmann et al. 2017). The transport of plasmids inside extracellular vesicles has been reported before (Gaudin et al. 2014; Altan-Bonnet 2016). What made this case so intriguing is that the plasmid itself encoded several putative membrane proteins, which were detected to be present in the isolated membrane vesicles. The resulting, plasmid-containing vesicles were more regularly shaped in comparison to 'typical' EVs of the same organism. The vesicles were termed 'plasmid vesicles' (PVs). Another exciting aspect of PVs was discovering that plasmid-free strains could be infected with isolated PVs. An infected host would then start the production of PVs, without an apparent lysis of the host cell, thus mimicking the viral lifestyle of a chronic infection. Others have subsequently speculated on the evolutionary link PVs might represent: Is it a plasmid that, over time, recruited membrane proteins in order to guarantee further dissemination? Or is it, in fact, a new family of viruses, similar to pleomorphic virus HHPV, which also infects halophilic archaea (Forterre et al. 2017)? In any case, it questioned the existing nomenclature and filled another hole within the space of replicators. Chapter 2 describes the metagenomic discovery of pR1SE-like elements, which were coined 'apHPVs', for archaeal plasmids of *haloarchaea* potentially transferred in plasmid vesicles. apHPVs seem to be distributed globally, and a general genetic organization is conserved, while the single proteins underwent strong diversification on a sequence level. Erdmann *et al.* demonstrated that pR1SE was not only able to infect a plasmid-free strain but also found multiple configurations of the plasmid where it had apparently picked up host material. In Chapter 2 we found additional evidence for such a highly recombination-dependent and 'transducing' lifestyle. *Haloarchaea* usually appear in hypersaline and sometimes psychrophile environments, which naturally reduces the diversity of the microbial community

since only very few specialists can thrive in such environments. For example, in the Antarctic 'Deep Lake', where pR1SE was first isolated from, the majority of the community is composed of very few *haloarchaea* species. apHPVs might, therefore, be a mechanism to introduce a higher diversity in an otherwise relatively homogenous community, improving its ability to adapt to changes. Others have already demonstrated widespread intergenera gene-exchange in the same community (Tschitschko et al. 2018). Whether apHPVs and apHPVs-like elements are the reason or the result of drastic genomic flexibility is not clear. Nevertheless, apHPVs are strong drivers of HGT *haloarchaea*.

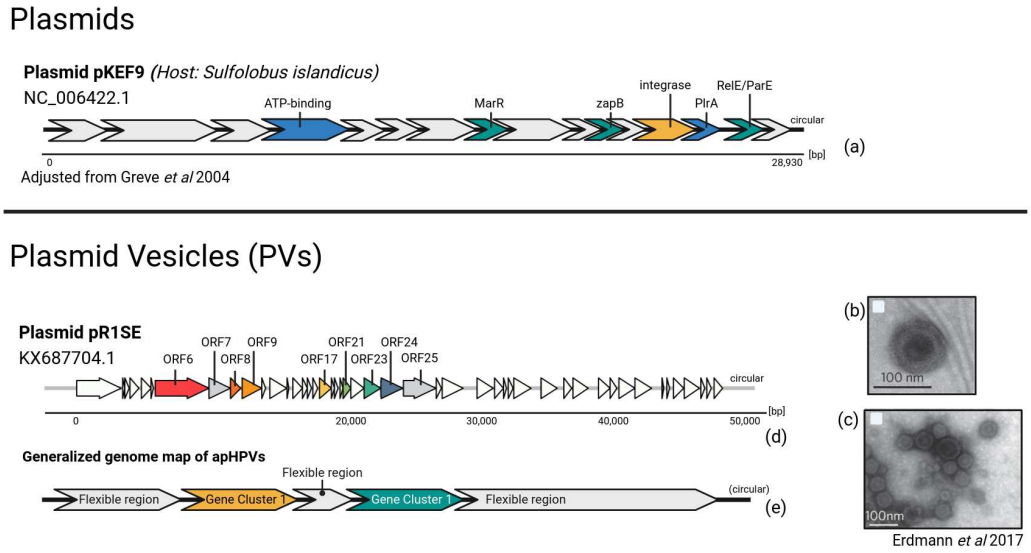


Figure 1.5: (a) Genome map of plasmid pKEF9 of *Sulfolobus islandicus*. (b-c) transmission electron micrographs of pR1SE plasmid vesicles (Erdmann et al. 2017). (d) Genome map of plasmid pR1SE of *Halorubrum lacusprofundi* and (e) the generalized genome map of apHPVs (Chapter 2). Figure created with `biorender.com`.

1.3 Interactions of Mobile Genetic Elements and HGT

Mobile genetic elements should be seen as agents with an interest in their own evolutionary success instead of mere traits or passengers of their hosts. This opens up an interesting perspective on the interactions of MGEs and interactions of MGEs and their environment (Horne et al. 2023): '*The enemy of my enemy is my friend*' becomes all too applicable in a world in which an MGE that provides an anti-viral defense system, will be kept by a host, even if the maintenance of the MGE comes with a certain cost. The possibilities of interactions are numerous (Koonin et al. 2020; Horne et al. 2023); therefore, we will focus on some key cases with fundamental implications specifically for HGT.

Transposons enable the packaging of DNA in extracellular vesicles. As described in subsection 1.2.4, extracellular vesicles can package and deliver DNA to a recipient cell. However, the exact packaging mechanism remains unclear. While there is some evidence that DNA closer to the origin of replication is packaged more frequently than other parts

of the genome (Biller et al. 2014), no apparent bias in packaging has been identified. In Chapter 4 we find that most of the DNA packaged inside EVs is either similar to transposons or genes associated with genomic islands. We hypothesize that transposons mobilize themselves and their surrounding genes, so-called passenger genes (Pál and Papp 2013), frequently making regions with high transposon activity stochastically more likely to be packaged. From an evolutionary perspective, the linkage of an HGT mechanism, such as EV-mediated DNA transport and the activity of MGEs is unsurprising, though notable. Both work seemingly hand in hand to expand the pangenome of a community without risking major changes to the conserved regions of a genome.

apHPVs could provide anti-viral defense systems to their host. The continuous competition between organisms and parasites is one of the most ancient struggles of life (Koonin and Dolja 2013). The inevitable emergence of such a competition is fundamental: Without parasites, there is no stable existence of life (Szathmáry and Demeter 1987; Koonin et al. 2017). A defenseless host, exposed to its inevitable parasites, is similarly unstable. Thus, the emergence of anti-parasitic defense systems must be an ancient development of life (Koonin 2016). The subsequent emergence of MGEs providing defense systems to their host, described by Koonin *et al.* as 'guns for hire', leads to interesting and diverse interactions between MGEs. In Chapter 2 we describe apHPVs, which provide a fully functioning CRISPR-cas system to their host. For example, the apHPV `'AOIO01000049.1'` of *Natrialab asiatica* transports Type I-D CRISPR module, making this plasmid the perfect 'mercenary' for the host. In theory, if a virus infects the population of *N. asiatica*, cells that contain the plasmid will have a significant fitness advantage. On the other hand, the moment the infection is dealt with, the plasmid becomes an energetic burden to the host. One can imagine how easily the interactions become interwoven and complex, even in such a simple three-actor network.

1.4 Bioinformatic Detection of Mobile Genetic Elements

1.4.1 Bioinformatic Detection and Analysis of Viruses

Viromics - or viral metagenomics - is an established field of research within modern microbiology that focuses on the comprehensive study of viral communities based on sequence data (Breitbart et al. 2002; Edwards and Rohwer 2005). Unlike traditional virology, which typically isolates and studies individual viruses, viromics seeks to understand the diversity, dynamics, and ecological roles of viruses by analyzing their genetic material within complex environmental samples. Depending on the experimental goals, scientists may enrich their sample in viruses, e.g. by size fractionation. In order to analyze the resulting pile of sequence data, virus-centric bioinformatic methods and programs were developed. This included tools for virus discovery in metagenomes (Ren et al. 2017, 2020; Guo et al. 2021), virus genome annotation (Bouras et al. 2023), host prediction (Gregory et al. 2020; Coutinho et al. 2021; Roux et al. 2023), lifestyle classification (McNair et al. 2012), virus taxonomy (Meier-Kolthoff and Göker 2017; Aiewsakun and Simmonds 2018; Bin Jang et al.

2019) and specialized assembly (Antipov et al. 2020) and binning tools (Johansen et al. 2022). Please note, that the tools listed here represent only a selection of tools. One of the fundamental goals of viromics is the *de novo* discovery of viruses within metagenomes. Over the years, standards and best practices for bioinformatic virus detection were established (Roux et al. 2019). Viromics led to the discovery of thousands of novel viral lineages in various environments and allowed ecological insights into communities which were previously inaccessible by traditional methods. In Chapter 3, the various approaches of state-of-the-art virus detection tools and potential future developments are discussed.

1.4.2 Bioinformatic Detection and Analysis of Plasmids

Similar to detecting novel viruses within metagenomic data, the reconstruction and detection of plasmids requires the development of specialized bioinformatic tools. The ever-growing number of large metagenomic datasets requires efficient, high-throughput workflows. Multiple approaches have been developed over time: reference-dependent workflows (Royer et al. 2018), secondary sequence attributes (Schwengers et al. 2020) or neural networks trained on k-mer distribution and other genomic signatures (Krawczyk et al. 2018). In Chapter 3 the different approaches and the workflows that apply them are described.

1.4.3 Detection of other Mobile Genetic Elements

The power of bioinformatic approaches for the detection of MGEs lies in their scalability. Searching billions of sequences for a specific sequence signature enables researchers to find 'the needle in the haystack' more often than not. However, not all MGEs are similar, so there is not one approach or tool to detect them. Most programs work for a specific MGE or a specific family of MGEs. Exotic or recently identified MGEs in particular, require a carefully curated custom workflow. However, even the sequences of extremely closely related MGEs can vary drastically. For example, the proteins of apHPVs share only around ∼25 % amino acid identity (AAI) (see Chapter 2). Therefore, scientists develop heuristics, evolution-aware sequence profiles and use metadata to detect their sequence of interest (Pfeifer et al. 2021; Alarcón-Schumacher et al. 2023; Chapter 2). Nevertheless, if the target sequence presumably 'looks' exactly like normal microbial DNA, all these approaches become inherently unsuitable. This was the challenge tackled in see Chapter 4: DNA transported in EVs looks just like host DNA, because it is host DNA. Detecting EV-mediated horizontal gene transfer events in metagenomics data, required a combination of specific data generation protocols and a novel approach. The development of this approach is discussed in Chapter 3, while the resulting workflow, the application of the workflow to large datasets and the resulting ecological implications are discussed in Chapter 4.

1.5 Aims of the Dissertation

Horizontal gene transfer is a fundamental feature of all forms of life. The different mechanisms and entities that drive the exchange of genetic material between organisms are manifold. In this work, I aimed to illuminate one particular case of an HGT-driving entity

(apHPVs) and one understudied and novel HGT-driving mechanism (EV-mediated transfer of DNA). The discovery of the plasmid vesicle pR1SE in Hrr. *lacusprofundi* and its unique set of characteristics filled a gap between viruses and plasmids. Therefore, the following objectives were the center of Chapter 2:

1. Detection of pR1SE-like elements in public and specifically generated datasets.
2. Establishing a core set of proteins shared between the detected relatives.
3. Illuminating the functions of the established core set of proteins.
4. Analyzing the life cycle of pR1SE-elements.
5. Disentangling the plasmid/virus-like nature of pR1SE.

While EV-mediated HGT is an established mechanism, the scale and prevalence of this transport mechanism is yet to be described. The following objectives are tackled in the second and third chapters:

1. Reviewing existing plasmid and virus identification tools (Chapter 3) .
2. possible approaches to circumvent the ostensible impossibility of deciphering EV-mediated HGT (Chapter 3).
3. Establishing the sequence space of the protected extracellular sequence space (Chapter 4).
4. Identifying suitable methods to generate datasets from protected extracellular DNA (Chapter 4).
5. Analyzing HGT coverage patterns (Chapter 4).
6. Semi-quantitative estimation of EV-mediated transfer events.

References

Acar Kirit, H., J. P. Bollback, and M. Lagator (2022): "The Role of the Environment in Horizontal Gene Transfer," *Molecular Biology and Evolution*, 39, msac220.

Aiewsakun, P. and P. Simmonds (2018): "The genomic underpinnings of eukaryotic virus taxonomy: creating a sequence-based framework for family-level virus classification," *Microbiome*, 6, 38.

Alarcón-Schumacher, T., D. Lücking, and S. Erdmann (2023): "Revisiting evolutionary trajectories and the organization of the Pleolipoviridae family," *PLOS Genetics*, 19, e1010998, publisher: Public Library of Science.

Alarcón-Schumacher, T., A. Naor, U. Gophna, and S. Erdmann (2022): "Isolation of a virus causing a chronic infection in the archaeal model organism Haloferax volcanii reveals antiviral activities of a provirus," *Proceedings of the National Academy of Sciences of the United States of America*, 119, e2205037119.

Almpanis, A., M. Swain, D. Gatherer, and N. McEwan (2018): "Correlation between bacterial G+C content, genome size and the G+C content of associated plasmids and bacteriophages," *Microbial Genomics*, 4, e000168.

Altan-Bonnet, N. (2016): "Extracellular vesicles are the Trojan horses of viral infection," *Current Opinion in Microbiology*, 32, 77–81.

Antipov, D., M. Raiko, A. Lapidus, and P. A. Pevzner (2020): "MetaviralSPAdes: assembly of viruses from metagenomic data," *Bioinformatics*, 36, 4126–4129.

Arnold, B. J., I.-T. Huang, and W. P. Hanage (2022): "Horizontal gene transfer and adaptive evolution in bacteria," *Nature Reviews Microbiology*, 20, 206–218, number: 4 Publisher: Nature Publishing Group.

Baxter, J. C. and B. E. Funnell (2014): "Plasmid Partition Mechanisms," *Microbiology Spectrum*, 2, 10.1128/microbiolspec.plas–0023–2014, publisher: American Society for Microbiology.

Bell, S. D. (2012): "Archaeal Orc1/Cdc6 Proteins," in *The Eukaryotic Replisome: a Guide to Protein Structure and Function*, ed. by S. MacNeill, Dordrecht: Springer Netherlands, Subcellular Biochemistry, 59–69.

Beltran, L. C., V. Cvirkaite-Krupovic, J. Miller, F. Wang, M. A. B. Kreutzberger, J. B. Patkowski, T. R. D. Costa, S. Schouten, I. Levental, V. P. Conticello, E. H. Egelman, and M. Krupovic (2023): "Archaeal DNA-import apparatus is homologous to bacterial conjugation machinery," *Nature Communications*, 14, 666, number: 1 Publisher: Nature Publishing Group.

Biet, F., Y. Cenatiempo, and C. Fremaux (2002): "Identification of a Replicon from pTXL1, a Small Cryptic Plasmid from Leuconostoc mesenteroides subsp. mesenteroides Y110, and Development of a Food-Grade Vector," *Applied and Environmental Microbiology*, 68, 6451–6456, publisher: American Society for Microbiology.

Biller, S. J., A. Coe, A. A. Arellano, K. Dooley, J. S. Gong, E. A. Yeager, J. W. Becker, and S. W. Chisholm (2022): "Environmental and taxonomic drivers of bacterial extracellular vesicle production in marine ecosystems," Pages: 2022.01.18.476865 Section: New Results.

Biller, S. J., L. D. McDaniel, M. Breitbart, E. Rogers, J. H. Paul, and S. W. Chisholm (2017): "Membrane vesicles in sea water: heterogeneous DNA content and implications for viral abundance estimates," *The ISME Journal*, 11, 394–404, number: 2 Publisher: Nature Publishing Group.

Biller, S. J., F. Schubotz, S. E. Roggensack, A. W. Thompson, R. E. Summons, and S. W. Chisholm (2014): "Bacterial Vesicles in Marine Ecosystems," *Science*, 343, 183–186, publisher: American Association for the Advancement of Science.

Bin Jang, H., B. Bolduc, O. Zablocki, J. H. Kuhn, S. Roux, E. M. Adriaenssens, J. R. Brister, A. M. Kropinski, M. Krupovic, R. Lavigne, D. Turner, and M. B. Sullivan (2019): "Taxonomic assignment of uncultivated prokaryotic virus genomes is enabled by gene-sharing networks," *Nature Biotechnology*, 37, 632–639, number: 6 Publisher: Nature Publishing Group.

Bitto, N. J., R. Chapman, S. Pidot, A. Costin, C. Lo, J. Choi, T. D'Cruze, E. C. Reynolds, S. G. Dashper, L. Turnbull, C. B. Whitchurch, T. P. Stinear, K. J. Stacey, and R. L. Ferrero (2017): "Bacterial membrane vesicles transport their DNA cargo into host cells," *Scientific Reports*, 7, 7072, number: 1 Publisher: Nature Publishing Group.

Bouras, G., R. Nepal, G. Houtak, A. J. Psaltis, P.-J. Wormald, and S. Vreugde (2023): "Pharokka: a fast scalable bacteriophage annotation tool," *Bioinformatics*, 39, btac776.

Breitbart, M., P. Salamon, B. Andresen, J. M. Mahaffy, A. M. Segall, D. Mead, F. Azam, and F. Rohwer (2002): "Genomic analysis of uncultured marine viral communities," *Proceedings of the National Academy of Sciences*, 99, 14250–14255, publisher: Proceedings of the National Academy of Sciences.

Brock, T. D., M. T. Madigan, J. M. Martinko, and J. Parker (2003): *Brock biology of microorganisms*, Upper Saddle River (NJ): Prentice-Hall, 2003.

Burmeister, A. R. (2015): "Horizontal Gene Transfer," *Evolution, Medicine, and Public Health*, 2015, 193–194.

Bárdy, P., T. Füzik, D. Hrebík, R. Pantůček, J. Thomas Beatty, and P. Plevka (2020): "Structure and mechanism of DNA delivery of a gene transfer agent," *Nature Communications*, 11, 3034, number: 1 Publisher: Nature Publishing Group.

Catchpole, R. J., V. Barbe, G. Magdelenat, E. Marguet, M. Terns, J. Oberto, P. Forterre, and V. Da Cunha (2023): "A self-transmissible plasmid from a hyperthermophile that facilitates genetic modification of diverse Archaea," *Nature Microbiology*, 8, 1339–1347.

Chen, J., N. Quiles-Puchalt, Y. N. Chiang, R. Bacigalupe, A. Fillol-Salom, M. S. J. Chee, J. R. Fitzgerald, and J. R. Penadés (2018): "Genome hypermobility by lateral transduction," *Science*, 362, 207–212, publisher: American Association for the Advancement of Science.

Chiang, Y. N., J. R. Penadés, and J. Chen (2019): "Genetic transduction by phages and chromosomal islands: The new and noncanonical," *PLoS Pathogens*, 15, e1007878.

Coutinho, F. H., A. Zaragoza-Solas, M. López-Pérez, J. Barylski, A. Zielezinski, B. E. Dutilh, R. Edwards, and F. Rodriguez-Valera (2021): "RaFAH: Host prediction for viruses of Bacteria and Archaea based on protein content," *Patterns*, 2, 100274.

Davidson, A. R. (2018): "A common trick for transferring bacterial DNA," *Science*, 362, 152–153, publisher: American Association for the Advancement of Science.

Deatherage, B. L. and B. T. Cookson (2012): "Membrane vesicle release in bacteria, eukaryotes, and archaea: a conserved yet underappreciated aspect of microbial life," *Infection and Immunity*, 80, 1948–1957.

Del Solar, G., R. Giraldo, M. J. Ruiz-Echevarría, M. Espinosa, and R. Díaz-Orejas (1998): "Replication and control of circular bacterial plasmids," *Microbiology and molecular biology reviews*, 62, 434–464, publisher: Am Soc Microbiol.

Diener, T. O. and W. B. Raymer (1967): "Potato Spindle Tuber Virus: A Plant Virus with Properties of a Free Nucleic Acid," *Science*, 158, 378–381, publisher: American Association for the Advancement of Science.

Dion, M. B., F. Oechslin, and S. Moineau (2020): "Phage diversity, genomics and phylogeny," *Nature Reviews Microbiology*, 18, 125–138, number: 3 Publisher: Nature Publishing Group.

Doolittle, W. F. (1999): "Phylogenetic Classification and the Universal Tree," *Science*, 284, 2124–2128, publisher: American Association for the Advancement of Science.

Díaz-Orejas, R., M. Espinosa, and C. C. Yeo (2017): "The Importance of the Expendable: Toxin–Antitoxin Genes in Plasmids and Chromosomes," *Frontiers in Microbiology*, 8, 1479.

Edwards, R. A. and F. Rohwer (2005): "Viral metagenomics," *Nature Reviews. Microbiology*, 3, 504–510.

Erdmann, S., B. Tschitschko, L. Zhong, M. J. Raftery, and R. Cavicchioli (2017): "A plasmid from an Antarctic haloarchaeon uses specialized membrane vesicles to disseminate and infect plasmid-free cells," *Nature Microbiology*, 2, 1446–1455, number: 10 Publisher: Nature Publishing Group.

Feiss, M. and D. A. Siegele (1979): "Packaging of the bacteriophage lambda chromosome: Dependence of cos cleavage on chromosome length," *Virology*, 92, 190–200.

Finkel, S. E. and R. Kolter (2001): "DNA as a Nutrient: Novel Role for Bacterial Competence Gene Homologs," *Journal of Bacteriology*, 183, 6288–6293, publisher: American Society for Microbiology.

Forterre, P., V. Da Cunha, and R. Catchpole (2017): "Plasmid vesicles mimicking virions," *Nature Microbiology*, 2, 1340–1341, number: 10 Publisher: Nature Publishing Group.

Friehs, K. (2004): "Plasmid copy number and plasmid stability," *Advances in Biochemical Engineering/Biotechnology*, 86, 47–82.

Gaudin, M., M. Krupovic, E. Marguet, E. Gauliard, V. Cvirkaite-Krupovic, E. Le Cam, J. Oberto, and P. Forterre (2014): "Extracellular membrane vesicles harbouring viral genomes," *Environmental Microbiology*, 16, 1167–1175.

Gill, S., R. Catchpole, and P. Forterre (2019): "Extracellular membrane vesicles in the three domains of life and beyond," *FEMS Microbiology Reviews*, 43, 273–303.

Gorbalenya, A. E., M. Krupovic, A. Mushegian, A. M. Kropinski, S. G. Siddell, A. Varsani, M. J. Adams, A. J. Davison, B. E. Dutilh, B. Harrach, R. L. Harrison, S. Junglen, A. M. Q. King, N. J. Knowles, E. J. Lefkowitz, M. L. Nibert, L. Rubino, S. Sabanadzovic, H. Sanfaçon, P. Simmonds, P. J. Walker, F. M. Zerbini, J. H. Kuhn, and International Committee on Taxonomy of Viruses Executive Committee (2020): "The new scope of virus taxonomy: partitioning the virosphere into 15 hierarchical ranks," *Nature Microbiology*, 5, 668–674, number: 5 Publisher: Nature Publishing Group.

Gregory, A. C., O. Zablocki, A. A. Zayed, A. Howell, B. Bolduc, and M. B. Sullivan (2020): "The Gut Virome Database Reveals Age-Dependent Patterns of Virome Diversity in the Human Gut," *Cell Host & Microbe*, 28, 724–740.e8, publisher: Elsevier.

Grogan, D. W. (1996): "Exchange of genetic markers at extremely high temperatures in the archaeon Sulfolobus acidocaldarius," *Journal of Bacteriology*, 178, 3207–3211, publisher: American Society for Microbiology.

Grüll, M. P., M. E. Mulligan, and A. S. Lang (2018): "Small extracellular particles with big potential for horizontal gene transfer: membrane vesicles and gene transfer agents," *FEMS Microbiology Letters*, 365, fny192.

Guo, J., B. Bolduc, A. A. Zayed, A. Varsani, G. Dominguez-Huerta, T. O. Delmont, A. A. Pratama, M. C. Gazitúa, D. Vik, M. B. Sullivan, and S. Roux (2021): "VirSorter2: a multi-classifier, expert-guided approach to detect diverse DNA and RNA viruses," *Microbiome*, 9, 37.

Hackl, T., R. Laurenceau, M. J. Ankenbrand, C. Bliem, Z. Cariani, E. Thomas, K. D. Dooley, A. A. Arellano, S. L. Hogle, P. Berube, G. E. Leventhal, E. Luo, J. M. Eppley, A. A. Zayed, J. Beaulaurier, R. Stepanauskas, M. B. Sullivan, E. F. DeLong, S. J. Biller, and S. W. Chisholm (2023): "Novel integrative elements and genomic plasticity in ocean ecosystems," *Cell*, 186, 47–62.e16.

Hall, J. P. J., J. Botelho, A. Cazares, and D. A. Baltrus (2021): "What makes a megaplasmid?" *Philosophical Transactions of the Royal Society B: Biological Sciences*, 377, 20200472.

Hall, R. J., F. J. Whelan, J. O. McInerney, Y. Ou, and M. R. Domingo-Sananes (2020): "Horizontal Gene Transfer as a Source of Conflict and Cooperation in Prokaryotes," *Frontiers in Microbiology*, 11.

Harrison, P. W., R. P. J. Lower, N. K. D. Kim, and J. P. W. Young (2010): "Introducing the bacterial 'chromid': not a chromosome, not a plasmid," *Trends in Microbiology*, 18, 141–148, publisher: Elsevier.

Horne, T., V. T. Orr, and J. P. Hall (2023): "How do interactions between mobile genetic elements affect horizontal gene transfer?" *Current Opinion in Microbiology*, 73, 102282.

Howard-Varona, C., K. R. Hargreaves, S. T. Abedon, and M. B. Sullivan (2017): "Lysogeny in nature: mechanisms, impact and ecology of temperate phages," *The ISME Journal*, 11, 1511–1520.

Johansen, J., D. R. Plichta, J. N. Nissen, M. L. Jespersen, S. A. Shah, L. Deng, J. Stokholm, H. Bisgaard, D. S. Nielsen, S. J. Sørensen, and S. Rasmussen

(2022): "Genome binning of viral entities from bulk metagenomics data," *Nature Communications*, 13, 965, number: 1 Publisher: Nature Publishing Group.

Johnsborg, O., V. Eldholm, and L. S. Håvarstein (2007): "Natural genetic transformation: prevalence, mechanisms and function," *Research in Microbiology*, 158, 767–778.

Johnston, C., B. Martin, G. Fichant, P. Polard, and J.-P. Claverys (2014): "Bacterial transformation: distribution, shared mechanisms and divergent control," *Nature Reviews Microbiology*, 12, 181–196, number: 3 Publisher: Nature Publishing Group.

Koonin, E. V. (2015): "Origin of eukaryotes from within archaea, archaeal eukaryome and bursts of gene gain: eukaryogenesis just made easier?" *Philosophical Transactions of the Royal Society B: Biological Sciences*, 370, 20140333, publisher: Royal Society.

——— (2016): "Viruses and mobile elements as drivers of evolutionary transitions," *Philosophical Transactions of the Royal Society B: Biological Sciences*, 371, 20150442, publisher: Royal Society.

Koonin, E. V. and V. V. Dolja (2013): "A virocentric perspective on the evolution of life," *Current Opinion in Virology*, 3, 546–557.

Koonin, E. V., V. V. Dolja, M. Krupovic, and J. H. Kuhn (2021): "Viruses Defined by the Position of the Virosphere within the Replicator Space," *Microbiology and Molecular Biology Reviews : MMBR*, 85, e00193–20.

Koonin, E. V., K. S. Makarova, Y. I. Wolf, and M. Krupovic (2020): "Evolutionary entanglement of mobile genetic elements and host defence systems: guns for hire," *Nature Reviews Genetics*, 21, 119–131, number: 2 Publisher: Nature Publishing Group.

Koonin, E. V., Y. I. Wolf, and M. I. Katsnelson (2017): "Inevitability of the emergence and persistence of genetic parasites caused by evolutionary instability of parasite-free states," *Biology Direct*, 12, 31.

Krawczyk, P., L. Lipinski, and A. Dziembowski (2018): "PlasFlow: predicting plasmid sequences in metagenomic data using genome signatures," *Nucleic acids research*, 46.

Krupovic, M., K. S. Makarova, Y. I. Wolf, S. Medvedeva, D. Prangishvili, P. Forterre, and E. V. Koonin (2019): "Integrated mobile genetic elements in Thaumarchaeota," *Environmental Microbiology*, 21, 2056–2078, _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/1462-2920.14564.

Kuhn, J., V. Dolja, M. Krupovic, E. Adriaenssens, F. Di Serio, B. Dutilh, R. Flores, B. Harrach, A. Mushegian, B. Owens, J. Randles, L. Rubino, S. Sabanadzovic, P. Simmonds, A. Varsani, F. Zerbini, and E. Koonin (2021): *Expand, amend, and emend the International Code of Virus Classification and Nomenclature (ICVCN; "the Code") and the Statutes to clearly define the remit of the ICTV*, .

Lang, A. S. and J. T. Beatty (2000): "Genetic analysis of a bacterial genetic exchange element: The gene transfer agent of Rhodobacter capsulatus," *Proceedings of the National Academy of Sciences*, 97, 859–864, publisher: Proceedings of the National Academy of Sciences.

Lang, A. S., A. B. Westbye, and J. T. Beatty (2017): "The Distribution, Evolution, and Roles of Gene Transfer Agents in Prokaryotic Genetic Exchange," *Annual Review of Virology*, 4, 87–104, _eprint: https://doi.org/10.1146/annurev-virology-101416-041624.

Lang, A. S., O. Zhaxybayeva, and J. T. Beatty (2012): "Gene transfer agents: phage-like elements of genetic exchange," *Nature reviews. Microbiology*, 10, 472–482.

Lindell, D., M. B. Sullivan, Z. I. Johnson, A. C. Tolonen, F. Rohwer, and S. W. Chisholm (2004): "Transfer of photosynthesis genes to and from Prochlorococcus viruses," *Proceedings of the National Academy of Sciences of the United States of America*, 101, 11013–11018.

Linney, M. D., J. M. Eppley, A. E. Romano, E. Luo, E. F. DeLong, and D. M. Karl (2022): "Microbial Sources of Exocellular DNA in the Ocean," *Applied and Environmental Microbiology*, 88, e02093–21, publisher: American Society for Microbiology.

Linney, M. D., C. R. Schvarcz, G. F. Steward, E. F. DeLong, and D. M. Karl (2021): "A method for characterizing dissolved DNA and its application to the North Pacific Subtropical Gyre," *Limnology and Oceanography: Methods*, 19, 210–221, _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/lom3.10415.

Lwoff, A. (1957): "The concept of virus," *Microbiology*, 17, 239–253, publisher: Citeseer.

Marrs, B. (1974): "Genetic Recombination in Rhodopseudomonas capsulata," *Proceedings of the National Academy of Sciences*, 71, 971–973, publisher: Proceedings of the National Academy of Sciences.

Matson, E. G., M. G. Thompson, S. B. Humphrey, R. L. Zuerner, and T. B. Stanton (2005): "Identification of genes of VSH-1, a prophage-like gene transfer agent of Brachyspira hyodysenteriae," *Journal of Bacteriology*, 187, 5885–5892.

McNair, K., B. A. Bailey, and R. A. Edwards (2012): "PHACTS, a computational approach to classifying the lifestyle of phages," *Bioinformatics (Oxford, England)*, 28, 614–618.

Meier-Kolthoff, J. P. and M. Göker (2017): "VICTOR: genome-based phylogeny and classification of prokaryotic viruses," *Bioinformatics*, 33, 3396–3404.

Pfeifer, E., J. A. Moura de Sousa, M. Touchon, and E. P. C. Rocha (2021): "Bacteria have numerous distinctive groups of phage–plasmids with conserved phage and variable plasmid gene repertoires," *Nucleic Acids Research*, 49, 2655–2673.

Prangishvili, D., S. V. Albers, I. Holz, H. P. Arnold, K. Stedman, T. Klein, H. Singh, J. Hiort, A. Schweier, J. K. Kristjansson, and W. Zillig (1998): "Conjugation in archaea: frequent occurrence of conjugative plasmids in Sulfolobus," *Plasmid*, 40, 190–202.

Pál, C. and B. Papp (2013): "From passengers to drivers," *Mobile Genetic Elements*, 3, e23617.

Québatte, M. and C. Dehio (2019): "Bartonella gene transfer agent: Evolution, function, and proposed role in host adaptation," *Cellular Microbiology*, 21, e13068, _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/cmi.13068.

Rao, V. B. and M. Feiss (2015): "Mechanisms of DNA Packaging by Large Double-Stranded DNA Viruses," *Annual Review of Virology*, 2, 351–378, _eprint: https://doi.org/10.1146/annurev-virology-100114-055212.

Raymann, K., C. Brochier-Armanet, and S. Gribaldo (2015): "The two-domain tree of life is linked to a new root for the Archaea," *Proceedings of the National Academy of Sciences*, 112, 6670–6675, publisher: Proceedings of the National Academy of Sciences.

Ren, J., N. A. Ahlgren, Y. Y. Lu, J. A. Fuhrman, and F. Sun (2017): "VirFinder: a novel k-mer based tool for identifying viral sequences from assembled metagenomic data," *Microbiome*, 5, 69.

Ren, J., K. Song, C. Deng, N. A. Ahlgren, J. A. Fuhrman, Y. Li, X. Xie, R. Poplin, and F. Sun (2020): "Identifying viruses from metagenomic data using deep learning," *Quantitative Biology*, 8, 64–77, publisher: Springer.

Roux, S., E. M. Adriaenssens, B. E. Dutilh, E. V. Koonin, A. M. Kropinski, M. Krupovic, J. H. Kuhn, R. Lavigne, J. R. Brister, A. Varsani, C. Amid, R. K. Aziz, S. R. Bordenstein, P. Bork, M. Breitbart, G. R. Cochrane, R. A. Daly, C. Desnues, M. B. Duhaime, J. B. Emerson, F. Enault, J. A. Fuhrman, P. Hingamp, P. Hugenholtz, B. L. Hurwitz, N. N. Ivanova, J. M. Labonté, K.-B. Lee, R. R. Malmstrom, M. Martinez-Garcia, I. K. Mizrachi, H. Ogata, D. Páez-Espino, M.-A. Petit, C. Putonti, T. Rattei, A. Reyes, F. Rodriguez-Valera, K. Rosario, L. Schriml, F. Schulz, G. F. Steward, M. B. Sullivan, S. Sunagawa, C. A. Suttle, B. Temperton, S. G. Tringe, R. V. Thurber, N. S. Webster, K. L. Whiteson, S. W. Wilhelm, K. E. Wommack, T. Woyke, K. C. Wrighton, P. Yilmaz, T. Yoshida, M. J. Young, N. Yutin, L. Z. Allen, N. C. Kyrpides, and E. A. Eloe-Fadrosh (2019): "Minimum Information about an Uncultivated Virus Genome (MIUViG)," *Nature Biotechnology*, 37, 29–37.

Roux, S., A. P. Camargo, F. H. Coutinho, S. M. Dabdoub, B. E. Dutilh, S. Nayfach, and A. Tritt (2023): "iPHoP: An integrated machine learning framework to maximize host prediction for metagenome-derived viruses of archaea and bacteria," *PLOS Biology*, 21, e3002083, publisher: Public Library of Science.

Royer, G., J. W. Decousser, C. Branger, M. Dubois, C. Médigue, E. Denamur, and D. Vallenet (2018): "PlaScope: a targeted approach to assess the plasmidome from genome assemblies at the species level," *Microbial Genomics*, 4, e000211.

Rozhon, W., E. Petutschnig, M. Khan, D. K. Summers, and B. Poppenberger (2010): "Frequency and diversity of small cryptic plasmids in the genus Rahnella," *BMC Microbiology*, 10, 56.

Schwengers, O., P. Barth, L. Falgenhauer, T. Hain, T. Chakraborty, and A. Goesmann (2020): "Platon: identification and characterization of bacterial plasmid contigs in short-read draft assemblies exploiting protein sequence-based replicon distribution scores," *Microbial Genomics*, 6, mgen000398.

Simmonds, P., M. J. Adams, M. Benkő, M. Breitbart, J. R. Brister, E. B. Carstens, A. J. Davison, E. Delwart, A. E. Gorbalenya, B. Harrach, R. Hull, A. M. Q. King, E. V. Koonin, M. Krupovic, J. H. Kuhn, E. J. Lefkowitz, M. L. Nibert, R. Orton, M. J. Roossinck, S. Sabanadzovic, M. B. Sullivan, C. A. Suttle, R. B. Tesh, R. A. van der Vlugt, A. Varsani, and F. M. Zerbini (2017): "Virus taxonomy in the age of metagenomics," *Nature Reviews Microbiology*, 15, 161–168, number: 3 Publisher: Nature Publishing Group.

Soucy, S. M., J. Huang, and J. P. Gogarten (2015): "Horizontal gene transfer: building the web of life," *Nature Reviews. Genetics*, 16, 472–482.

Spang, A., J. H. Saw, S. L. Jørgensen, K. Zaremba-Niedzwiedzka, J. Martijn, A. E. Lind, R. van Eijk, C. Schleper, L. Guy, and T. J. G. Ettema (2015): "Complex archaea that bridge the gap between prokaryotes and eukaryotes," *Nature*, 521, 173–179, number: 7551 Publisher: Nature Publishing Group.

Szathmáry, E. and L. Demeter (1987): "Group selection of early replicators and the origin of life," *Journal of Theoretical Biology*, 128, 463–486.

Tazzyman, S. J. and S. Bonhoeffer (2015): "Why There Are No Essential Genes on Plasmids," *Molecular Biology and Evolution*, 32, 3079–3088.

Thierauf, A., G. Perez, and a. S. Maloy (2009): "Generalized Transduction," in *Bacteriophages: Methods and Protocols, Volume 1: Isolation, Characterization, and Interactions*, ed. by M. R. Clokie and A. M. Kropinski, Totowa, NJ: Humana Press, Methods in Molecular Biology™, 267–286.

Toyofuku, M., S. Schild, M. Kaparakis-Liaskos, and L. Eberl (2023): "Composition and functions of bacterial membrane vesicles," *Nature Reviews Microbiology*, 21, 415–430, number: 7 Publisher: Nature Publishing Group.

Tschitschko, B., S. Erdmann, M. Z. DeMaere, S. Roux, P. Panwar, M. A. Allen, T. J. Williams, S. Brazendale, A. M. Hancock, E. A. Eloe-Fadrosh, and R. Cavicchioli (2018): "Genomic variation and biogeography of Antarctic haloarchaea," *Microbiome*, 6, 113.

van der Gulik, P., W. Hoff, and D. Speijer (2017): "In defence of the three-domains of life paradigm," *BMC Evolutionary Biology*, 17, 218.

van Wolferen, M., A. Pulschen, B. Baum, S. Gribaldo, and S.-V. Albers (2022): "The Cell Biology of Archaea," *Nature microbiology*, 7, 1744–1755.

Wang, H., N. Peng, S. A. Shah, L. Huang, and Q. She (2015): "Archaeal Extrachromosomal Genetic Elements," *Microbiology and Molecular Biology Reviews*, 79, 117–152, publisher: American Society for Microbiology.

Woese, C. R. (1998): "Default taxonomy: Ernst Mayr's view of the microbial world," *Proceedings of the National Academy of Sciences*, 95, 11043–11046, publisher: Proceedings of the National Academy of Sciences.

Woese, C. R., O. Kandler, and M. L. Wheelis (1990): "Towards a natural system of organisms: proposal for the domains Archaea, Bacteria, and Eucarya." *Proceedings of the National Academy of Sciences*, 87, 4576–4579, publisher: Proceedings of the National Academy of Sciences.

Zhao, Y., K. Wang, C. Budinoff, A. Buchan, A. Lang, N. Jiao, and F. Chen (2009): "Gene transfer agent (GTA) genes reveal diverse and dynamic Roseobacter and Rhodobacter populations in the Chesapeake Bay," *The ISME Journal*, 3, 364–373, number: 3 Publisher: Nature Publishing Group.

Zinder, N. D. and J. Lederberg (1952): "Genetic exchange in salmonella," *Journal of Bacteriology*, 64, 679–699, publisher: American Society for Microbiology.

Chapter 2

Distribution and implications of haloarchaeal plasmids disseminated in self-encoded plasmid vesicles

Dominik Lücking[1],
Tomàs Alarcón-Schumacher[1]
and Susanne Erdmann[1]

Manuscript submitted and currently under review at *Microorganisms*.
Supplementary material in Appendix A.

Author Contributions:
SE designed and led the study and collected the original samples from Australian salt lakes. DL performed the laboratory work and the bioinformatic analysis. TAS carried out the analysis of ORF6 and the assembly of Australian salt lake metagenomes. DL and SE performed the primary writing of the manuscript. All authors participated in the analysis and interpretation of the data and contributed to the writing of the manuscript.

[1]Max-Planck Institute for Marine Microbiology, Bremen, Germany

microorganisms

MDPI

Article

Distribution and implications of haloarchaeal plasmids disseminated in self-encoded plasmid vesicles

Dominik Lücking[1], Tomàs Alarcón-Schumacher[1] and Susanne Erdmann [1]*

[1] Max-Planck Institute for Marine Microbiology, 28359 Bremen, Germany

* Correspondence: Susanne Erdmann, Max-Planck-Institute for Marine Microbiology, Celsiusstrasse 1, 28359 Bremen, Germany, Email: serdmann@mpi-bremen.de

Abstract: Even though viruses and plasmids are both drivers of horizontal gene transfer, they differ fundamentally in their mode of transfer. Virus genomes are enclosed in virus capsids and are not dependent on cell-cell contacts for their dissemination. In contrast, transfer of plasmids most often requires the physical contact between cells. However, plasmid pR1SE of *Halorubrum lacusprofundi* is disseminated between cells, independent of cell-cell contacts, in specialized membrane vesicles that contain plasmid proteins. In this study we searched for pR1SE-like elements in public databases and a metagenomics dataset from Australian salt lakes, and identified 40 additional pR1SE-like elements in hypersaline environments worldwide. Herein, these elements are named apHPVs (archaeal plasmids of haloarchaea potentially transferred in plasmid vesicles). They share two set of closely related proteins with conserved synteny, strongly indicating an organization into different functional cluster. We find that apHPVs, besides transferring themselves, have the potential to transfer large fragments of DNA between host cells, including virus defense systems. Most interestingly, apHPVs likely play an important role for the evolution of viruses and plasmids in haloarchaea, as they appear to recombine with both of them. This further supports that plasmids and viruses are not distinct but closely related mobile genetic elements.

Keywords: archaea, metagenomics, mobile genetic elements, plasmids, haloarchaea

Citation: To be added by editorial staff during production.

Academic Editor: Firstname Lastname

Received: date
Revised: date
Accepted: date
Published: date

1. Introduction

The genomes of bacteria and archaea are subjected to constant change. Vertical gene transfer and mutations allow the continuous adaptation of organisms to their environment. Horizontal gene transfer (HGT) events introduce additional flexibility by transferring larger segments of genetic material between organisms, sometimes even across domain borders [1, 2]. These events are most often driven by mobile genetic elements (MGEs). A plethora of MGEs such as viruses, plasmids, transposons,retrotransposons and gene transfer agents (GTAs) shape microbial communities at a fundamental level. Traditionally, these MGEs have been described as distinct entities. However, the discovery of elements that break traditional definitions, lead to the view of a more continuous sequence space [3].

In 2017 Erdmann *et al* isolated and characterized the ~50 kbp MGE, plasmid pR1SE of the halophilic archaeon (haloarchaeon) *Halorubrum lacusporufundi* R1S1 [4]. Plasmids are usually transferred between cells as unprotected DNA or by conjugation, requiring cell-to-cell contact. However, pR1SE is transferred by a mechanism that is more similar to the dissemination of viruses. *Hrr. lacusprofundi* produces extracellular vesicles (EVs) as do other haloarchaea [5]. Plasmid pR1SE, if present, is enclosed into EV-like structures, which then are taken up by plasmid-free strains, thus transferring pR1SE without direct cell-to-cell contact. While transfer of plasmids in EVs was observed previously in archaea, these plasmid-containing vesicles were found to contain exclusively host proteins [6]. However, pR1SE-containing EVs were shown to include pR1SE-encoded proteins that are

proposed to be involved in the formation of these vesicles. We consider them plasmid vesicles (PVs), as they can be used to infect a plasmid-free strain that subsequently produces PVs without cell lysis, thus mimicking the life cycle of a chronic viral infection [7]. Plasmid proteins included in PVs were not found to be homologous to any known virus capsid protein, and also did not show similarities to proteins involved in conjugation [4]. The ambivalent set of characteristics of pR1SE make it an intriguing case for the evolutionary trajectory of archaeal plasmids and viruses.

Plasmid pR1SE was also found to integrate into the host genome, and incorporate and transfer large fragments of host genomic DNA [4], indicating that it could play a major role for the remarkable HGT described for the haloarchaeal host organisms [8]. Haloarchaea are known to exhibit secondary chromosomes and megaplasmids [9], and it has been hypothesized that pR1SE could be involved in the emergence of these additional replicons [4]. However, only one element with similarities to pR1SE, was detected in public databases by BlastP searches upon the discovery of pR1SE in 2017. Therefore, in this study, we used a hidden Markov model (HMM) based approach to discover pR1SE-like plasmids. By analyzing metagenomes from Australian salt lakes and searching public databases, we identified a set of pR1SE-like elements. This enabled the description of the general genomic organization of the plasmid, leading to a better understanding of the functional roles of conserved proteins. We describe the core proteins in detail and offer hypotheses on the function of two distinct core gene clusters as well as the ecological and evolutionary impact of PV-mediated HGT.

2. Materials and Methods

2.1. Sampling of Australian salt lakes, DNA extraction and sequencing

From December 2018 to January 2019 the sediment salt crust of eleven hypersaline lakes was sampled under the permission from the Department for Environment and Water, South Australia (Permission number: U26817-1) and the Department of Environment, Land, Water Planning, Victoria (Permission number: 1008945). Subsamples of 1 g dry-weight sediment for each lake were used for DNA extraction, using the FastDNA™ SPIN Kit for Soils. DNA sequencing libraries were prepared (FS DNA Library, NEBNext® Ultra™) and sequenced (2 x 250 bp paired end, Illumina HiSeq2500, Rapid Mode) at the Max Planck Genome-Centre Cologne (Germany). The data have previously been described in Alarcón-Schumacher *et al* [10] and raw data are available at ENA-EMBL under project number PRJEB61734. Reads were trimmed with Cutadapt (v4.6) [11], removing low quality, short or unpaired reads (parameters: -q 30 -m 30). The remaining reads were assembled with metaSPAdes v3.13.1 [12].

2.2. Creation of pR1SE protein clusters

Hidden-Markov-Model (HMM) profiles were generated for 6 open reading frames of pR1SE (ORF6, ORF8, ORF17, ORF21, ORF23, ORF24) as follows, also described in Figure S1. Input proteins (Table S1) were iteratively compared by BLASTp against the nr database of non-redundant proteins (accessed 18.04.2023) using diamond blastp [13]. Proteins with a bitscore above 50 and an evalue below 10^{-5} were downloaded. Sequences were aligned using mafft [14] (--localpair –reorder) and upon manual inspection of the alignment, proteins that introduced larger gaps in the alignment or extended the alignment on either side were removed. A coverage threshold was established (Table S1). Sequences that were removed in this step were added to an 'accession blacklist' and were not considered in further alignments. The following steps were done iteratively until no further proteins could be added to an ORF cluster: First, multiple sequence alignments were generated using mafft [14] with the –localpair –reorder flags. The resulting alignment was visually inspected. Proteins that extended over the edges of the original alignment or introduced large gaps within the alignment were removed and black-listed, in order to avoid the artificial inflation of the search space. Then, HMM profiles of the

alignments were generated using hmmbuild, part of the hmmer tool package (v3.3.2) [15]. The resulting profiles were then searched against nr using hmmsearch of the same package. Finally, hits above an ORF-dependent bitscore and coverage threshold (Table S1) were retrieved and added to the ORF cluster. After three (ORF6, ORF17, ORF21, ORF23, ORF24) or 5 (ORF8) iterations no further proteins could be retrieved and a final HMM profile was created.

2.3. Detection of pR1SE-like elements (apHPVs) in public databases and metagenomes from Australian salt lakes

Using the pR1SE protein cluster HMM profiles, the Australian salt lake contigs were searched for pR1SE relatives (hmmsearch, score according to Table S1). Contigs longer than 10,000 bp that showed at least three hits were selected for detailed analysis (N = 18). Similarly, all genomes associated with the proteins retrieved from the nr database when generating the HMM profiles were downloaded from NCBI, using the rentrez library (David Winter, https://CRAN.R-project.org/package=rentrez) and searched using the HMM profiles (hmmsearch, –score 50). Again, contigs longer than 10,000 bp that showed at least three hits were considered for detailed analysis (N = 63). Similarly, the IMG/VR database (v4.1 release Dec. 2022) [16] was searched for potential candidates. Contigs with at least five of six core genes were considered "complete" pR1SE-like elements. In total, 40 pR1SE-like elements, in addition to the original pR1SE plasmid, were retrieved and from now on termed apHPVs (see Results 3.2) (Table S3).

Host information of NCBI apHPVs was retrieved from NCBI. Host prediction for the apHPVs from Australian salt lakes was done using iPhoP (https://bitbucket.org/srouxjgi/iphop/src/main/, [17]) in predict mode. Circularity information of the contigs was either directly retrieved from NCBI or determined manually by searching for terminal repeats longer than 30 bp at both ends of the contig. Direct, palindromic or interspersed direct repeats were detected using the 'Find repeats' function in Geneious (Build 2022-08-18). Repeats >25 bp with a maximum of 1 mismatch were considered positive identifications.

2.4. Identification of the apHPV core genes

In order to identify core genes of apHPVs, we selected ten genes upstream of ORF6, all genes between and including ORF6 and ORF24, and the 10 genes downstream of ORF24 of each contig, if present. This yielded a total of 1784 protein sequences, which were clustered at 15 % sequence similarity, using psi-cd-hit with the parameter -c 0.15 [18]. Protein clusters which recruited proteins from at least 37 plasmids (>90 % of the 41 apHPVs), were considered "core" genes. Clusters that recruited from between 50 % and 90 % of plasmids (at least 21) were considered "shell genome" and clusters that recruited from between 10 % and 50 % (at least 5) were considered "cloud genome". Core, shell, and cloud genome form, together, the pangenome. Protein alignments to calculate the pairwise amino acid identity of apHPV homologs of core proteins were computed using MUSCLE [19] with default parameters.

2.5. Phylogenetic analysis

A distance matrix of the 41 apHPVs was calculated using the presence and absence of the pangenome using the 'hclust' function with the 'average' method in R. Based on this matrix, a phylogenetic tree was constructed using the 'ggtree' R library [20]. A second tree was calculated based on the host genomes of the plasmids. This was done by calculating an all genome alignment using the 'gtdbtk classify_wf' command, part of GTDB-Tk v2 suite [21] with standard parameters, which was converted into a distance matrix using the IQ-TREE webserver [22] with the parameters '-bb 1000 -alrt 1000'. The tree was visualized using 'ggtree'.

147 *2.6. Gene sharing network of apHPVs, archaeal viruses and archaeal plasmids*

148 First, the following search query was used in order to retrieve archaeal plasmids from
149 NCBI: *'"archaeal"[Organism] OR "Archaea"[Organism]) AND "plasmid"[Filter] NOT*
150 *"shotgun"[All Fields] NOT "MAG"[All Fields]' (date of retrieval: 22 Sep 2023).* Plasmids
151 between 20 kbp and 70 kbp were selected for detailed analysis. This yielded 134 archaeal
152 plasmids, which were used as the input for vConTACT2 [23], together with all complete
153 apHPVs. Parameters were set to '--rel-mode 'Diamond' --db 'ArchaeaViralRefSeq207-
154 Merged' --pcs-mode MCL --vcs-mode ClusterONE'. The network was visualized with
155 flourish.studio (https://flourish.studio/). Sequences in the cluster that contained both viral
156 and plasmid-like elements (see Figure 4, top left), were analyzed using VirSorter2 [24]
157 with default parameters. Sequences that scored >0.9 and showed more than 3 viral
158 hallmark genes were considered viral.

159 *2.7. Functional Annotation of apHPV genes*

160 Open reading frames were predicted using prodigal
161 (https://github.com/hyattpd/Prodigal, [25]) using the metagenomic flag -p meta.
162 Functional annotation was done using DRAM
163 (https://github.com/WrightonLabCSU/DRAM, [26]) with the following parameters: –
164 *prodigal-mode meta –min-contig-size 10000*. Protein sequences were searched against the
165 NCBI nr database of non-redundant proteins, using diamond blastp
166 (https://github.com/bbuchfink/diamond, [13]) and the following parameters: *--max-target-*
167 *seqs 3 --evalue 0.00001 --fast.* Lastly, proteins were also searched against a set of databases
168 (CDD, COILS, Gene3D, HAMAP, MOBIDB, PANTHER, Pfam, PIRSF, PRINTS, PROSITE,
169 SFLD, SMART, SUPERFAMILY and NCBIFAM) using InterProScan [27] with default
170 parameters.
171 In addition to the aforementioned annotation steps, the protein structures of seven
172 representatives of ORF8, ORF9, ORF17, ORF21, ORF23 and ORF24 were predicted and
173 searched against protein structure databases, using the Foldseek web server [28] (Table
174 S2). Since the web server only allows for proteins shorter than 400 amino acids (aa), in
175 cases where the protein was longer than 400 aa, it was manually split into an N-terminal
176 and an C-terminal part, each consisting of 400 aa. Protein structures were searched in
177 3Di/AA mode. For each protein, the two best hits (ranked by a low evalue and high
178 Foldseek probability value) with an available annotation were selected, preferably from
179 the AFDB50 database. Structures of ORF7 and ORF25, and their homologs, were
180 downloaded from the AlphaFold protein structure database (https://alphafold.ebi.ac.uk/).
181 For ORF6 homologs, the protein structures were predicted using Alphafold 2 [29] as
182 protein monomers. Protein structures were visualized with open source PyMOL(TM)
183 Molecular Graphics System, Version 2.2.0 (L. DW. The PyMOL molecular graphics
184 system. http://www.pymol.org/. 2002). Structural predictions were manually inspected
185 and low-confidence regions were removed (pLDDT < 50). Subsequently, selected models
186 were used for structural similarity comparison with DALI [30].

187 *2.8. Identification of Cluster of orthologous groups (COGs)*

188 The functional potential of two sets of genes were compared against each other: Set 1 was
189 comprised of apHPV genes *outside* of the core-region (from ORF6 to ORF25), and set 2
190 contained genes of 993 whole haloarchaeal genomes, retrieved from NCBI (NCBI genome
191 site, filtered for taxon '183963', Halobacteria). Genes were compared by BLASTp against
192 the COG database (version 2020, access 10.01.23) [31] using diamond blastp with an evalue
193 cutoff of 10^{-5}. Each gene was assigned to one or more COG categories, resulting in
194 functional profiles for the two sets of genes. For each category a relative abundance was
195 calculated as follows:

196 $$rel(set)_{cat} = \frac{n_{hc}}{n_{tot(set)}},$$

where n_c is the number of hits per category and $n_{tot(set)}$ is the total number of hits for this category. Then, the difference of relative abundances was calculated per category:

$$\Delta rel_{cat} = rel(apHPV) - rel(halo).$$

2.9. Detection of antiviral defense systems and anti-CRISPR proteins in apHPVs

Antiviral defense systems were identified using the PADLOC webserver with the corresponding PADLOC database (v2.0.0) [32]. Detected defense systems for each apHPV are listed in Table S3. Anti-CRISPR proteins (arcs) were detected by diamond blastp searches of the apHPV proteins against the Anti-CRISPRdb (version 2.2) [33]. Additionally, the core regions (between ORF6 and ORF25) of all apHPVs were searched for matches against the IPhoP CRISPR-spacer database with blastn with default parameters. Hits with >97 % identity and length >30 were considered to be positives.

2.10. Number of plasmids per class

In order to calculate the number of plasmids per haloarchaeal class, NCBI was searched with the following search-query: *""archaeal"[Organism] OR "Archaea"[Organism]) AND "plasmid"[Filter] NOT "shotgun"[All Fields] NOT "MAG"[All Fields]"*. This yielded 1134 plasmids for which the taxonomy was fetched. The number of genomes on GTDB for each class was retrieved through the built-in tree-browser (https://gtdb.ecogenomic.org/tree?r=p__Halobacteriota). On the day of retrieval (22 Sep 2023), the number of genomes were: Archaeoglobi (71), Halobacteria (750), Methanobacteria (559), Methanococci (61), Methanomicrobia (311), Thermococci (141), Thermoplsamata (801), Thermoprotei (122).

3. Results

3.1 Selection of core proteins of pR1SE

In order to detect pR1SE-like elements in public databases and in our own dataset of metagenomes from Australian salt lakes, HMM profiles of six proteins of the original pR1SE were created. The proteins selected for this search - ORF6, ORF8, ORF17, ORF21, ORF23, ORF24 - were detected in pR1SE-containing vesicles (PVs) and were suggested to be structural proteins of PVs. For each of the six ORFs a homolog was detected in *Halorubrum saccharovorum*, *Haloterrigena turkmenica* and *Halopiger xanaduensis*, resulting in a protein set that was used to create the respective HMM profiles as described in the methods section 2.2..

3.2 Identification of forty complete pR1SE-like mobile genetic elements using an HMM based approach

Using the HMM profiles of the core proteins, 40 complete (at least five proteins present) pR1SE-like entities were detected. Additionally, 38 'incomplete' pR1SE-like elements were detected with a minimum of 3 of the 6 proteins detected (Table S3). Of the 40 complete pR1SE-like entities, 35 were retrieved from NCBI, while 5 additional contigs were detected in metagenomes of Australian salt lakes [10], and none could be detected in the IMG/VR database. We will refer to the detected pR1SE-like entities as apHPVs, for archaeal plasmids of haloarchaea potentially transferred in plasmid vesicles, while 'H' can be replaced by the abbreviation for the host (e.g. HR for *Halorubrum*; apHRPV1 for pR1SE) as common for virus names.. The 40 apHPVs originate in hypersaline environments from all around the world, e.g. salt lakes, salt mines, salted food or marine solar salterns. More than half (26) were found as part of plasmids with a size below 100,000 bp, twelve on plasmids between 100,000 bp and 500,000 bp and two were integrated in secondary chromosomes (766 kbp and 596 kbp respectively) and one was integrated into a main chromosome (3.7 mbp). For 23 out of 41 apHPVs the respective contigs were circular,

either by NCBI annotation or by our analysis (see Methods section 2.3.). The average GC content ranged from 57.6 % to 64.8 %. The apHPVs and incomplete relatives are listed in Table S3.

3.3. apHPVs exhibit two highly conserved gene clusters interspersed by a variable region

All complete apHPVs are organized into two gene clusters (Figure 1). Both gene clusters contained proteins used for the search and additional ORFs conserved across the majority of apHPVs, that we define as core proteins of apHPVs. ORF numbers are referring to the original pR1SE annotation (accession KX687704.1). Gene cluster 1 contains ORF6-9 and gene cluster 2 contains ORF17, ORF21, ORF23-25. The synteny of genes within a cluster and the order of the two gene clusters was conserved in all identified complete apHPVs. However, the region between the two clusters (between ORF9 and ORF17) was highly diverse, consisting of three to eleven genes that were rarely shared between different apHPVs (Figure S2). Interestingly, all of the 38 `incomplete' pR1SE-like elements, detected in our search, contained either genes from cluster 1 (34) or genes from cluster 2 (3), with only one contig containing ORFs 6 to 8 and ORF17 (Gairdner_NODE_4298_length_11003_cov_3.762240).

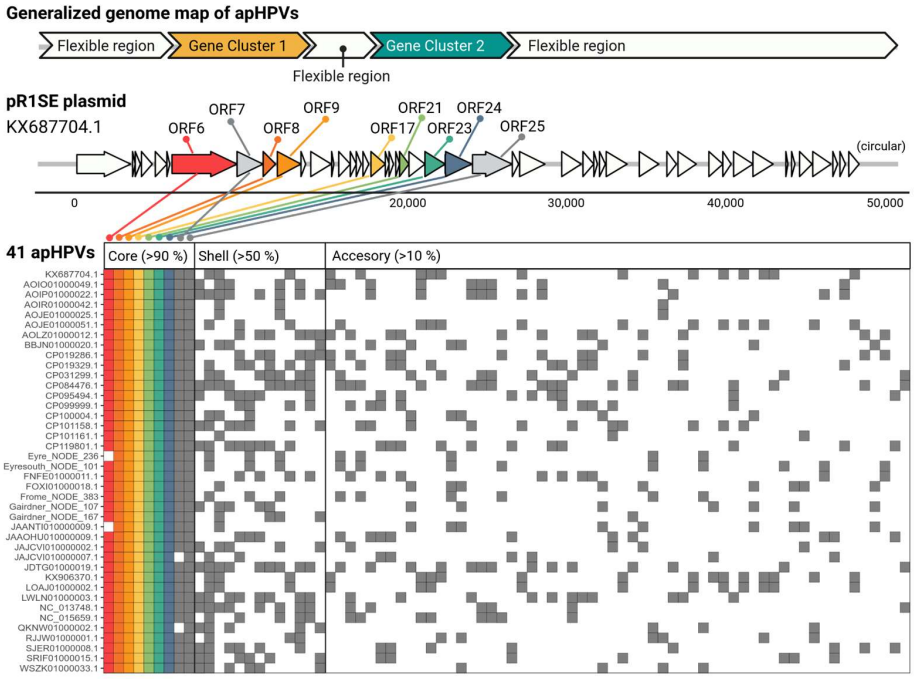


Figure 1. Generalized genome map and cluster analysis of proteins belonging to the core, shell and cloud genome of apHPVs. On top, a generalized genome of apHPV is shown, with a flexible region between two conserved gene clusters and flexible regions surrounding this core region. Below, the genome of the original pR1SE (KX687704.1) is drawn, where the core proteins, identified in >90% of all apHPVs, are highlighted. Below a presence-absence plot of protein clusters (columns) in different apHPVs (rows) is presented. Note, that the homologs of ORF6, ORF8, ORF9, ORF21 and ORF24 did not cluster at the selected threshold (15 % AAI), but were split into multiple clusters and were collapsed for better visualization. An expanded view of the 'core proteins' is given in Figure S3.

3.4. Phylogeny and host association of apHPVs

All apHPVs were identified in species belonging to the class *Halobacteria*, of the order *Natrialbales* (n = 18), *Haloferacales* (n = 9) and *Halobacteriales* (n = 11). Detailed taxonomic information is given in Table S3 and family rank is indicated in Figure 2. Host information was retrieved directly from NCBI if possible. For metagenome derived sequences, for which host information was not available, hosts were predicted using IPhoP. For three of the five metagenomic sequences, a host could not be predicted (Table S3). A cluster analysis of the apHPVs was calculated based on the presence and absence of genes in the cloud genome (present in more than 10 % of all plasmids) of each apHPV (Figure 2). Interestingly, we found two examples of 2 different apHPVs co-existing in the same host, *Halorubrum saccharovorum* DSM 1137 and *Halorubellus salinus* GX3 (Table S3).

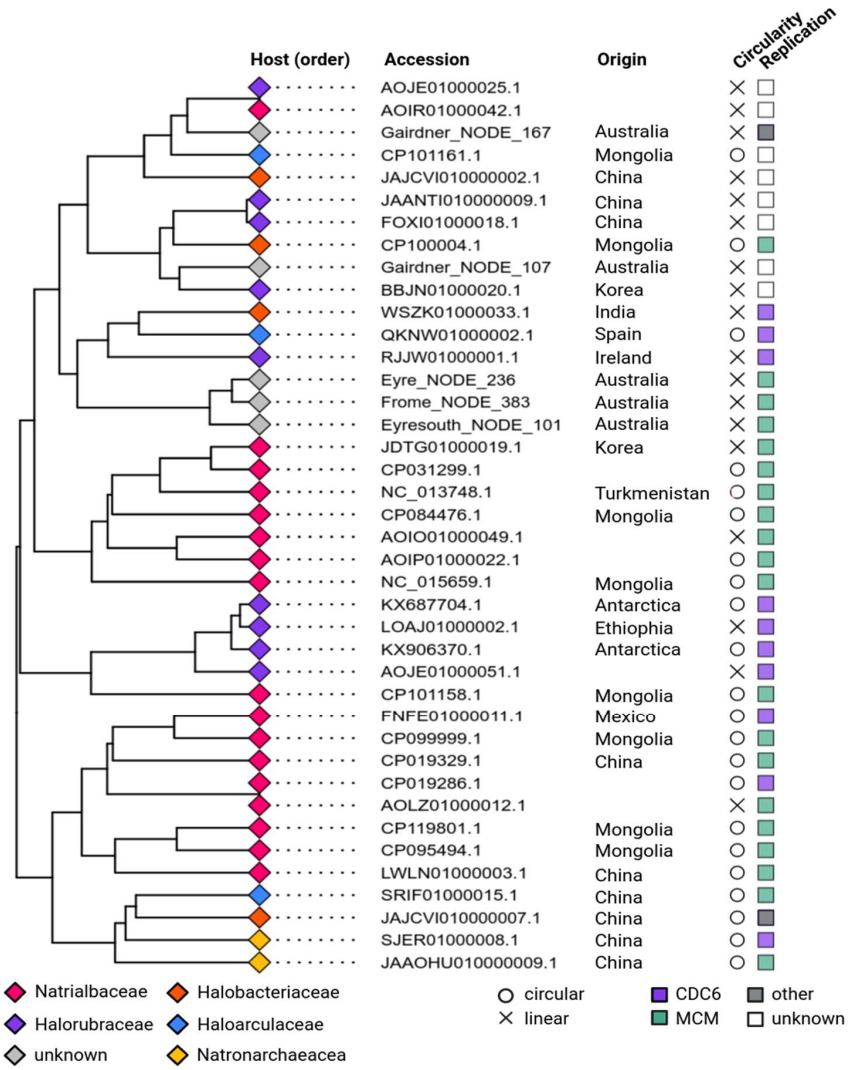


Figure 2. Cluster analysis and additional features of complete apHPVs. The tips of the branches are labeled according to the host associated with the plasmid (taxonomic rank: family). Country of origin is indicated in the second column. Circularity and the identified putative replication gene are shown in the columns to the right. CDC6 - CDC6/ORC1-like protein, MCM - Minichromosome maintenance complex.

291 *3.5 Replication of apHPVs*

292 The plasmid pR1SE contains two proteins that might be responsible for plasmid
293 replication. ORF1 encodes a predicted RepH plasmid replication protein and ORF29 an
294 CDC6/ORC1-like protein that is commonly involved in initiation of replication of
295 haloarchaeal chromosomes and plasmids [34]. We searched the newly discovered
296 apHPVs for proteins involved in replication. For 33 of 41 plasmids a potential replication
297 protein was found close (+/-10 ORFs up- or downstream) to the core region (Figure 2).
298 Most commonly, proteins belonging to the minichromosome maintenance complex
299 (MCM) were identified (n = 20). A CDC6/ORC1-like protein was detected on ten apHPVs.
300 In two cases, a HNH endonuclease and a recombinase was found close to the core region.
301 For seven apHPVs no obvious candidate for a replication protein was detected. Potential
302 origins of replication (direct repeats, interspersed direct repeats or palindromic repeats of
303 at least 25 bp) were detected in 16 plasmids, either directly before ORF6 or just after
304 ORF25. None of the newly identified apHPVs exhibits a homolog to ORF1 of pR1SE.
305 About ~100 aa (aa340-440) of the predicted tertiary structure of ORF1 shows good
306 alignments with DNA binding proteins (transcriptional regulators, helicase, chromosome
307 partition protein, CDC6/Orc1-like), indicating that it could indeed be DNA binding and
308 involved in plasmid replication.

309 *3.6. Core genes and their predicted functions*

310 A total of ten genes were conserved in nearly all (>90 %) plasmids which were therefore
311 considered as the core genome of apHPVs (Table 1 and Figure 1). All of the protein
312 products of the core genes, except one (ORF25), were detected as potentially structural
313 proteins in PVs. However, their role in the generation of PVs remains unknown. ORF7
314 was consistently (amongst all apHPVs) annotated as a S8 family serine peptidase based
315 on conserved domains as reported previously [4]. However, all other ORFs showed no
316 similarity to known proteins on amino acid sequence level, or very inconsistent results
317 across homologs of different apHPVs. Therefore, we used the predicted tertiary structures
318 [29] and searched for related structures in public databases [25, 27] (see method section
319 2.7.). This approach confirmed the annotation of ORF7 as a subtilisin-like serine protease,
320 because the structures of ORF7 and apHPV homologs consistently hit subtilisin-like serine
321 protease homologs. ORF8 and its homologs in other apHPVs, best aligned (~aa 1 - 180, see
322 Figure S4) with the structure of Necrosis and Ethylene inducing Protein 1 (NEP1).
323 Predicted structures of ORF9 homologs were very poorly resolved (pair-wise AAI of ~18
324 %), leading to inconsistent hits, however, best hits included S-layer proteins (~aa 30 - 200)
325 and Spore coat protein CotH (IPR014867). The N-terminus of ORF23 (aa 50-150) and its
326 homologs consists of eight beta strands forming two putative beta sheets, and hits putative
327 pilus assembly proteins (Type II and Type III secretion system) and TonB receptor
328 proteins. The predicted structures of ORF25 and its homologs align (aa 135 – 565) best
329 (RMSD values between 4-5) with components of the Type VI secretion system (VirB4) and
330 the bacterial conjugation protein TRWB, but also with HerA (RMSD 5.0) [35] and genome
331 packaging NTPase of Sulfolobus turret icosahedral virus 2 [36] (RMSD 2.7). Different
332 apHPV homologs of ORF17 yielded inconsistent hits. Best alignments were against Ricin
333 B-type lectin, PKD, SH3 and UNBV_ASPIC domain-containing proteins. For ORF21 and
334 ORF24 no homologs, neither on a sequence nor on a structural level were found.
335
336 ORF6, the largest conserved protein, was previously proposed to be one of the major
337 structural proteins in PVs. Surprisingly, ORF6 proteins presented a relatively low
338 sequence similarity (pairwise AAI ~18.3%) and high variability. While ORF6 of pR1SE has
339 a WD40 domain (IPR001680) at the N-terminus (Figure S5), this domain was present in
340 only ~56% of the predicted structures of ORF6 homologs. Meanwhile, the other fraction
341 presented variable N-terminal domains, often in the form of extended arrays of
342 antiparallel beta sheets (Figure S6). On the other hand, the C-terminal domain of ORF6

343 homologs was highly conserved, composed of multiple arrays of antiparallel beta sheets
344 facing each other (Figure S7), that likely form the hydrophobic core of the protein.
345 Homology search against protein structure databases using Foldseek and DALI, revealed
346 this array to be unique to ORF6-like proteins, as no matches with similar folds to the whole
347 array were obtained. Significant matches to smaller subsections of the C-terminal domain
348 include a chromosome condensation regulator RCC1 (A0A2Z4FI68; score 141) and a pair
349 of lipoproteins of bacterial origin (UniProt F3ZI24 and A0A1H8ZS84; score 130 and 128
350 respectively). Despite the high degree of structural conservation among C-termini of
351 ORF6 homologs, their function remains unknown.

352 **Table 1. Core genes of apHPVs.** A detailed list of structural homologs is given in Table S2.

ORF No.[1]	Pairwise AAI[2]	Best hit	Method of annotation
ORF6	18.3 %	-	AlphaFold/DALI
ORF7	29.1 %	subtilisin-like serine protease	foldseek/AlphaFold/ DALI
ORF8	23.8 %	Necrosis and ethylene inducing protein 1	foldseek
ORF9	17.5 %	S-layer / Big6 / CotH	foldseek
ORF17	34.2 %	PKD/SH3/Ricin B-type lectin / UnbV_ASPIC	foldseek
ORF21	26 %	-	foldseek
ORF23	31.5 %	pilus assembly proteins / TonB receptor proteins	foldseek
ORF24	23.9 %	-	foldseek
ORF25	41 %	VirB4 / helicase / genome packaging ATPase	foldseek/ AlphaFold/DALI

353 [1] According to the original annotation of pR1SE (accession KX687704.1) and homologs in other
354 apHPVs.
355 [2] Based on a protein alignment of all homologs of a respective ORF in detected complete apHPVs.

356 *3.7. Transfer of host genetic material by apHPVs*

357 Plasmid pR1SE was shown to integrate into host chromosomes or other replicons. Upon
358 excision, it often included host genomic material, resulting in pR1SE derivatives with
359 variable extensions [4]. Three of the novel apHPVs were also found integrated into long
360 replicons (>500 kbp), while the remaining 38 entities were found on plasmids ranging
361 from 44 kbp up to 271 kbp. The core region (gene cluster 1 and 2, and the flexible region
362 in between) of apHPVs ranges from a minimum of 18 kbp to 44 kbp (Figure 1, Figure S2),
363 the plasmids carrying the core region are up to eight times larger than the respective core
364 region. We analyzed the additional genetic material and found that it often matched host
365 sequences with up to 100% identity. Comparing the functional profile of transported (non-
366 core) regions of apHPVs in comparison to haloarchaeal genomes, revealed that apHPVs
367 contain significantly higher amounts of genes associated with replication, recombination
368 and repair, defense mechanisms, the mobilome (including transposases and viral-like
369 genes), transcription, cell motility and cell cycle and control (Figure 3). In contrast, genes
370 related to metabolism and translation are present much rarer in apHPV associated
371 genomic information.
372

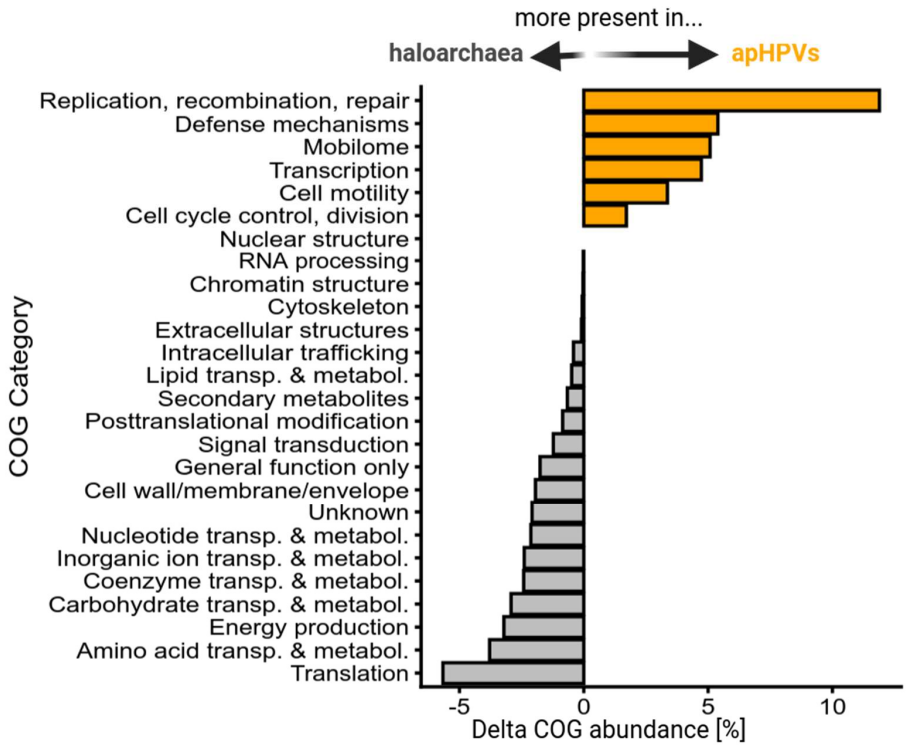


Figure 3. Functional profile of apHPV-transported genes in comparison to host genes. The difference between the percentages of genes associated with one COG category, for two sets of genes: Genes from 993 halobacterial genomes and genes from 38 apHPV non-core regions, excluding 3 apHPVs which were integrated into larger replicons (see methods section 2.8.). Positive values indicate an increased relative abundance of this COG category in apHPVs, while negative values indicate the increased relative abundance of this COG category in haloarchaeal genomes.

3.8. Antiviral defense systems and anti-CRISPR proteins on apHPVs

In sixteen of the 38 complete, non-integrated apHPVs a putative antiviral defense system was identified. The defense systems were located adjacent or very close (+/- 10 ORFs up- or downstream) to the core region (Table S3). Among the systems detected were CRISPR-Cas systems [37], cyclic GMP-AMP type systems [36, 37], a complete phosphorothioation modification module [40, 41] and other defense systems (e.g. restriction modification systems) [30, 38]. Only one apHPV (AOID01000019) contained an anti-CRISPR protein (AcrIIA7, WP_081009421.1), notably within the core region of that apHPV. Conversely, searching the core-regions for CRISPR-spacer hits revealed that seven out of 41 apHPVs are targeted by CRISPR-systems of haloarchaea. Most often (6 out of 7) apHPVs were targeted by a close taxonomic relative of their host organism, and in one case the host was targeting its own intracellular apHPV.

3.9. Clustering of apHPVs with archaeal plasmids and archaeal viruses

While the pR1SE plasmid showed a virus-like life style, being transferred in virus-like particles, no virus-like genes, e.g. genes encoding for virus capsid proteins, could be identified on the plasmid, not even through analysis of predicted tertiary structures. To determine the positioning of apHPVs between viruses and plasmid, we generated a gene sharing network between archaeal plasmids, archaeal viruses and the discovered apHPVs (Figure 3). This resulted in clusters composed of either only viruses, only plasmids or only apHPVs, in all but one case. One cluster containing archaeal virus BJ1 and *Halorubrum*

phage CGphi46 amongst others also contained sequences labeled as plasmids in public databases. However, we found these sequences to be of viral origin using viral prediction tools (see method section 2.6., score >0.9 with at least 3 viral hallmark genes). apHPVs clustered into a single large cluster, with connections to both a cluster of haloarchaeal viruses and a cluster of haloarchaeal plasmids. The genes shared between apHPVs and the other sequences were found to be outside the core region.

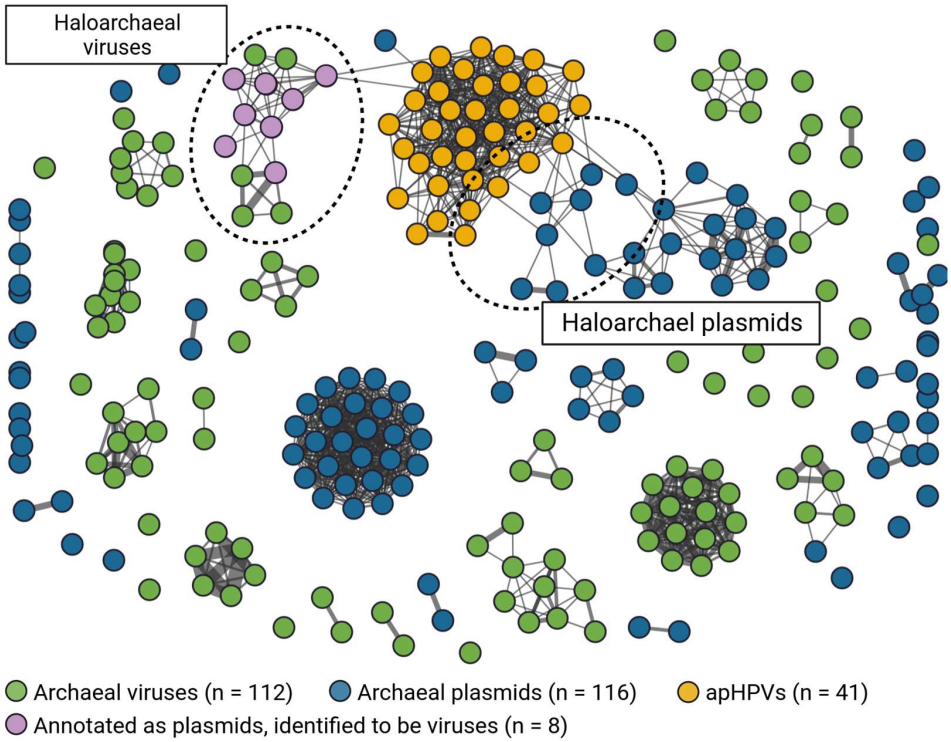


Figure 4. Gene sharing network between archaeal plasmids, archaeal viruses and apHPVs. The mixed cluster of viruses and plasmids (top left), was found to be exclusively viral with some virus genomes mistakenly classified as plasmids. apHPVs share non-core genes with both a viral cluster and haloarchaeal plasmids.

4. Discussion

Upon discovery of pR1SE only one other organisms with similar proteins could be detected in public databases [4]. Our database search using HMM profiles of six pR1SE proteins revealed at least 41 full length apHPVs (including pR1SE) and a number of contigs that represent partial apHPVs, evident of pR1SE being more widespread rather than being a single case of a non-conjugative, vesicle forming plasmid. The apHPVs seem to be restricted to archaea belonging to the class of *Halobacteriales* and were found in hypersaline environments all over the world.

The genetic organization of the 'core region' in two syntenic gene clusters is conserved among all complete apHPVs. The core region contains all proteins that were detected as potential structural proteins in PVs, indicating that other apHPVs could also be capable of PV formation. However, whether all of the detected plasmids indeed disseminate via self-encoded plasmid vesicles will require experimental verification. We also identified a significant number of apHPV-like entities that contain only one of the clusters with the majority containing the first cluster (ORF6-ORF9). These could be remnants of inactive

integrated apHPVs. Similar to integrated defective proviruses they likely have an advantage for the host, otherwise the regions would not be maintained.

The region between the two conserved gene clusters of the core is very flexible and differs between apHPVs, not only in size but also gene content. Both virus and plasmid genomes are often organized in modules that have different functions [43, 44]. Plasmid genomes usually have modules that are responsible for plasmid replication and maintenance, and modules that are required for plasmid transfer. Both gene clusters of the core region contain proteins that were identified as structural proteins in PVs by mass spectrometry, therefore, we previously suggested that both clusters are involved in PV formation and plasmid dissemination [4]. However, the clear organization into two clusters, interspersed by a variable region, might indicate that the two clusters have distinct functions.

A conserved module for plasmid replication could not be identified on apHPVs. Instead, the majority of the apHPVs encode for either a CDC6/ORC1 or a MCM helicase, that have most likely been acquired from a host genome. Haloarchaea are known to exhibit multiple origins of replication and of CDC6/ORC1 proteins within one genome, and those can likely be transferred between different haloarchaea [34]. Direct repeats of ~25 bp length in proximity to the core region were detected in nearly all apHPVs and could serve as the ORI of each plasmid, possibly allowing replication even if no plasmid encoded CDC6/ORC1 is present. Replication of haloarchaeal megaplasmids or secondary chromosomes is driven by CDC6/ORC1 and associated ORIs, and they rarely reach copy numbers that are significantly higher than those of the main chromosome (up to 2-fold) [45, 46]. Plasmid pR1SE was shown to reach up to 15 copies per host chromosome copy [7]. It remains to be elucidated whether this is due to ORF1 activity, or whether other apHPVs without ORF1 can also reach higher copy numbers.

The functional prediction of the identified core proteins remained to be challenging, despite the advancement of tools for tertiary structure prediction and structural comparison. However, structural analysis revealed conserved domains in some of the proteins. ORF6 was thought to be a central structural and potentially coat-forming protein in PVs, recruiting other structural proteins and the cargo. The prediction was based on the N-terminal WD40 domain, which is also present in coat proteins of eukaryotic intracellular vesicles [47]. However, our analysis of ORF6 in other apHPVs revealed that the N-terminus is highly diverse and only half of the other homologs contain a WD40 domain. In contrast, the structure of the C-terminus of the protein appears to be highly conserved, but no structural homologs could be identified in public databases. Thus, the role of ORF6 in PV formation remains unresolved. ORF7, for which the annotation as a subtilisin-like serine protease was supported by structural predictions, was detected in all size fractions of PVs separated on polyacrylamide gels [4], indicating a strong interaction with other PV proteins. Therefore, we suggest that ORF7 could potentially be involved in the maturation of different pR1SE proteins, resembling viral proteases involved in maturation [40, 48]. The structure of ORF8 was conserved amongst different apHPVs and showed good alignments with the structure of NEP1-like proteins (Necrosis and ethylene inducing proteins), NLPs. NLPs proteins contain NPP1 domains (necrosis-inducing Phytophthora protein), are secreted by phytopathogenic bacteria and fungi, and contribute to host infection by plasma membrane permeabilization [49]. Proteins with NPP1 domains have been identified in extracellular vesicles of phytopathogenic fungi [50]. While the detailed molecular function of NLPs for the producing organisms is still unknown, it is known that NLPs interact with the plant membrane and are involved in membrane permeabilization. We, therefore, assume that ORF8 is membrane interacting

476 and may even have membrane remodeling functions. ORF9 was predicted to build a
477 stabilizing outer coat of PVs, and our reanalysis did not reveal any other potential
478 function. Predicted structures of ORF17, ORF21 and ORF24 homologs were inconsistent
479 and did not allow the identification of structural homologs in public databases, thus
480 prohibiting a reasonable prediction of function. The N-terminus of ORF23 is highly
481 conserved and shows similarity to domains within pilus assembly proteins and TonB
482 receptor proteins. However, the function of ORF23 still remains elusive, because the
483 function of this particular domain is unknown. The structure of ORF25 is highly conserved
484 across all apHPVs, and similar to both a viral DNA-packaging ATPase as well as ATPases
485 involved in plasmid recruitment during conjugation. ORF25 was not detected in PVs,
486 therefore, we predict that ORF25 is responsible for recruitment of apHPVs into PVs. Both
487 ORF23 and ORF25 showed similarities to components of Type II or type IV secretion
488 systems, respectively, which could indicate that they are acting together to either secrete
489 or to pull the cargo into PVs. Considering the genomic organization into two cluster
490 (ORF6-9 and ORF17, ORF21-ORF25), and the fact that proteins of cluster 1, and ORF17,
491 were more abundant in PVs [4], we suggest that cluster 1 could play a major role in the
492 formation and the structure of PVs, while cluster 2 could be responsible for plasmid
493 packaging. ORF17 was highly diverse amongst different apHPVs, and could possibly be
494 responsible for host cell recognition.

495 While we have no experimental evidence that apHPVs other than pR1SE are transferred
496 via PVs, it is evident from our data that other apHPVs also integrate into host
497 chromosomes or secondary replicons, because we found examples of integration.
498 Additionally, the size of the majority of apHPVs exceeds significantly the size of the core
499 region, indicating that they have recruited a substantial amount of genomic information
500 from earlier integration events (Figure 5B). Integrases were detected on most apHPVs,
501 however, these integrases could have also been retrieved from the host, as in the case of
502 pR1SE, without being responsible for the active integration of apHPVs. Integration,
503 excision and uptake of host genomic information was shown for pR1SE [4], and we infer
504 that other apHPVs exhibit the same capability (Figure 5A). Considering the sizes of the
505 newly discovered apHPVs, we suggest that apHPVs are responsible for the horizontal
506 transfer of very large fragments of genomic DNA. Notably, this function coincides with
507 the high general genomic flexibility of haloarchaea and the significantly higher number of
508 plasmids per genome in this class (Figure 5C). This could indicate that elements like
509 apHPVs are either responsible for the high genomic flexibility, or that apHPVs could only
510 evolve because recombination events were already happening at an increased rate.

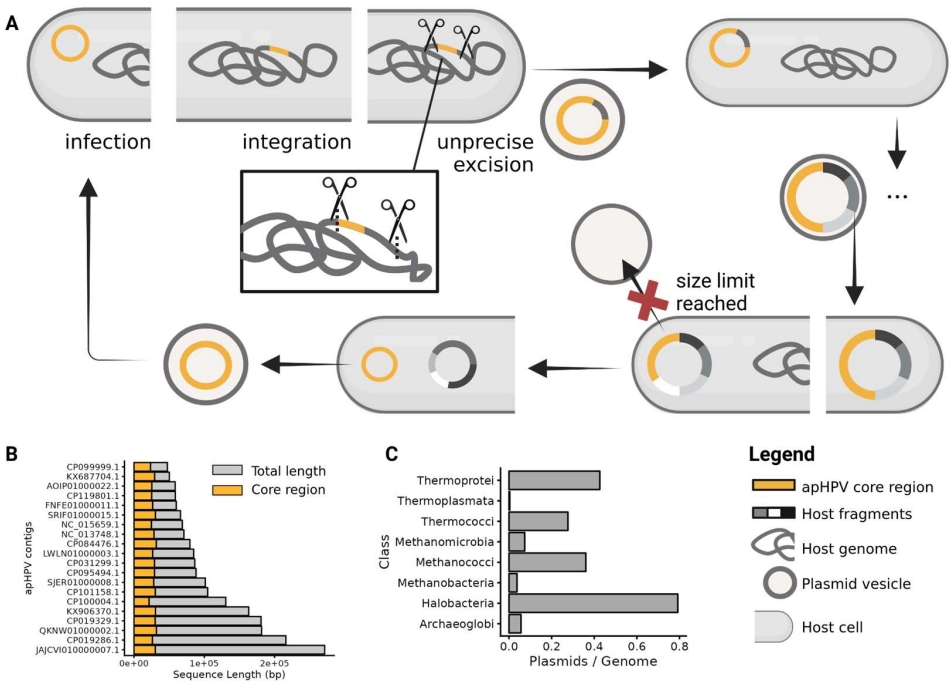


Figure 5. Proposed replication cycle of apHPVs, based on the replication cycle of pR1SE [4]**. A** - Clockwise replication cycle of apHPVs: A vesicle containing an apHPV plasmid infects a plasmid-free host. The plasmid integrates into the main chromosome, secondary replicons or replicates independently. During excision out of host replicons, genomic fragments of the host genome are co-excised and incorporated into the plasmid. This infection-integration-excision cycle potentially occurs multiple times, leading to an increased size of the plasmid, containing genetic material of multiple hosts. At some point, the plasmid becomes too large to be packaged into PVs (red cross), only the core-region of apHPVs excises and the additional genetic material remains within the host. If an ORI is present, the remains could replicate as secondary replicon. **B** - Length (base pairs) of the core region and the total length of apHPVs. Only circular, non-integrated apHPVs are shown (n = 20). **C -** Number of plasmids on NCBI per archaeal class, divided by the number of genomes of this class on GTDB.

The genomic composition of the non-core regions of apHPVs is similar to the flexible regions of the host genome [8, 51], enriched in genes associated with replication, recombination and repair, defense mechanisms, motility, transcription and the general mobilome. Additionally, we found virus defense systems on almost all apHPVs. Therefore, we suggest that apHPVs could play an important role in transfer of virus defense systems between cells, and hosts could strongly benefit from apHPVs which support their defense against viruses. While we did find CRISPR spacers against the core region of apHPVs, only a few were targeted, indicating that apHPVs are either often not recognized as invading genetic elements, or interfere with the uptake of new spacers by the CRISPR system. One apHPVs was found in host organisms encoding a spacer against the same apHPV, and one anti-CRISPR protein was identified, suggesting that apHPVs also counteract CRISPR interference. We found two cases of two different apHPVs co-existing in one host, demonstrating that they do not exclude each other as it is known for some viruses (superinfection exclusion).

Strikingly, the gene sharing network of archaeal plasmids, viruses and apHPVs revealed that apHPVs potentially exchange genetic material not only with host genomes and plasmids, but also with viruses. However, it remains to be elucidated whether apHPVs pick up regions of integrated viruses or recombine with actively replicating viruses.

Considering that the life cycle of apHPVs already places them at the interface between viruses and plasmids, their ability to recombine with both plasmids and viruses further strengthens the emerging connection between viruses and plasmids that have long been regarded as very distinct and unrelated MGEs.

5. Conclusions

In our study we identified 40 pR1SE-related elements, apHPVs, in haloarchaea, showing that this class of elements are more widespread than previously thought. All apHPVs share a very conserved genetic organization, including the majority of proteins that were detected in PVs of pR1SE. It thus is conceivable that all apHPVs could potentially be disseminated in PVs. However, even modeling the tertiary structures of these core proteins often did not allow us to draw conclusions about their function, pointing towards a unique mechanism. The potential of pR1SE for horizontal gene transfer was already demonstrated by Erdmann *et al* [4], and the analysis of the additional apHPVs detected in this study revealed even more complex interactions. The apHPVs appear to exchange genetic material not only with the host but also with archaeal plasmids and viruses, possibly playing an important role for the evolution of both, plasmids and viruses.

Supplementary Materials: The following supporting information can be downloaded at: www.mdpi.com/xxx/s1, Figure S1-S7, Table S1-S3.

Author Contributions: SE designed and led the study and collected the original samples from Australian salt lakes. DL performed the laboratory work and the bioinformatic analysis. TAS carried out the analysis of ORF6 (method section 2.8.) and the assembly of Australian salt lake metagenomes. DL and SE performed the primary writing of the manuscript. All authors participated in the analysis and interpretation of the data and contributed to the writing of the manuscript.

Funding: Funding was provided by the Volkswagen Foundation (reference 98 190) and the Max Planck Society (Munich, Germany) to S.E.

Data Availability Statement: All scripts, plots, intermediate results and pipelines are deposited on github: https://github.com/dluecking/pR1SE_australian_lakes.

Acknowledgments: We thank Daniela Thies and Ingrid Kunze (MPI for Marine Microbiology, Bremen, Germany) for assistance with the experiments, Lauren Queiss for helpful discussions and support, Prof. Dr. A. Probst, Dr. Luis 'Coto' Orellana and Dr. A. Zayed for insightful discussions and suggestions. We are very grateful to Dr. Friedhelm Pfeiffer who critically reviewed the manuscript several times. Finally, we want to thank the Max-Planck-Institute for Marine Microbiology and the Max-Planck-Society for continuous support.

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

References

[1] C. Hall, S. Brachat, and F. S. Dietrich, "Contribution of Horizontal Gene Transfer to the Evolution of Saccharomyces cerevisiae," *Eukaryot. Cell*, vol. 4, no. 6, pp. 1102–1115, Jun. 2005, doi: 10.1128/EC.4.6.1102-1115.2005.

[2] S. M. Soucy, J. Huang, and J. P. Gogarten, "Horizontal gene transfer: building the web of life," *Nat. Rev. Genet.*, vol. 16, no. 8, pp. 472–482, Aug. 2015, doi: 10.1038/nrg3962.

[3] E. V. Koonin, V. V. Dolja, M. Krupovic, and J. H. Kuhn, "Viruses Defined by the Position of the Virosphere within the Replicator Space," *Microbiol. Mol. Biol. Rev. MMBR*, vol. 85, no. 4, pp. e00193-20, 2021, doi: 10.1128/MMBR.00193-20.

[4] S. Erdmann, B. Tschitschko, L. Zhong, M. J. Raftery, and R. Cavicchioli, "A plasmid from an Antarctic haloarchaeon uses specialized membrane vesicles to disseminate and infect plasmid-free cells," *Nat. Microbiol.*, vol. 2, no. 10, Art. no. 10, Oct. 2017, doi: 10.1038/s41564-017-0009-2.

[5] J. Mills *et al.*, "Extracellular vesicles of Euryarchaeida: precursor to eukaryotic membrane trafficking." bioRxiv, p.

2023.03.03.530948, Mar. 03, 2023. doi: 10.1101/2023.03.03.530948.

[6] M. Gaudin *et al.*, "Hyperthermophilic archaea produce membrane vesicles that can transfer DNA," *Environ. Microbiol. Rep.*, vol. 5, no. 1, pp. 109–116, 2013, doi: 10.1111/j.1758-2229.2012.00348.x.

[7] L. J. Gebhard, Z. Vershinin, T. Alarcón-Schumacher, J. Eichler, and S. Erdmann, "Influence of N-Glycosylation on Virus–Host Interactions in Halorubrum lacusprofundi," *Viruses*, vol. 15, no. 7, Art. no. 7, Jul. 2023, doi: 10.3390/v15071469.

[8] M. Z. DeMaere *et al.*, "High level of intergenera gene exchange shapes the evolution of haloarchaea in an isolated Antarctic lake," *Proc. Natl. Acad. Sci.*, vol. 110, no. 42, pp. 16939–16944, Oct. 2013, doi: 10.1073/pnas.1307090110.

[9] P. Forterre, M. Krupovic, K. Raymann, and N. Soler, "Plasmids from Euryarchaeota," *Microbiol. Spectr.*, vol. 2, no. 6, p. 10.1128/microbiolspec.plas-0027–2014, Nov. 2014, doi: 10.1128/microbiolspec.plas-0027-2014.

[10] T. Alarcón-Schumacher, D. Lücking, and S. Erdmann, "Revisiting evolutionary trajectories and the organization of the Pleolipoviridae family," *PLOS Genet.*, vol. 19, no. 10, p. e1010998, Oct. 2023, doi: 10.1371/journal.pgen.1010998.

[11] M. Martin, "Cutadapt removes adapter sequences from high-throughput sequencing reads," *EMBnet.journal*, vol. 17, no. 1, Art. no. 1, May 2011, doi: 10.14806/ej.17.1.200.

[12] S. Nurk, D. Meleshko, A. Korobeynikov, and P. A. Pevzner, "metaSPAdes: a new versatile metagenomic assembler," *Genome Res.*, vol. 27, no. 5, pp. 824–834, May 2017, doi: 10.1101/gr.213959.116.

[13] B. Buchfink, K. Reuter, and H.-G. Drost, "Sensitive protein alignments at tree-of-life scale using DIAMOND," *Nat. Methods*, vol. 18, no. 4, Art. no. 4, Apr. 2021, doi: 10.1038/s41592-021-01101-x.

[14] K. Katoh, K. Misawa, K. Kuma, and T. Miyata, "MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform," *Nucleic Acids Res.*, vol. 30, no. 14, pp. 3059–3066, Jul. 2002, doi: 10.1093/nar/gkf436.

[15] S. R. Eddy, "Accelerated Profile HMM Searches," *PLOS Comput. Biol.*, vol. 7, no. 10, p. e1002195, Oct. 2011, doi: 10.1371/journal.pcbi.1002195.

[16] A. P. Camargo *et al.*, "IMG/VR v4: an expanded database of uncultivated virus genomes within a framework of extensive functional, taxonomic, and ecological metadata," *Nucleic Acids Res.*, vol. 51, no. D1, pp. D733–D743, Jan. 2023, doi: 10.1093/nar/gkac1037.

[17] S. Roux *et al.*, "iPHoP: An integrated machine learning framework to maximize host prediction for metagenome-derived viruses of archaea and bacteria," *PLOS Biol.*, vol. 21, no. 4, 2023, doi: 10.1371/journal.pbio.3002083.

[18] L. Fu, B. Niu, Z. Zhu, S. Wu, and W. Li, "CD-HIT: accelerated for clustering the next-generation sequencing data," *Bioinforma. Oxf. Engl.*, vol. 28, no. 23, pp. 3150–3152, Dec. 2012, doi: 10.1093/bioinformatics/bts565.

[19] R. C. Edgar, "MUSCLE: multiple sequence alignment with high accuracy and high throughput," *Nucleic Acids Res.*, vol. 32, no. 5, pp. 1792–1797, Mar. 2004, doi: 10.1093/nar/gkh340.

[20] G. Yu, *Data Integration, Manipulation and Visualization of Phylogenetic Trees*. CRC Press, 2022.

[21] P.-A. Chaumeil, A. J. Mussig, P. Hugenholtz, and D. H. Parks, "GTDB-Tk v2: memory friendly classification with the genome taxonomy database," *Bioinforma. Oxf. Engl.*, vol. 38, no. 23, pp. 5315–5316, Nov. 2022, doi: 10.1093/bioinformatics/btac672.

[22] J. Trifinopoulos, L.-T. Nguyen, A. von Haeseler, and B. Q. Minh, "W-IQ-TREE: a fast online phylogenetic tool for maximum likelihood analysis," *Nucleic Acids Res.*, vol. 44, no. W1, pp. W232–W235, Jul. 2016, doi: 10.1093/nar/gkw256.

[23] H. Bin Jang *et al.*, "Taxonomic assignment of uncultivated prokaryotic virus genomes is enabled by gene-sharing networks," *Nat. Biotechnol.*, vol. 37, no. 6, Art. no. 6, Jun. 2019, doi: 10.1038/s41587-019-0100-8.

[24] J. Guo, D. Vik, A. A. Pratama, S. Roux, and M. Sullivan, "Viral sequence identification SOP with VirSorter2," Jul. 2021, Accessed: Mar. 21, 2023. [Online]. Available: https://www.protocols.io/view/viral-sequence-identification-sop-with-virsorter2-bwm5pc86

[25] D. Hyatt, G.-L. Chen, P. F. LoCascio, M. L. Land, F. W. Larimer, and L. J. Hauser, "Prodigal: prokaryotic gene recognition and translation initiation site identification," *BMC Bioinformatics*, vol. 11, no. 1, p. 119, Mar. 2010, doi: 10.1186/1471-2105-11-119.

[26] M. Shaffer *et al.*, "DRAM for distilling microbial metabolism to automate the curation of microbiome function," *Nucleic Acids Res.*, vol. 48, no. 16, pp. 8883–8900, Sep. 2020, doi: 10.1093/nar/gkaa621.

[27] P. Jones *et al.*, "InterProScan 5: genome-scale protein function classification," *Bioinforma. Oxf. Engl.*, vol. 30, no. 9,

642 pp. 1236–1240, May 2014, doi: 10.1093/bioinformatics/btu031.
643 [28] M. van Kempen *et al.*, "Fast and accurate protein structure search with Foldseek," *Nat. Biotechnol.*, pp. 1–4, May
644 2023, doi: 10.1038/s41587-023-01773-0.
645 [29] J. Jumper *et al.*, "Highly accurate protein structure prediction with AlphaFold," *Nature*, vol. 596, no. 7873, Art. no.
646 7873, Aug. 2021, doi: 10.1038/s41586-021-03819-2.
647 [30] L. Holm, A. Laiho, P. Törönen, and M. Salgado, "DALI shines a light on remote homologs: One hundred
648 discoveries," *Protein Sci.*, vol. 32, no. 1, p. e4519, 2023, doi: 10.1002/pro.4519.
649 [31] M. Y. Galperin, Y. I. Wolf, K. S. Makarova, R. Vera Alvarez, D. Landsman, and E. V. Koonin, "COG database
650 update: focus on microbial diversity, model organisms, and widespread pathogens," *Nucleic Acids Res.*, vol. 49,
651 no. D1, pp. D274–D281, Jan. 2021, doi: 10.1093/nar/gkaa1018.
652 [32] L. J. Payne *et al.*, "PADLOC: a web server for the identification of antiviral defence systems in microbial
653 genomes," *Nucleic Acids Res.*, vol. 50, no. W1, pp. W541–W550, Jul. 2022, doi: 10.1093/nar/gkac400.
654 [33] C. Dong *et al.*, "Anti-CRISPRdb v2.2: an online repository of anti-CRISPR proteins including information on
655 inhibitory mechanisms, activities and neighbors of curated anti-CRISPR proteins," *Database*, vol. 2022, p. baac010,
656 Jan. 2022, doi: 10.1093/database/baac010.
657 [34] Z. Wu, H. Liu, J. Liu, X. Liu, and H. Xiang, "Diversity and evolution of multiple orc/cdc6-adjacent replication
658 origins in haloarchaea," *BMC Genomics*, vol. 13, no. 1, p. 478, Sep. 2012, doi: 10.1186/1471-2164-13-478.
659 [35] N. J. Rzechorzek, J. K. Blackwood, S. M. Bray, J. D. Maman, L. Pellegrini, and N. P. Robinson, "Structure of the
660 hexameric HerA ATPase reveals a mechanism of translocation-coupled DNA-end processing in archaea," *Nat.*
661 *Commun.*, vol. 5, no. 1, Art. no. 1, Nov. 2014, doi: 10.1038/ncomms6506.
662 [36] L. J. Happonen, E. Oksanen, L. Liljeroos, A. Goldman, T. Kajander, and S. J. Butcher, "The Structure of the
663 NTPase That Powers DNA Packaging into Sulfolobus Turreted Icosahedral Virus 2," *J. Virol.*, vol. 87, no. 15, pp.
664 8388–8398, Aug. 2013, doi: 10.1128/jvi.00831-13.
665 [37] K. S. Makarova *et al.*, "Evolutionary classification of CRISPR–Cas systems: a burst of class 2 and derived
666 variants," *Nat. Rev. Microbiol.*, vol. 18, no. 2, Art. no. 2, Feb. 2020, doi: 10.1038/s41579-019-0299-x.
667 [38] D. Cohen *et al.*, "Cyclic GMP–AMP signalling protects bacteria against viral infection," *Nature*, vol. 574, no. 7780,
668 Art. no. 7780, Oct. 2019, doi: 10.1038/s41586-019-1605-5.
669 [39] A. Millman, S. Melamed, G. Amitai, and R. Sorek, "Diversity and classification of cyclic-oligonucleotide-based
670 anti-phage signalling systems," *Nat. Microbiol.*, vol. 5, no. 12, Art. no. 12, Dec. 2020, doi: 10.1038/s41564-020-0777-
671 y.
672 [40] L. Tong, "Viral Proteases," *Chem. Rev.*, vol. 102, no. 12, pp. 4609–4626, Dec. 2002, doi: 10.1021/cr010184f.
673 [41] T. Xu, F. Yao, X. Zhou, Z. Deng, and D. You, "A novel host-specific restriction system associated with DNA
674 backbone S-modification in Salmonella," *Nucleic Acids Res.*, vol. 38, no. 20, pp. 7133–7141, Nov. 2010, doi:
675 10.1093/nar/gkq610.
676 [42] A. Millman *et al.*, "An expanded arsenal of immune systems that protect bacteria from phages," *Cell Host Microbe*,
677 vol. 30, no. 11, pp. 1556-1569.e5, Nov. 2022, doi: 10.1016/j.chom.2022.09.017.
678 [43] G. Lima-Mendez, A. Toussaint, and R. Leplae, "A modular view of the bacteriophage genomic space:
679 identification of host and lifestyle marker modules," *Res. Microbiol.*, vol. 162, no. 8, pp. 737–746, Oct. 2011, doi:
680 10.1016/j.resmic.2011.06.006.
681 [44] C. M. Thomas, "Paradigms of plasmid organization," *Mol. Microbiol.*, vol. 37, no. 3, pp. 485–491, 2000, doi:
682 10.1046/j.1365-2958.2000.02006.x.
683 [45] C. Mercier, D. Thies, L. Zhong, M. J. Raftery, and S. Erdmann, "Characterization of an archaeal virus-host system
684 reveals massive genomic rearrangements in a laboratory strain," *Front. Microbiol.*, vol. 14, p. 1274068, 2023, doi:
685 10.3389/fmicb.2023.1274068.
686 [46] M. Hawkins, S. Malla, M. J. Blythe, C. A. Nieduszynski, and T. Allers, "Accelerated growth in the absence of
687 DNA replication origins," *Nature*, vol. 503, no. 7477, Art. no. 7477, Nov. 2013, doi: 10.1038/nature12650.
688 [47] A. Eugster, G. Frigerio, M. Dale, and R. Duden, "The α- and β'-COP WD40 Domains Mediate Cargo-selective
689 Interactions with Distinct Di-lysine Motifs," *Mol. Biol. Cell*, vol. 15, no. 3, pp. 1011–1023, Mar. 2004, doi:
690 10.1091/mbc.E03-10-0724.
691 [48] H. Cheng, N. Shen, J. Pei, and N. V. Grishin, "Double-stranded DNA bacteriophage prohead protease is
692 homologous to herpesvirus protease," *Protein Sci.*, vol. 13, no. 8, pp. 2260–2269, 2004, doi: 10.1110/ps.04726004.

693 [49] C. Ottmann *et al.*, "A common toxin fold mediates microbial attack and plant defense," *Proc. Natl. Acad. Sci.*, vol.
694 106, no. 25, pp. 10359–10364, Jun. 2009, doi: 10.1073/pnas.0902362106.
695 [50] D. Rutter, "Extracellular vesicles in phytopathogenic fungi," *Extracell. Vesicles Circ. Nucleic Acids*, vol. 4, no. 1, pp.
696 90–106, Mar. 2023, doi: 10.20517/evcna.2023.04.
697 [51] B. Tschitschko *et al.*, "Genomic variation and biogeography of Antarctic haloarchaea," *Microbiome*, vol. 6, no. 1, p.
698 113, Jun. 2018, doi: 10.1186/s40168-018-0495-3.

Chapter 3

Bioinformatic Analysis of Drivers of Horizontal Gene Transfer

Dominik Lücking[1] and Susanne Erdmann[1]

Manuscript in preparation.

Author Contributions:
SE and DL designed and conceptualized the study. Manuscript prepared by DL, currently in preparation and potential subject to change in future iterations.

[1]Max-Planck Institute for Marine Microbiology, Bremen, Germany

3.1 The mobilome of prokaryotes in the era of big-data next-generation sequencing

Horizontal gene transfer (HGT) and mobile genetic elements (MGEs) are closely connected. Nevertheless, not all HGT events are mediated by MGEs, and not all MGEs are actively involved in HGT (see Figure 3.1). The genomes of all organisms are subjected to the constant assault of MGEs, ranging from a few base pairs (e.g. MITEs) to a few hundred kilobase pairs (plasmids or prophages) (Frost et al. 2005; Durrant et al. 2020). By mobilizing themselves and segments of the host genome, they induce genomic flexibility. Mobilization either occurs within a genome or between different hosts, the latter being termed horizontal gene transfer. When analyzing microbial sequence data in order to understand evolutionary and ecological dynamics, the identification of MGEs and the detection of HGT events is crucial. Tools for the identification and classification of viruses (see subsection 3.2.1), plasmids (see subsection 3.2.2), integrative and conjugative elements (ICEs), transposons (Lerat 2010; Jiang et al. 2015) and other MGEs have been added to the bioinformatic toolkit available to researchers. Similarly, tools for identifying genome recombination events (Darling et al. 2010) or identifying genomic islands (Bertelli et al. 2017), in other words, the detection of intragenomic mobilization and HGT events, independent of MGEs, were developed. Together, they allow culture-independent analysis of the genomic flexibility of microbial communities and their mobilome, the sum of all mobile genetic elements of a community. However, not all mechanisms of HGT can be detected with existing bioinformatic tools and pipelines. Extracellular vesicles have the potential to transfer DNA and induce heritable change (Bitto et al. 2017; Biller et al. 2017), and recent studies demonstrated that EV-mediated HGT events are more prevalent than previously thought (Hackl et al. 2023; Lücking et al. 2023). Similar to transformation events, this form of HGT is independent of MGEs, however, it requires the formation of membrane-derived vesicles by the host. Furthermore, the DNA transported is ostensibly indistinguishable from host DNA, which poses a unique challenge for the bioinformatic detection of EV-mediated HGT (Lücking et al. 2023). This review presents tools and their general approach(es) for identifying viruses and plasmids, as well as commonly used methods for detecting subclasses of MGEs (e.g., a specific viral family or an exotic archaeal plasmid). The challenges posed by the bioinformatic analysis of EV-mediated HGT are also discussed. Lastly, a brief outlook of future developments and challenges in the field is given.

3.2 Bioinformatic identification of mobile genetic elements

3.2.1 Detection of viral sequences in metagenomic datasets

Viral metagenomics or 'viromics' has emerged as an important field to study the diversity and abundance of viruses. Detecting viruses in mixed microbial samples has led to the discovery of thousands of viral populations (or viral species) in all kinds of environments, revealing the extraordinary diversity and abundance of viruses (Gregory et al. 2019; Malathi

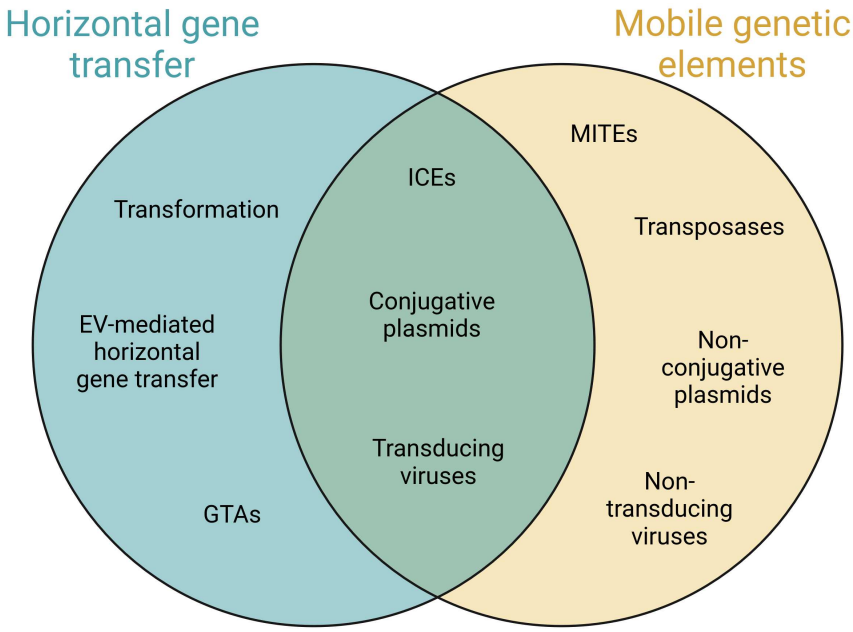


Figure 3.1: Venn diagram of Horizontal gene transfer mechanisms and mobile genetic elements. GTAs - gene transfer agents, ICEs - integrative and conjugative elements, MITEs - Miniature Inverted-repeat Transposable Elements, EV - extracellular vesicles. Figure created with `biorender.com`.

and Devi 2019; Edgar et al. 2022; Dominguez-Huerta et al. 2023 and Figure 3.2).

In order to fully leverage the potential of cultivation-independent approaches, the correct identification of viral sequences among non-viral sequences is essential. This recovery rate depends not only on the correct bioinformatic method but also on the sample preparation (Kosmopoulos et al. 2023). Either, the sample is immediately sequenced or enriched for virus-like particles (VLPs). This enrichment is accomplished with methods that exploit the physical properties of VLPs in comparison to microbial cells, such as sequential size filtration, ultracentrifugation, tangential flow filtration or flow cytometry (Li and Dickie 1985; Van der pol et al. 2010; Hillebrandt et al. 2020; McNamara and Dittmer 2020). Additionally, amplification steps may be used in order to enrich less abundant sequences. Amplification increases recovery rates of rare virus species. However, amplification potentially introduces unwanted biases, as shown elsewhere (Kim and Bae 2011; Duhaime et al. 2012; Parras-Moltó et al. 2018). The resulting sequences are then analyzed for signals of viral origin. In 2002, Breitbart *et al.* pioneered the analysis of viral communities by comparing the gene and protein content to existing databases (Breitbart et al. 2002), and the development of intricate bioinformatic pipelines and programs for the detection of viral sequences followed. In the diverse landscape of viral metagenomics, specialized programs often exhibit a narrow focus tailored to specific experiments targeting distinct viruses or environments (Fouts 2006; Roux et al. 2013). However, a new generation of versatile and user-friendly tools has emerged, capable of detecting highly diverse viruses from different environments. Generally, these programs apply one of two overarching approaches. i) Marker-based searches that detect homologous sequences of known viruses or ii) alignment-free and pattern-based approaches (see Figure 3.3). Marker-based approaches

align unknown sequences to curated databases of viral sequences, identifying genes or proteins that are indicative of viral sequences, e.g. the presence of a major capsid protein (Fouts 2006; Roux et al. 2015; Starikova et al. 2020). Alignment-free approaches classify unknown sequences by applying deep learning structures to identify viral features outside of normal sequence homologies, e.g. gene density (Ren et al. 2017, 2020; Sukhorukov et al. 2022). While these models are still trained on known viral sequences, they can detect viral sequences for which homologs have yet to be identified. Both approaches have inherent advantages and disadvantages, so the newest generation of viral detection tools apply hybrid approaches. In one of the most recently published viral identifiers, the two approaches are run simultaneously on different branches (Camargo et al. 2023). One branch predicts open-reading frames (ORFs) and compares predicted protein sequences to a highly curated dataset of marker genes, calculating likelihood scores for each input sequence. At the same time, the other branch gives the one-hot encoded nucleotide sequence to an IGLOO neural network, which assigns a likelihood score for each sequence to be of viral (or plasmid or chromosomal) origin. The results of both branches are then aggregated, and a final score is given for each sequence.

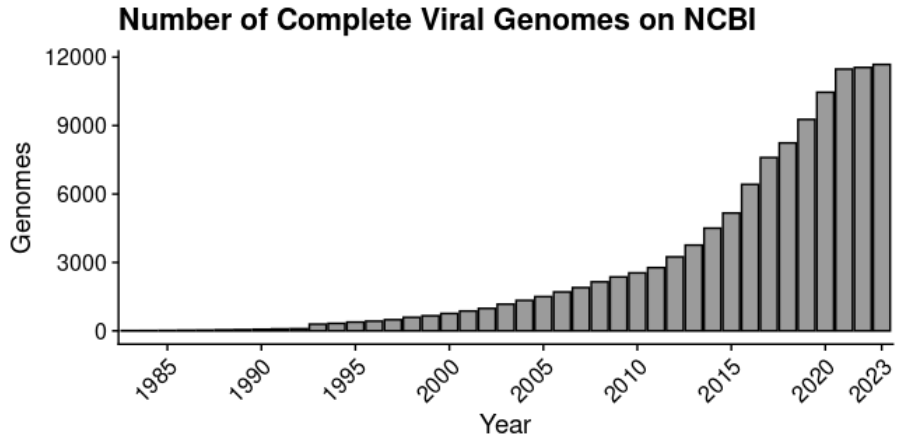


Figure 3.2: Number of complete viral genomes deposited on NCBI (access: 19.11.2023).

3.2.2 Detection of plasmids in metagenomic datasets

Plasmids carry secondary, non-essential genes yet have the ability to provide antibiotic resistance and alter metabolisms or virulence to their host. The correct identification of them is therefore crucial from a medical (Johnson et al. 2012), agronomic (Pemberton and Don 1981; Sen et al. 2011) and ecological (Thomas and Nielsen 2005) perspective. Similar to virus identification, bioinformatic tools for the detection of plasmids in metagenomic samples leverage the gene content (equivalent to marker-based approaches for viruses) and sequence composition (pattern-based) of plasmids (see Figure 3.3). Comparing unknown sequences to a database of known plasmids or known plasmid-associated genes (Gomi et al. 2021) is a simple and effective way to identify plasmids. For example, `PlasmidFinder` (Carattoli et al. 2014) compares input sequences to a set of plasmid-associated replicons. Plasmids often have significantly different GC content, k-mer frequencies and increased

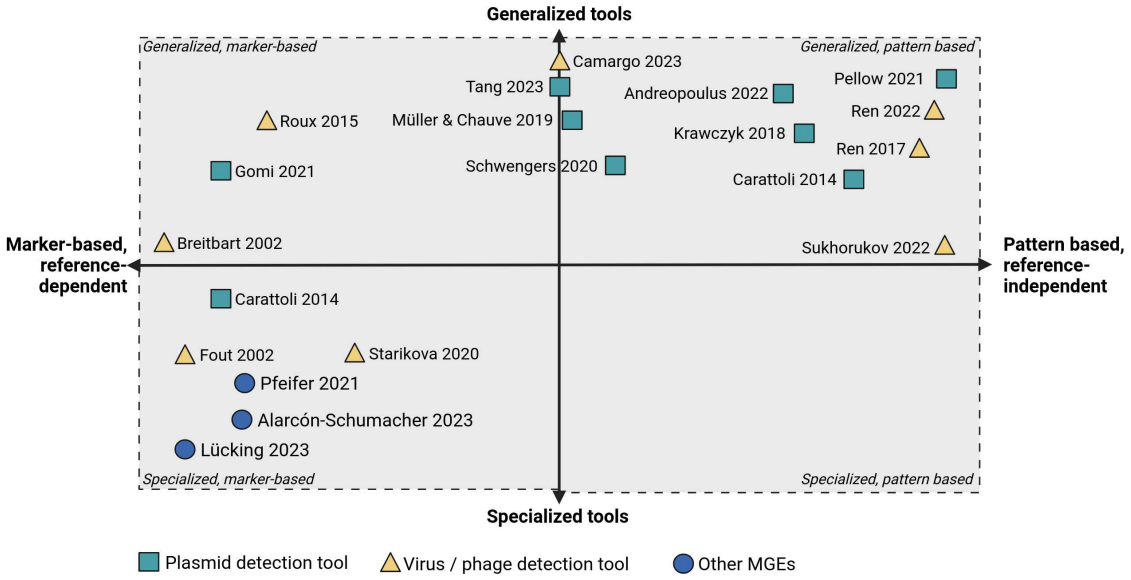


Figure 3.3: Conceptual diagram that places MGE-prediction tools along two axes of study design: Pattern-based versus marker-based approaches (x-axis) and tools with a generalized versus specialized detection range (y-axis). Position of tools are approximate. Note that no tool is placed in the 'specialized but pattern-based / reference-independent' quadrant (bottom right). First author and year is shown, references are noted in the text. Figure created with `biorender.com`.

coverage compared to chromosomal DNA. These features can be identified during assembly (`metaplasmidSPAdes`, `HyAsp`, `SCAPP`) (Antipov et al. 2016; Müller and Chauve 2019; Pellow et al. 2021) or afterward (`Plasflow`) (Krawczyk et al. 2018). `Platon` calculates a replicon distribution score (RDS) for each contig, a metric based on a bias in protein-coding genes' replicon distribution (Schwengers et al. 2020). Most modern tools apply a hybrid approach by combining pattern-based and marker-based approaches and incorporating machine and deep learning structures (Andreopoulos et al. 2022; Tang et al. 2023).

3.2.3 Detection of specific subclasses of MGEs

In order to detect specific, uncommon or even exotic MGEs among other sequences, existing tools might need to be more sensitive. Therefore, custom pipelines have to be developed. While the specific pipelines and approaches differ, one established method is the creation of Hidden-Markov-Models of protein alignments for more sensitive detection of homologs in comparison to strictly protein sequence-based searches (Pfeifer et al. 2021; Alarcón-Schumacher et al. 2023). In order to create meaningful HMM profiles, several homologs of the protein in question need to be known. The sequences are aligned, and subsequently, a HMM profile is calculated for the alignment. In short, HMM profiles of sequence alignments represent probabilistic models that capture the conserved patterns and variability in biological sequences, offering a statistical framework to analyze and compare the likelihood of observing specific positions across diverse sequences (Eddy 2011). Such profiles are generated for a particular, manually selected set of marker proteins and can be searched in a given dataset. In addition to the sheer presence or absence of proteins, the

genetic context, e.g. conserved synteny or contig length, is considered. Such an approach is implemented in Chapter 2, in order to detect pR1SE-like elements.

3.3 Detection of Extracellular Vesicle-mediated HGT

3.3.1 The ostensible impossibility of deciphering EV-mediated HGT

The detection of EV-mediated horizontal gene transfer poses unique challenges. DNA transported in EVs is indistinguishable from regular microbial DNA in cells based on the sequence itself (Biller et al. 2014; Bitto et al. 2017). Therefore, the stringent removal of cells and all other sources of non-EV-mediated microbial DNA is crucial to identify EV-mediated DNA correctly. The reliable removal of free, environmental DNA (eDNA) can be assured by DNase treatment of the sample (Sorensen et al. 2021). In order to remove microbial cells from the sample, the physical properties of EVs in comparison to cells can be exploited. The laboratory methods to achieve EV isolation are identical to the methods used for VLP enrichment, namely sequential size filtration, ultracentrifugation, tangential flow filtration or flow cytometry (Li and Dickie 1985; Van der pol et al. 2010; Hillebrandt et al. 2020; McNamara and Dittmer 2020). However, these methods only remove microbial cells, while separating EVs from other small particles is not easily achievable. That secondary separation step is necessary since the DNA transported in EVs is indiscernible from DNA transduced by viruses. Yet, EVs, virus particles, gene transfer agents and other potentially present MGEs have overlapping physical properties; i) their size ranges from 20 to 300 nm (excluding larger giantviruses), ii) their mass ranges from femtograms to picograms. Therefore, isolation techniques like size filtration fail to separate co-occurring small particles from each other. This separation might be possible in a controlled experimental setup, e.g., if the absence of a viral infection and, therefore, the absence of viral particles can be guaranteed. However, EV isolation becomes challenging for mixed microbial samples, where multiple infections occur simultaneously. One promising technique that successfully separates the different entities is using density gradients. Here, the minor differences between the mass of a viral particle and EVs are leveraged to concentrate them in different fractions of the gradient (Linney et al. 2021). However, this works best for uniform populations of EVs and other particles. Particularly EVs can vary in size and mass, making a strict separation near impossible for mixed environmental samples, even with the use of density gradients (Linney et al. 2022 and Chapter 4). Ultimately, in most cell-free and eDNA-free environmental samples, there will be a mixture of GTAs, viruses and EVs. DNA sequenced from this mix of entities traditionally has been called 'virome' or 'viral metagenome', however, new terminology for this sequence space has been proposed: protected extracellular DNA (peDNA) (Chapter 4 and Figure 3.4).

3.3.2 Leveraging host information to resolve EV-mediated HGT

In Chapter 4, a host-centric approach for the analysis of peDNA was developed, specifically to resolve the seeming impossibility of discerning EV-mediated DNA from DNA transduced by viruses or GTAs. While the approach is explained in detail in Chapter 4, a brief overview

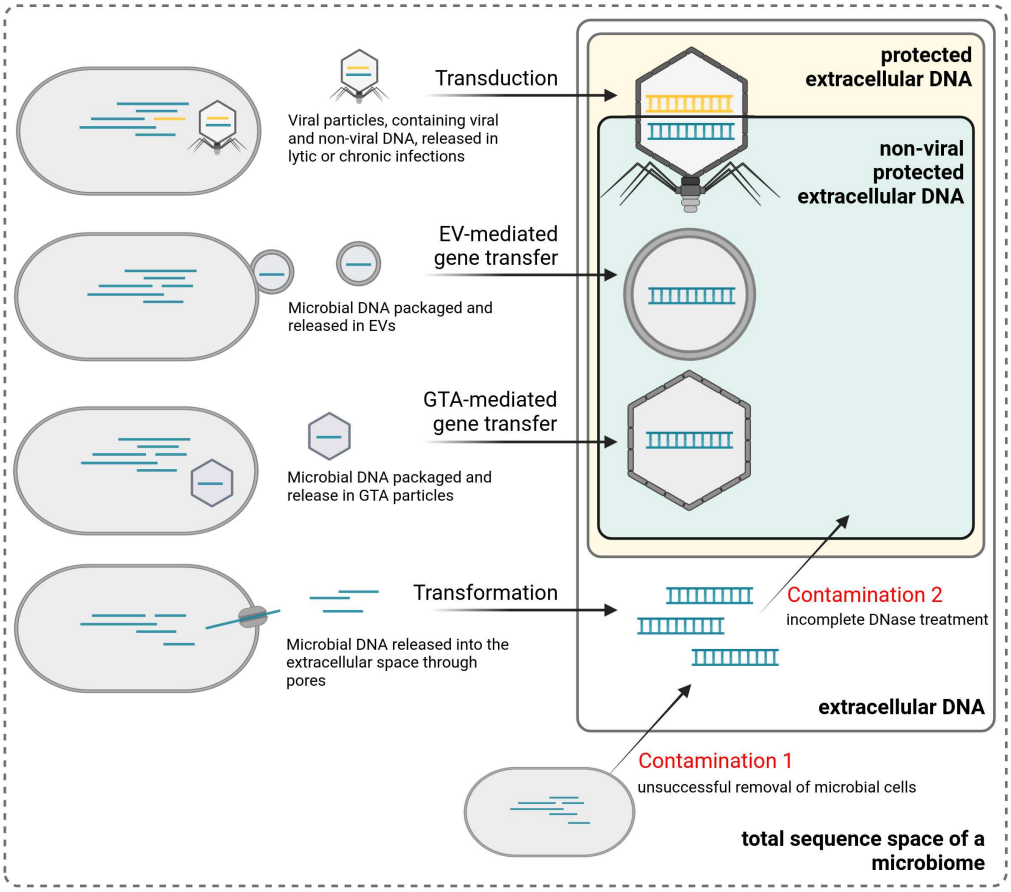


Figure 3.4: Composition of the sequence space of peDNA, traditionally referred to as a 'virome' and potential origins of contamination when analyzing non-viral peDNA. Figure created with `biorender.com`.

of the workflow and a number of design decisions will be discussed here. In general, the approach relies on two conditions. First, the peDNA dataset at hand must be as cell- and contamination-free as possible, i.e. only DNase-treated and size-filtered samples should be considered. Second, genomes of the potential hosts of the peDNA have to be available. Often, high-quality metagenomic assembled genomes (MAGs) are more readily available and, if the set of MAGs was generated from sequence data sampled parallel to the peDNA dataset, are potentially the best representation of the microbial community. Then, by mapping peDNA reads to the potential hosts, most peDNA-producing hosts were identified. Lastly, a primary transport mechanism (GTA transduced, virus transduced, EV-mediated) was identified for each host. This categorization of primary transport mechanisms allowed a semi-quantitative assessment of the different HGT mechanisms in the given dataset.

3.3.3 Key design decisions and limitations

The first critical decision is the use of MAGs instead of complete genomes. MAGs have the advantage that they could be assembled from the same exact location as the peDNA. Future EV-focussed experiments could sample from a specific environment, filter the sample, retrieve MAGs from the microbial fraction, and purify the corresponding peDNA fraction.

The resulting MAGs would very well, depending on sequencing depth, represent the microbial community. The disadvantage of MAGs in this case is that incomplete assembly makes detecting integrated MGEs more difficult. Especially highly flexible regions on a genome, in which viruses or other mobile elements tend to integrate and excise challenge assemblers. A careful selection of a set of highly curated MAGs that best represent the local community is therefore crucial. A second important design decision was the usage of multiple virus prediction tools. Instead of relying on a single predictor, a combination of marker-based and pattern-based predictors was used, maximizing the sensitivity of the workflow. Missing integrated viruses would lead to an underestimation of transduced peDNA compared to peDNA transported in EVs. Naturally, completely unknown viruses will still be missed by current predictors, limiting the workflow's predictive power. Future iterations could implement the most current virus prediction tools, e.g. `geNomad` (Camargo et al. 2023), to increase sensitivity and improve viral detection. Similarly, the detection of GTA gene clusters is limited to known GTA systems and, therefore, limits the workflow's predictive power. While most of the workflow has been implemented in `snakemake`, making it a portable, scalable and automatic pipeline, a manual inspection step is still required. Here, transduction patterns need to be manually analyzed in order to detect the occurrence of active transduction. While this step certainly could be automated in future iterations of the workflow, it currently limits the scalability.

3.4 Conclusion

The growing metagenomic toolset, developed over the last decades, has enabled a mobilome-centric analysis of microbial communities, allowing for a more holistic view of microbiomes. Tools for virus and plasmid discovery and subsequent analysis are well established compared to tools for EV-mediated HGT and - unsurprisingly - other exotic mechanisms and MGEs. Nevertheless, for all aspects of metagenomic analysis, several developments need to be tackled to use the full potential of computer-driven analysis of this specific sequence space. Most developments in MGE identification rely on growing and improving reference databases. Therefore, the following key points should be the focus of future work and are already been addressed in recent publications: i) standardized reporting of novel virus-like sequences (Roux et al. 2019) and other MGEs, ii) improved representation of undersampled and highly diverse environments (Trubl et al. 2020). Lower costs and improvements in sequencing techniques (e.g. PacBio or MinION sequencing) will help to address these key points by enabling the analysis of rare or highly flexible MGEs. For well-described MGEs, benchmarking existing tools in terms of sensitivity, precision and runtime is essential. Additionally, benchmarking different workflows and reference databases for samples of different biomes is similarly crucial since it allows users to select the optimal tool for the specific samples (Paganini et al. 2021; Wu et al. 2023). Even for machine learning and pattern-based approaches, choosing the correct training set is crucial for high accuracy when identifying e.g. viruses in undersampled environments or samples with low viral abundance (Ponsero and Hurwitz 2019). Even the best bioinformatic tools rely on suitable datasets. Therefore, carefully considering the selected methods for sample preparation,

DNA extraction, library preparation, and sequencing is essential. This is particularly true for the detection of EV-mediated HGT since the ad-hoc differentiation between true signals and contamination is impossible. However, for carefully curated datasets, the detection of EV-mediated HGT was successfully demonstrated in Chapter 4. Improvements to the existing pipeline, e.g. the automation of steps that for now require manual curation, would allow for better scalability. However, the results already indicate that EV-mediated HGT is widespread and a major driver of HGT in the oceans. Applying the developed method on a bigger scale and to different environments, such as soils, gut systems and freshwater, should be the focus of future studies. Linking EV-mediated HGT to other MGEs, as demonstrated by Hackl *et al.* (Hackl et al. 2023), is a similarly promising approach for future studies in order to unravel the intricate interactions of MGEs and all mechanisms of HGT that fundamentally shape all biomes on earth.

References

Alarcón-Schumacher, T., D. Lücking, and S. Erdmann (2023): "Revisiting evolutionary trajectories and the organization of the Pleolipoviridae family," *PLOS Genetics*, 19, e1010998, publisher: Public Library of Science.

Andreopoulos, W. B., A. M. Geller, M. Lucke, J. Balewski, A. Clum, N. N. Ivanova, and A. Levy (2022): "Deeplasmid: deep learning accurately separates plasmids from bacterial chromosomes," *Nucleic Acids Research*, 50, e17.

Antipov, D., N. Hartwick, M. Shen, M. Raiko, A. Lapidus, and P. A. Pevzner (2016): "plasmidSPAdes: assembling plasmids from whole genome sequencing data," *Bioinformatics*, 32, 3380–3387.

Bertelli, C., M. R. Laird, K. P. Williams, S. F. U. R. C. Group, B. Y. Lau, G. Hoad, G. L. Winsor, and F. S. Brinkman (2017): "IslandViewer 4: expanded prediction of genomic islands for larger-scale datasets," *Nucleic acids research*, 45, W30–W35, publisher: Oxford University Press.

Biller, S. J., L. D. McDaniel, M. Breitbart, E. Rogers, J. H. Paul, and S. W. Chisholm (2017): "Membrane vesicles in sea water: heterogeneous DNA content and implications for viral abundance estimates," *The ISME Journal*, 11, 394–404, number: 2 Publisher: Nature Publishing Group.

Biller, S. J., F. Schubotz, S. E. Roggensack, A. W. Thompson, R. E. Summons, and S. W. Chisholm (2014): "Bacterial Vesicles in Marine Ecosystems," *Science*, 343, 183–186, publisher: American Association for the Advancement of Science.

Bitto, N. J., R. Chapman, S. Pidot, A. Costin, C. Lo, J. Choi, T. D'Cruze, E. C. Reynolds, S. G. Dashper, L. Turnbull, C. B. Whitchurch, T. P. Stinear, K. J. Stacey, and R. L. Ferrero (2017): "Bacterial membrane vesicles transport their DNA cargo into host cells," *Scientific Reports*, 7, 7072, number: 1 Publisher: Nature Publishing Group.

Breitbart, M., P. Salamon, B. Andresen, J. M. Mahaffy, A. M. Segall, D. Mead, F. Azam, and F. Rohwer (2002): "Genomic analysis of uncultured marine viral communities," *Proceedings of the National Academy of Sciences*, 99, 14250–14255, publisher: Proceedings of the National Academy of Sciences.

Camargo, A. P., S. Roux, F. Schulz, M. Babinski, Y. Xu, B. Hu, P. S. G. Chain, S. Nayfach, and N. C. Kyrpides (2023): "Identification of mobile genetic elements with geNomad," *Nature Biotechnology*, 1–10, publisher: Nature Publishing Group.

Carattoli, A., E. Zankari, A. García-Fernández, M. Voldby Larsen, O. Lund, L. Villa, F. Møller Aarestrup, and H. Hasman (2014): "In Silico Detection and Typing of Plasmids using PlasmidFinder and Plasmid Multilocus Sequence Typing," *Antimicrobial Agents and Chemotherapy*, 58, 3895–3903.

Darling, A. E., B. Mau, and N. T. Perna (2010): "progressiveMauve: multiple genome alignment with gene gain, loss and rearrangement," *PloS one*, 5, e11147, publisher: Public Library of Science San Francisco, USA.

Dominguez-Huerta, G., J. M. Wainaina, A. A. Zayed, A. I. Culley, J. H. Kuhn, and M. B. Sullivan (2023): "The RNA virosphere: How big and diverse is it?" *Environmental Microbiology*, 25, 209, publisher: Wiley-Blackwell.

Duhaime, M. B., L. Deng, B. T. Poulos, and M. B. Sullivan (2012): "Towards quantitative metagenomics of wild viruses and other ultra-low concentration DNA samples: a rigorous assessment and optimization of the linker amplification method," *Environmental Microbiology*, 14, 2526–2537, _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1462-2920.2012.02791.x.

Durrant, M. G., M. M. Li, B. A. Siranosian, S. B. Montgomery, and A. S. Bhatt (2020): "A Bioinformatic Analysis of Integrative Mobile Genetic Elements Highlights Their Role in Bacterial Adaptation," *Cell Host & Microbe*, 27, 140–153.e9.

Eddy, S. R. (2011): "Accelerated Profile HMM Searches," *PLOS Computational Biology*, 7, e1002195, publisher: Public Library of Science.

Edgar, R. C., J. Taylor, V. Lin, T. Altman, P. Barbera, D. Meleshko, D. Lohr, G. Novakovsky, B. Buchfink, and B. Al-Shayeb (2022): "Petabase-scale sequence alignment catalyses viral discovery," *Nature*, 602, 142–147, publisher: Nature Publishing Group UK London.

Fouts, D. E. (2006): "Phage_Finder: automated identification and classification of prophage regions in complete bacterial genome sequences," *Nucleic acids research*, 34, 5839–5851, publisher: Oxford University Press.

Frost, L. S., R. Leplae, A. O. Summers, and A. Toussaint (2005): "Mobile genetic elements: the agents of open source evolution," *Nature Reviews Microbiology*, 3, 722–732, number: 9 Publisher: Nature Publishing Group.

Gomi, R., K. L. Wyres, and K. E. Holt (2021): "Detection of plasmid contigs in draft genome assemblies using customized Kraken databases," *Microbial Genomics*, 7, 000550.

Gregory, A. C., A. A. Zayed, N. Conceicao-Neto, B. Temperton, B. Bolduc, A. Alberti, M. Ardyna, K. Arkhipova, M. Carmichael, C. Cruaud, C. Dimier, G. Domínguez-Huerta, J. Ferland, S. Kandels, Y. Liu, C. Marec, S. Pesant, M. Picheral, S. Pisarev, J. Poulain, J.-E. Tremblay, D. Vik, S. G. Acinas, M. Babin, P. Bork, E. Boss, C. Bowler, G. Cochrane, C. de Vargas, M. Follows, G. Gorsky, N. Grimsley, L. Guidi, P. Hingamp, D. Iudicone, O. Jaillon, S. Kandels-Lewis, L. Karp-Boss, E. Karsenti, F. Not, H. Ogata, S. Pesant, N. Poulton, J. Raes, C. Sardet, S. Speich, L. Stemmann, M. B. Sullivan, S. Sunagawa, P. Wincker, M. Babin, C. Bowler, A. I. Culley, C. de Vargas, B. E. Dutilh, D. Iudicone, L. Karp-Boss, S. Roux, S. Sunagawa, P. Wincker, and M. B. Sullivan (2019): "Marine DNA Viral Macro- and Microdiversity from Pole to Pole," *Cell*, 177, 1109–1123.e14.

Hackl, T., R. Laurenceau, M. J. Ankenbrand, C. Bliem, Z. Cariani, E. Thomas, K. D. Dooley, A. A. Arellano, S. L. Hogle, P. Berube, G. E. Leventhal, E. Luo, J. M. Eppley, A. A. Zayed, J. Beaulaurier, R. Stepanauskas, M. B. Sullivan, E. F. DeLong, S. J. Biller, and S. W. Chisholm (2023): "Novel integrative elements and genomic plasticity in ocean ecosystems," *Cell*, 186, 47–62.e16.

Hillebrandt, N., P. Vormittag, N. Bluthardt, A. Dietrich, and J. Hubbuch (2020): "Integrated Process for Capture and Purification of Virus-Like Particles: Enhancing Process Performance by Cross-Flow Filtration," *Frontiers in Bioengineering and Biotechnology*, 8.

Jiang, C., C. Chen, Z. Huang, R. Liu, and J. Verdier (2015): "ITIS, a bioinformatics tool for accurate identification of transposon insertion sites using next-generation sequencing data," *BMC bioinformatics*, 16, 1–8, publisher: BioMed Central.

Johnson, T. J., C. M. Logue, J. R. Johnson, M. A. Kuskowski, J. S. Sherwood, H. J. Barnes, C. DebRoy, Y. M. Wannemuehler, M. Obata-Yasuoka, L. Spanjaard, and L. K. Nolan (2012): "Associations Between Multidrug Resistance, Plasmid Content, and Virulence Potential Among Extraintestinal Pathogenic and Commensal Escherichia coli from Humans and Poultry," *Foodborne Pathogens and Disease*, 9, 37–46, publisher: Mary Ann Liebert, Inc., publishers.

Kim, K.-H. and J.-W. Bae (2011): "Amplification Methods Bias Metagenomic Libraries of Uncultured Single-Stranded and Double-Stranded DNA Viruses," *Applied and Environmental Microbiology*, 77, 7663–7668.

Kosmopoulos, J. C., K. M. Klier, M. V. Langwig, P. Q. Tran, and K. Anantharaman (2023): "Viromes vs. mixed community metagenomes: choice of method dictates interpretation of viral community ecology," Pages: 2023.10.15.562385 Section: New Results.

Krawczyk, P., L. Lipinski, and A. Dziembowski (2018): "PlasFlow: predicting plasmid sequences in metagenomic data using genome signatures," *Nucleic acids research*, 46.

Lerat, E. (2010): "Identifying repeats and transposable elements in sequenced genomes: how to find your way through the dense forest of programs," *Heredity*, 104, 520–533, publisher: Nature Publishing Group.

Li, W. K. W. and P. M. Dickie (1985): "Growth of bacteria in seawater filtered through 0.2 um Nuclepore membranes: implications for dilution experiments," *Marine Ecology Progress Series*, 26, 245–252, publisher: Inter-Research Science Center.

Linney, M. D., J. M. Eppley, A. E. Romano, E. Luo, E. F. DeLong, and D. M. Karl (2022): "Microbial Sources of Exocellular DNA in the Ocean," *Applied and Environmental Microbiology*, 88, e02093–21, publisher: American Society for Microbiology.

Linney, M. D., C. R. Schvarcz, G. F. Steward, E. F. DeLong, and D. M. Karl (2021): "A method for characterizing dissolved DNA and its application to the North Pacific Subtropical Gyre," *Limnology and Oceanography: Methods*, 19, 210–221, _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/lom3.10415.

Lücking, D., C. Mercier, T. Alarcón-Schumacher, and S. Erdmann (2023): "Extracellular vesicles are the main contributor to the non-viral protected extracellular sequence space," *ISME Communications*, 3, 1–10, number: 1 Publisher: Nature Publishing Group.

Malathi, V. G. and P. R. Devi (2019): "ssDNA viruses: key players in global virome," *VirusDisease*, 30, 3, publisher: Springer.

McNamara, R. P. and D. P. Dittmer (2020): "Modern Techniques for the Isolation of Extracellular Vesicles and Viruses," *Journal of neuroimmune pharmacology : the official journal of the Society on NeuroImmune Pharmacology*, 15, 459–472.

Müller, R. and C. Chauve (2019): "HyAsP, a greedy tool for plasmids identification," *Bioinformatics*, 35, 4436–4439.

Paganini, J. A., N. L. Plantinga, S. Arredondo-Alonso, R. J. L. Willems, and A. C. Schürch (2021): "Recovering Escherichia coli Plasmids in the Absence of Long-Read Sequencing Data," *Microorganisms*, 9, 1613, number: 8 Publisher: Multidisciplinary Digital Publishing Institute.

Parras-Moltó, M., A. Rodríguez-Galet, P. Suárez-Rodríguez, and A. López-Bueno (2018): "Evaluation of bias induced by viral enrichment and random amplification protocols in metagenomic surveys of saliva DNA viruses," *Microbiome*, 6, 119.

Pellow, D., A. Zorea, M. Probst, O. Furman, A. Segal, I. Mizrahi, and R. Shamir (2021): "SCAPP: an algorithm for improved plasmid assembly in metagenomes," *Microbiome*, 9, 144.

Pemberton, J. M. and R. H. Don (1981): "Bacterial plasmids of agricultural and environmental importance," *Agriculture and Environment*, 6, 23–32.

Pfeifer, E., J. A. Moura de Sousa, M. Touchon, and E. P. C. Rocha (2021): "Bacteria have numerous distinctive groups of phage–plasmids with conserved phage and variable plasmid gene repertoires," *Nucleic Acids Research*, 49, 2655–2673.

Ponsero, A. J. and B. L. Hurwitz (2019): "The Promises and Pitfalls of Machine Learning for Detecting Viruses in Aquatic Metagenomes," *Frontiers in Microbiology*, 10.

Ren, J., N. A. Ahlgren, Y. Y. Lu, J. A. Fuhrman, and F. Sun (2017): "VirFinder: a novel k-mer based tool for identifying viral sequences from assembled metagenomic data," *Microbiome*, 5, 69.

Ren, J., K. Song, C. Deng, N. A. Ahlgren, J. A. Fuhrman, Y. Li, X. Xie, R. Poplin, and F. Sun (2020): "Identifying viruses from metagenomic data using deep learning," *Quantitative Biology*, 8, 64–77, publisher: Springer.

Roux, S., E. M. Adriaenssens, B. E. Dutilh, E. V. Koonin, A. M. Kropinski, M. Krupovic, J. H. Kuhn, R. Lavigne, J. R. Brister, A. Varsani, C. Amid, R. K. Aziz, S. R. Bordenstein, P. Bork, M. Breitbart, G. R. Cochrane, R. A. Daly, C. Desnues, M. B. Duhaime, J. B. Emerson, F. Enault, J. A. Fuhrman, P. Hingamp, P. Hugenholtz, B. L. Hurwitz, N. N. Ivanova, J. M. Labonté, K.-B. Lee, R. R. Malmstrom, M. Martinez-Garcia, I. K. Mizrachi, H. Ogata, D. Páez-Espino, M.-A. Petit, C. Putonti, T. Rattei, A. Reyes, F. Rodriguez-Valera, K. Rosario, L. Schriml, F. Schulz, G. F. Steward, M. B. Sullivan, S. Sunagawa, C. A. Suttle, B. Temperton, S. G. Tringe, R. V. Thurber, N. S. Webster, K. L. Whiteson, S. W. Wilhelm, K. E. Wommack, T. Woyke, K. C. Wrighton, P. Yilmaz, T. Yoshida, M. J. Young, N. Yutin, L. Z. Allen, N. C. Kyrpides, and E. A. Eloe-Fadrosh (2019): "Minimum Information about an Uncultivated Virus Genome (MIUViG)," *Nature Biotechnology*, 37, 29–37.

Roux, S., F. Enault, G. Bronner, D. Vaulot, P. Forterre, and M. Krupovic (2013): "Chimeric viruses blur the borders between the major groups of eukaryotic single-stranded DNA viruses," *Nature Communications*, 4, 2700, number: 1 Publisher: Nature Publishing Group.

Roux, S., S. J. Hallam, T. Woyke, and M. B. Sullivan (2015): "Viral dark matter and virus-host interactions resolved from publicly available microbial genomes," *eLife*, 4, e08490.

Schwengers, O., P. Barth, L. Falgenhauer, T. Hain, T. Chakraborty, and A. Goesmann (2020): "Platon: identification and characterization of bacterial plasmid contigs in short-read draft assemblies exploiting protein sequence-based replicon distribution scores," *Microbial Genomics*, 6, mgen000398.

Sen, D., G. A. Van der Auwera, L. M. Rogers, C. M. Thomas, C. J. Brown, and E. M. Top (2011): "Broad-Host-Range Plasmids from Agricultural Soils Have IncP-1 Backbones with Diverse Accessory Genes," *Applied and Environmental Microbiology*, 77, 7975–7983.

Sorensen, J. W., L. A. Zinke, A. M. ter Horst, C. Santos-Medellín, A. Schroeder, and J. B. Emerson (2021): "DNase Treatment Improves Viral Enrichment in Agricultural Soil Viromes," *mSystems*, 6, 10.1128/msystems.00614–21, publisher: American Society for Microbiology.

Starikova, E. V., P. O. Tikhonova, N. A. Prianichnikov, C. M. Rands, E. M. Zdobnov, E. N. Ilina, and V. M. Govorun (2020): "Phigaro: high-throughput prophage sequence annotation," *Bioinformatics*, 36, 3882–3884.

Sukhorukov, G., M. Khalili, O. Gascuel, T. Candresse, A. Marais-Colombel, and M. Nikolski (2022): "VirHunter: A Deep Learning-Based Method for Detection of Novel RNA Viruses in Plant Sequencing Data," *Frontiers in Bioinformatics*, 2.

Tang, X., J. Shang, Y. Ji, and Y. Sun (2023): "PLASMe: a tool to identify PLASMid contigs from short-read assemblies using transformer," *Nucleic Acids Research*, 51, e83.

Thomas, C. M. and K. M. Nielsen (2005): "Mechanisms of, and Barriers to, Horizontal Gene Transfer between Bacteria," *Nature Reviews Microbiology*, 3, 711–721, number: 9 Publisher: Nature Publishing Group.

Trubl, G., P. Hyman, S. Roux, and S. T. Abedon (2020): "Coming-of-Age Characterization of Soil Viruses: A User's Guide to Virus Isolation, Detection within Metagenomes, and Viromics," *Soil Systems*, 4, 23, number: 2 Publisher: Multidisciplinary Digital Publishing Institute.

Van der pol, E., A. G. Hoekstra, A. Sturk, C. Otto, T. G. Van leeuwen, and R. Nieuwland (2010): "Optical and non-optical methods for detection and characterization of microparticles and exosomes," *Journal of Thrombosis and Haemostasis*, 8, 2596–2607.

Wu, L.-Y., N. Pappas, Y. Wijesekara, G. J. Piedade, C. P. D. Brussaard, and B. E. Dutilh (2023): "Benchmarking Bioinformatic Virus Identification Tools Using Real-World Metagenomic Data across Biomes," Pages: 2023.04.26.538077 Section: New Results.

Chapter 4

Extracellular vesicles are the main contributor to the non-viral protected extracellular sequence space

Dominik Lücking[1],
Coraline Mercier[1],
Tomás Alarcon-Schumacher[1]
and Susanne Erdmann[1]

Manuscript accepted and published in *ISME Communications*.

Supplementary material in Appendix B.

[1]Max-Planck Institute for Marine Microbiology, Bremen, Germany

ISME
www.nature.com/ismecomms

ARTICLE OPEN

Check for updates

Extracellular vesicles are the main contributor to the non-viral protected extracellular sequence space

Dominik Lücking[1], Coraline Mercier[1], Tomas Alarcón-Schumacher[1] and Susanne Erdmann[1]✉

Environmental virus metagenomes, commonly referred to as "viromes", are typically generated by physically separating virus-like particles (VLPs) from the microbial fraction based on their size and mass. However, most methods used to purify VLPs, enrich extracellular vesicles (EVs) and gene transfer agents (GTAs) simultaneously. Consequently, the sequence space traditionally referred to as a "virome" contains host-associated sequences, transported via EVs or GTAs. We therefore propose to call the genetic material isolated from size-fractionated (0.22 µm) and DNase-treated samples *protected environmental DNA (peDNA)*. This sequence space contains viral genomes, DNA transduced by viruses and DNA transported in EVs and GTAs. Since there is no genetic signature for peDNA transported in EVs, GTAs and virus particles, we rely on the successful removal of contaminating remaining cellular and free DNA when analyzing peDNA. Using marine samples collected from the North Sea, we generated a thoroughly purified peDNA dataset and developed a bioinformatic pipeline to determine the potential origin of the purified DNA. This pipeline was applied to our dataset as well as existing global marine "viromes". Through this pipeline, we identified known GTA and EV producers, as well as organisms with actively transducing proviruses as the source of the peDNA, thus confirming the reliability of our approach. Additionally, we identified novel and widespread EV producers, and found quantitative evidence suggesting that EV-mediated gene transfer plays a significant role in driving horizontal gene transfer (HGT) in the world's oceans.

ISME Communications; https://doi.org/10.1038/s43705-023-00317-6

INTRODUCTION

The presence of extracellular entities strongly shapes microbial communities. Particles of various origins mediate the transport of genetic material from one cell to another, thus playing a crucial role in horizontal gene transfer (HGT) [1, 2]. Most prominently, viruses are highly abundant and diverse drivers of ecological and evolutionary interactions within a community [3–6]. However, due to the limited culturability of their hosts, viruses often escape traditional culture-based approaches [7], leading to the development of culture-independent techniques to study their fundamental impact on microbial communities, global biogeochemical cycles and their effect on climate change. Similar to metagenomic studies, researchers have sequenced and analyzed the genetic content of the viral fraction on a community level, leading to the advent of viral metagenomics or "viromics" [8]. This approach traditionally relies on the physical, pre-sequencing separation of virus-like particles (VLPs) from microbial cells. Methods like sequential size filtration, ultracentrifugation, tangential flow filtration, and flow cytometry exploit the distinct physical properties of VLPs when compared to microbes [9–12]. Additionally, bioinformatic methods were developed to identify virus-like sequences among microbial sequences [13–16]. This led to the discovery of many diverse viruses, fulfilling crucial functions in their respective microbial community [17]. Interestingly, even after the most thorough removal of microbial cells, non-viral genes were shown to be present in "viromes" generated from many diverse environments [7, 18]. The contamination of "viromes" with microbial sequences originating from remaining intact cells and free extracellular DNA has been reported in several studies. Consequently, tools have been developed to estimate the proportion of true viral DNA in a given virome, using the abundance of reads with homologs in available prokaryotic databases or a specific set of microbial marker genes (e.g., 16S rRNA gene) [18–20]. These tools fall short of assessing the true degree of contamination of a "virome", because the abundance of prokaryotic-, non-virus-like genes is not necessary due to microbial contaminations, but can be the result of horizontal gene transfer processes.

Long before the development of modern "viromics", studies showed that viruses carry and distribute random microbial genes [21, 22] or specific "auxiliary metabolic genes" (AMGs) in addition to bona fide viral genes (e.g., genes necessary for particle assembly or viral genome replication), in a well-described process termed "transduction". Here, either genetic material adjacent to the integrated viral genome (specialized transduction) or random snippets of the host genome (general transduction) are packaged into the viral particles [1]. AMGs have been shown to fundamentally alter the metabolism of microbes by providing genes otherwise unavailable to their host [8, 23], further demonstrating the need for a good understanding of non-viral DNA in viromes.

Viromes are traditionally generated by separating VLPs from cells. However, methods that enrich VLPs by removing larger and heavier microbial cells also enrich entities similar in size and mass to VLPs. Most prominently, gene transfer agents (GTAs) and

[1]Max-Planck Institute for Marine Microbiology, Celsiusstraße 1, 28359 Bremen, Germany. ✉email: serdmann@mpi-bremen.de

Received: 25 May 2023 Revised: 28 September 2023 Accepted: 5 October 2023
Published online: 17 October 2023

extracellular vesicles (EVs, also referred to as membrane vesicles, MVs, or outer membrane vesicles, OMVs) are particles with similar physical properties and both have been shown to be involved in HGT, thus contributing to the presence of non-viral DNA in "viromes".

GTAs are particles transporting host DNA from one cell to another. They likely derived from defective prophages, and retained functional genes for the head and tail components of a head-tailed virus particle, including the genes for DNA packaging. Therefore, mass and size (40–60 µm) of GTA particles are very similar to head-tailed viruses, making it hard to differentiate them from viruses solely based on morphology. Notably and in contrast to true viruses, GTAs do not specifically package the GTA-producing gene cluster into the particle, but transport short segments of the host genome. Up to this date, several distinct gene clusters have been identified that produce GTAs [24–28].

Prokaryotic EVs are small (10–300 nm) spherical structures derived from the cell membrane [29]. EVs represent compartments that protect their cargo from degradation and are used for the transport of a variety of different components across the extracellular space. This includes the transport of nutrients, toxins, antigens, lipids, proteins, RNA, and DNA [30–36]. Recent studies showed an abundance of EVs in marine environments of up to 10^6 vesicles per milliliter [33], produced across diverse taxa. EVs, produced by highly abundant marine heterotrophs and autotrophs, such as Pelagibacter, Marinobacter, and Prochlorococcus, have been shown to transport fragments of chromosomal and plasmid DNA [37–41], thus contributing to the fraction of non-viral DNA within viromes.

In this study, we aimed to explore the non-viral sequence space of viromics datasets. First, we generated our own dataset, carefully avoiding possible contaminations. Then we categorized the sequences from this dataset and publicly available viromics datasets as virus- or non-virus-derived. Subsequently, we explore the non-virus-derived sequence space to detect the extent of non-viral DNA potentially being horizontally transferred between cells. We then identify the means of transport (GTA-, EV-, or virus-driven) for the sequences by linking the datasets to existing microbial metagenomes and genomes. We identify potential novel EV- and GTA producers and metagenomics-assembled genomes (MAGs) with an actively transducing virus. We propose using the term "protected extracellular DNA" (peDNA) for DNA sequence data derived from appropriately purified environmental fractions <0.2 µm, so far referred to as viromics datasets.

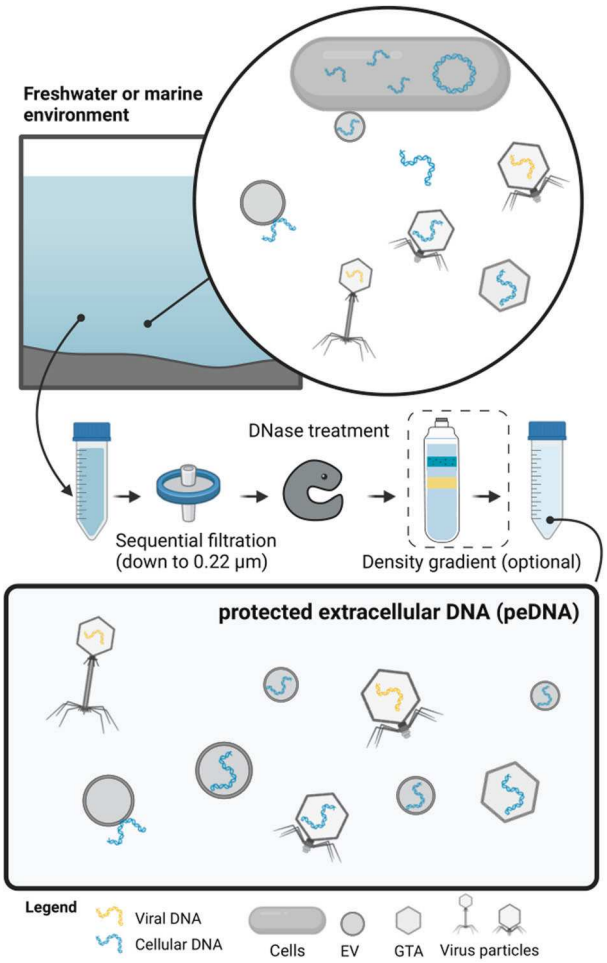


Fig. 1 Conceptual composition of protected extracellular DNA. The top panel depicts microbial entities present in a water body: microbial cells, viruses containing viral and microbial genetic material, gene transfer agents and extracellular vesicles containing host DNA. After size filtration (0.22 µm) and DNase treatment and, if applicable, purification via density gradients, microbial cells and free DNA are removed (middle panel). The remaining DNA makes up the sequence space of protected extracellular DNA, peDNA (bottom panel).

RESULTS AND DISCUSSION

Viromics datasets represent the sequence space of protected extracellular DNA (peDNA)

The majority of samples prepared for "viromics" include GTAs and EVs, in addition to virus particles. All three entities are small protein- and/or lipid-containing particles that can enclose cellular DNA [3] or were found to bind cellular DNA on their surface [42], thus inflating the sequence space that traditionally has been described as a "virome". In contrast to free extracellular DNA, DNA that is enclosed in or tightly associated with particles, or DNA that is tightly enclosed in protein/DNA or DNA/RNA complexes, is protected against degradation by extracellular nucleases occurring in the environment or nucleases used to clean samples from free extracellular DNA. Hence, we propose the term "protected extracellular DNA" (peDNA) to describe the entirety of DNA transported by viruses, GTAs and EVs (Fig. 1). We will use this term throughout this work.

Purification of environmental samples for the generation of peDNA datasets is essential to explore of the entire dataset

Previously, the percentage of 16S/18S rRNA-mapping reads was used as a proxy for host contamination in virome datasets [19]. However, it has since been shown that GTAs and EVs enclose host DNA randomly, including 16S/18S rRNA genes [43]. We calculated SSU rRNA alignment rates for two highly purified samples: DNA extracted from virus isolates, purified by sequential plaque assays and 0.2 um size filtration [44] and DNA extracted from EVs purified from culture supernatants of *Prochlorococcus* [33] (Fig. 2). While the alignment rates were low for virus isolates (mean = 0.000437%), the percentage of 16S/18S-mapping reads was five orders of magnitude higher in DNA extracted from purified EVs (1.81%), even exceeding the mean alignment rate of publicly available microbial environmental metagenomes (mean = 0.078%, [19]). For these samples, Biller et al. confirmed the absence of microbial cells using electron microscopy. Thus, the presence of 16/18S rRNA-mapping reads neither proves or disproves contamination in peDNA samples. The only way to exclude contaminations with cellular DNA or extracellular free DNA is the rigorous purification of the sample before sequencing.

For this purpose, we generated a dataset from rigorously purified samples using several sequential filtration steps, DNase treatment and density gradient purification (Methods, Supplementary Fig. 1), resulting in cell-free samples containing virus-like particles, GTA particles and EVs. Subsequently, we compared the

SSU rRNA alignment rates of our dataset with one metagenomic peDNA dataset and two viromic datasets: Density gradient-purified EVs isolated from seawater samples ("Environmental EV enrichment") [33]; the "Tara Oceans virome" dataset [17], purified by size filtration and DNase treatment; and the "GDOCB virome" dataset [45]. GDOCB viromes are purified by flow cytometry. The process excludes free DNA and microbial cells with physical properties outside of the analyzed size spectrum, which makes it possible to assume that these viromes are free from contamination. Both datasets showed increased SSU rRNA alignment rates (Fig. 2), with some samples even exceeding the microbial metagenome alignment rate of 0.078%. Likewise, our dataset, while on average showing lower alignment rates (mean = 0.066%), contained samples exceeding that threshold. Lastly, Tara Oceans viromes mostly showed very low SSU rRNA alignment rates, with very few exceptions (mean = 0.031%). Overall, even thoroughly purified and confirmed contamination-free datasets show highly variable SSU rRNA alignment rates. We concluded that the majority of SSU rRNA hits in these datasets are enclosed in VLPs, GTA's or EVs rather than in contaminating microbial cells, and therefore included the datasets in the subsequent analysis.

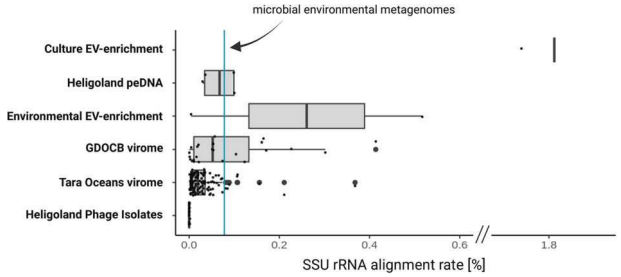


Fig. 2 Comparison of SSU rRNA alignment rates of diverse viromes and EV preparations. Each dot represents the percentage of reads aligning to either 16S or 18S rRNA genes. The cyan line indicates the average alignment rate for publicly available metagenomes from various environments [19]. "Culture EV enrichment": a cell-free preparation of EVs from *Prochlorococcus* cultures [33]. "Heligoland peDNA": dataset generated in this study from a highly purified (filtration, DNase treatment, gradient purification) <0.2 µm Heligoland water fraction. "Environmental EV enrichment": Density gradient-purified EVs isolated from seawater samples [33]. "GDOCB virome": <0.2 µm fraction enriched for VLPs by flow cytometry [42]. "Tara Oceans virome": <0.2 µm fraction purified by size filtrations and DNase treatment [17]. "Heligoland phage Isolates": DNA extracted from virus isolates, purified by 0.2 µm size filtration [41].

Separation of non-viral (nvpeDNA) from viral protected extracellular DNA (vpeDNA) indicates that EVs and GTAs could be very abundant entities in the ocean

The sequence space of protected extracellular DNA (peDNA) consists of virus genomes (viral protected extracellular DNA, vpeDNA) and non-viral, microbial DNA (non-viral extracellular DNA, nvpeDNA), deriving from transducing viruses, GTAs and EVs. nvpeDNA represents the sequence space that is potentially horizontally transferred between cells and therefore has major implications on the ecology and evolution of the organism in this environment and the environment itself.

In order to separate non-viral peDNA from viral peDNA, we developed a bioinformatic pipeline that, in brief, identifies virus sequences, separates those from non-viral sequences and calculates a non-viral to viral peDNA ratio in a given dataset (Fig. S1 and Methods). This pipeline was first applied to isolated viruses and purified EVs (Fig. 3A). As expected, vpeDNA made up >99% of the DNA of purified Heligoland phage isolates and nvpeDNA made up 98% of the DNA transported in purified EVs of a pure *Prochlorococcus* culture, verifying that the pipeline is reliably separating vpeDNA from nvpeDNA.

Consequently, we applied the pipeline to the entire Helgoland peDNA dataset and the other datasets that we verified earlier to be reasonably contamination-free peDNA datasets (Fig. 3A). Helgoland peDNA contained 39% non-viral reads. In the Tara Oceans Viromes, nvpeDNA made up, on average, 40% of all reads (105 samples). While viruses are considered the most abundant nucleic acid-containing biological entities in the ocean [3], these findings clearly indicate that EVs and GTAs, transferring cellular DNA, are likely very abundant entities as well. Surprisingly, in GDOCB viromes, the proportion of nvpeDNA to vpeDNA was even

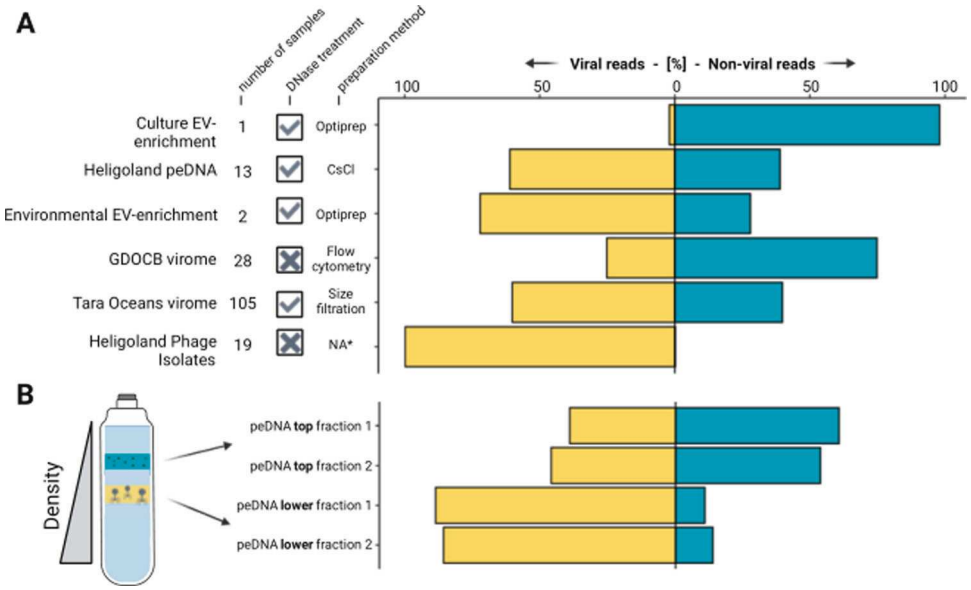


Fig. 3 Non-viral to viral peDNA ratios. A Non-viral to viral peDNA ratio across different studies. Each bar represents the percentage of read pairs mapping to contigs classified to be non-viral or viral for viromes, peDNA enrichments, EV enrichments and pure phage isolates. Sample size and purification methods are indicated for each sample. **B** Non-viral to viral peDNA ratio in different fractions of CsCl gradients. Left, schematic view of seawater samples running through CsCl gradients (adapted from [43]). Right, non-viral to viral peDNA ratio for top and bottom fractions in CsCl gradients for the Heligoland peDNA sample.

higher (75% nvpeDNA). However, this may be due to the comparably low sequencing depth, resulting in incomplete assembly and therefore hindering a reliable identification of virus contigs. Additionally, since the samples were not DNase treated, particles with the same size as the sorted VLPs might carry free DNA attached to the surface of the particles, therefore inflating nvpeDNA. In order to avoid any artificially introduced biases, this dataset was excluded from further downstream work.

We analyzed individual fractions of density gradients that were used to purify the Heligoland peDNA, because it was shown previously that density gradients can separate VLPs from EVs. VLPs, also including GTAs, were found to be more abundant in lower fractions of the gradients, while EV-like particles were more abundant in upper fractions [33, 46]. Indeed, in the upper gradient fractions of the Helgoland peDNA, nvpeDNA made up 60 and 54% of all reads in two biological replicates, while in the lower fraction nvpeDNA made up only 14 and 11% (Fig. 3B), confirming the previous observations. Thus the proportion of nvpeDNA is much higher in the upper fraction, that enriches EVs additionally to some VLPs, compared to the lower fraction, that enriches mainly virus particles and GTA's. This indicates that EVs are likely contributing significantly more to the nvpeDNA sequence space than GTA's and viruses.

Identifying the origin of non-viral protected extracellular DNA reveals that EVs could be the main driver of horizontal gene transfer in the oceans

For contamination-free peDNA samples, we consider three major possible origins of nvpeDNA: DNA transduced by viruses, DNA transported in GTA particles and DNA associated with EVs (Fig. 4A). While we cannot exclude that some of the nvpeDNA could originate from very stable protein/DNA or DNA/RNA complexes, we assume that the proportion of these complexes is rather small in comparison with DNA enclosed in particles.

It is inherently difficult to differentiate the origin of nvpeDNA based on sequence content. To this date, there are no reports on specific sequence signatures (e.g., marker genes) for DNA transported in EVs or GTAs, making DNA transported in EVs indistinguishable from DNA transported in GTA's or virus particles. Therefore, we developed a bioinformatic approach that tackles this differentiation from a different perspective. First, each read in the nvpeDNA fraction was linked to a given potential microbial host (Fig. S2). Then, the 20 most nvpeDNA recruiting MAGs per sample were selected. The main mechanism, which is most likely used to transport its DNA into the extracellular space was predicted, thus linking each read to either EV-, GTA- or transduction-associated transport. We confirmed that the abundance of these organisms (MAGs) in peDNA datasets does not correlate ($R^2 < 0.01$) with their abundance in the corresponding metagenomes (Fig. S3), and that none of the organisms identified are known to produce particularly small cells that could pass 0.2 µm filters. This additionally indicates that the high abundance of their genomic DNA in the nvpeDNA fraction is not due to cellular contamination with cells, but indeed the active transport of genomic DNA into the extracellular space via EVs, GTAs or virus particles. This approach was applied to nine Tara Oceans viromes which were linked to 2307 MAGs [44] and the entire Heligoland peDNA dataset which was linked to 457 MAGs sequenced from seawater coming from the same sampling station on Heligoland [47]. This yielded 200 MAGs (180 from Tara Oceans plus 20 from Heligoland), subsequently categorized as either GTA- or EV-producer or containing an actively transducing provirus. For details on the categorization approach, refer to the following sections, in brief: MAGs that contained an active (increased coverage in provirus region) provirus were labeled as "transducer". Similarly, the respective MAG was labeled as "GTA producer" if a complete or nearly complete GTA cluster could be identified. If neither an active provirus or a GTA cluster was identified, the MAG

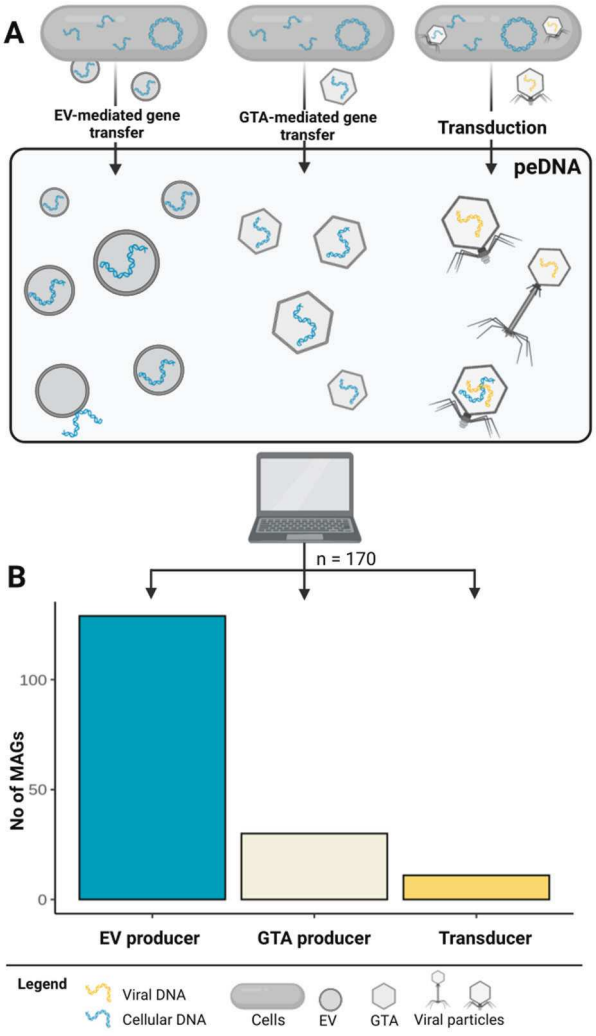


Fig. 4 Mechanisms of horizontal gene transfer that contribute to peDNA. A The origin of sequences comprising peDNA can be either EV-mediated or GTA-mediated gene transfer or via transduction. **B** Number of MAGs assigned to each mechanism. Of 200 analyzed metagenomic assembled MAGs, 170 could be assigned to predominantly use one of the three mechanisms in order to transport their genetic material into the extracellular space: 129 EV producers, 30 GTA producers and 11 genomes, with an actively transducing phage

was labeled as "EV producer" (Figs. S2 and S4). All labels were manually checked and verified by scrutinizing coverage plots (Fig. S5), prophage regions and GTA clusters.

Among the 200 top peDNA-recruiting MAGs 170 could be assigned unambiguously to one of the three categories (Fig. 4B). Most importantly, the majority was identified as EV producers, confirming that EVs that have been shown to be very abundant in the marine environment [38], are not just abundant entities but also significantly contribute to the peDNA sequence space and thereby are likely one of the most important drivers of horizontal gene transfer.

Identification of four novel GTA producers with RcGTA-like clusters

Of the 170 unambiguously assigned MAGs, 30 were identified to contain a functional (>10 GTA-associated genes, core genes present) GTA cluster. For most identified GTA producers, the presence of a GTA cluster has been described elsewhere: *Roseobacter sp.* ($n = 16$), *Sulfitobacter sp.* ($n = 8$) and *Roseovarius sp.* ($n = 1$) are known GTA producers of the order *Rhodobacterales*

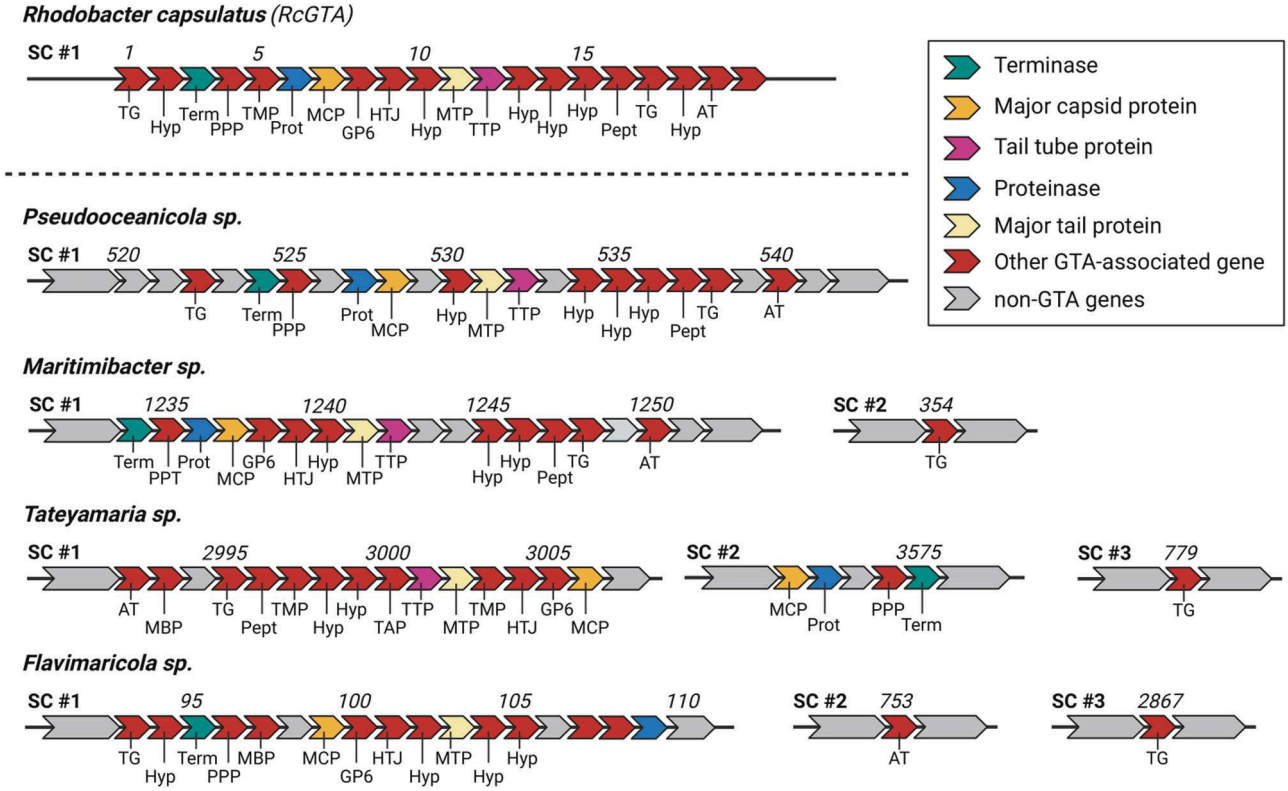


Fig. 5 Genome maps of four novel GTA producers. Organization of genes for the four identified, potentially novel GTAs, compared to the GTA cluster of *Rhodobacter capsulatus*. ORF number is given above the map, encoded protein function is indicated below. Non-GTA encoding genes are gray, GTA-associated genes are shown in red. In addition, core GTA encoding genes are colored accordingly. Encoding protein function is given below: AT acetyltransferase, MBP membrane bound protein, TG transglycosylase, Pept phage cell wall peptidase, TMP transmembrane protein, Hyp hypothetical, TAP tail assembly protein, TTP phage tail tube protein, MTP phage major tail protein, HTJ head-tail joining protein, GP6 gp6-like protein, MCP major capsid protein, Prot proteinase, PPP phage portal protein.

[48–50], indicating the efficiency of our approach. Additionally, we detected a functional GTA cluster in four more species: *Tateyamaria sp.*, *Pseudooceanicola sp.*, *Maritimibacter sp.* and *Flavimaricola sp.* all contained RcGTA homologs, including genes encoding the major capsid protein, a terminase, a proteinase and proteins associated with tail assembly (Fig. 5). The GTA cluster on the genome of the *Tateyamaria* species was distributed more widely over the genome in partial subclusters, as it has been described elsewhere [26]. As far as we know, these four species' clusters are not described elsewhere. We suggest that they are complete and functional because the core genes necessary to form GTA particles are present [51]; however, laboratory experiments are necessary to confirm their full functionality.

Only a few transducing proviruses could be identified with confidence in the peDNA sequence space

In order to label a MAG as containing an actively transducing provirus and therefore as a transducer, we relied on a combination of virus prediction tools, manual analysis of coverage plots and functional annotation of the proviral regions (see "Methods"—"Identification of potential transducers extracellular vesicle- and gene transfer agent producers"). The virus genome itself is present in all viral particles produced, in contrast to the transduced microbial DNA, which could be a randomly selected host DNA fragment (general transduction) or a specifically selected region (specialized transduction). In both cases, the coverage over the proviral region should be increased compared to the surrounding non-viral regions (Fig. 6C). Thus, only MAGs which showed the expected coverage profiles were labeled as transducers. Contrastingly, if a region recruited no reads from the peDNA fraction, the region was considered absent and the MAG was labeled as an EV producer (compare Fig. 6A *Haliea sp.* Station 158 SRF). In some (17 out of 200) cases, a clear assignment was not possible due to inconclusive coverage profiles or contradicting GTA and prophage predictions. These MAGs were labeled "unclear" and removed from further analysis. We identified 11 MAGs that carried an integrated and actively transducing provirus. Interestingly, a *Haliea sp.* with a proviral region was identified that recruited peDNA reads coming from one sampling station but none from the other. However, the non-viral part of the genome recruited high amounts of reads in both stations, albeit with lower coverage in the station where the provirus was absent (Fig. 6B). We hypothesize that there are two separate, distinct populations of the same *Haliea* species at the two stations: one, with the provirus integrated and actively transducing and one without the provirus. The fact that DNA from the population without the provirus is present in the peDNA fraction indicates that *Haliea sp.* transports its DNA into the extracellular space differently. In the absence of a GTA cluster, we hypothesize that *Haliea* sp. is capable of EV production and EV-mediated gene transfer. This demonstrates that transduction and EV-mediated gene transfer are not exclusive mechanisms of HGT but can overlap.

Identification of known and novel EV producers reveals that EV production is common amongst abundant marine bacteria

We identified 129 MAGs as EV producers. Most identified genera are known EV producers: *Marinobacter* ($n = 19$), *Alcanivorax* [18], *Flavobacteria* [9], *Thalassospira* [8], *Rheinheimera* [6], and *Polaribacter* [3], are known to produce high amounts of EVs [38, 52]. The fact that most organisms we labeled as EV producers are

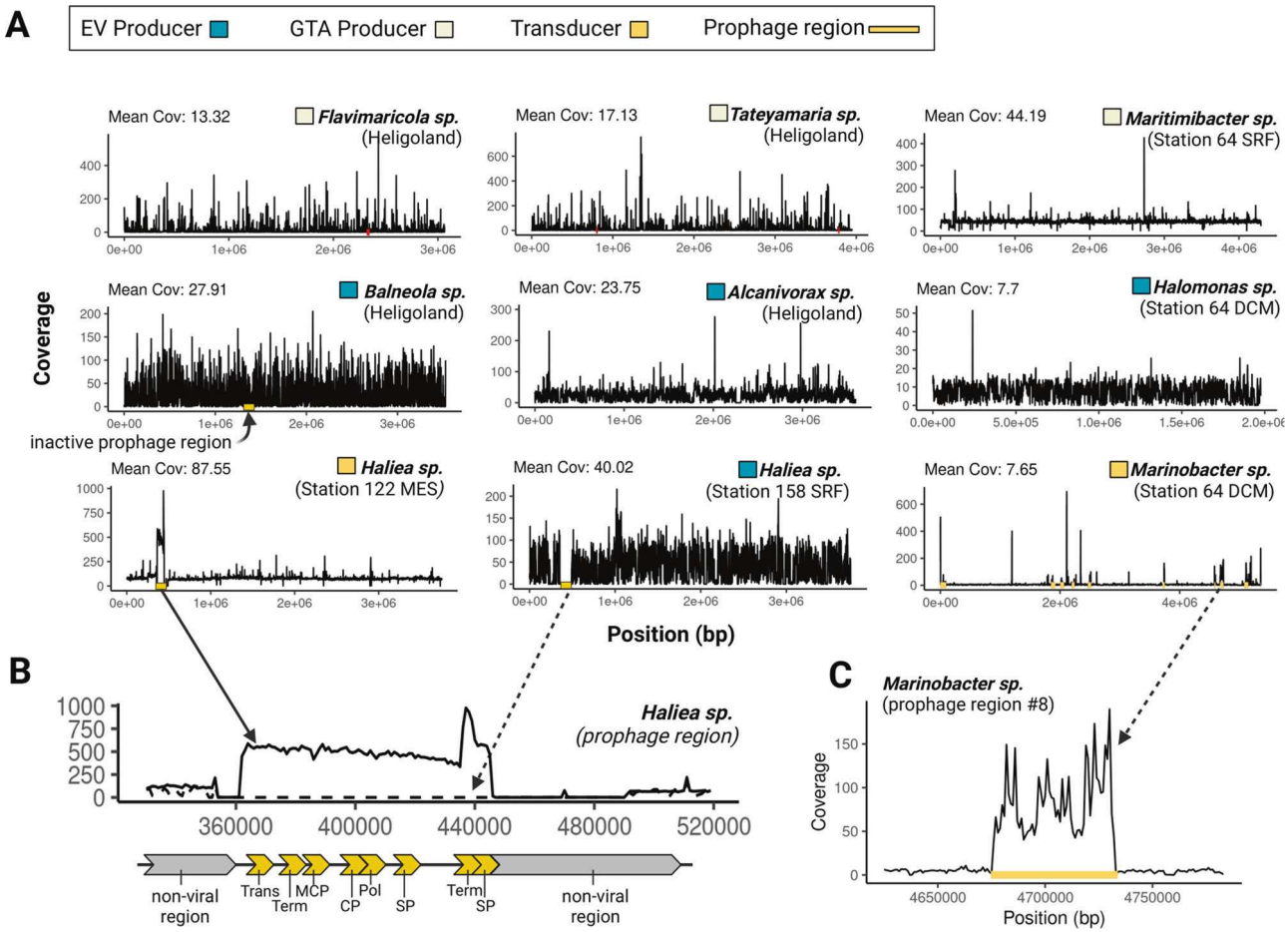


Fig. 6 Coverage plots of identified GTA- and EV-Producers and MAGs containing an actively transducing phage (transducer). A Coverage plots of 4 EV-producer, 2 genomes with an actively transducing phage and 3 GTA-producer (coverage blots of all analyzed MAGs see Fig. S12). Active prophage regions are indicated with yellow bars. **B** Genome map and detailed coverage plot of identified prophage region. On top, detailed coverage plot of the prophage region in two different samples. Coverage of reads from 122_MES (solid line) and Station 158_SRF (dotted) differs for this specific region. Below, schematic genome map for the genes identified in the prophage region and their approximate positions. Trans transposase, Term terminase, MCP major capsid protein, CP coat protein, Pol polymerase, SP shaft protein. **C** Coverage plot of an actively transducing phage. Close up of the coverage of prophage region #8 of *Marinobacter sp.* and surrounding non-viral regions.

already known EV producers again supports the efficiency of the approach. Interestingly, multiple MAGs of the genera *Haliea* ($n = 16$) and *Idiomarina* ($n = 8$) were identified to be EV producers. So far, EV production by either genus has not been described elsewhere and experimental confirmation is needed. However, this supports the observation by Biller et al., that many marine heterotrophs are actively producing EVs. While EVs could be used as a nutrient source [33, 38], or facilitate horizontal gene transfer (HGT) within microbial communities [37] potentially contributing to the evolution and adaptation of marine microbial populations, future research should aim to clarify the ecological and evolutionary role of EVs in the ocean.

The functional profile of peDNA links EV production to transposon induced gene mobilization

The functional profile of "viromes" or peDNA has been assessed previously [20, 53]. However, since these studies analyzed this sequence space from a virus perspective, mainly focusing on auxiliary metabolic genes (AMGs), they often excluded genes not directly associated with viral genomes. Here, we analyzed the functional profile of peDNA for each mode of transportation, EV- and GTA-mediated gene transfer, and transduction. In brief, each peDNA read that mapped to either an EV-, GTA producer or a transducer was classified into a cluster of orthologous groups (COG), and the resulting profile was then normalized with the

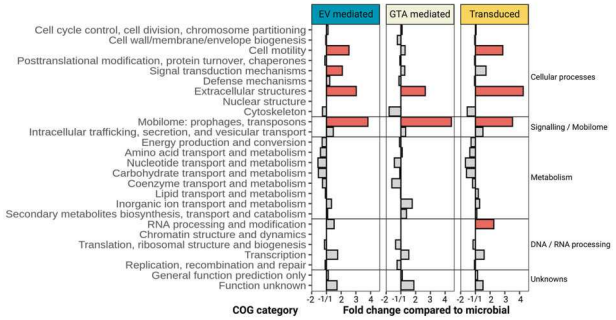


Fig. 7 Frequency of overrepresentation of clusters of orthologous groups (COG) categories for peDNA assigned to EV- and GTA producer and transducer. Bars represent the frequency of over-representation of genes assigned to each category, for each type of MAG (GTA producer, EV producer, transducer). Red bars indicate that a higher percentage of genes belonging to this category showed increased (two times standard deviation above mean) recruitment rates of peDNA reads.

profile of microbial reads (corresponding metagenome) mapping the respective sample (Fig. 7). The COG category "Mobilome (Prophages, transposons)" was overrepresented in all three groups. For transducers, this overrepresentation is mainly due to

an actively transducing provirus present in the peDNA. However, the overrepresentation of the COG category "Mobilome" in GTA and EV derived peDNA, is significant and we hypothesize, this is due to the presence of transposons. In fact, 76% of all reads assigned to the category "Mobilome" were assigned to a COG cluster containing the term "transposase" (Fig. S8). We suggest that transposon activity is also the reason for the overrepresentation of the other COG categories: "Extracellular structures", "Signal transduction mechanisms", and "Cell motility". All three categories are associated with the adaptation of the organism to a changing environment. Genes of these categories are often found on "genomic islands" (GIs), highly variable and mobile regions on the genome [54, 55]. At the same time, the occurrence of transposons on GIs is well-documented. Transposons have been shown to mobilize not only themselves but also adjacent "passenger genes", genes that are located in proximity to transposons and are therefore co-mobilized by transposons [56]. Evidence shows that environmental stressors increase the activity of transposons [57]. as well as the production of EVs [58], and the induction of proviruses [59] and GTAs [27]. Increased transposon activity increases the intracellular mobilization of genes surrounding transposons and therefore could lead to an increased uptake into EVs, GTAs or virus particles. Our data suggest that these two stress-induced mobilization mechanisms may be linked in a way that enhances the community's adaptability to the environment, by increasing genetic transfer between individual cells. However, whether the transposons and associated genes are indeed transferred in their complete active form or as fragments will require experimental evidence.

CONCLUSION AND OUTLOOK

In this study, we propose the term "protected extracellular DNA" (peDNA) to refer to genetic material obtained from size-filtered and DNase-treated samples, thereby accommodating non-viral, EV- or GTA-transported DNA in that sequence space. So far, there is no known sequence marker to distinguish horizontally transferred DNA from cellular DNA, therefore the removal of contaminating cells and free DNA is crucial when analyzing peDNA samples. The level of contamination however, should not be assessed using the presence of ribosomal subunit-mapping reads, since EVs have been shown to transport 16/18S rRNA genes.

In our study, we analyzed a carefully purified marine sample of peDNA and existing global marine datasets. We were able to link peDNA sequences to potential hosts and identify their primary mode of DNA transfer. Among the identified GTA and EV producers, most were shown to produce the respective particles in previous studies, confirming the validity of our approach, however, new potential GTA and EV producers were also identified. Overall, EV-mediated gene transfer was the most common mechanism and we hypothesize that EVs are a main driver of HGT in the ocean. Lastly, our findings suggest that EV-mediated gene transfer and transposon induced gene mobilization potentially work together and enhance the ability of microbial communities to adapt to a changing environment. Given the considerable ecological stressors imposed by climate change, comprehensive investigations into the role of EVs, GTAs and viruses for HGT is essential to understand genetic adaptability in marine microbes. This study highlights the need for further research into HGT mechanisms, and peDNA in general, since the community composition and function of marine microbes, and therefore the global oceans, is strongly shaped by the abundance of protected viral and non-viral extracellular DNA.

METHODS

Sampling and filtration

A visual overview of the sampling and filtration methods is given in Fig. S6.

Three seawater samples (G, H, I) of 100 liters were collected off the shore of Helgoland at the sampling station "Kabeltonne" (54°11'02.4'N 7°53'49.2'E). Each sample was sequentially filtered through 10, 3, 0.8, 0.45 and 0.22 µm (polyethersulfone filters, Merck Millipore, Burlington, MA, US). Filters were immediately stored at −20 °C for later DNA extraction. Flow through of the 0.22 µm filters was subsequently concentrated using tangential flow filtration with a 100 kDa cassette (Sartorius Stedim). The concentrated samples were stored at 4 °C, until further concentration down to 0.5 ml, using Amicon filter centrifugation (1 MDa AmiCon tube filters, 2500 × *g*). Finally, the concentrated sample was diluted with purified seawater (flow-through from the tangential flow filtration) to 2 ml. Two aliquots were created for each sample á 0.5 ml, one treated with DNase before gradient purification, one treated afterwards (see "Purification of peDNA samples").

Purification of peDNA samples

In order to remove free DNA, half of the samples were incubated with 100 U/ml DNase I (Thermo Scientific), supplied with the buffer provided with the enzyme, at 37 °C for 10 min, while the other half were DNase-treated after density gradient purification. EDTA was added to a final concentration of 5 mM and the enzyme was deactivated at 75 °C for 10 min. CsCl density gradients were prepared as following: Five CsCl solutions were prepared with 25, 30, 35, 40 and 60% CsCl solved in artificial sea water (480 mM NaCl, 27 mM $MgCl_2$, 2.8 mM $MgSO_4$, 9 mM KCl, 6 mM $NaHCO_3$, 10 mM $CaCl_2$) [60]. For each gradient, 1 ml of each solution was carefully layered on top of each other and stored at 4 °C overnight in order to establish the gradient. 0.5 ml of sample was carefully placed on top of the samples, before ultracentrifugation (20 h, 38,000 × *g*, 4 °C). For each gradient, individual 0.5 ml fractions were carefully extracted and incubated with 40% PEG 6000 (final concentration 10%) overnight at 4 °C. Particles were precipitated by centrifugation (13,000 × *g*, 45 min, 4 °C) and particle pellets dissolved in 1 ml artificial sea water. The other half of the samples were DNase-treated at this time point. A detailed overview of the samples is given in Table S1 and a visual overview in Fig. S7.

DNA extraction

Frozen polycarbonate filters (3, 0.8, 0.45 and 0.22 µm) were placed in a 50 ml tube together with 13.5 ml of extraction buffer (100 mM Tris-HCl pH 8.0, 100 mM EDTA pH 8.0, 100 mM Na-Phosphate buffer pH 8.0, 1.5 M NaCl, 1% CTAB). peDNA samples were processed directly. DNA was extracted as described elsewhere [61]. In brief, samples were treated with 10 mg/ml Proteinase K and incubated at 37 °C for 30 min on a shaker. Then, 1/10 vol of 20% SDS was added, before incubating again at 65 °C for 2 h on a shaker. After centrifugation (53,000 × *g*, 10 min, RT), the samples were transferred into a new tube and 1 vol of chloroform/isoamylalcohol was added and samples were thoroughly mixed, before centrifuging at 4000 × *g* for 20 min at RT. The aqueous phase upper phase was collected and transferred into a new tube. This step was repeated until no protein/polysaccharide layer was visible. DNA was then precipitated by adding 0.6 vol isopropanol and incubation for 1 h at room temperature. DNA was pelleted at 53,000 × *g* for 10 min at RT washed with 1 ml cold (4 °C) 80% ethanol and resuspended in 60 µl 1x TE buffer overnight. DNA concentration was assessed using a spectrophotometer (DS-11 FX+ by DeNovix®, Wilmington, DE, US), see Table S1.

Sequencing

DNA samples were pooled according to Table S1, assuring enough DNA content per sample for successful sequencing. Library preparation (FS DNA Library, NEBNext® Ultra™, Ipswitch, MA, US) and sequencing (Illumina HiSeq2500 by Illumina, San Diego, CA, US, 2 × 250 bp for peDNA samples and Illumina HiSeq3000, 2 × 150 bp for Filter DNA) was performed at the Max Planck Genome Centre Cologne (MP-GC).

Read trimming and assembly

Paired-end reads from Heligoland EV enrichments were trimmed using Trimmomatic [Bolger 2014] in paired-end mode, with the parameters LEADING:8 TRAILING:8 SLIDINGWINDOW:5:24 MINLEN:50. Paired-end reads from EV Enrichments [33] and paired-end reads from GDOCB [45] were trimmed using bbduk.sh, part of the BBTools suite [62] with the following parameters: bbduk.sh qtrim = rl trimq = 20 maq = 20 minlen = 30 ordered t = 8 ref = adapters.fa, where adapters.fa were fasta files containing adapters identified to be present in the reads using FastQC [63]. Reads from Heligoland EV enrichments were assembled using metaSPAdes [64] with default parameters.

Handling of external data

External datasets were downloaded from public servers. An overview of external datasets used in this study, with SRR, ERS and DRR accessions, is given in Table S1. Reads from Tara Ocean viromes [17] were already trimmed. Reads from EV enrichments [33] and GDOCB viromes [45] were assembled using metaSPAdes [64] with default parameters. For Tara Oceans viromes, assembled contigs were downloaded from https://www.ebi.ac.uk/. Tara Oceans MAGs were published elsewhere by Tully et al. [65], accessions are listed in Table S1.

Calculation of SSU alignment rates

SSU alignment rates were calculated using ViromeQC [19], which maps input reads against 16S and 16S rRNA subunits. This was done for all Tara Ocean viromes, Heligoland peDNA, EV enrichments (see Table S1 for an overview of all samples used).

Calculation of the percentage of non-viral associated reads

The percentage of non-viral peDNA in viromic samples was calculated with a pipeline of bioinformatic tools. An overview is given in Fig. S1. Paired-end, trimmed input reads were assembled. Contigs shorter than 2000 bp were removed from downstream analysis. Then, contigs were subject to two viral-prediction steps: viral sequences were predicted (1) using a combination of VirSorter2 and CheckV as described previously [66] and (2) DeepVirFinder [16]. The results of both steps were summarized using a custom script (https://github.com/dluecking/peDNA_custom_scripts/) and each contig was labeled as either "viral" or "non-viral". Then, the initial input reads were mapped against labeled contigs, using bbmap.sh, part of the BBTools suite [62] with default parameters. Then the number of non-viral-contig mapping reads was divided by the number of total reads mapping against viral or non-viral contigs. This ratio of non-viral/viral reads is referred to as "percentage of non-viral to viral associated reads" or "nvpeDNA/peDNA read ratio" in this study.

Identification of potential transducers, extracellular vesicle- and gene transfer agent producers

In order to identify potential EV producers, GTA producers and MAGs with an actively transducing virus, a second bioinformatic pipeline was developed (see Fig. S5). First, MAGs were filtered by removal of MAGs shorter than 100,000 bp. Then, reads from the corresponding viromes/peDNA samples were mapped against the MAGs, using bbmap.sh with default parameters. For each sample (in total 9 samples from Tara Oceans, 1 combined Heligoland sample) the 20 most recruiting MAGs were selected for further downstream work. VirSorter2 (default parameters) was used in order to predict potential integrated proviruses. GTA clusters were predicted by searching for homologs of proteins of known GTA clusters using *diamond blastp* with default parameters (evalue $\leq 10^{-5}$, pident > 50%). Finally, a customscript (https://github.com/dluecking/peDNA_custom_scripts/) summarized the results and an automated label was given. Additionally, each label was manually curated and each MAG was labeled as either EV producer, GTA producer or an organism with an actively transducing virus (see Figs. S5 and S6).

Annotation of GTA producers and viral regions

Open reading frames were predicted using prodigal with the metagenome flag (*prodigal -i < fasta-file > -d < genes-out > -a <protein-out > -p meta*). Each ORF was then annotated using the InterProScan API (https://github.com/ebi-wp/webservice-clients-generator) with default parameters [67] and additionally checked manually. For the prophage region, shown in Fig. 6, the DNA sequence was submitted and annotated in PHASTER [68, 69]. Genome maps and presence-absence plots were generated using ggplot [70] and BioRender.com.

Identification of cluster of orthologous groups

In order to assess the functional profile of peDNA reads, each read was mapped to the 170 MAGs for which the primary transport mechanism was identified using bbmap, part of the BBTools suite [62] with minid = 95 and otherwise default parameters. This resulted in three sets of reads: EV-mediated, GTA-mediated, VLP-mediated. For each read partial ORFs were predicted using FragGeneScan [71] with the parameters -complete = 0, -train = illumina_5 and otherwise default parameters. The partial ORFs were then blasted against the COG database [72] using the diamond tool set [73] with the following parameters: *-f 6 –max-target-seqs 1 –query-cover 80 –subject-cover 10*. Each read was then assigned a COG cluster and consequently a COG category. For each category, the relative abundance was calculated using:

$$freq_{cat} = \frac{n_{hc}}{n_{tot}}$$

where *n_hc* is the number of reads in category *cat* that show high coverage and *n_tot* is the total number of reads assigned by this label. The same procedure was done for metagenome reads of the corresponding metagenome samples (see Supplementary Table S1—Sample overview). The fold change between categories was calculated pairwise with the formula:

$$fold\ change\frac{e_{cat-label} = freq_{cat-label}}{fre} q_{cat-microbial}$$

where *cat* refers to a specific COG category, label to either EV, GTA or virally transduced and microbial to the microbial counterpart of that sample. For visualization reasons, fold changes smaller 1 were calculated with the reversed formula:

$$fold\ change_{cat-label} = fre\frac{q_{cat-microbial}}{fre} q_{cat-label}$$

Fold changes between −1 and 1 are therefore not possible and this area is excluded from the plot.

In order to get a detailed resolution of EV-mediated reads belonging to COG category × -Mobilome a subset of 10 million reads for each sample (9 Tara Ocean stations and 1 sample from heligoland) were selected at random. From these, ~11 M protein fragments were predicted and blasted against nr with an *e*-value threshold of 10^{-5}, query coverage >80%, subject coverage >10%, resulting in a total of 34,826 assigned reads. The results were visualized in R.

Identification of transposable elements on EV producing genomes

Putative EV producing MAGs were annotated using DRAM (Distilled and Refined Annotation of Metabolism, https://github.com/WrightonLabCSU/DRAM), [74]. Transposases were then detected using the regex term "IS\\d\\+*|Tn\\d\\+*|attTn\\d\\+*|transposase|Transposase" among all annotations found.

Coverage plots, genome maps and schematic figures

Coverage plots of potential transducers, EV- and GTA producers were created using the R package ggplot2 [70]. Genome maps of potential GTA producers were created using the R package gggenes (https://github.com/wilkox/gggenes). Schematic genome maps and additional elements in figures were created with BioRender.com.

DATA AVAILABILITY

Heligoland metagenome and peDNA reads are available at the European Nucleotide Archive (ENA) under BioProject PRJEB60526. Custom scripts are available at https://github.com/dluecking/peDNA_custom_scripts/.

REFERENCES

1. Soucy SM, Huang J, Gogarten JP. Horizontal gene transfer: building the web of life. Nat Rev Genet. 2015;16:472–82.
2. Arnold BJ, Huang IT, Hanage WP. Horizontal gene transfer and adaptive evolution in bacteria. Nat Rev Microbiol. 2022;20:206–18.
3. Fuhrman JA. Marine viruses and their biogeochemical and ecological effects. Nature. 1999;399:541–8.
4. Wommack KE, Colwell RR. Virioplankton: viruses in aquatic ecosystems. Microbiol Mol Biol Rev. 2000;64:69–114.
5. Suttle CA. Viruses in the sea. Nature. 2005;437:356–61.
6. Breitbart M, Thompson LR, Suttle CA, Sullivan MB. Exploring the vast diversity of marine viruses. Oceanography. 2007;20:135–9.
7. Edwards RA, Rohwer F. Viral metagenomics. Nat Rev Microbiol. 2005;3:504–10.
8. Breitbart M, Salamon P, Andresen B, Mahaffy JM, Segall AM, Mead D, et al. Genomic analysis of uncultured marine viral communities. Proc Natl Acad Sci. 2002;99:14250–5.

9. Li WKW, Dickie PM. Growth of bacteria in seawater filtered through 0.2 µm Nuclepore membranes: implications for dilution experiments. Mar Ecol Prog Ser. 1985;26:245–52.
10. Van der Pol E, Hoekstra AG, Sturk A, Otto C, Van leeuwen TG, Nieuwland R. Optical and non-optical methods for detection and characterization of microparticles and exosomes. J Thromb Haemost. 2010;8:2596–607.
11. McNamara RP, Dittmer DP. Modern techniques for the isolation of extracellular vesicles and viruses. J Neuroimmune Pharmacol. 2020;15:459–72.
12. Hillebrandt N, Vormittag P, Bluthardt N, Dietrich A, Hubbuch J. Integrated process for capture and purification of virus-like particles: enhancing process performance by cross-flow filtration. Front Bioeng Biotechnol. 2020;8. https://doi.org/10.3389/fbioe.2020.00489.
13. Duhaime MB, Sullivan MB. Ocean viruses: rigorously evaluating the metagenomic sample-to-sequence pipeline. Virology. 2012;434:181–6.
14. Roux S, Hallam SJ, Woyke T, Sullivan MB. Viral dark matter and virus-host interactions resolved from publicly available microbial genomes. eLife. 2015;4:e08490.
15. 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.
16. Ren J, Song K, Deng C, Ahlgren NA, Fuhrman JA, Li Y, et al. Identifying viruses from metagenomic data using deep learning. Quant Biol. 2020;8:64–77.
17. Gregory AC, Zayed AA, Conceição-Neto N, Temperton B, Bolduc B, Alberti A, et al. Marine DNA viral macro- and microdiversity from pole to pole. Cell. 2019;177:1109–23.e14.
18. Roux S, Enault F, Robin A, Ravet V, Personnic S, Theil S, et al. Assessing the diversity and specificity of two freshwater viral communities through metagenomics. PLoS ONE. 2012;7:e33641.
19. Zolfo M, Pinto F, Asnicar F, Manghi P, Tett A, Bushman FD, et al. Detecting contamination in viromes using ViromeQC. Nat Biotechnol. 2019;37:1408–12.
20. Roux S, Krupovic M, Debroas D, Forterre P, Enault F. Assessment of viral community functional potential from viral metagenomes may be hampered by contamination with cellular sequences. Open Biol. 2013;3:130160.
21. Jiang S, Paul J, et al. Gene transfer by transduction in the marine environment. App Environ Microbiol. 1998;64:2780–7.
22. Zinder ND, Lederberg J. Genetic exchange in salmonella. J Bacteriol. 1952;64:679–99.
23. Lindell D, Sullivan MB, Johnson ZI, Tolonen AC, Rohwer F, Chisholm SW. Transfer of photosynthesis genes to and from Prochlorococcus viruses. Proc Natl Acad Sci USA. 2004;101:11013–8.
24. Matson EG, Thompson MG, Humphrey SB, Zuerner RL, Stanton TB. Identification of genes of VSH-1, a prophage-like gene transfer agent of Brachyspira hyodysenteriae. J Bacteriol. 2005;187:5885–92.
25. Krupovic M, Forterre P, Bamford DH. Comparative analysis of the mosaic genomes of tailed archaeal viruses and proviruses suggests common themes for virion architecture and assembly with tailed viruses of bacteria. J Mol Biol. 2010;397:144–60.
26. Lang AS, Zhaxybayeva O, Beatty JT. Gene transfer agents: phage-like elements of genetic exchange. Nat Rev Microbiol. 2012;10:472–82.
27. Québatte M, Christen M, Harms A, Körner J, Christen B, Dehio C. Gene transfer agent promotes evolvability within the fittest subpopulation of a bacterial pathogen. Cell Syst. 2017;4:611–21.e6.
28. Tomasch J, Wang H, Hall ATK, Patzelt D, Preusse M, Petersen J, et al. Packaging of Dinoroseobacter shibae DNA into gene transfer agent particles is not random. Genome Biol Evol. 2018;10:359–69.
29. Deatherage BL, Cookson BT. Membrane vesicle release in bacteria, eukaryotes, and archaea: a conserved yet underappreciated aspect of microbial life. Infect Immun. 2012;80:1948–57.
30. Dorward DW, Garon CF, Judd RC. Export and intercellular transfer of DNA via membrane blebs of Neisseria gonorrhoeae. J Bacteriol. 1989;171:2499–505.
31. Kadurugamuwa JL, Beveridge TJ. Bacteriolytic effect of membrane vesicles from Pseudomonas aeruginosa on other bacteria including pathogens: conceptually new antibiotics. J Bacteriol. 1996;178:2767–74.
32. Yaron S, Kolling GL, Simon L, Matthews KR. Vesicle-mediated transfer of virulence genes from Escherichia coli O157:H7 to other enteric bacteria. Appl Environ Microbiol. 2000;66:4414–20.
33. Biller SJ, Schubotz F, Roggensack SE, Thompson AW, Summons RE, Chisholm SW. Bacterial vesicles in marine ecosystems. Science. 2014;343:183–6.
34. Biller SJ, Berube PM, Lindell D, Chisholm SW. Prochlorococcus: the structure and function of collective diversity. Nat Rev Microbiol. 2015;13:13–27.
35. Schwechheimer C, Kuehn MJ. Outer-membrane vesicles from Gram-negative bacteria: biogenesis and functions. Nat Rev Microbiol. 2015;13:605–19.
36. Schatz D, Rosenwasser S, Malitsky S, Wolf SG, Feldmesser E, Vardi A. Communication via extracellular vesicles enhances viral infection of a cosmopolitan alga. Nat Microbiol. 2017;2:1485–92.
37. Domingues S, Nielsen KM. Membrane vesicles and horizontal gene transfer in prokaryotes. Curr Opin Microbiol. 2017;38:16–21.
38. Biller SJ, Coe A, Arellano AA, Dooley K, Gong JS, Yeager EA, et al. Environmental and taxonomic drivers of bacterial extracellular vesicle production in marine ecosystems. bioRxiv. 2022. https://doi.org/10.1101/2022.01.18.476865v2.
39. Biller SJ, McDaniel LD, Breitbart M, Rogers E, Paul JH, Chisholm SW. Membrane vesicles in sea water: heterogeneous DNA content and implications for viral abundance estimates. ISME J. 2017;11:394–404.
40. Hackl T, Laurenceau R, Ankenbrand MJ, Bliem C, Cariani Z, Thomas E, et al. Novel integrative elements and genomic plasticity in ocean ecosystems. Cell. 2023;186:47–62.e16.
41. Linney MD, Eppley JM, Romano AE, Luo E, DeLong EF, Karl DM. Microbial sources of exocellular DNA in the ocean. Appl Environ Microbiol. 2022;88:e02093–21.
42. Bitto NJ, Chapman R, Pidot S, Costin A, Lo C, Choi J, et al. Bacterial membrane vesicles transport their DNA cargo into host cells. Sci Rep. 2017;7:7072.
43. Ricci V, Carcione D, Messina S, Colombo GI, D'Alessandra Y. Circulating 16S RNA in biofluids: extracellular vesicles as mirrors of human microbiome? Int J Mol Sci. 2020;21:8959.
44. Bartlau N, Wichels A, Krohne G, Adriaenssens EM, Heins A, Fuchs BM, et al. Highly diverse flavobacterial phages isolated from North Sea spring blooms. ISME J. 2022;16:555–68.
45. De Corte D, Martínez JM, Cretoiu MS, Takaki Y, Nunoura T, Sintes E, et al. Viral communities in the global deep ocean conveyor belt assessed by targeted viromics. Front Microbiol. 2019;10:1801.
46. Linney MD, Schvarcz CR, Steward GF, DeLong EF, Karl DM. A method for characterizing dissolved DNA and its application to the North Pacific Subtropical Gyre. Limnol Oceanogr Methods. 2021;19:210–21.
47. Orellana LH, Ben Francis T, Krüger K, Teeling H, Müller MC, Fuchs BM, et al. Niche differentiation among annually recurrent coastal Marine Group II Euryarchaeota. ISME J. 2019;13:3024–36.
48. Zhao Y, Wang K, Budinoff C, Buchan A, Lang A, Jiao N, et al. Gene transfer agent (GTA) genes reveal diverse and dynamic Roseobacter and Rhodobacter populations in the Chesapeake Bay. ISME J. 2009;3:364–73.
49. Ankrah NYD, Lane T, Budinoff CR, Hadden MK, Buchan A. Draft genome sequence of Sulfitobacter sp. CB2047, a member of the roseobacter clade of marine bacteria, isolated from an emiliania huxleyi bloom. Genome Announc. 2014;2:e01125–14.
50. McDaniel LD, Young E, Delaney J, Ruhnau F, Ritchie KB, Paul JH. High frequency of horizontal gene transfer in the oceans. Science. 2010;330:50.
51. Sherlock D, Leong JX, Fogg PCM. Identification of the first gene transfer agent (GTA) small terminase in rhodobacter capsulatus and its role in GTA production and packaging of DNA. J Virol. 2019;93:e01328–19.
52. Schuster AK. Production of extracellular DNA (eDNA) of the γ-proteobacterium Rheinheimera sp. F8 in biofilms. Germany: Technische Universitaet Berlin; 2017.
53. Dinsdale EA, Edwards RA, Hall D, Angly F, Breitbart M, Brulc JM, et al. Functional metagenomic profiling of nine biomes. Nature. 2008;452:629–32.
54. Coleman ML, Sullivan MB, Martiny AC, Steglich C, Barry K, DeLong EF, et al. Genomic islands and the ecology and evolution of Prochlorococcus. Science. 2006;311:1768–70.
55. Thompson AW, Huang K, Saito MA, Chisholm SW. Transcriptome response of high- and low-light-adapted Prochlorococcus strains to changing iron availability. ISME J. 2011;5:1580–94.
56. Pál C, Papp B. From passengers to drivers. Mob Genet Elem. 2013;3:e23617.
57. Capy P, Gasperi G, Biémont C, Bazin C. Stress and transposable elements: coevolution or useful parasites? Heredity. 2000;85:101–6.
58. MacDonald IA, Kuehn MJ. Stress-induced outer membrane vesicle production by Pseudomonas aeruginosa. J Bacteriol. 2013;195:2971–81.
59. Choi J, Kotay SM, Goel R. Various physico-chemical stress factors cause prophage induction in Nitrosospira multiformis 25196–an ammonia oxidizing bacteria. Water Res. 2010;44:4550–8.
60. Carini P, Steindler L, Beszteri S, Giovannoni SJ. Nutrient requirements for growth of the extreme oligotroph 'Candidatus Pelagibacter ubique' HTCC1062 on a defined medium. ISME J. 2013;7:592–602.
61. Zhou J, Bruns MA, Tiedje JM. DNA recovery from soils of diverse composition. Appl Environ Microbiol. 1996;62:316–22.
62. Bushnell B. BBMap: a fast, accurate, splice-aware aligner. Report No.: LBNL-7065E. Berkeley, CA (United States): Lawrence Berkeley National Lab. (LBNL); 2014. https://www.osti.gov/biblio/1241166.
63. Andrews S, et al. FastQC: a quality control tool for high throughput sequence data. Cambridge, United Kingdom: Babraham Bioinformatics, Babraham Institute; 2010.
64. Nurk S, Meleshko D, Korobeynikov A, Pevzner PA. metaSPAdes: a new versatile metagenomic assembler. Genome Res. 2017;27:824–34.
65. Tully BJ, Graham ED, Heidelberg JF. The reconstruction of 2631 draft metagenome-assembled genomes from the global oceans. Sci Data. 2018;5:170203.

66. Guo J, Vik D, Pratama AA, Roux S, Sullivan M. Viral sequence identification SOP with VirSorter2. 2021. https://www.protocols.io/view/viral-sequence-identification-sop-with-virsorter2-bwm5pc86.
67. Jones P, Binns D, Chang HY, Fraser M, Li W, McAnulla C, et al. InterProScan 5: genome-scale protein function classification. Bioinformatics. 2014;30:1236–40.
68. 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.
69. Zhou Y, Liang Y, Lynch KH, Dennis JJ, Wishart DS. PHAST: a fast phage search tool. Nucleic Acids Res. 2011;39:W347–52.
70. Wickham H. ggplot2: elegant graphics for data analysis. New York: Springer-Verlag; 2016. https://ggplot2.tidyverse.org.
71. Rho M, Tang H, Ye Y. FragGeneScan: predicting genes in short and error-prone reads. Nucleic Acids Res. 2010;38:e191.
72. Galperin MY, Wolf YI, Makarova KS, Vera Alvarez R, Landsman D, Koonin EV. COG database update: focus on microbial diversity, model organisms, and widespread pathogens. Nucleic Acids Res. 2021;49:D274–81.
73. Buchfink B, Reuter K, Drost HG. Sensitive protein alignments at tree-of-life scale using DIAMOND. Nat Methods. 2021;18:366–8.
74. 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.

ACKNOWLEDGEMENTS

We thank the sampling crew of the Heligoland sampling station "Kabeltonne" for their help obtaining the sea water samples and the AWI research station on heligoland for providing laboratory workspace. We are grateful to Luis "Coto" Orellana and Isabella Maria Wilkie for providing access to their data. Finally, we thank A. Probst, A. Zayed and B. Fuchs for insights, discussions and feedback. We thank Daniela Thies and Ingrid Kunze (MPI for Marine Microbiology, Bremen, Germany) for assistance with some of the experiments. Finally, we want to thank the Max-Planck-Institute for Marine Microbiology and the Max-Planck-Society for continuous support.

AUTHORS CONTRIBUTIONS

DL performed the majority of the experimental laboratory and bioinformatic work. CM and TAS supported general laboratory work, sampling and purification of sea water samples. SE conceived and led the study. DL and SE performed the primary writing of the manuscript. All authors participated in the analysis and interpretation of the data and contributed to the writing of the manuscript.

FUNDING

This study was funded by the Max Planck Society (Munich, Germany) as part of the Max Planck Research Group Archaeal Virology to SE. Open Access funding enabled and organized by Projekt DEAL.

COMPETING INTERESTS

The authors declare no competing interests.

ADDITIONAL INFORMATION

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s43705-023-00317-6.

Correspondence and requests for materials should be addressed to Susanne Erdmann.

Reprints and permission information is available at http://www.nature.com/reprints

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

Chapter 5

Discussion and Outlook

This thesis contributes to ongoing efforts to comprehend the mechanisms that shape the intricate web of life. In Chapter 2, we successfully identified 40 novel pR1SE-like plasmids, subsequently termed archaeal plasmids of *Haloarchaea* potentially transferred in plasmid vesicles (apHPVs). We analyzed their genetic organization and found hints that reinforce and expand the model of their life cycle. Lastly, through the use of gene-sharing networks, we established that apHPVs connect archaeal viruses and plasmids and are, therefore, important drivers of genomic flexibility in *Haloarchaea*. In Chapter 3, I summarized the current state of the bioinformatic virus and plasmid identification and the challenges associated with identifying EV-mediated HGT. These challenges were then successfully tackled in Chapter 4, where we developed a bioinformatic method to identify EV-mediated HGT in carefully curated metagenomic datasets. In this pioneering work, we established the term protected extracellular DNA (peDNA) to describe the sequence space traditionally termed 'virome'. This shift in nomenclature was supported by the finding that the sequence space of peDNA contains strong signals of HGT outside of virus-induced transduction. In the environments analyzed, DNA transfer in extracellular vesicles appeared as a major driver of HGT.

5.1 Evolutionary and ecological implications of apHPVs

The first definition-breaking entity in this thesis is a type of plasmid that showed a hybrid set of plasmid- and virus-like characteristics. The plasmid pR1SE of the halophilic archaeon *Halorubrum lacusprofundi* was first discovered by Erdmann *et al.* (Erdmann et al. 2017). In Chapter 2, we searched public databases in addition to metagenomes of Australian salt lakes for pR1SE-like elements, named apHPVs. Erdmann *et al.* proposed a pR1SE life cycle of repeated insertion and excision events (Erdmann et al. 2017). In Chapter 2 we suggest that the discovered apHPVs follow a similar lifestyle and are capable of transferring significant (up to ∼250 kbp) amounts of host DNA. We describe the genomic organization of apHPVs into a core region and an extended, hyperflexible region of assimilated host material. The core region is organized into two conserved gene clusters (1 and 2). Two key questions follow these observations. First, are the non-core, hyperflexible regions of apHPVs recruited in a single integration and excision event, or are they the result of multiple sequential replication cycles? Second, do all apHPVs that are present at the same time

in a given environment look alike, or are they in fact, in different 'stages' of their cycles? It seems very unlikely that all different apHPVs within a population have picked up the same genetic host material at a given point in time. Therefore, we have to assume that the recovered apHPV just represents the most common configuration of all apHPVs. Unsurprisingly, we detected cases where assembly tools failed to connect gene cluster 1 and gene cluster 2 based on short read data, potentially a direct result of a broad spectrum of different apHPV configurations co-occurring. At the same time, the coverage of successfully assembled apHPVs showed no irregular or strikingly uneven patterns, though we could only verify this for a few examples (unpublished data). While we could speculate further and conduct additional analysis on the existing short-read data, using long-read data would significantly improve our ability to answer the questions at hand. In fact, for some datasets from which apHPVs were recovered, long-read PacBio sequences are publicly available. For example, when sequencing the samples of Australian salt lakes, additional long-read data was generated for Lake Gairdner and Lake Tyrell (unpublished). The analysis of the data at hand seems self-evident and could guide future resequencing efforts. However, even PacBio reads (average read length of 4.8 Mbp and 5.6 Mbp for Lake Eyre and Lake Tyrell) might prove to be too short to provide a comprehensive overview of a population of apHPVs. Therefore, using ultra-long Oxford Nanopore©sequencing techniques with read lengths exceeding the average length of apHPVs might be required. This would circumvent most assembly-related problems and enable a 'one-read-one-apHPV' representation of the apHPV population at hand. Another consequence of the proposed apHPV life cycle is the shedding of large fragments of host DNA once the length of the entire plasmid exceeds a size threshold. It is unclear which specific mechanism limits the packageable size, however, the discovered non-integrated apHPVs did not exceed ∼270 Mbp, Transmission electron micrographs suggest a relatively narrow size range of PVs (∼100 nm), especially compared to EVs (Erdmann et al. 2017; Deatherage and Cookson 2012). This could limit the size of the packageable DNA molecule, similar to viral capsids (Bárdy et al. 2020). Irrespective of the limiting factor, the frequent shedding of large fragments of DNA might be a potential explanation for the increased number of plasmids in *Haloarchaea* in general. However, as discussed in 2, the direction of the hypothesized causality is unclear, as it seems conversely plausible that MGEs like apHPVs would preferably emerge in an genomic context which provides multiple insertion, excision and recombination opportunities. The genomic organization of apHPVs into two disjunct gene clusters is discussed in 2. We propose that gene cluster 1 might be primarily involved in plasmid vesicle (PV) formation, while gene cluster 2 could be responsible for packaging. Strikingly, we found many occurrences of contigs that contained only cluster 1 and only a single instance of a contig with cluster 2, for which we could rule out incomplete assembly. The reason for this apparent disbalance is unclear. Potentially, gene cluster 1 can act as an independent and stable MGE. It seems possible that gene cluster 2 only improves the efficiency of packaging and, therefore, dissemination of the plasmid. However, without strong experimental data, this remains speculation. A potential experiment could be the infection of an apHPV-free strain with a synthesized or otherwise obtained version of each gene cluster, which then could prove the stability of the

single clusters.

5.2 The platinum age of bioinformatic MGE detection

In Chapter 3, current tools for the detection of mobile genetic elements and forms of horizontal gene transfer are reviewed. Additionally, the challenges of detecting EV-mediated gene transfer are elaborated. Despite all the advances in the field, the detection of MGEs, for which no homologs in annotated databases are currently available, i.e. completely unknown types of MGEs[1], will continue to pose a challenge. While pattern-based detection approaches *per se* are reference-independent, they must be trained on known sequences. Therefore, detecting MGEs for which no relatives are described and whose sequence characteristics do not match currently known patterns still seems impossible. Innovative use of host information and experimental setups, e.g. as described in two recent studies, allows the bioinformatic discovery of entirely novel MGEs (Durrant et al. 2020; Hackl et al. 2023). Even though these workflows are not traditional 'detection tools', their potential to detect the first representatives of a novel class of MGEs is immense. Paired with ever-growing sequence databases, improved generalized and specialized detection tools and growing computational power, the future of bioinformatic MGE and HGT detection seems bright. Despite the power of bioinformatic analyses, bioinformatics alone is not enough to get the complete picture. The continuous accumulation of in vivo results from rigorous and meticulous experimentation is absolutely fundamental to supporting bioinformatic findings. Nevertheless, as A. Babaian (University of Toronto) said at the Gairdner Science Week International Symposium 2023,

» *We might just be entering the 'platinum age' of bioinformatic virus discovery*[2] «,

and I am highly inclined to agree and expand that notion to other types of MGE and HGT events in general.

5.3 Insights on horizontal gene transfer mechanisms in the ocean

In contrast to the described tools for the detection of viruses, plasmids and other MGEs, the detection of EV-mediated HGT demands conceptually different approaches, as described in Chapter 4. We showed that by linking transported sequences to their potential host and predicting the main mechanism of HGT for this host (transduction, EV-mediated and GTA-mediated HGT), the impossibility of EV-mediated HGT detection could be circumvented. We applied the developed pipeline to large marine datasets and demonstrated the

[1] Often, sequences for which no homologs are currently available, but still are just out there, are referred to as 'microbial/viral dark matter' (Krishnamurthy and Wang 2017; Santiago-Rodriguez and Hollister 2022). However, I agree with Prof. Dr. Erens explanation why this is, while certainly catchy, a very poor term. In physics, dark matter refers to a categorically different type of matter, matter that's unlike anything known and not just undiscovered, but for all we know very common sequences (https://merenlab.org/2017/06/22/microbial-dark-matter/).

[2] The Future of AI in Science and Medicine - Gairdner Science Week International Symposium 2023 - https://www.youtube.com/watch?v=mPUM-adhXd8

prevalence of DNA exchange via extracellular vesicles in the global oceans. While the developed approach certainly has its biases and pitfalls, it pioneered a bioinformatic perspective to EV-mediated HGT and introduced a new nomenclature, by questioning the established term of 'virome' and proposed protected extracellular DNA (peDNA) instead. One significant limitation of the approach remains the dependence on contamination-free peDNA datasets. Undegraded free DNA and DNA from lysed host cells introduce a significant bias into the amount of peDNA in the environment. Furthermore, the pipeline analyzes the top peDNA-recruiting hosts, or in other words, it selects hosts for further analysis, that seemingly contribute a lot of DNA to the protected extracellular space. Therefore, the predictive power of the approach for assessing the composition of HGT mechanisms within a given microbiome diminishes with the degree of contamination. Conceptually, the presented approach uses a process of elimination when assigning HGT mechanisms to hosts. This means that any underestimation of transduction and GTA-mediated HGT will lead to an overestimation of EV-mediated HGT. Similarly, any unknown but active HGT mechanism will skew the results towards EV-mediated HGT. However, until a more complete set of EV formation associated genes has been established, which would allow for a direct approach, the presented results represent the best estimation of the composition of HGT-mechanisms. A more tangible, short-term improvement to the pipeline would be automating the manual curation steps, e.g. the automatic detection of transduction patterns would greatly improve scalability and applicability. Future studies could then analyze larger datasets and datasets from understudied environments, e.g. freshwater systems, in regards to EV-mediated HGT. This would reduce the bias introduced by selecting only top peDNA-producing hosts since this would allow the analysis of all peDNA-producing hosts. Additionally, by comparing the HGT-mechanism composition across biomes or even gradients within the same biomes, we could validate the findings by cross-referencing the observed trends with known drivers of e.g. community composition (Gregory et al. 2019). For example, EV-mediated HGT could become relatively more abundant along a stress gradient, as Biller *et al.* demonstrated that EV formation correlates with stressors (Biller et al. 2022). The results presented in this work, based on the presented approach, already indicate the importance of EV-mediated HGT, however, the discussed improvements would significantly enhance the predictive power of the approach.

5.4 Blurring the lines of entities and revisiting nomenclature

The main findings of this thesis demonstrated the need for the continuous evolution of existing nomenclature and classification schemes. Identifying extracellular vesicles as abundant and an impactful driver of HGT required the introduction of a novel term in order to fully describe the sequence space of protected extracellular DNA (peNA, Chapter 4). Discovering pR1SE-like elements widespread in haloarchaea resulted in a novel class of MGEs with a hybrid set of viral- and plasmid-like characteristics. It remains unclear where apHPVs should be placed in a binary either-or classification scheme (virus or plasmid) and reinforces the need for spectrum-like classification schemes, as visualized in Figure 1.3 (Koonin et al. 2021). Indeed, one could easily imagine tracing the evolutionary trajectory

of an MGE through this replicator space. This hypothetical MGE could recruit a viral capsid protein, thus entering the perivirosphere. Subsequently, the same MGE could be recombining with parts of a plasmid, again moving through the replicator space into e.g. the definition set of phage-plasmids (Pfeifer et al. 2021). In fact, while the literal tracing of a specific MGE remains hypothetical, as the replicator space remains an abstract conceptual space for now, scientists unraveling the evolutionary history of an MGE trace its trajectory through this abstract space. MGEs are particularly susceptible to drastic changes for at least two tangible reasons. First, due to their limited number of genes, HGT events provide a relatively more significant impact. Second, the complexity of MGE genetics, i.e., the number of interactions of each protein, is significantly lower compared to cellular organisms, and complexity has been shown to negatively impact the likelihood of successful HGT events (Burch et al. 2023), making MGEs the perfect target for HGT events. It is, therefore, unsurprising that the functional potential of MGEs is leveraged by host organisms and vice versa. The existence of GTAs and apHPVs demonstrates this bidirectional exchange. GTAs putatively resemble defective or domesticated viruses, whose gene transfer capabilities were recruited by the host in order to provide genomic flexibility (Lang et al. 2012). Fittingly, the evolution of different GTA clusters putatively was not based on vertical transmission but appears to result from horizontal gene transfer events. Conversely, it is hypothesized that apHPVs have recruited the hosts' capabilities to produce protected vesicles to increase the efficiency of their own transmission (Erdmann et al. 2017; Lücking et al. 2023). However, in both cases, the subject remains a matter of perspective: It is similarly reasonable to rephrase the emergence of apHPVs as 'haloarchaea improved the efficiency of EV-mediated HGT by recruiting pR1SE-like plasmids, in order to increase their genomic flexibility'. In doing so, apHPVs moved within the replicator space, while GTAs can no longer be considered as independent replicators.

5.5 Blurring the boundaries of life

Equivalent to the evolutionary trajectories of MGEs, the evolution of organisms can be understood as a movement on a fitness landscape, an abstract space of high and low reproductive success (Lobkovsky et al. 2011). Similarly to the placement of an MGE within the replicator space, the placement of a given taxonomic unit is not defined by discrete positions within that landscape but instead can be seen as a continuum of possible states. Particularly, species with a high genomic flexibility and an expanded pangenome, e.g. species of *Haloarchaea*, are challenging to place at a single distinct position in this abstract space (Papke et al. 2015; Tschitschko et al. 2018). In fact, organisms (and organelles) can be placed into the same replicator space, even though they are positioned distinctly outside the spaces discussed so far (Koonin et al. 2021). The complete replicator space therefore comprises spaces for organisms, organelles and all types of MGEs. Furthermore, recent studies described genetic elements for which an exact placement in either the extended virosphere or the sphere of living organisms is challenging. Giantviruses express traits typically associated with living organisms (Schulz et al. 2022), Borgs are exceptionally large linear plasmids in *Methanopredens* that provide crucial metabolic pathways to their

host (Al-Shayeb et al. 2022), and megaplasmids carry large amounts of accessory genes and possess selfish traits, bridging the gap between regular plasmids and chromids or secondary replicons (Hall et al. 2021). These elements blur the boundaries between organisms and parasitic elements, living and non-living things, and, therefore, life itself. Hence, it is crucial for scientists describing the vast complexity of life to develop helpful abstract concepts and practical nomenclature, and at the same time, be aware of the underlying fluidity and definition-breaking character of life.

References

Al-Shayeb, B., M. C. Schoelmerich, J. West-Roberts, L. E. Valentin-Alvarado, R. Sachdeva, S. Mullen, A. Crits-Christoph, M. J. Wilkins, K. H. Williams, J. A. Doudna, and J. F. Banfield (2022): "Borgs are giant genetic elements with potential to expand metabolic capacity," *Nature*, 610, 731–736, number: 7933 Publisher: Nature Publishing Group.

Biller, S. J., A. Coe, A. A. Arellano, K. Dooley, J. S. Gong, E. A. Yeager, J. W. Becker, and S. W. Chisholm (2022): "Environmental and taxonomic drivers of bacterial extracellular vesicle production in marine ecosystems," Pages: 2022.01.18.476865 Section: New Results.

Burch, C. L., A. Romanchuk, M. Kelly, Y. Wu, and C. D. Jones (2023): "Empirical Evidence That Complexity Limits Horizontal Gene Transfer," *Genome Biology and Evolution*, 15, evad089.

Bárdy, P., T. Füzik, D. Hrebík, R. Pantůček, J. Thomas Beatty, and P. Plevka (2020): "Structure and mechanism of DNA delivery of a gene transfer agent," *Nature Communications*, 11, 3034, number: 1 Publisher: Nature Publishing Group.

Deatherage, B. L. and B. T. Cookson (2012): "Membrane vesicle release in bacteria, eukaryotes, and archaea: a conserved yet underappreciated aspect of microbial life," *Infection and Immunity*, 80, 1948–1957.

Durrant, M. G., M. M. Li, B. A. Siranosian, S. B. Montgomery, and A. S. Bhatt (2020): "A Bioinformatic Analysis of Integrative Mobile Genetic Elements Highlights Their Role in Bacterial Adaptation," *Cell Host & Microbe*, 27, 140–153.e9.

Erdmann, S., B. Tschitschko, L. Zhong, M. J. Raftery, and R. Cavicchioli (2017): "A plasmid from an Antarctic haloarchaeon uses specialized membrane vesicles to disseminate and infect plasmid-free cells," *Nature Microbiology*, 2, 1446–1455, number: 10 Publisher: Nature Publishing Group.

Gregory, A. C., A. A. Zayed, N. Conceicao-Neto, B. Temperton, B. Bolduc, A. Alberti, M. Ardyna, K. Arkhipova, M. Carmichael, C. Cruaud, C. Dimier, G. Domínguez-Huerta, J. Ferland, S. Kandels, Y. Liu, C. Marec, S. Pesant, M. Picheral, S. Pisarev, J. Poulain, J.-E. Tremblay, D. Vik, S. G. Acinas, M. Babin, P. Bork, E. Boss, C. Bowler, G. Cochrane, C. de Vargas, M. Follows, G. Gorsky, N. Grimsley, L. Guidi, P. Hingamp, D. Iudicone, O. Jaillon, S. Kandels-Lewis, L. Karp-Boss, E. Karsenti, F. Not, H. Ogata, S. Pesant, N. Poulton, J. Raes, C. Sardet, S. Speich, L. Stemmann, M. B. Sullivan, S. Sunagawa, P. Wincker, M. Babin, C. Bowler, A. I. Culley, C. de Vargas, B. E. Dutilh, D. Iudicone, L. Karp-Boss, S. Roux, S. Sunagawa, P. Wincker, and M. B. Sullivan (2019): "Marine DNA Viral Macro- and Microdiversity from Pole to Pole," *Cell*, 177, 1109–1123.e14.

Hackl, T., R. Laurenceau, M. J. Ankenbrand, C. Bliem, Z. Cariani, E. Thomas, K. D. Dooley, A. A. Arellano, S. L. Hogle, P. Berube, G. E. Leventhal, E. Luo, J. M. Eppley, A. A. Zayed, J. Beaulaurier, R. Stepanauskas, M. B. Sullivan, E. F. DeLong, S. J. Biller, and S. W. Chisholm (2023): "Novel integrative elements and genomic plasticity in ocean ecosystems," *Cell*, 186, 47–62.e16.

Hall, J. P. J., J. Botelho, A. Cazares, and D. A. Baltrus (2021): "What makes a megaplasmid?" *Philosophical Transactions of the Royal Society B: Biological Sciences*, 377, 20200472.

Koonin, E. V., V. V. Dolja, M. Krupovic, and J. H. Kuhn (2021): "Viruses Defined by the Position of the Virosphere within the Replicator Space," *Microbiology and Molecular Biology Reviews : MMBR*, 85, e00193–20.

Krishnamurthy, S. R. and D. Wang (2017): "Origins and challenges of viral dark matter," *Virus Research*, 239, 136–142.

Lang, A. S., O. Zhaxybayeva, and J. T. Beatty (2012): "Gene transfer agents: phage-like elements of genetic exchange," *Nature reviews. Microbiology*, 10, 472–482.

Lobkovsky, A. E., Y. I. Wolf, and E. V. Koonin (2011): "Predictability of Evolutionary Trajectories in Fitness Landscapes," *PLOS Computational Biology*, 7, e1002302, publisher: Public Library of Science.

Lücking, D., C. Mercier, T. Alarcón-Schumacher, and S. Erdmann (2023): "Extracellular vesicles are the main contributor to the non-viral protected extracellular sequence space," *ISME Communications*, 3, 1–10, number: 1 Publisher: Nature Publishing Group.

Papke, R. T., P. Corral, N. Ram-Mohan, R. R. de la Haba, C. Sánchez-Porro, A. Makkay, and A. Ventosa (2015): "Horizontal Gene Transfer, Dispersal and Haloarchaeal Speciation," *Life*, 5, 1405–1426.

Pfeifer, E., J. A. Moura de Sousa, M. Touchon, and E. P. C. Rocha (2021): "Bacteria have numerous distinctive groups of phage–plasmids with conserved phage and variable plasmid gene repertoires," *Nucleic Acids Research*, 49, 2655–2673.

Santiago-Rodriguez, T. M. and E. B. Hollister (2022): "Unraveling the viral dark matter through viral metagenomics," *Frontiers in Immunology*, 13.

Schulz, F., C. Abergel, and T. Woyke (2022): "Giant virus biology and diversity in the era of genome-resolved metagenomics," *Nature Reviews Microbiology*, 20, 721–736, number: 12 Publisher: Nature Publishing Group.

Tschitschko, B., S. Erdmann, M. Z. DeMaere, S. Roux, P. Panwar, M. A. Allen, T. J. Williams, S. Brazendale, A. M. Hancock, E. A. Eloe-Fadrosh, and R. Cavicchioli (2018): "Genomic variation and biogeography of Antarctic haloarchaea," *Microbiome*, 6, 113.

Appendix A

Distribution and implications of haloarchaeal plasmids disseminated in self-encoded plasmid vesicles

Table of contents

Supplementary Material

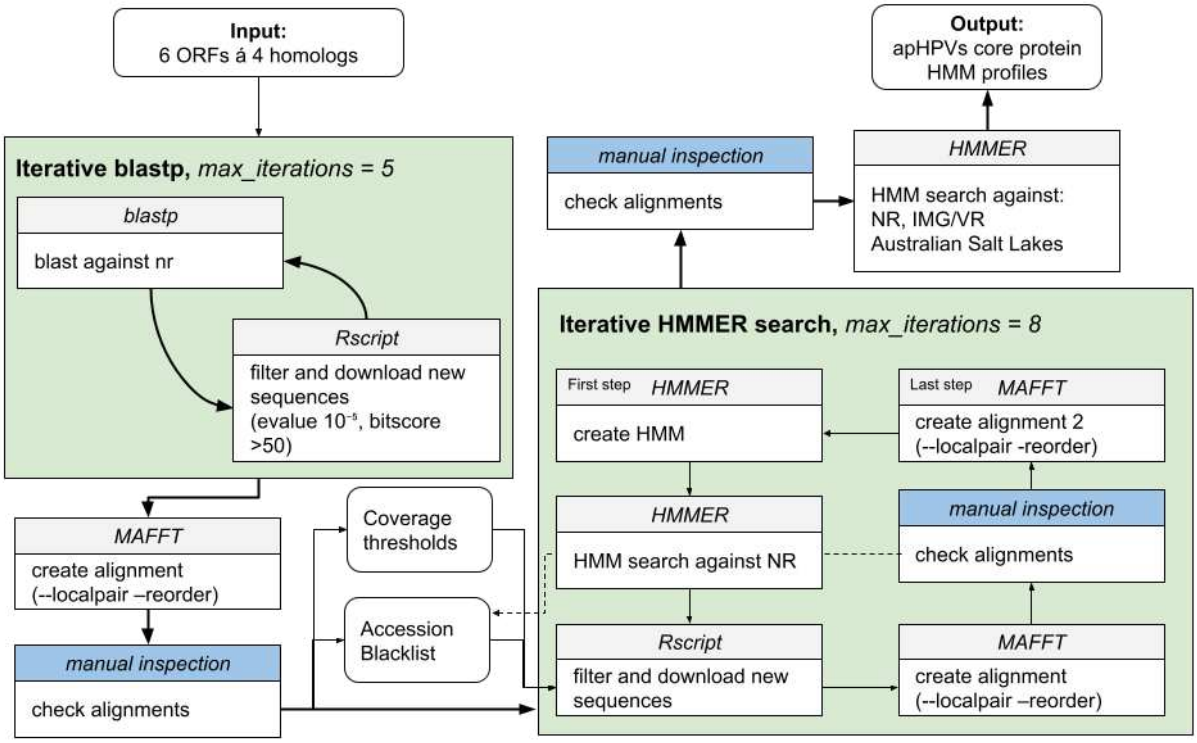


Figure S1: Overview of HMM profile generation for each ORF.

Details are given in method section 2.2. In brief, each ORF and homologs were iteratively blasted against NR, an alignment was created and curated. Then an iterative HMM search against NR and Australian Salt Lake proteomes was performed, inflating the HMM profiles. Each alignment was manually inspected before the next iteration. Sequences that were removed in one of the inspections, were added to an ‘accession blacklist’, which excluded them from being added again at a later iteration.

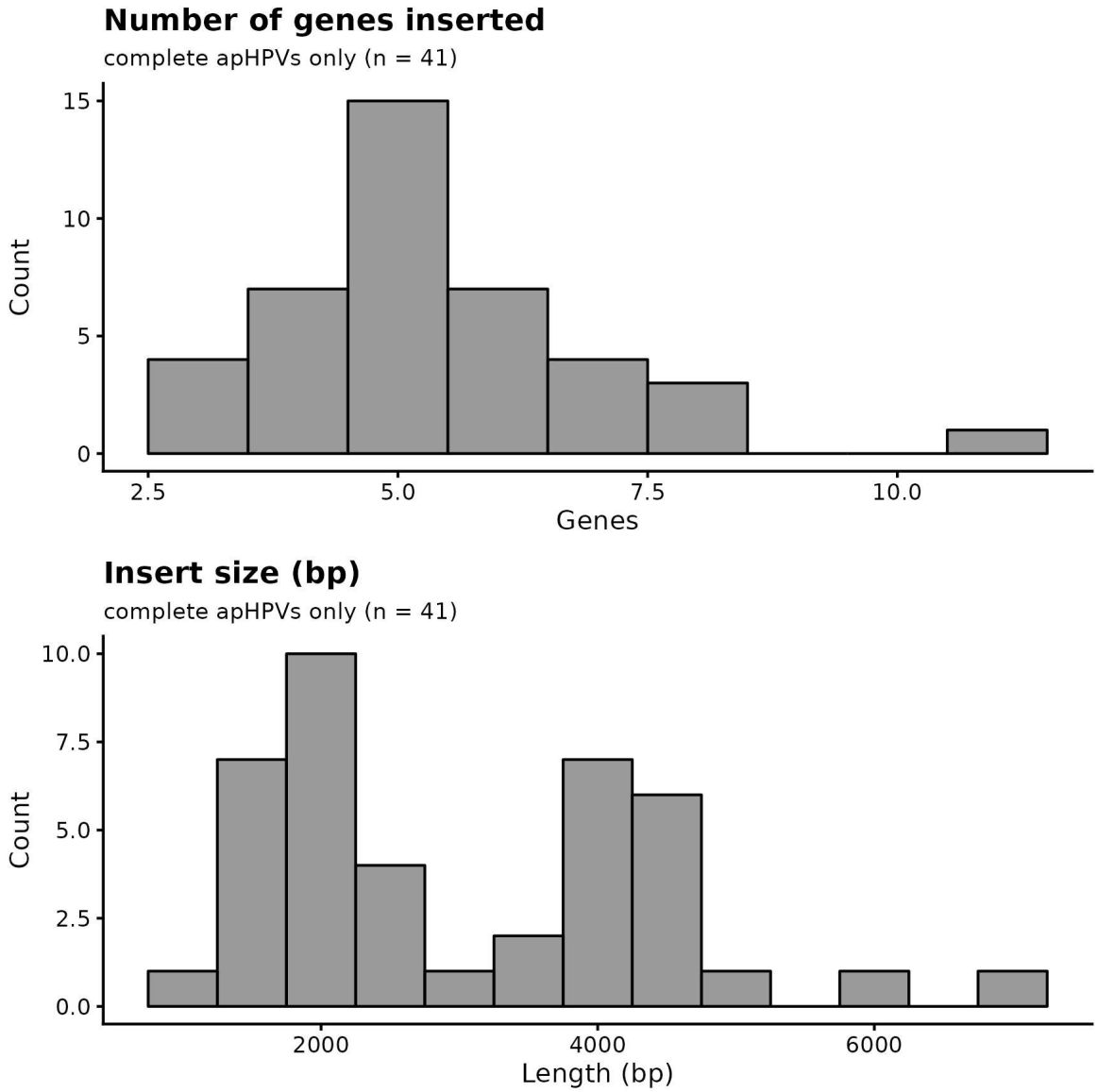


Figure S2: Flexible region of apHPVs.

Number of inserted genes and size (bucket size 500 bp) of the insert between cluster 1 (ORF6-9) and cluster 2 (ORF17-25) of 41 apHPVs.

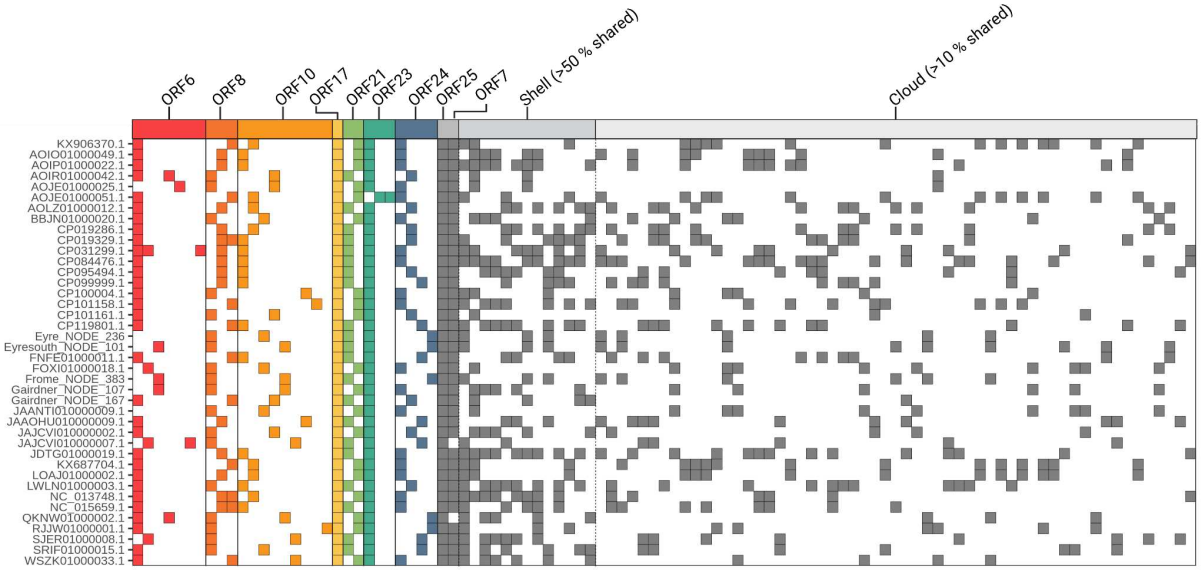


Figure S3: Expanded view of core protein clusters.

Presence-absence plot of protein clusters (columns) in different apHPVs (rows), in contrast to Figure 1 in the manuscript, the protein clusters of ORF6, 8, 10, 21, 23 and 24 are expanded.

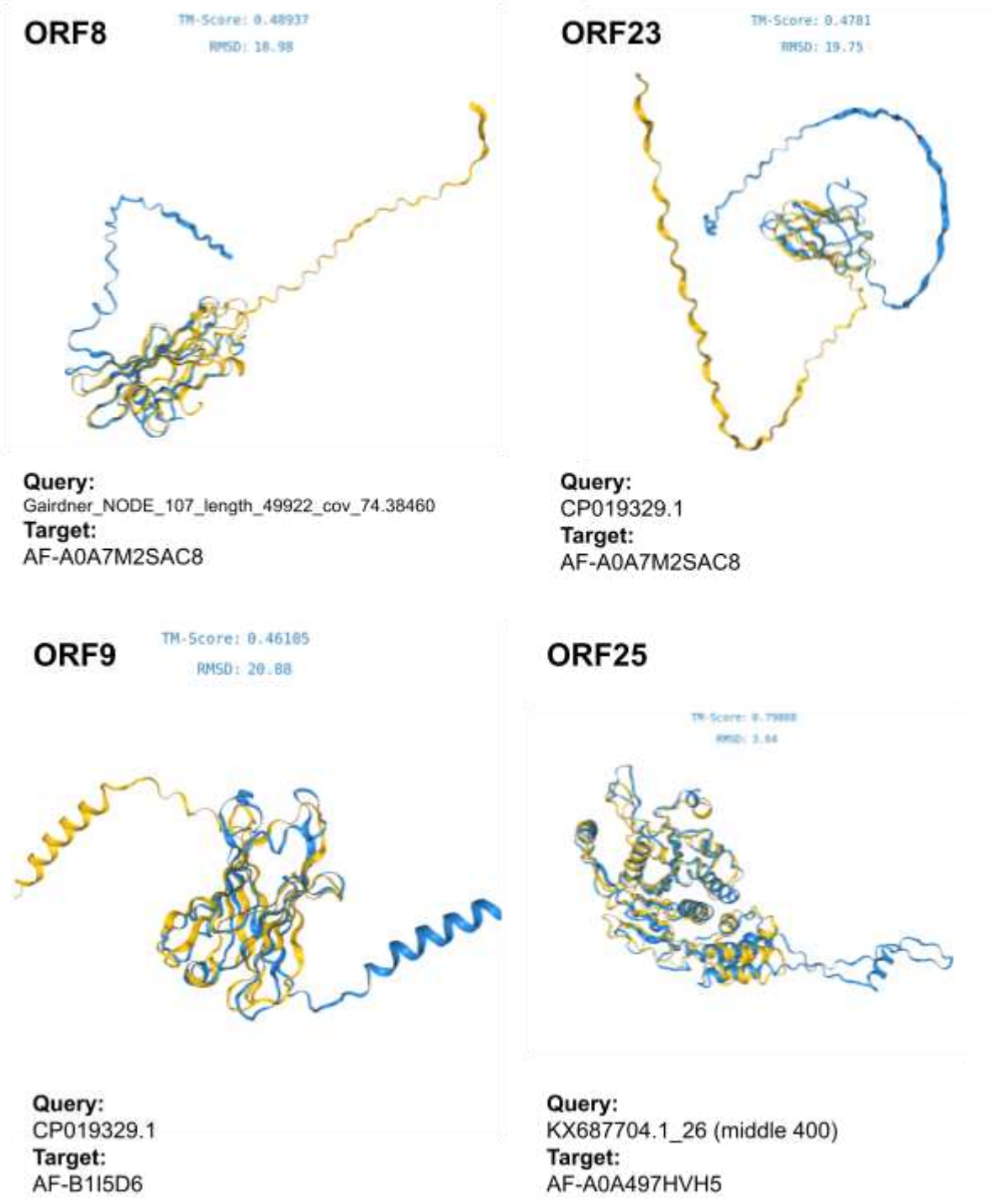


Figure S4: Foldseek alignments of selected core-proteins of apHPVs.

Selected examples of foldseek alignments of core proteins (blue) with publicly available protein structures (yellow). RMSD values and TM-scores are indicated on top. Detailed search results are documented in Table S2.

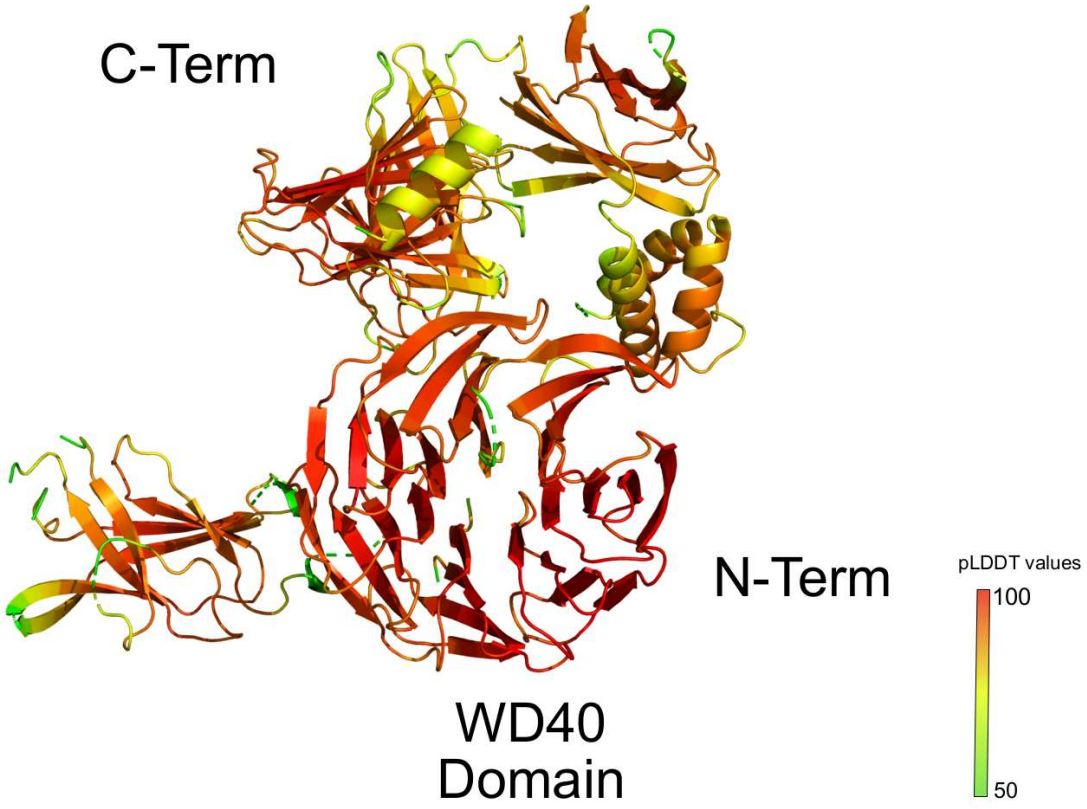


Figure S5: Predicted structure of ORF6 of pR1SE

Visible is a WD40 domain (IPR001680) at the N-terminus, found in ~56 % of all ORF6 homologs. Prediction was done using AlphaFold, color indicates pLDDT values.

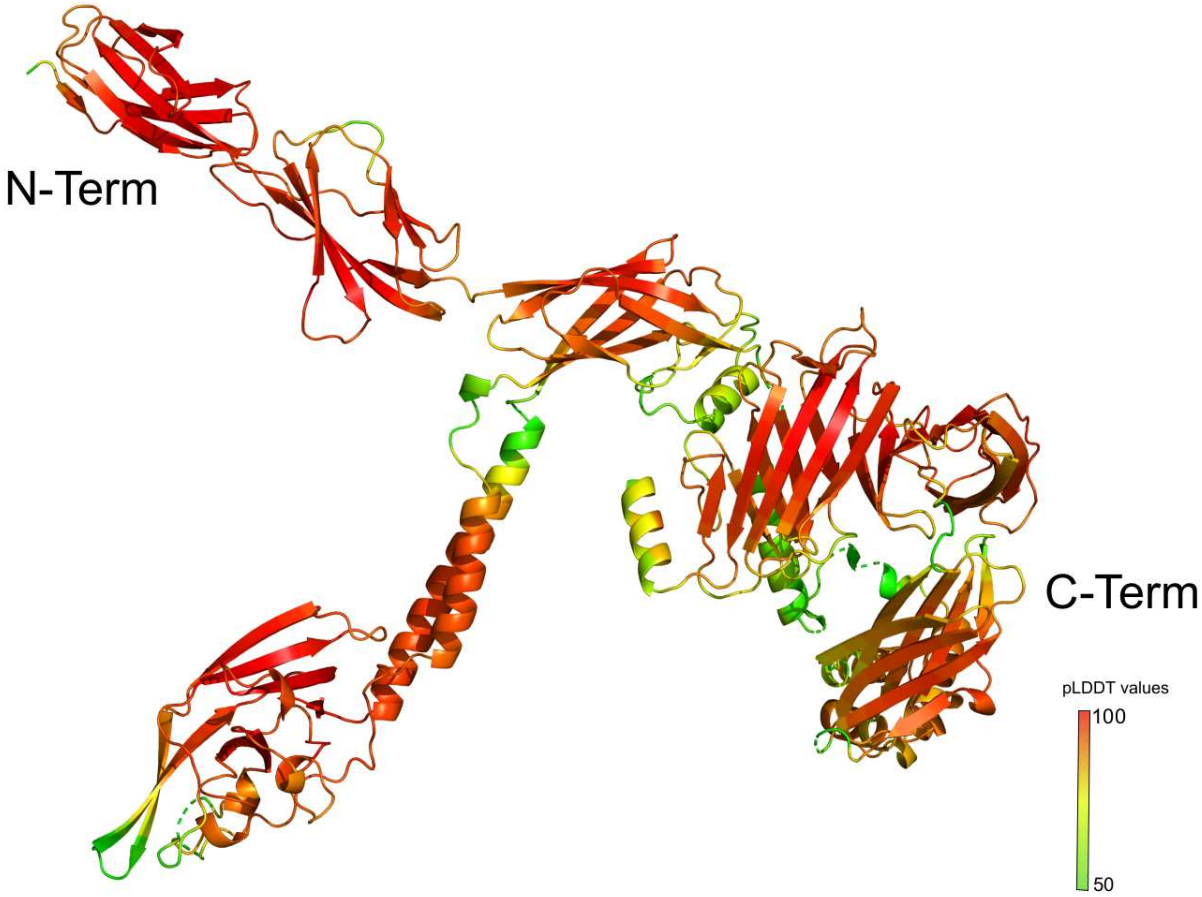


Figure S6: Extended arrays of antiparallel beta sheets in ORF6

Visible is a WD40 domain (IPR001680) at the N-terminus of the ORF6 homolog of apHPV BBJN01000020.1. Prediction was done using AlphaFold, color indicates pLDDT values.

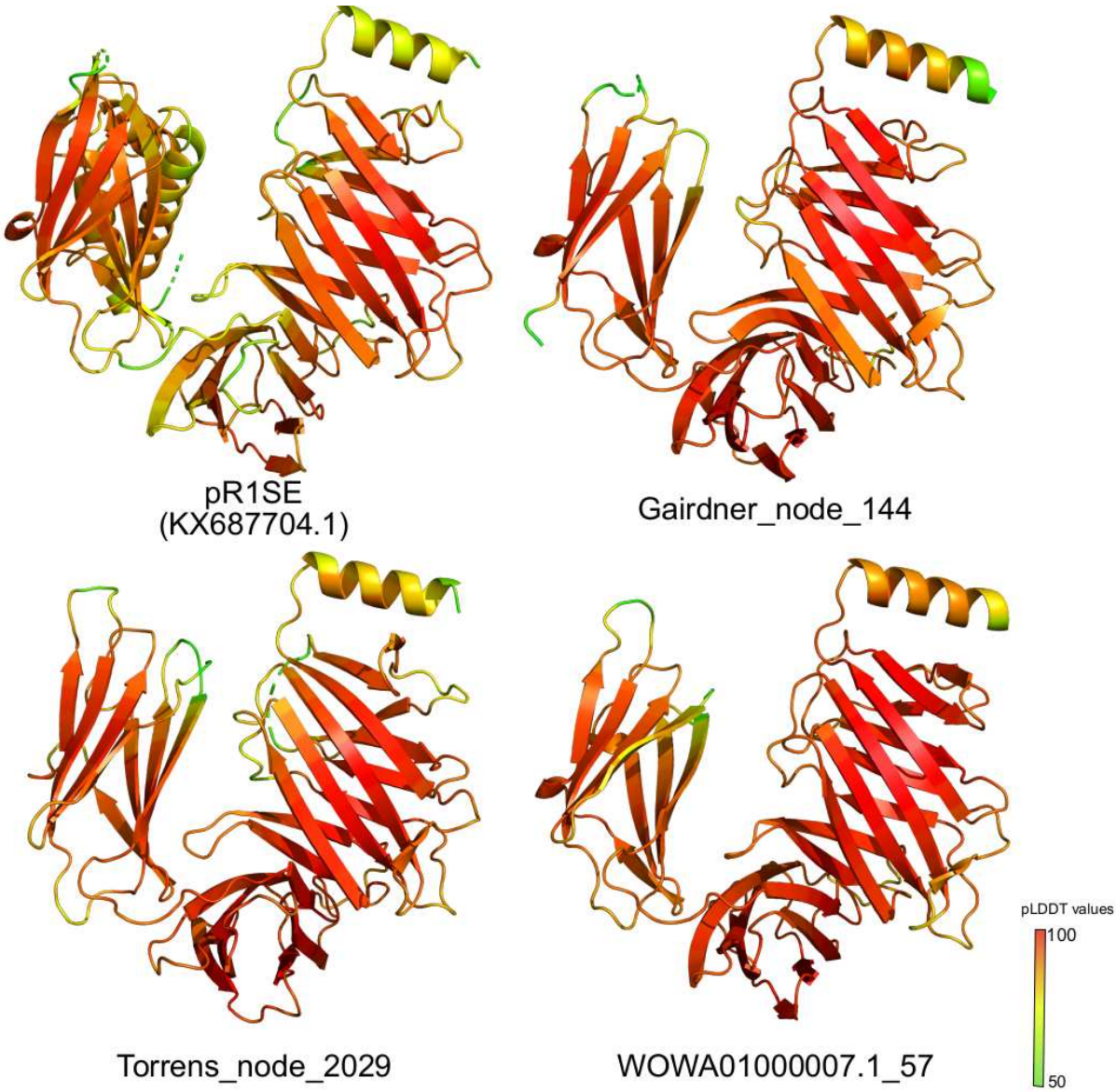


Figure S7: C-Terminus of four selected ORF6 homologs

Four examples of the C-terminus of ORF6 homologs, composed of multiple arrays of antiparallel beta sheets facing each other. Prediction was done using AlphaFold, color indicates pLDDT values.

Additional Supplementary Material

Table S1 - Accessions of proteins used for the initial search of homologs, minimum score and alignment length for the searches for each ORF
Table S2 - Foldseek results of ORF8, 9, 17, 21, 23 and 24.
Table S3 - apHPV overview, overview of all complete and incomplete discovered apHPVs, including: accessions, taxonomy, GC content, lengths, core-region position, circularity, defense systems.

Table S1 - Accessions of proteins used for the initial search of homologs. and their corresponding minimum scores and alignment lengths

ORF	pR1SE	Halorubrum saccharovor	Haloterrigena turkmenica	Halopiger xanaduensis	Min Score (hmmer)	Min Alignment Length (blastp)
ORF6	AQM75270.1	WP_152421865.1	WP_012946144.1	WP_013876082.1	50	300
ORF8	AQM75272.1	WP_004048522.1	WP_012946146.1	WP_013876080.1	80	206
ORF17	AQM75281.1	WP_004048507.1	WP_012946156.1	WP_013876072.1	80	120
ORF21	AQM75285.1	WP_004048502.1	WP_012946160.1	WP_013876068.1	55	80
ORF23	AQM75287.1	WP_004048500.1	WP_012946162.1	WP_049895738.1	50	192
ORF24	AQM75288.1	WP_004048499.1	WP_012946163.1	WP_013876066.1	50	245

Table S2 - Foldseek results, one sheet holds the foldseek results of one ORF

contig	part	sequence	foldseek_result	foldseek_proba	result_link	foldseek_result2	foldseek_proba	result_link2
AOJE01000025.	front	MGSDRARIPQG	Putative vacuola	1	https://www.alpha	NPP1 family prot	1	https://www.alphafold.ebi.ac.uk/entry/AF-A0A7M2SAC8
CP019329.1	front	MQDQLTRRQLL	Necrosis and eth	1	https://www.alpha	Vacuolar protein	1	https://www.alphafold.ebi.ac.uk/entry/AF-A0A200Q450
Eyre_NODE_23(	front	MVTRRRVLAGI/	Necrosis and eth	1	https://www.alpha	NPP1 family prot	1	https://www.alphafold.ebi.ac.uk/entry/AF-A0A7M2SAC8
Eyresouth_NODI	front	MTRQRQRTVSF	NPP1 family prot	1	https://www.alpha	Putative secretec	1	https://www.alphafold.ebi.ac.uk/entry/AF-Q0RHM4
Frome_NODE_3	front	MTAPRQRGVSF	NPP1 family prot	1	https://www.alpha	Necrosis inducing	1	https://www.alphafold.ebi.ac.uk/entry/AF-A0A1Y5WW06
Gairdner_NODE_	front	MDRRTFCKTAA	NPP1 family prot	1	https://www.alpha	Necrosis inducing	1	https://www.alphafold.ebi.ac.uk/entry/AF-A0A1Y5WW06
KX687704.1	front	MTRTRRQFLAS	PrsW family intra	1	https://www.alpha	NPP1-domain-co	1	https://www.alphafold.ebi.ac.uk/entry/AF-G4UYM1
NC_013748.1	front	MTEHTLTRRDA	NPP1 family prot	0,98	https://www.alpha	Vacuolar sorting-	0,98	https://www.alphafold.ebi.ac.uk/entry/AF-C0NCK1

contig	part	sequence	foldseek_result	foldseek_proba	result_link	foldseek_result2	foldseek_proba	result_link2
AOJE01000025.	front	MSILNINRSSVIS	Signal peptidase	1	https://www.alpha	Fibronectin type-	1	https://www.alphafold.ebi.ac.uk/entry/AF-A0A3N2BDN8
CP019329.1	front	MTTTFNRVATLF	Sporulation and s	1	https://www.alpha	S-layer domain p	1	https://www.alpha many good hits, could dive more into this
Eyre_NODE_23(	front	MNFNRSQLAVV	Signal peptidase	1	https://www.alpha	SusE outer mem	0,98	https://www.alpha many more hits
Eyresouth_NODI	front	MSRFTRLSGVA	Signal peptidase	1	https://www.alpha	S-layer homology	1	https://www.alpha also more hits
Frome_NODE_3	front	MSRFTRLSGVA	Por_Secre_tail d	1	https://www.alpha	Spore coat protei	1	https://www.alpha there were more hits of por_secre before in others!
Gairdner_NODE_	front	MTSKTKRLSSIA	Serine/threonine	1	https://www.alpha	Fibronectin type-	0,99	https://www.alphafold.ebi.ac.uk/entry/AF-A0A673BS56
KX687704.1	front	MFDTNTNSTTIA	Big_6 domain-co	1	https://www.alpha	DUF2808 domain	1	https://www.alphafold.ebi.ac.uk/entry/AF-M0LUY6
KX687704.1	back	EVNISGVENPHI	VWA domain-cor	0,99	https://www.alpha	Gliding motility-a	0,98	https://www.alphafold.ebi.ac.uk/entry/AF-A0A4U3KYA7
NC_013748.1	front	MRNFTKRVATL	Fibronectin type-	1	https://www.alpha	DUF4625 domain	1	https://www.alphafold.ebi.ac.uk/entry/AF-A0A520C251

contig	part	sequence	foldseek_result	foldseek_proba	result_link	foldseek_result2	foldseek_proba	result_link2
AOJE01000025.	front	MNDRALTILLGI	UnbV_ASPIC do	0,99	https://www.alph	Alpha amylase c	0,96	https://www.alphafold.ebi.ac.uk/entry/AF-F4LQA5
CP019329.1	front	MSRKKGNNGPI	-	-	-	-	-	
Eyre_NODE_23	front	MALLSREELSW	Ricin B-type lecti	0,99	https://www.alph	-	-	
Eyresouth_NOD	front	MGLLSRTDISWI	Ricin B-type lecti	0,99	https://www.alphafold.ebi.ac.uk/entry/AF-A0A7J0BYG8			
Frome_NODE_3	front	MALLSRTDISWI	PKD domain-con	0,99	https://www.alph	Ricin B-type lecti	0,99	https://www.alphafold.ebi.ac.uk/entry/AF-A0A7J0BYG8
Gairdner_NODE_	front	MTTRVRPAVLA	SH3 domain-con	0,99	https://www.alph	Metal-independe	1	https://www.alphafold.ebi.ac.uk/entry/AF-A0A7J3ZMY9
KX687704.1	front	MHQHTFRALLA	Curli production	0,99	https://www.alph	Signal peptidase	0,92	https://www.alphafold.ebi.ac.uk/entry/AF-A0A495WZR5
NC_013748.1	front	MGSLTTPRVLFI	SH3b domain-co	0,97	https://www.alphafold.ebi.ac.uk/entry/AF-A0A0G1DMA0			

contig	part	sequence	foldseek_result	foldseek_proba	result_link	foldseek_result2	foldseek_proba	result_link2
AOJE01000025.	front	MIERNAEELRSE	-	-	-	-	-	-
CP019329.1	front	MSTYINRQRKQ	-	-	-	-	-	-
Eyre_NODE_236	front	MNNLDRHGTTE	-	-	-	-	-	-
Eyresouth_NODI	front	MNNLDRHGTTE	WRKY transcripti	0,91	https://www.alpha	-	-	-
Frome_NODE_3	front	MNNLDRHGTTE	-	-	-	-	-	-
Gairdner_NODE_	front	MTSQTQQQPPF	-	-	-	-	-	-
KX687704.1	front	MSTSNTPESIDL	-	-	-	-	-	-
NC_013748.1	front	MRHTSTTEAQC	-	-	-	-	-	-

contig	part	sequence	foldseek_result	foldseek_proba	result_link	foldseek_result2	foldseek_proba	result_link2
AOJE01000025.	front	MLVVCLIGAGV(	TonB-dependent	0,99	https://www.alph	Ig-like domain-co	0,99	https://www.alphafold.ebi.ac.uk/entry/AF-A0A7C6JRN5
CP019329.1	front	MSKPTPHRVAT	T2SS-T3SS_pil_	1	https://www.alph	Pilus assembly p	1	https://www.alphafold.ebi.ac.uk/entry/AF-A0A2N3LUC0
Eyre_NODE_23(	front	MRGTFQHVAV\	Type II and III se	1	https://www.alph	TonB-dependent	1	https://www.alphafold.ebi.ac.uk/entry/AF-W4PEA6
Eyresouth_NODI	front	MRGTIQHAAVV	Pilus assembly p	1	https://www.alph	TonB-dependent	1	https://www.alphafold.ebi.ac.uk/entry/AF-W4PEA6
Frome_NODE_3	front	MRGTLKHTAVV	Pilus assembly p	1	https://www.alph	TonB-dependent	1	https://www.alphafold.ebi.ac.uk/entry/AF-W4PEA6
Gairdner_NODE_	front	MKRVVLAALLVI	Peptidase C39 d	1	https://www.alph	S-layer protein	1	https://www.alphafold.ebi.ac.uk/entry/AF-W7KNZ5
KX687704.1	front	MNIRAILAALLVI	Peptidase_M6 dc	0,96	https://search.fol	-	-	
NC_013748.1	front	MRRAFALLAVC	T2SS-T3SS_pil_	1	https://www.alph	Pilus formation p	1	https://www.alphafold.ebi.ac.uk/entry/AF-A0A1G5NBI9

contig	part	sequence	foldseek_result	foldseek_proba	comment	result1	result2
AOJE01000025.	front	MNPRDVVKKLV	transmembrane	1		https://gmgc.embl.de/search.cgi?search_id=GMGC10.045_962_180.UNKNOWN_trun_0	
AOJE01000025.	back	TGSKKREARTA	-	-			
CP019329.1	front	MRSRFSILDRAI	-	-			
CP019329.1	back	EQELDAIRAELI	Transcriptional a	1	also: DNA-directe	https://www.alpha	https://www.alphafold.ebi.ac.uk/entry/AF-A0A059LCG6
Eyre_NODE_23	front	MALAFVSELVS	-	-			
Eyre_NODE_23	back	VADVDVDVEDL	-	-			
Eyresouth_NODI	front	MALAFVSEVVS	-	-			
Eyresouth_NODI	back	SPDDVDVDIDD	-	-			
Frome_NODE_3	front	MALAFVSEVVS	-	-			
Frome_NODE_3	back	APDDVDVDIDD	-	-			
Gairdner_NODE	front	MNAEIIRSLLLD	-	-			
Gairdner_NODE	back	LAAAETGYAPA	-	-			
KX687704.1	front	MTDKRNDHGRI	transmembrane	0,9		https://gmgc.embl.de/search.cgi?search_id=GMGC10.045_962_180.UNKNOWN_trun_0	
KX687704.1	back	SRAATKAASRT	B-block_TFIIIC d	0,96		https://www.alphafold.ebi.ac.uk/entry/AF-A0A3M6Y5K4	
NC_013748.1	front	MEMPDTPDNDI	-	-			
NC_013748.1	back	KRPDLDRCERV	-	-			

Table S3 - Overview of all complete and incomplete discovered apHPVs

Only a selection of columns is shown. The full table is available online.

contig	contig_category	name	type	environment	country	kingdom	class	order	family	genus	% GC	genome_segmen	antiviral_defens	anti-CRISPR_proteins
AOID01000019.1	complete	Natrinema versiforme JCM 10478_JCM 10478	culture			Archaea	Halobacteria	Natrialbales	Natrialbaceae	Natrinema	60.8%	circular		AcrIIA7
AOIO01000049.1	complete	Natrialba asiatica DSM 12278_DSM 12278	culture			Archaea	Halobacteria	Natrialbales	Natrialbaceae	Natrialba	56.6%	linear	cas_type_I-D (complete)	
AOIP01000022.1	complete	Natrialba aegyptia DSM 13077_DSM 13077	culture			Archaea	Halobacteria	Natrialbales	Natrialbaceae	Natrialba	56.9%	circular		
AOIR01000042.1	complete	Natrinema thermotolerans DSM 11552_DSM 11552	culture			Archaea	Halobacteria	Natrialbales	Natrialbaceae	Natrinema	63.2%	linear		
AOJE01000025.1	complete	Halorubrum saccharovorum DSM 1137_DSM 1137	culture			Archaea	Halobacteria	Haloferacales	Halorubraceae	Halorubrum	61.1%	linear		
AOJE01000051.1	complete	Halorubrum saccharovorum DSM 1137_DSM 1137	culture			Archaea	Halobacteria	Haloferacales	Halorubraceae	Halorubrum	61.4%	linear		
AOLZ01000012.1	complete	Halobiforma lacisalsi AJ5_AJ5	culture			Archaea	Halobacteria	Natrialbales	Natrialbaceae	Halobiforma	64.7%	linear		
BBJN01000020.1	complete	Halolamina rubra_CBA1107	isolate	solar salt	Korea	Archaea	Halobacteria	Haloferacales	Halorubraceae	Halolamina	62.5%	linear	PDC-S11 (incomplete), PDC-S45 (incomplete)	
CP019286.1	complete	Halobiforma lacisalsi AJ5_AJ5	culture			Archaea	Halobacteria	Natrialbales	Natrialbaceae	Halobiforma	56.4%	circular	cas_type_I-B1 (complete), Tiamat (?), DMS_other (?)	
CP019329.1	complete	Natronorubrum daqingense_JX313	isolate	salt soil	China	Archaea	Halobacteria	Natrialbales	Natrialbaceae	Natronorubrum	56.0%	circular		
CP031299.1	complete	Natrinema thermotolerans_A29	culture			Archaea	Halobacteria	Natrialbales	Natrialbaceae	Natrinema	62.3%	circular	cas_type_other (incomplete)	
CP084476.1	complete	Natrinema sp. DC36_DC36	isolate	salt lake	Mongolia	Archaea	Halobacteria	Natrialbales	Natrialbaceae	Natrinema	60.9%	circular		
CP095494.1	complete	Halosolutus gelatinilyticus_SQT-29-1	isolate	salt lake	Mongolia	Archaea	Halobacteria	Natrialbales	Natrialbaceae	Halosolutus	55.9%	circular	cas_type_I-B1 (complete)	
CP099999.1	complete	Halovivax sp. AGai3-5_AGai3-5	isolate	salt lake	Mongolia	Archaea	Halobacteria	Natrialbales	Natrialbaceae	Halovivax	60.9%	circular		
CP100004.1	complete	Halorussus sp. XZYJT10_XZYJT10	isolate	salt lake	Mongolia	Archaea	Halobacteria	Halobacteriales	Halobacteriaceae	Halorussus	59.7%	circular	PDC-M50 (incomplete)	
CP101158.1	complete	Saliphagus sp. WLHSJ1_WLHSJ1	isolate	salt lake	Mongolia	Archaea	Halobacteria	Natrialbales	Natrialbaceae	Saliphagus	54.8%	circular	cbass_type_I (complete), RM_type_I (complete)	
CP101161.1	complete	Halosegnis sp. ZY10_ZY10	isolate	salt lake	Mongolia	Archaea	Halobacteria	Halobacteriales	Haloarculaceae	Halorarius	68.4%	circular	8 different systems, completeness questionable	
CP119801.1	complete	Halovivax sp. TS33_TS33	isolate	salt lake	Mongolia	Archaea	Halobacteria	Natrialbales	Natrialbaceae	Halovivax	58.8%	circular		
Eyre_NODE_236_length_33155_cov_19.932	complete		metagenomic	hypersaline lake	Australia	unclear					62.2%	linear		
Eyresouth_NODE_101_length_48437_cov_7.17035	complete		metagenomic	hypersaline lake	Australia	unclear					62.8%	linear		
FNFE01000011.1	complete	Natronorubrum texcoconense_B4,CECT 8067,JCM 17497	isolate	hypersaline lake	Mexico	Archaea	Halobacteria	Natrialbales	Natrialbaceae	Natronorubrum	57.6%	circular		
FOXI01000018.1	complete	Halolamina pelagica_CGMCC 1.10329	isolate	marine solar salt	China	Archaea	Halobacteria	Haloferacales	Halorubraceae	Halolamina	66.5%	linear		
Frome_NODE_383_length_24936_cov_4.316346	complete		metagenomic	hypersaline lake	Australia	unclear					64.0%	linear		
Gairdner_NODE_107_length_49922_cov_74.384603	complete		metagenomic	hypersaline lake	Australia	Archaea	Halobacteria	Halobacteriales	Haloarculaceae		60.6%	linear		
Gairdner_NODE_167_length_41210_cov_9.888325	complete		metagenomic	hypersaline lake	Australia	Archaea	Halobacteria	Halobacteriales	Haloarculaceae	Haloarcula	58.6%	linear		
JAANTI010000009.1	complete	Halolamina sp. R1-12_R1-12	isolate	marine solar salt	China	Archaea	Halobacteria	Haloferacales	Halorubraceae	Halolamina	68.0%	linear		
JAAOHU010000009.1	complete	Halostella sp. JP-L12_JP-L12	isolate	marine solar salt	China	Archaea	Halobacteria	Halobacteriales	Natronoarchaea	Halostella	61.1%	circular		
JAJCVI010000002.1	complete	Halorubellus salinus_GX3	isolate	marine solar salt	China	Archaea	Halobacteria	Halobacteriales	Halobacteriaceae	Halorubellus	69.1%	linear	PD-T7-4 (complete ?)	
JAJCVI010000007.1	complete	Halorubellus salinus_GX3	isolate	marine solar salt	China	Archaea	Halobacteria	Halobacteriales	Halobacteriaceae	Halorubellus	62.7%	circular	DMS_other (complete), PDC-S27/70, DMS_other	
JDTG01000019.1	complete	Natrinema thermotolerans A29_A29	isolate	salt-fermented fc	Korea	Archaea	Halobacteria	Natrialbales	Natrialbaceae	Natrinema	62.3%	linear	cas_type_other (complete)	
KX687704.1 (pR1SE)	complete	Halorubrum lacusprofundi_R1S1	isolate	hypersaline lake	Antarctica	Archaea	Halobacteria	Haloferacales	Halorubraceae	Halorubrum	58.9%	circular		
KX906370.1	complete	Halorubrum lacusprofundi_HLS1	isolate	hypersaline lake	Antarctica	Archaea	Halobacteria	Haloferacales	Halorubraceae	Halorubrum	58.5%	circular	PDC-S28, PDC-S70	
LOAJ01000002.1	complete	Halorubrum aethiopicum_SAH-A6	isolate	rock salt	Ethopia	Archaea	Halobacteria	Haloferacales	Halorubraceae	Halorubrum	58.0%	linear	cbass_type_I	
LWLN01000003.1	complete	Natrinema saccharevitans_AB14	isolate	salt lake	China	Archaea	Halobacteria	Natrialbales	Natrialbaceae	Natrinema	61.2%	circular		
NC_013748.1	complete	Haloterrigena turkmenica DSM 5511_DSM 5511	isolate	sulfate saline soil	Turkmenistan	Archaea	Halobacteria	Natrialbales	Natrialbaceae	Haloterrigena	60.0%	circular		
NC_015659.1	complete	Halopiger xanaduensis SH-6_SH-6	isolate	salt lake	Mongolia	Archaea	Halobacteria	Natrialbales	Natrialbaceae	Halopiger	62.2%	circular		
QKNW01000002.1	complete	Halosegnis longus_F12-1	isolate	hypersaline lake	Spain	Archaea	Halobacteria	Halobacteriales	Haloarculaceae	Halosegnis	57.6%	circular	SoFic, DMS_other, PifA, Tiamat	
RJJW01000001.1	complete	Halorubrum sp. CSM-61_CSM-61	isolate	salt mine	Ireland	Archaea	Halobacteria	Haloferacales	Halorubraceae	Halorubrum	64.8%	linear	HEC-06 (incomplete), PT_DndABCDE (complete)	
SJER01000008.1	complete	Halostella litorea_DLLS-108	isolate	salted food	China	Archaea	Halobacteria	Halobacteriales	Natronoarchaea	Halostella	62.8%	circular	PDC-S70	
SRIF01000015.1	complete	Halomicroarcula amylolytica_LR21	isolate	salt mine	China	Archaea	Halobacteria	Halobacteriales	Haloarculaceae	Halomicroarcula	60.1%	circular		
WSZK01000033.1	complete	Halomarina oriensis_JCM 16495	isolate	saltpan	India	Archaea	Halobacteria	Halobacteriales	Halobacteriaceae	Halomarina	63.1%	linear		
Gairdner_NODE_1649_length_16492_cov_7.650423	incomplete										63.0%			
Gairdner_NODE_217_length_37248_cov_7.812599	incomplete										61.6%			
Gairdner_NODE_2803_length_13236_cov_7.077763	incomplete										65.9%			
Gairdner_NODE_4298_length_11003_cov_3.762240	incomplete										61.1%			
Torrens_NODE_2029_length_12012_cov_23.663126	incomplete										64.7%			
AOMA01000019.1	incomplete	Halobiforma nitratireducens JCM 10879_JCM 10879				Archaea	Halobacteria	Natrialbales	Natrialbaceae	Halobiforma	63.5%			
AP017558.1	incomplete	Halopenitus persicus_CBA1233				Archaea	Halobacteria	Haloferacales	Halorubraceae	Halopenitus	66.6%			
CP019330.1	incomplete	Natronorubrum daqingense_JX313T				Archaea	Halobacteria	Natrialbales	Natrialbaceae	Natronorubrum	57.7%			
CP026314.1	incomplete	Salinigranum rubrum_GX10				Archaea	Halobacteria	Haloferacales	Halorubraceae	Salinigranum	63.2%			
CP040639.1	incomplete	Natrinema pallidum_BOL6-1				Archaea	Halobacteria	Natrialbales	Natrialbaceae	Natrinema	60.5%			
CP058909.1	incomplete	Halosimplex pelagicum_R2				Archaea	Halobacteria	Halobacteriales	Haloarculaceae	Halosimplex	67.1%			
CP081958.1	incomplete	Halobaculum magnesiiphilum_NBRC 109044				Archaea	Halobacteria	Haloferacales	Halorubraceae	Halobaculum	69.5%			
CP100355.1	incomplete	Saliphagus sp. WLHS1_WLHS1				Archaea	Halobacteria	Natrialbales	Natrialbaceae	Saliphagus	63.8%			
CP104003.1	incomplete	Salinirubellus salinus_ZS-35-S2				Archaea	Halobacteria	Halobacteriales	Haloarculaceae	Salinirubellus	69.0%			
Eyresouth_NODE_218_length_34093_cov_9.73377	incomplete										64.2%			
FNFC01000002.1	incomplete	Halovenus aranensis_IBRC-M10015				Archaea	Halobacteria	Halobacteriales	Halobacteriaceae	Halovenus	60.8%			
FOIS01000007.1	incomplete	Natrinema salifodinae_CGMCC 1.12284				Archaea	Halobacteria	Natrialbales	Natrialbaceae	Natrinema	59.1%			
FOXI01000019.1	incomplete	Halolamina pelagica_CGMCC 1.10329				Archaea	Halobacteria	Haloferacales	Halorubraceae	Halolamina	61.0%			
Frome_NODE_208_length_31991_cov_42.105993	incomplete										62.9%			
Gairdner_NODE_118_length_47723_cov_5.388856	incomplete										61.6%			
Gairdner_NODE_144_length_43966_cov_7.457380	incomplete										49.5%			
Gairdner_NODE_200_length_38460_cov_9.422054	incomplete										56.5%			

Gairdner_NODE_74_length_59309_cov_8.317565	incomplete								63.8%
JABUQZ010000003.1	incomplete	Haloterrigena gelatinilytica_SYSU A558-1	Archaea	Halobacteria	Natrialbales	Natrialbaceae	Haloterrigena		62.0%
JAKRVX010000010.1	incomplete	Natronocalculus amylovorans_AArc-St2	Archaea	Halobacteria	Haloferacales	Halorubraceae	Natronocalculus		47.9%
JAQLTW010000001.1	incomplete	Halorubrum ezzemoulense_AMM088	Archaea	Halobacteria	Haloferacales	Halorubraceae	Halorubrum		69.7%
Kenyon_NODE_596_length_16493_cov_7.34451	incomplete								60.2%
NXNI01000003.1	incomplete	Natrinema ejinorense_JCM 13890	Archaea	Halobacteria	Natrialbales	Natrialbaceae	Natrinema		61.0%
QPLN01000011.1	incomplete	Haloarcula sp. Atlit-120R_Atlit-120R	Archaea	Halobacteria	Halobacteriales	Haloarculaceae	Haloarcula		60.6%
Tyrell_NODE_42_length_52436_cov_43.400050	incomplete								62.4%
WOWA01000007.1	incomplete	Haloarcula argentinensis_pws5	Archaea	Halobacteria	Halobacteriales	Haloarculaceae	Haloarcula		61.3%
WPCC01000021.1	incomplete	Natronomonas sp. CBA1123_CBA1123	Archaea	Halobacteria	Halobacteriales	Haloarculaceae	Natronomonas		59.2%
CP025066.1	incomplete	Halalkaliarchaeum desulfuricum_AArc-Sl	Archaea	Halobacteria	Haloferacales	Halorubraceae	Halalkaliarchaeu	m	63.1%
CP058334.1	incomplete	Natronomonas halophila_C90	Archaea	Halobacteria	Halobacteriales	Haloarculaceae	Natronomonas		64.4%
BMCI01000010.1	incomplete	Haloferax sulfurifontis_CCM 7217	Archaea	Halobacteria	Haloferacales	Haloferacaceae	Haloferax		63.0%
BMCI01000011.1	incomplete	Haloferax sulfurifontis_CCM 7217	Archaea	Halobacteria	Haloferacales	Haloferacaceae	Haloferax		64.7%
JAJNEG010000027.1	incomplete	Halorubrum lacusprofundi_HL1RP11	Archaea	Halobacteria	Haloferacales	Halorubraceae	Halorubrum		62.4%
JAJNEG010000028.1	incomplete	Halorubrum lacusprofundi_HL1RP11	Archaea	Halobacteria	Haloferacales	Halorubraceae	Halorubrum		60.2%

Appendix B

Extracellular vesicles are the main contributor to the non-viral protected extracellular sequence space

Table of contents

Supplementary Material

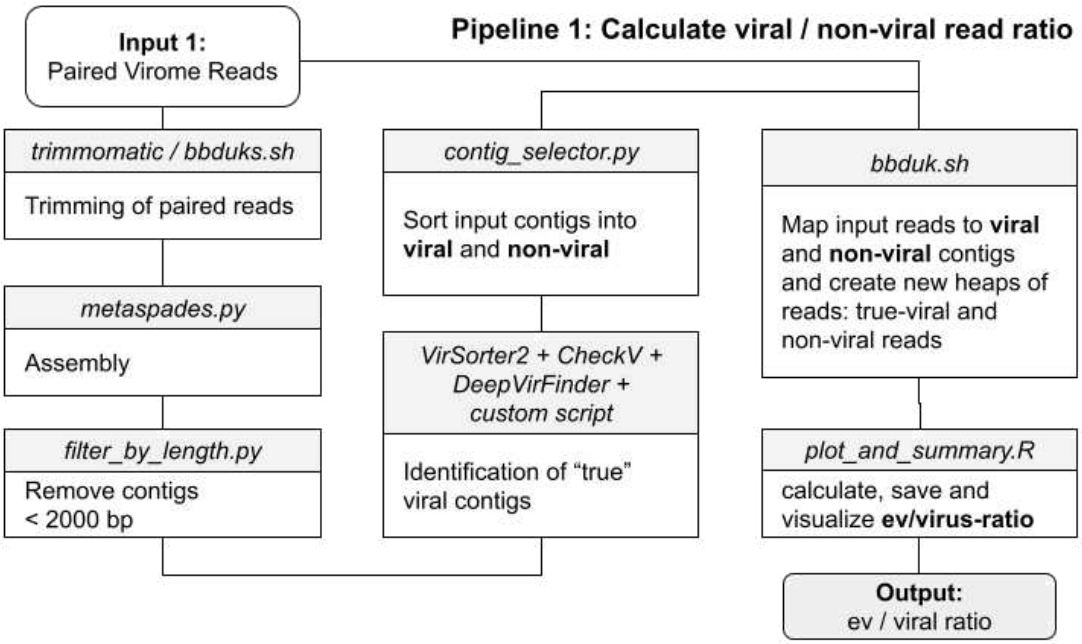


Figure S1: Pipeline 1 - Calculation of "non-viral / viral read ratio" or "percentage of non-viral associated reads".

Schematic overview of the bioinformatic pipeline in order to calculate the percentage of non-viral associated reads within a given virome / EV enrichment. Reads were trimmed, assembled into contigs and short contigs removed. Each contig was labeled as "viral" or "non-viral" based on the results of virus prediction tools. Finally, the input reads were mapped against the contigs and the ratio between viral-mapping and non-viral-mapping reads was calculated.

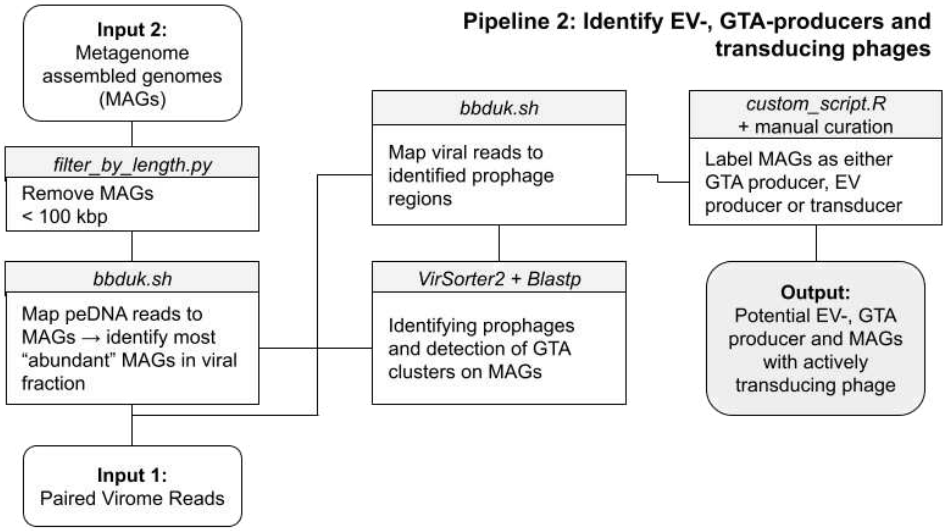


Figure S2: Pipeline 2 - Identification of potential EV-, GTA producers and microbes with actively transducing virus.

MAGs shorter than 100 kbp were removed. peDNA reads were mapped against each MAG and the 20 most peDNA-recruiting MAGs were selected for further analysis. Each MAG was subsequently scrutinized for the presence of a prophage or a GTA cluster and subsequently labeled as either GTA producer, EV producer or transducer using an in-house custom script and manual curation (decision making logic, see Figure S6).

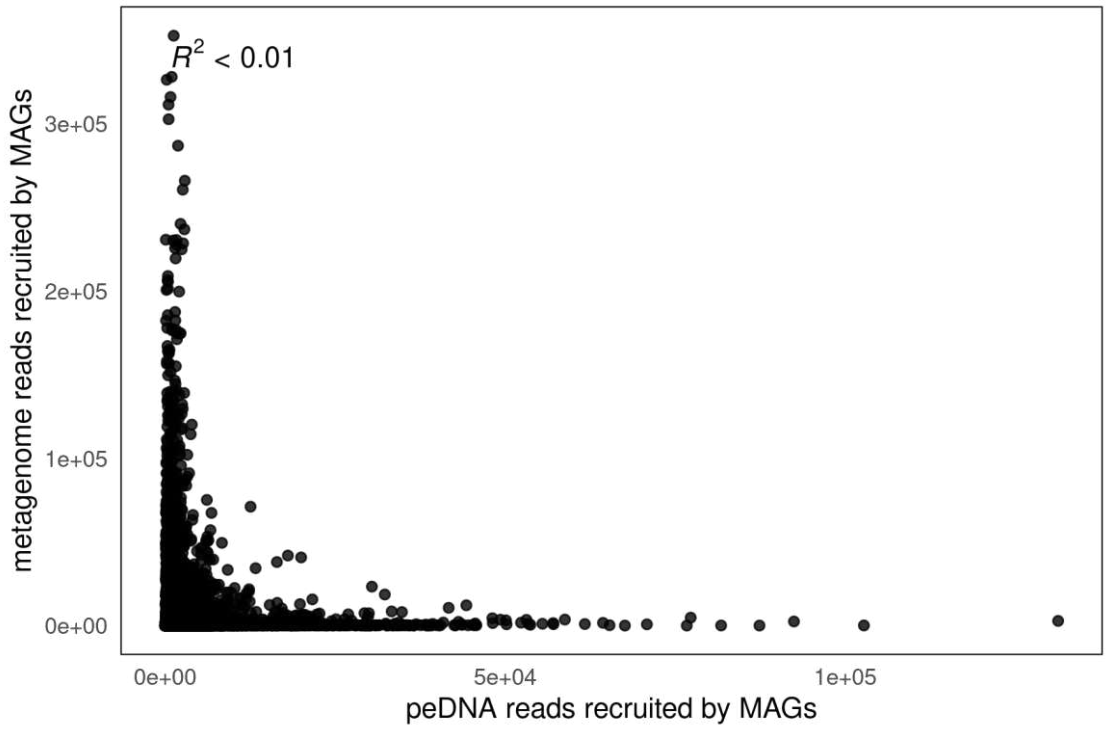


Figure S3: MAG abundance in metagenomes versus virome/EV-enrichment.

Number of virome reads recruited on the x-axis, number of metagenome reads recruited on the y-axis, per MAG. R^2 = 0.003391 calculated using linear regression.

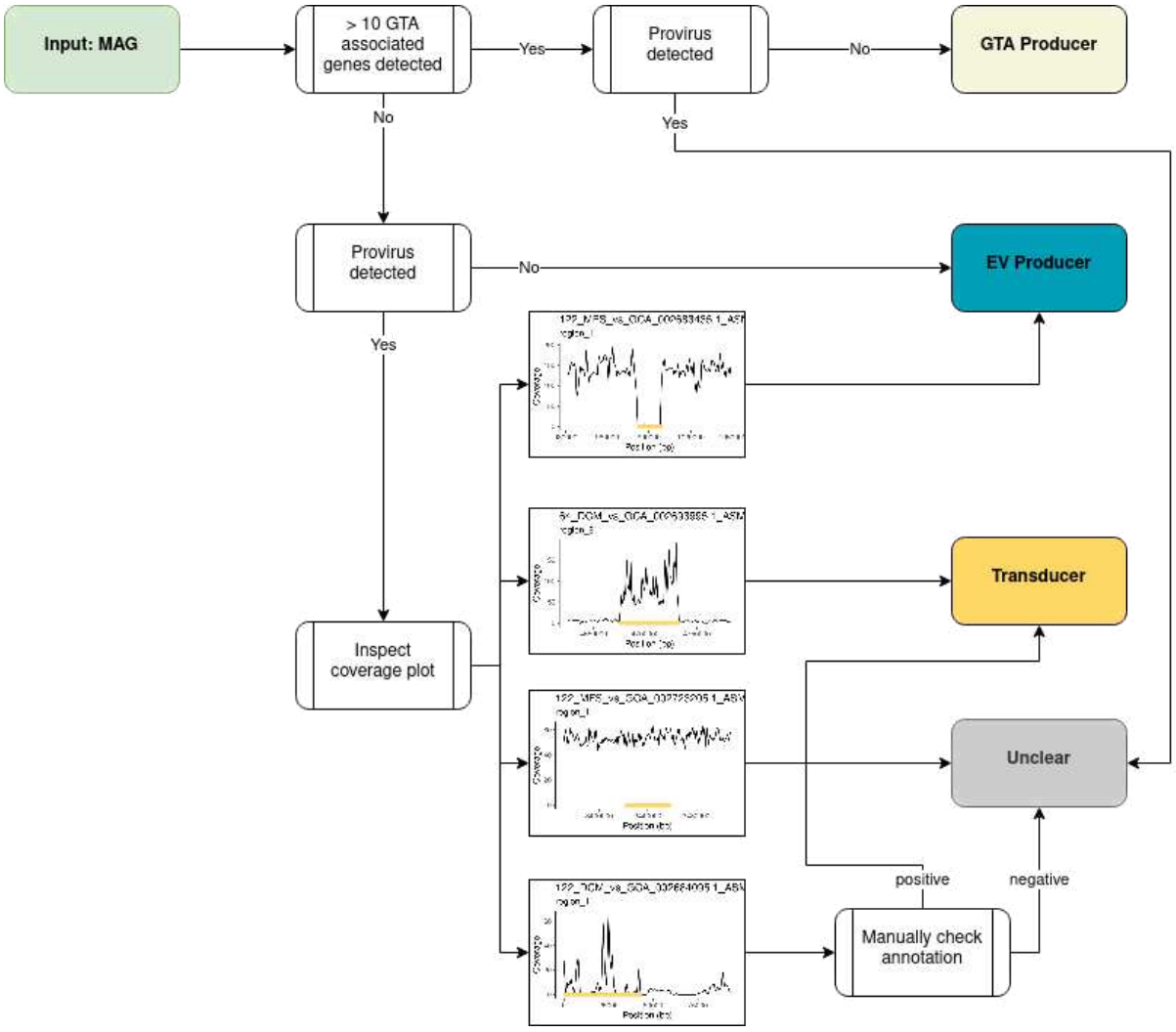


Figure S4: Decision making logic for the identification of potential EV-, GTA producers and microbes with actively transducing virus.

A MAG was labeled as “GTA producer” if >10 GTA associated genes were detected and no provirus was found. If <10 GTA genes and no provirus was found, the MAG was labeled “EV producer”. If both a provirus was detected and >10 GTA genes, the MAG was labeled “unclear”. If a provirus was detected, the coverage plot was scrutinized for manual inspection and decision making. If the viral region was “absent”, the MAG was labeled “EV producer”. If the viral region showed an increased coverage above the region, it was labeled as “Transducer”. For unclear cases, the exact annotation of the viral region was analyzed and manually curated and combined with the coverage plot. Based on this information, the MAG was labeled either as “unclear” or “Transducer”.

See file: Figure S5 - Coverage Plots

Figure S5: Coverage plots (Separate pdf file)

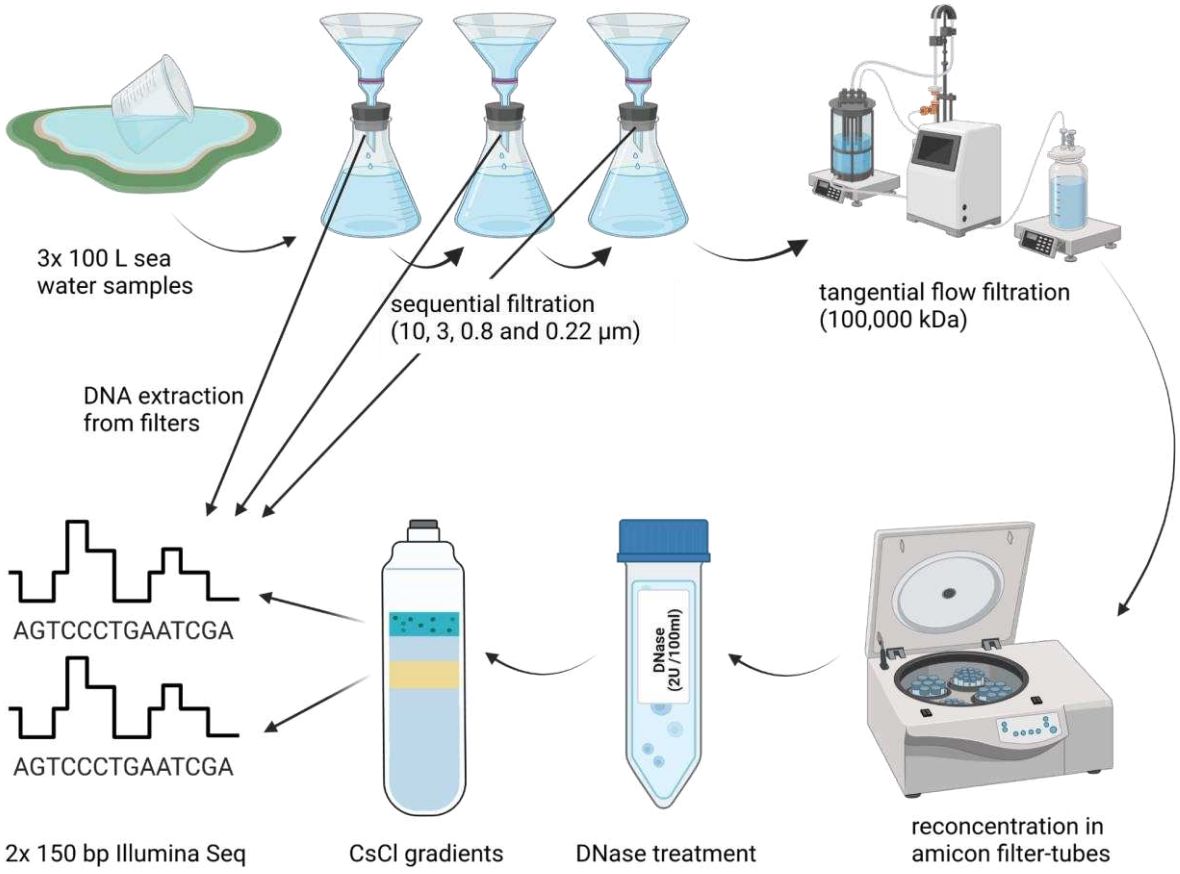


Figure S6: Overview of the sampling and purification workflow.

For details refer to the section “Methods - Sampling and Filtration”. In brief: Samples were sequentially filtered, concentrated using tangential flow filtration, further concentrated using amicon filter tubes, DNase treated, gradient purified using CsCl gradients. DNA from bands/fractions was extracted and sequenced.

DNA from the 3 µm, 0.8 µm and 0.22 µm filters was extracted and sequenced as well.

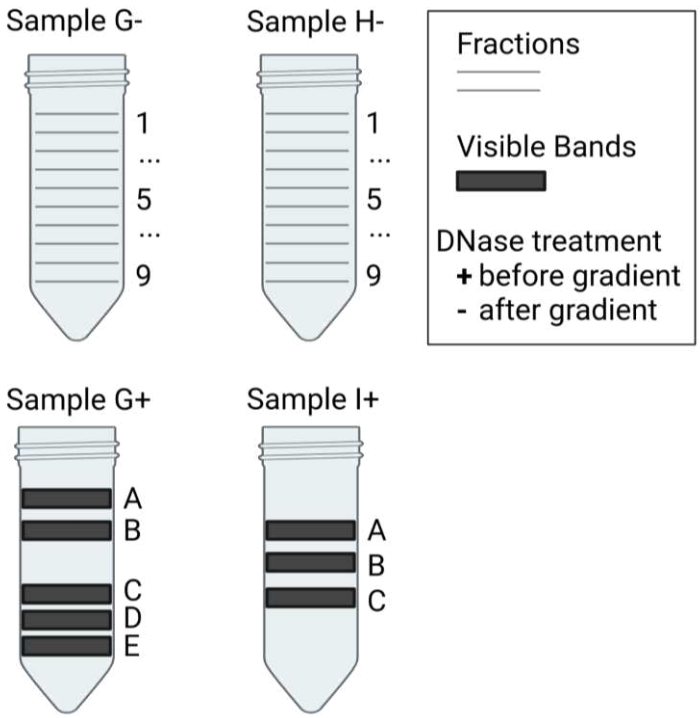


Figure S7: Overview of sequencing effort of CsCl-gradients.

Schematic overview of the samples used in this study. Samples marked with a + are treated with DNase, samples marked with a - are untreated. Sample names G, H and I refer to 3 different biological replicates (see Sampling and Filtration chapter). If bands were visible, bands were extracted and sequenced. If none were visible, fractions of 0.5 ml were extracted and sequenced.

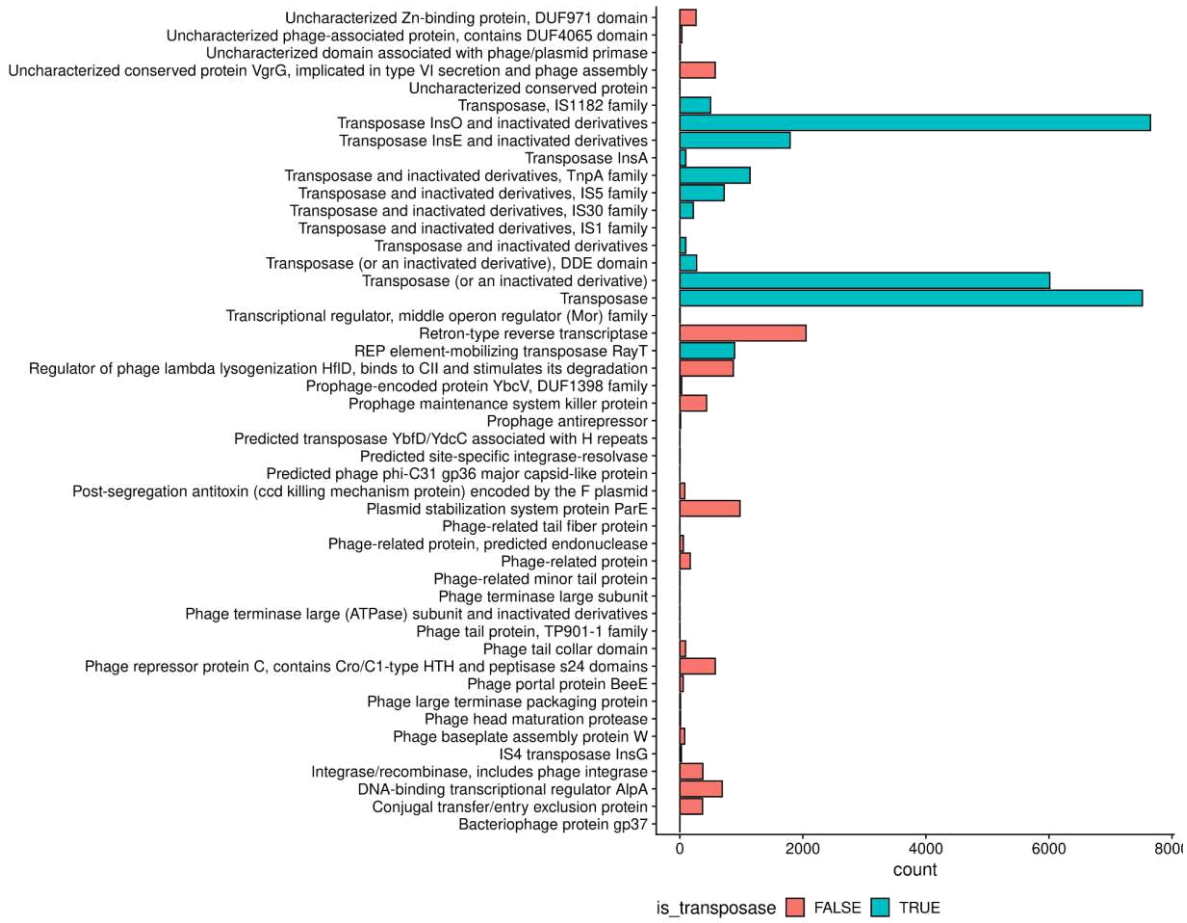


Figure S8: Detailed resolution of EV-mediated reads belonging to COG category X - Mobilome

Resolving the cluster affiliation of single ev-mediated peDNA reads. Clusters are colored cyan, if their cluster description contains the term “transposase”. Non-transposase clusters are colored red. This is based on 200 M reads of which ~12.7 M mapped to MAGs identified as EV-producing. For these, ~11 M protein fragments were predicted and blasted against *nr* with an e-value threshold of 10^{-5}, query coverage > 80%, subject coverage >10%, resulting in a total of 34826 assigned reads.

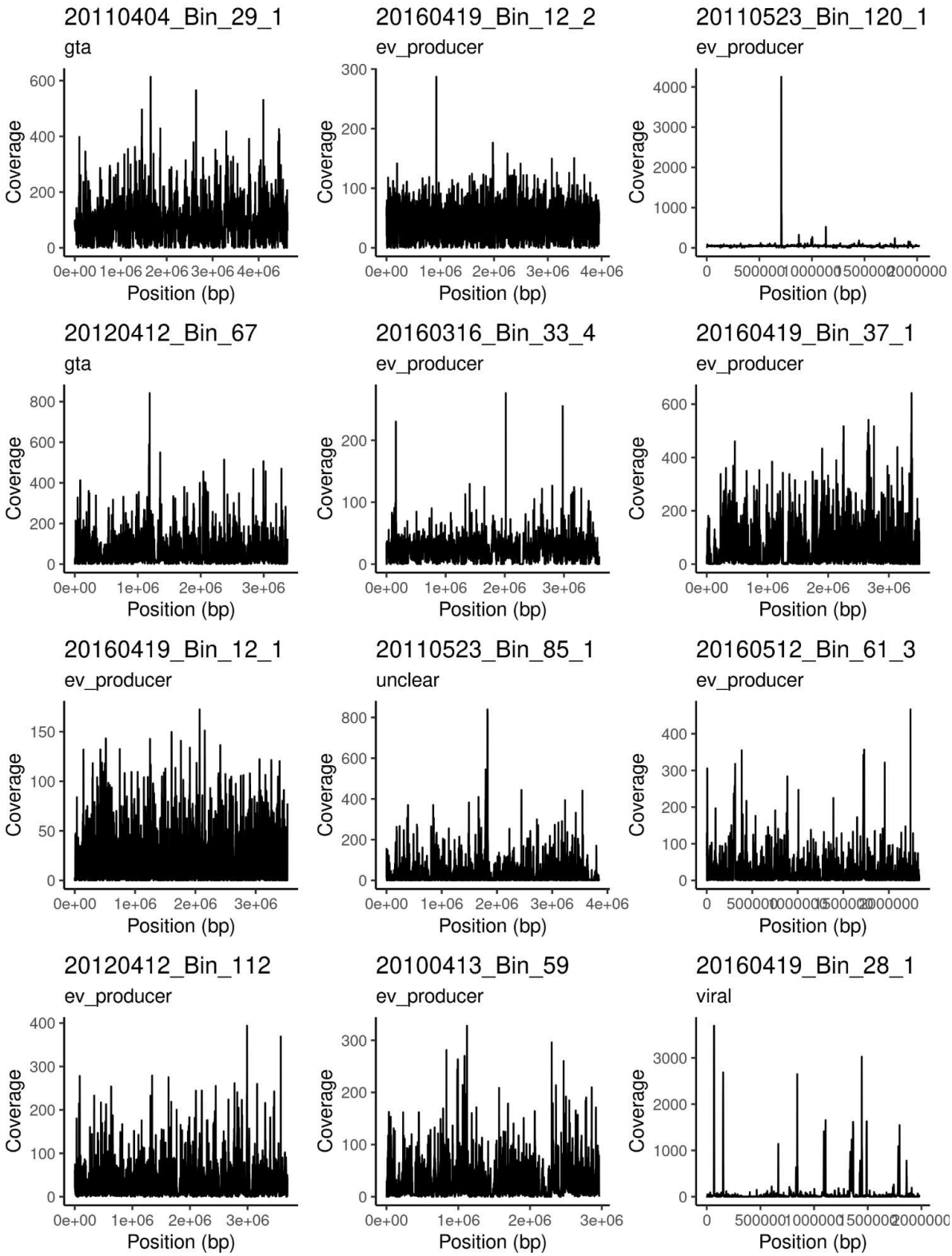
Table S1: File - Table S1.xlsx

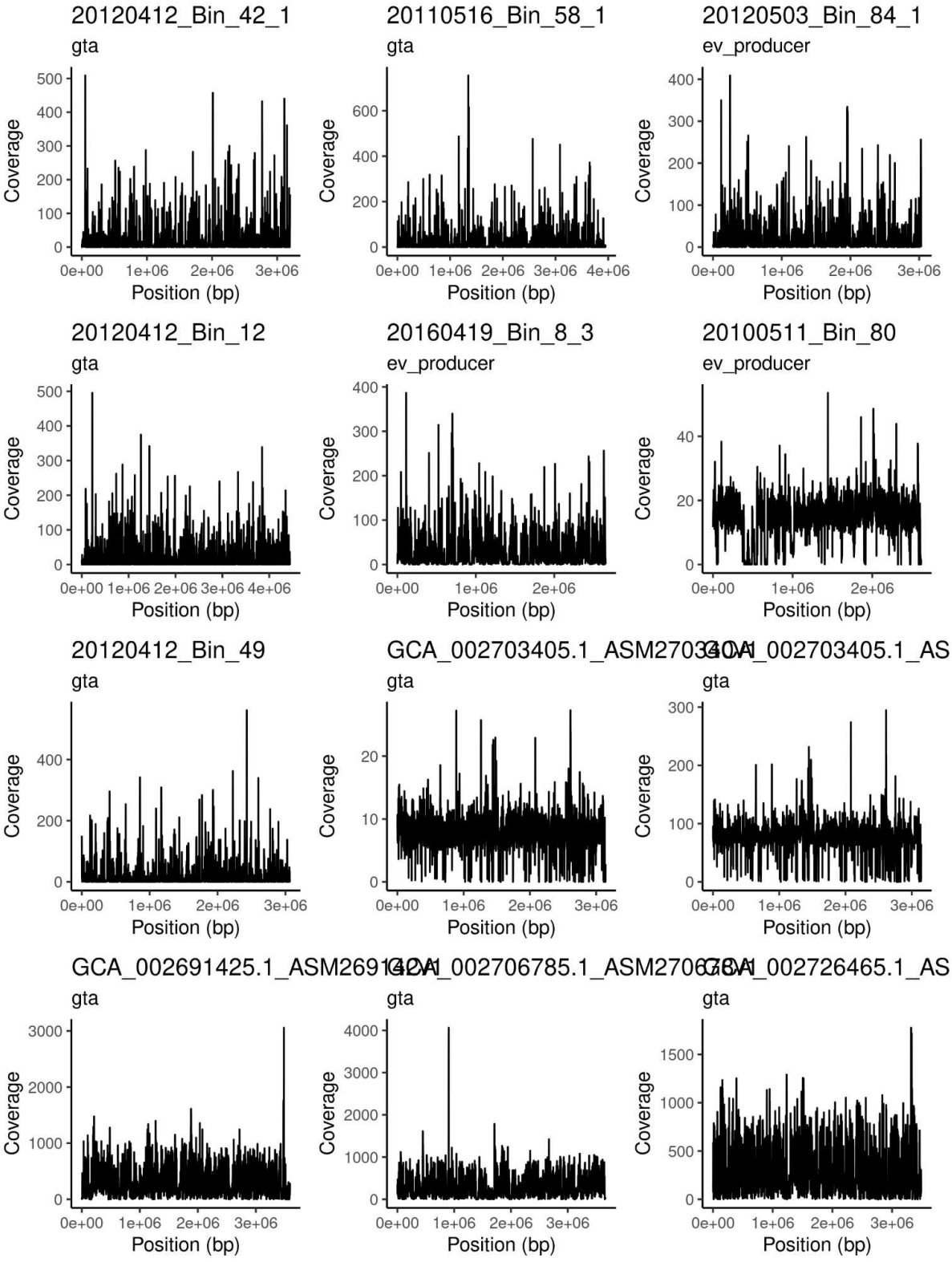
Overview of sheets given in Table S1:

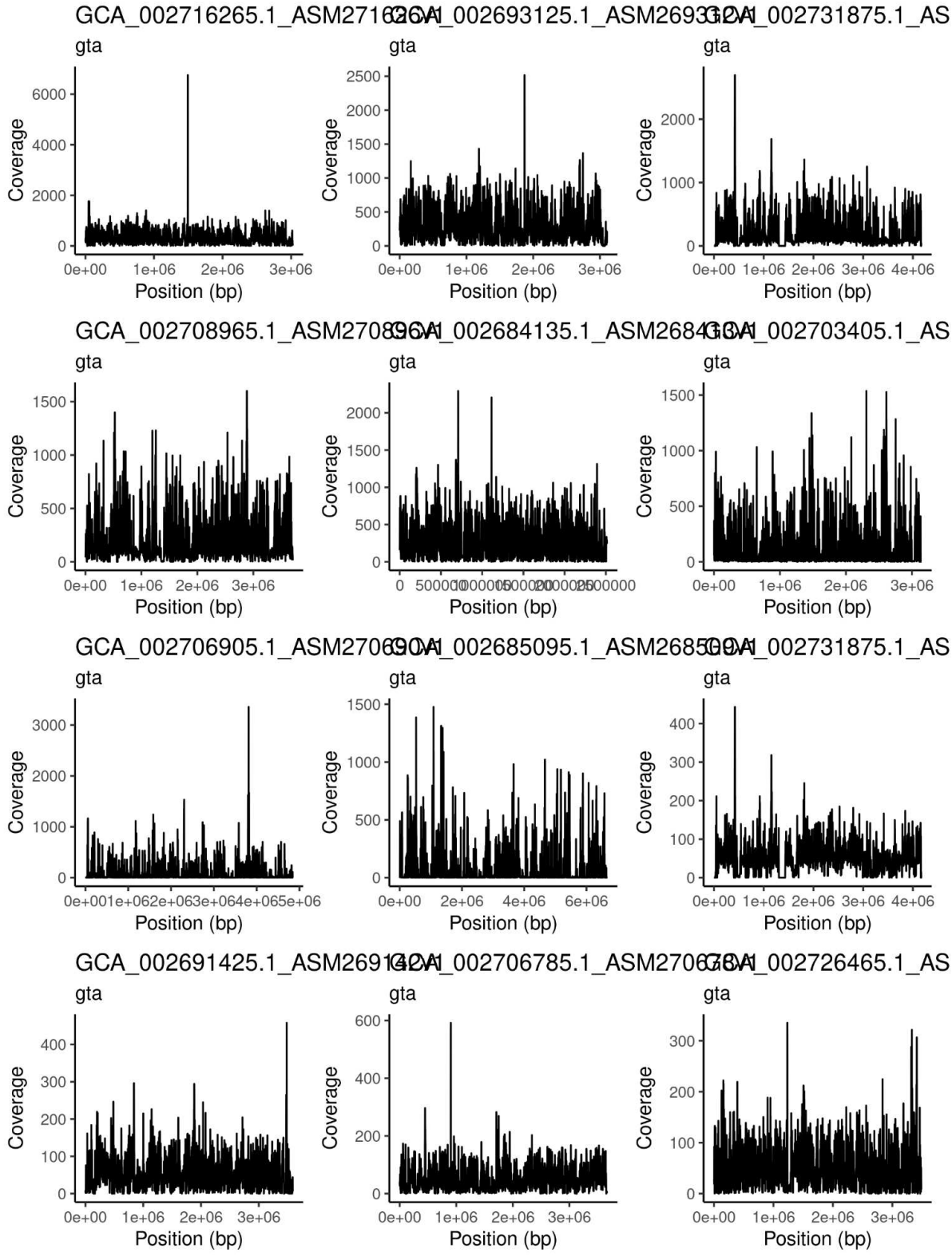
Sheet	**Description**
"Sample Overview"	Overview of all sea water samples analyzed. Information on sample processing and sequencing result
"peDNA/virome Datasets"	Run/Sample accessions for all 4 external peDNA/virome datasets used in this study
"Tully MAGs"	Overview of MAGs used in this study coming from Tully et al 2018
"Orellana MAGs"	Overview of MAGs used in this study coming from Orellana et al 2019
"top200 MAGs"	Summary of the top 200 peDNA-recruiting MAGs

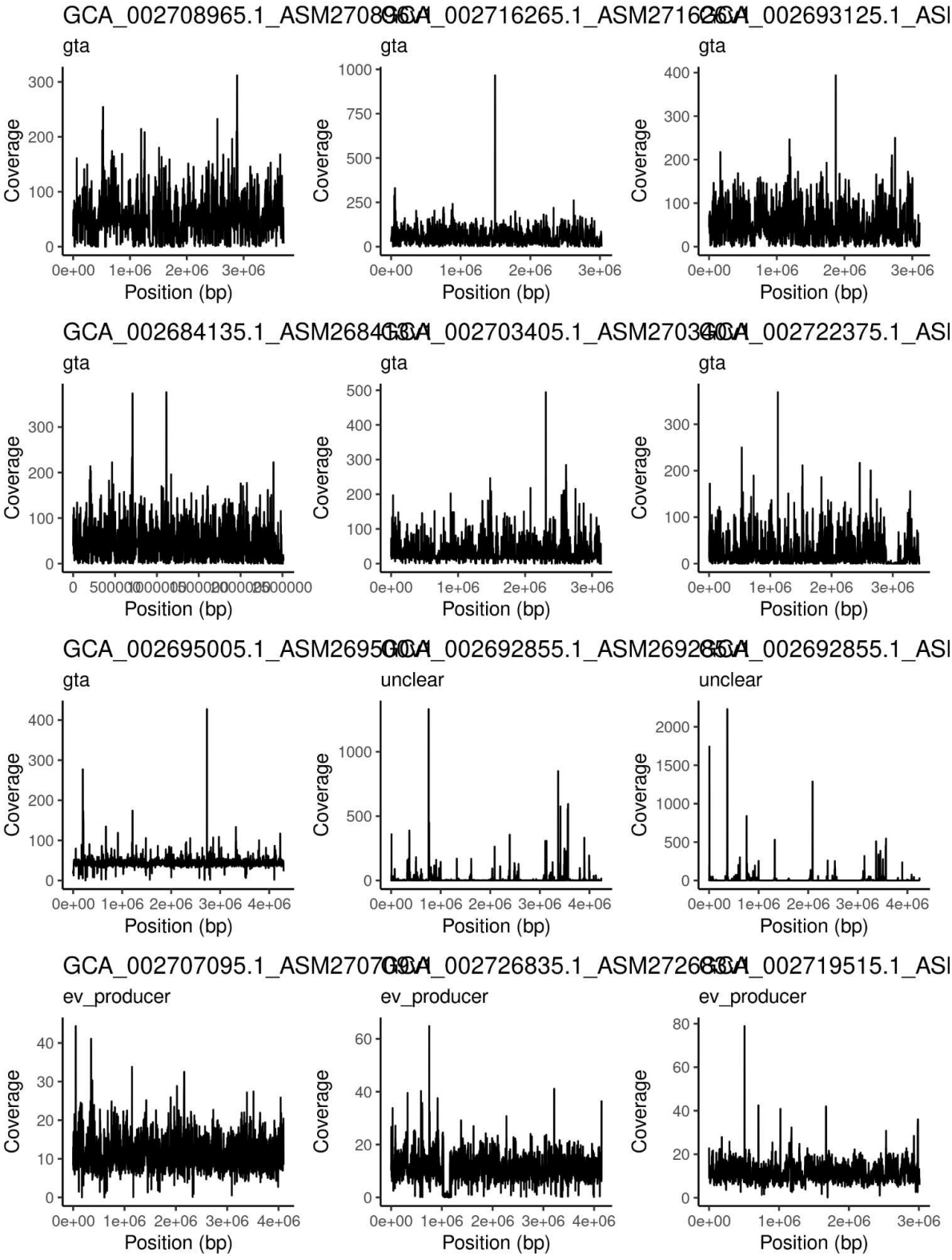
Figure S5 - Coverage Plots

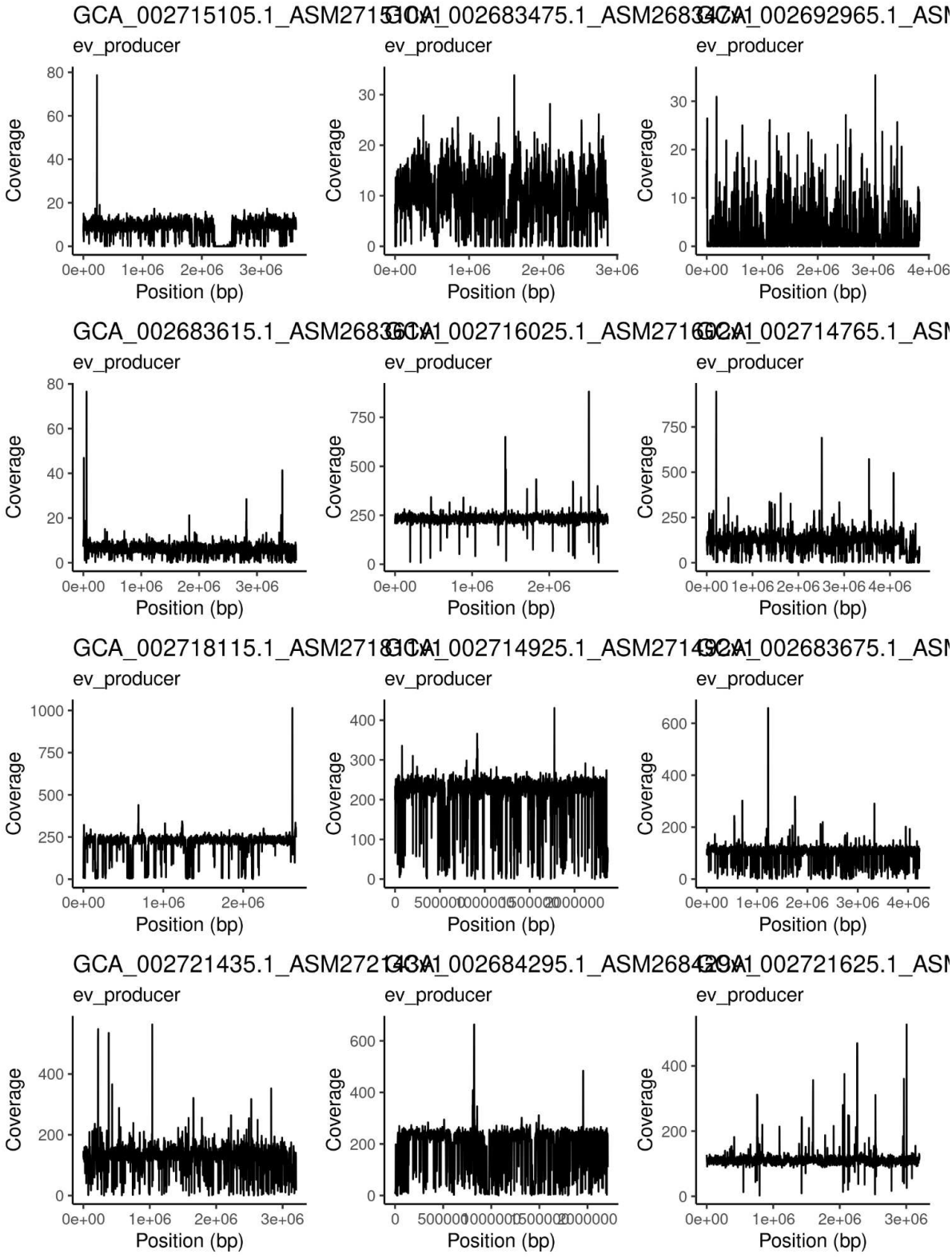
Binned (bin size = 1000 bp) coverage plots of all analyzed MAGs.

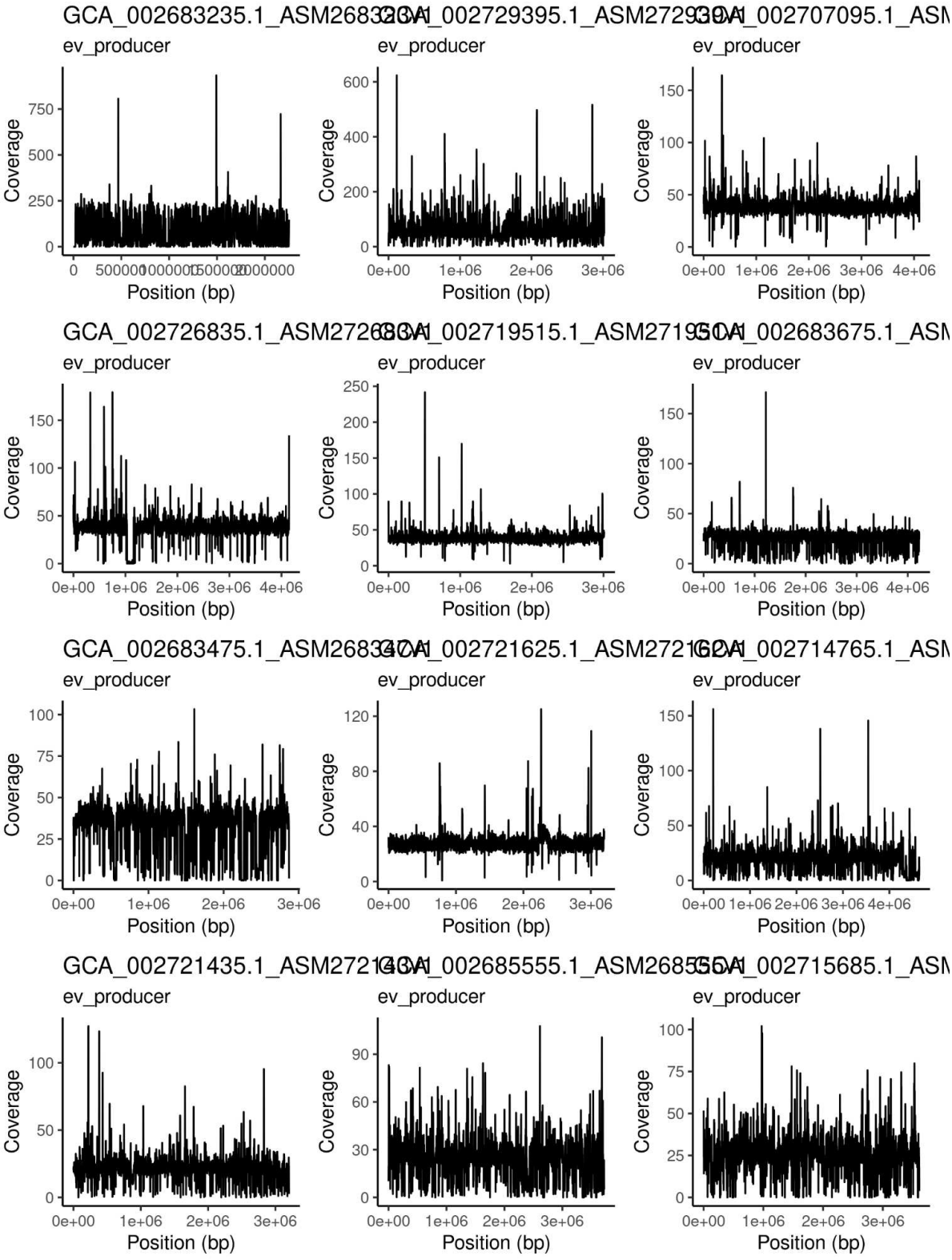

20110404_Bin_29_1
gta
20160419_Bin_12_2
ev_producer
20110523_Bin_120_1
ev_producer
20120412_Bin_67
gta
20160316_Bin_33_4
ev_producer
20160419_Bin_37_1
ev_producer
20160419_Bin_12_1
ev_producer
20110523_Bin_85_1
unclear
20160512_Bin_61_3
ev_producer
20120412_Bin_112
ev_producer
20100413_Bin_59
ev_producer
20160419_Bin_28_1
viral
Coverage
Position (bp)

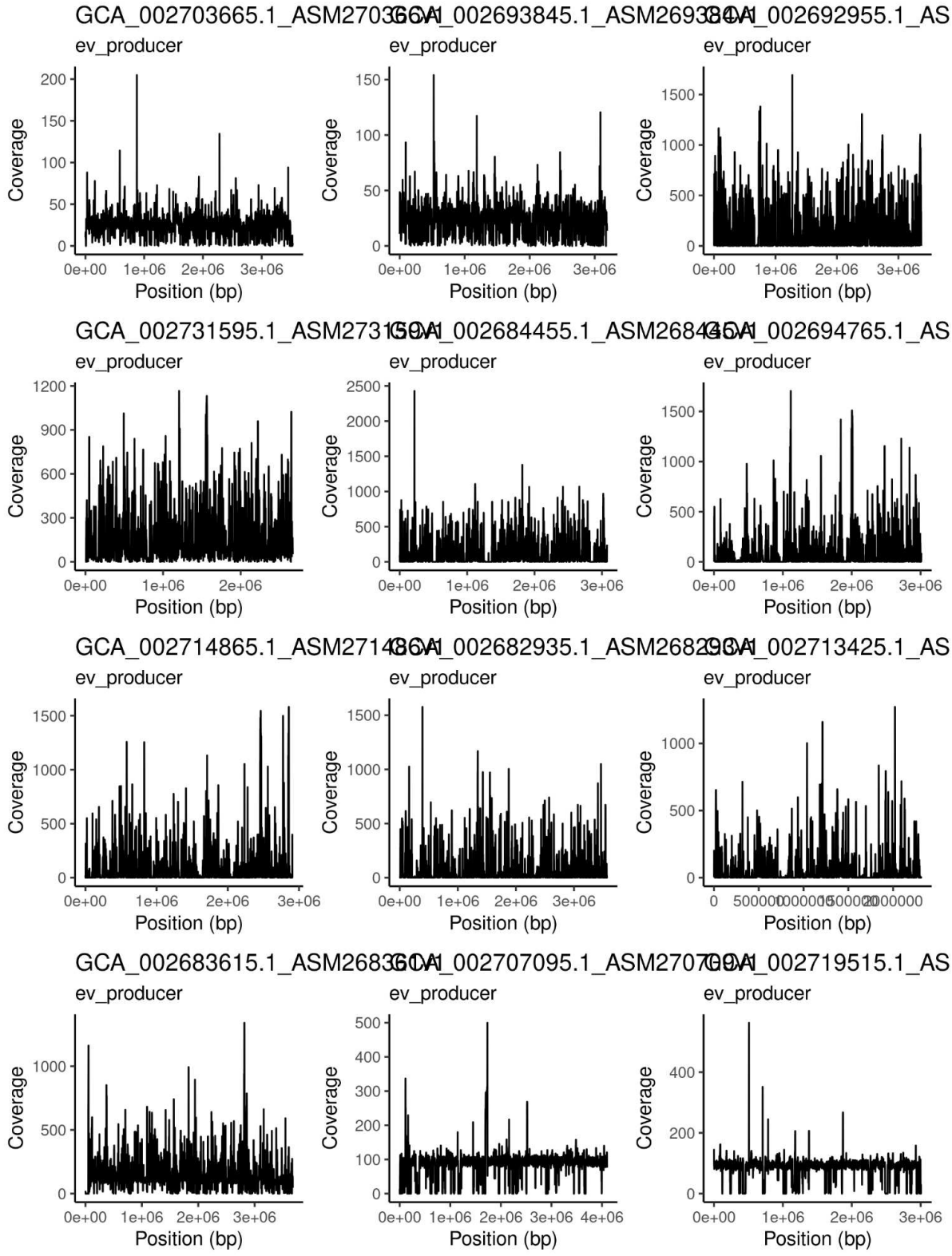

20120412_Bin_42_1
gta
20110516_Bin_58_1
gta
20120503_Bin_84_1
ev_producer
20120412_Bin_12
gta
20160419_Bin_8_3
ev_producer
20100511_Bin_80
ev_producer
20120412_Bin_49
gta
GCA_002703405.1_ASM270340v1_002703405.1_AS
gta
gta
GCA_002691425.1_ASM269142v1_002706785.1_ASM270678v1_002726465.1_AS
gta
gta
gta
Coverage
Position (bp)

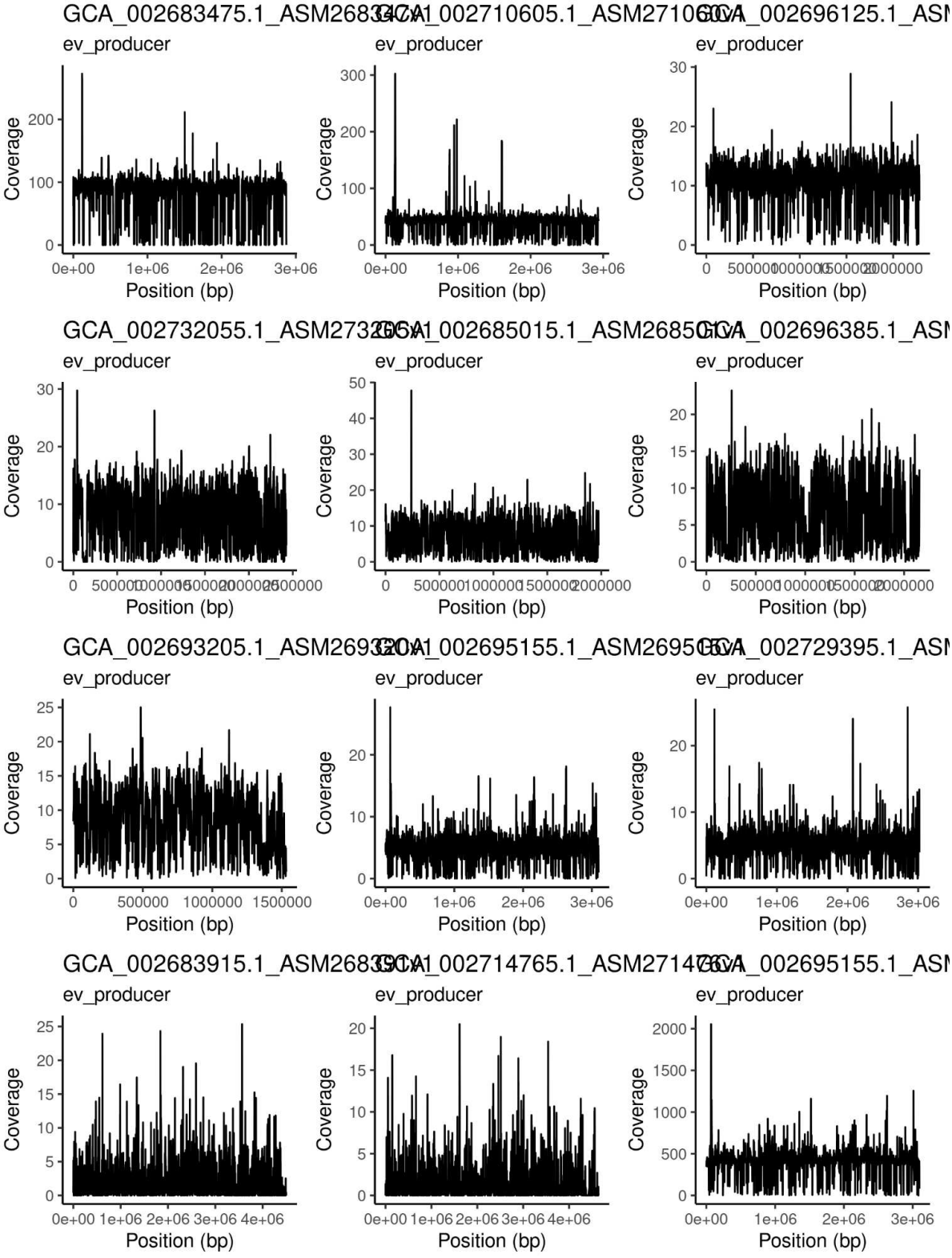

GCA_002716265.1_ASM271626v1_002693125.1_ASM269312v1_002731875.1_AS
gta
Coverage
Position (bp)
GCA_002708965.1_ASM270896v1_002684135.1_ASM268413v1_002703405.1_AS
gta
Coverage
Position (bp)
GCA_002706905.1_ASM270690v1_002685095.1_ASM268509v1_002731875.1_AS
gta
Coverage
Position (bp)
GCA_002691425.1_ASM269142v1_002706785.1_ASM270678v1_002726465.1_AS
gta
Coverage
Position (bp)

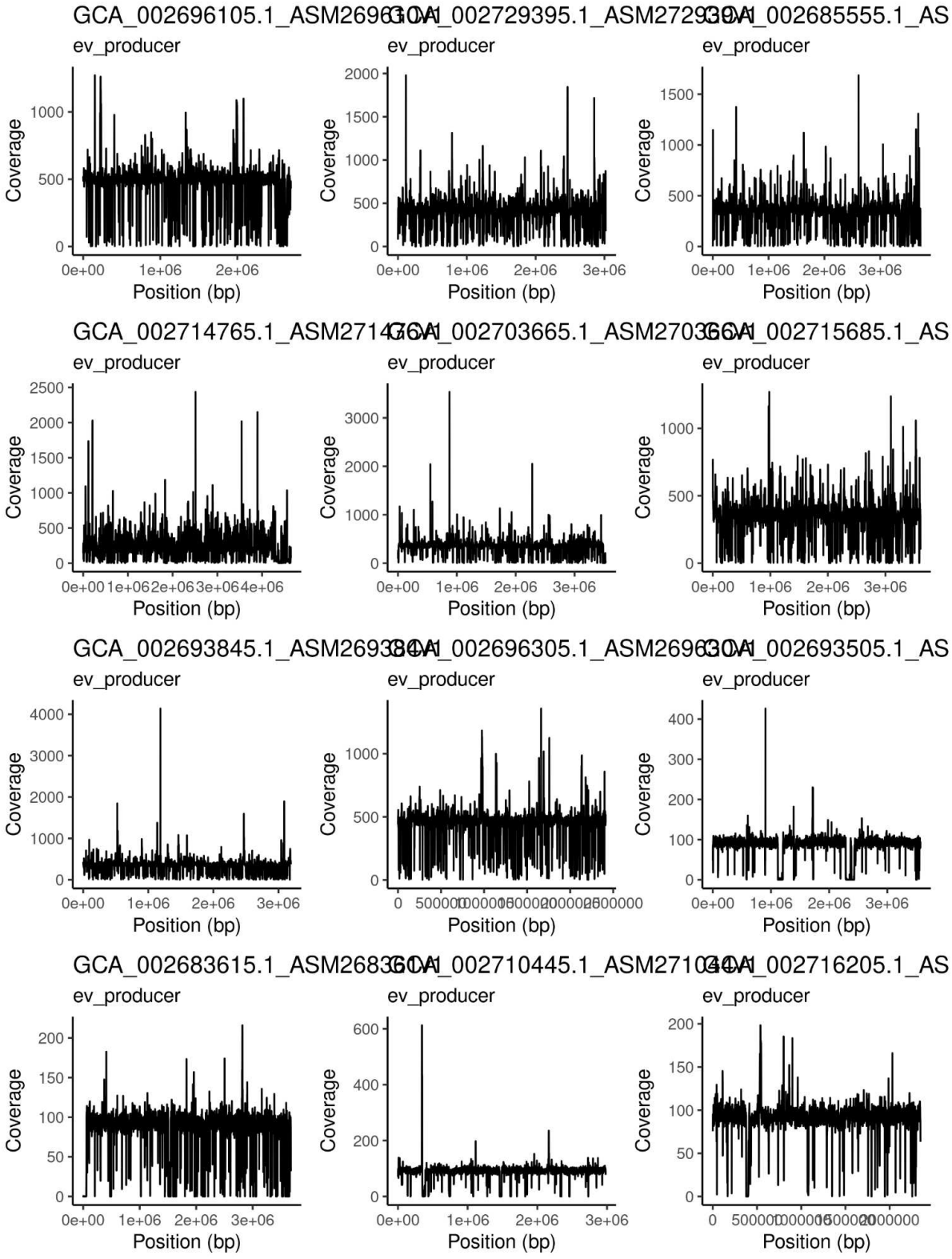

GCA_002708965.1_ASM2708965v1
gta
GCA_002716265.1_ASM2716265v1
gta
GCA_002693125.1_ASM2693125v1
gta
GCA_002684135.1_ASM2684135v1
gta
GCA_002703405.1_ASM2703405v1
gta
GCA_002722375.1_ASM2722375v1
gta
GCA_002695005.1_ASM2695005v1
gta
GCA_002692855.1_ASM2692855v1
unclear
GCA_002692855.1_ASM2692855v1
unclear
GCA_002707095.1_ASM2707095v1
ev_producer
GCA_002726835.1_ASM2726835v1
ev_producer
GCA_002719515.1_ASM2719515v1
ev_producer
Coverage
Position (bp)

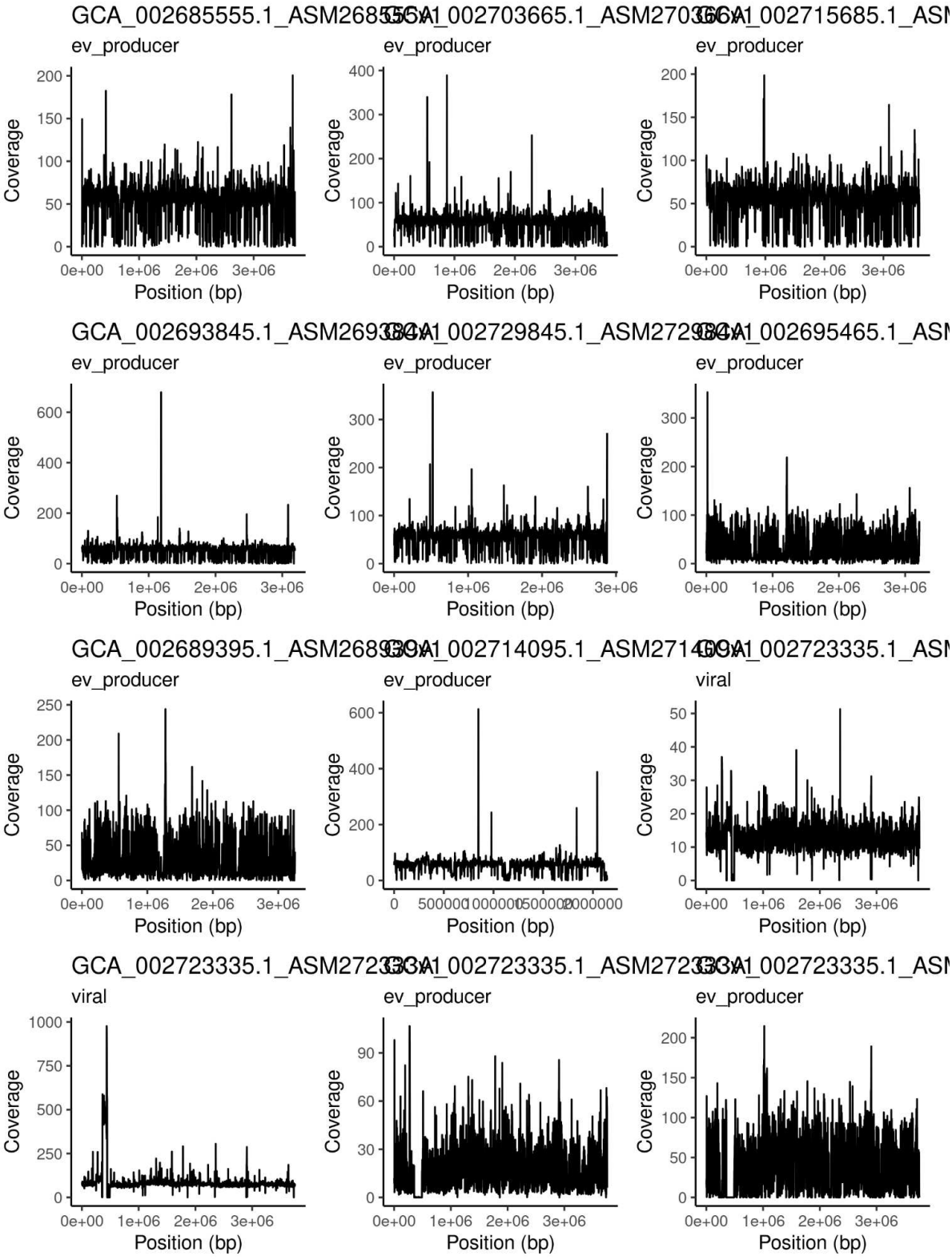

GCA_002715105.1_ASM271510v1
GCA_002683475.1_ASM268347v1
GCA_002692965.1_ASM
GCA_002683615.1_ASM268361v1
GCA_002716025.1_ASM271602v1
GCA_002714765.1_ASM
GCA_002718115.1_ASM271811v1
GCA_002714925.1_ASM271492v1
GCA_002683675.1_ASM
GCA_002721435.1_ASM272143v1
GCA_002684295.1_ASM268429v1
GCA_002721625.1_ASM
ev_producer
Coverage
Position (bp)

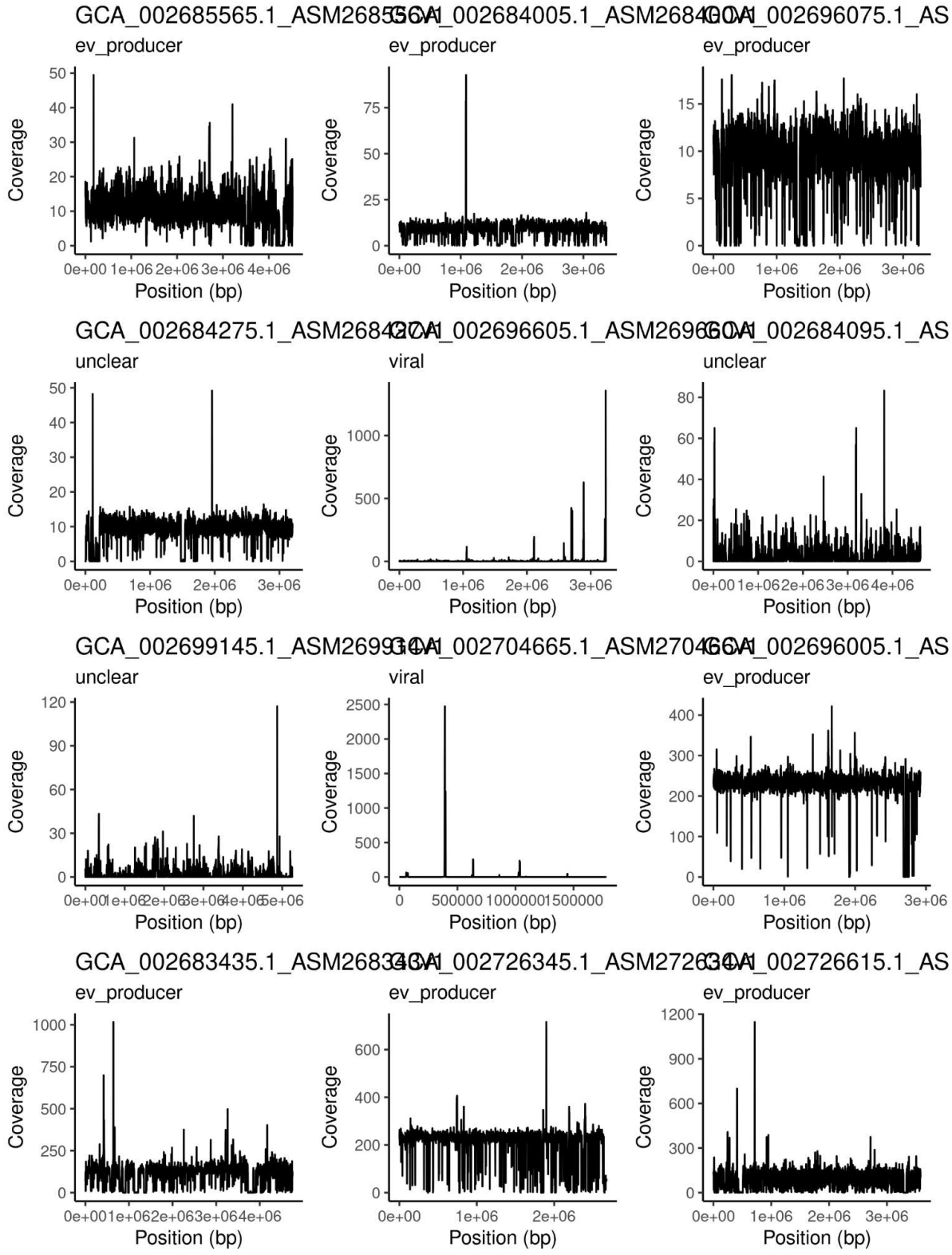

GCA_002683235.1_ASM268323v1
GCA_002729395.1_ASM272939v1
GCA_002707095.1_ASM
GCA_002726835.1_ASM272683v1
GCA_002719515.1_ASM271951v1
GCA_002683675.1_ASM
GCA_002683475.1_ASM268347v1
GCA_002721625.1_ASM272162v1
GCA_002714765.1_ASM
GCA_002721435.1_ASM272143v1
GCA_002685555.1_ASM268555v1
GCA_002715685.1_ASM
ev_producer
Coverage
Position (bp)

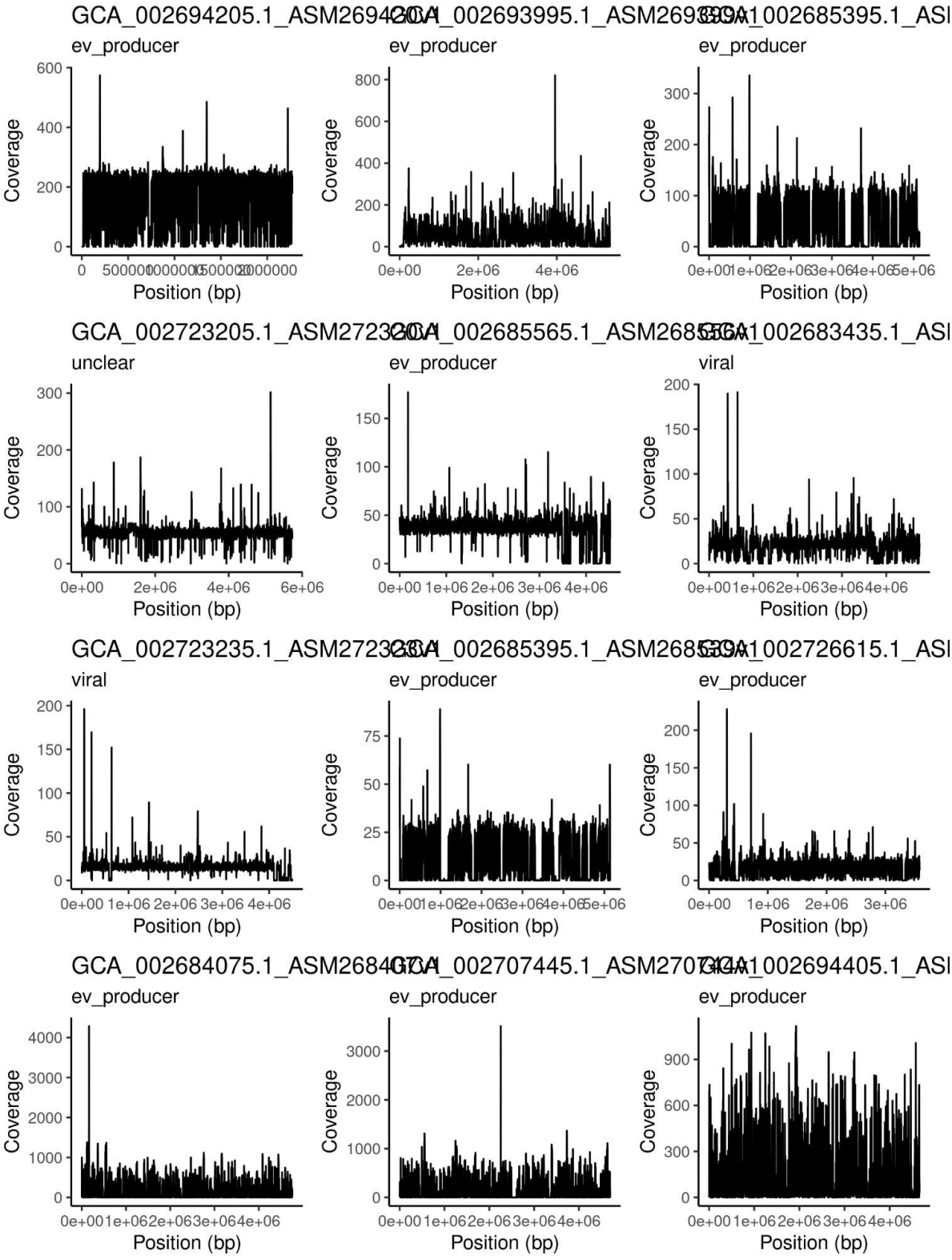

GCA_002703665.1_ASM270366v1
GCA_002693845.1_ASM269384v1
GCA_002692955.1_AS
GCA_002731595.1_ASM273159v1
GCA_002684455.1_ASM268445v1
GCA_002694765.1_AS
GCA_002714865.1_ASM271486v1
GCA_002682935.1_ASM268293v1
GCA_002713425.1_AS
GCA_002683615.1_ASM268361v1
GCA_002707095.1_ASM270709v1
GCA_002719515.1_AS
ev_producer
Coverage
Position (bp)

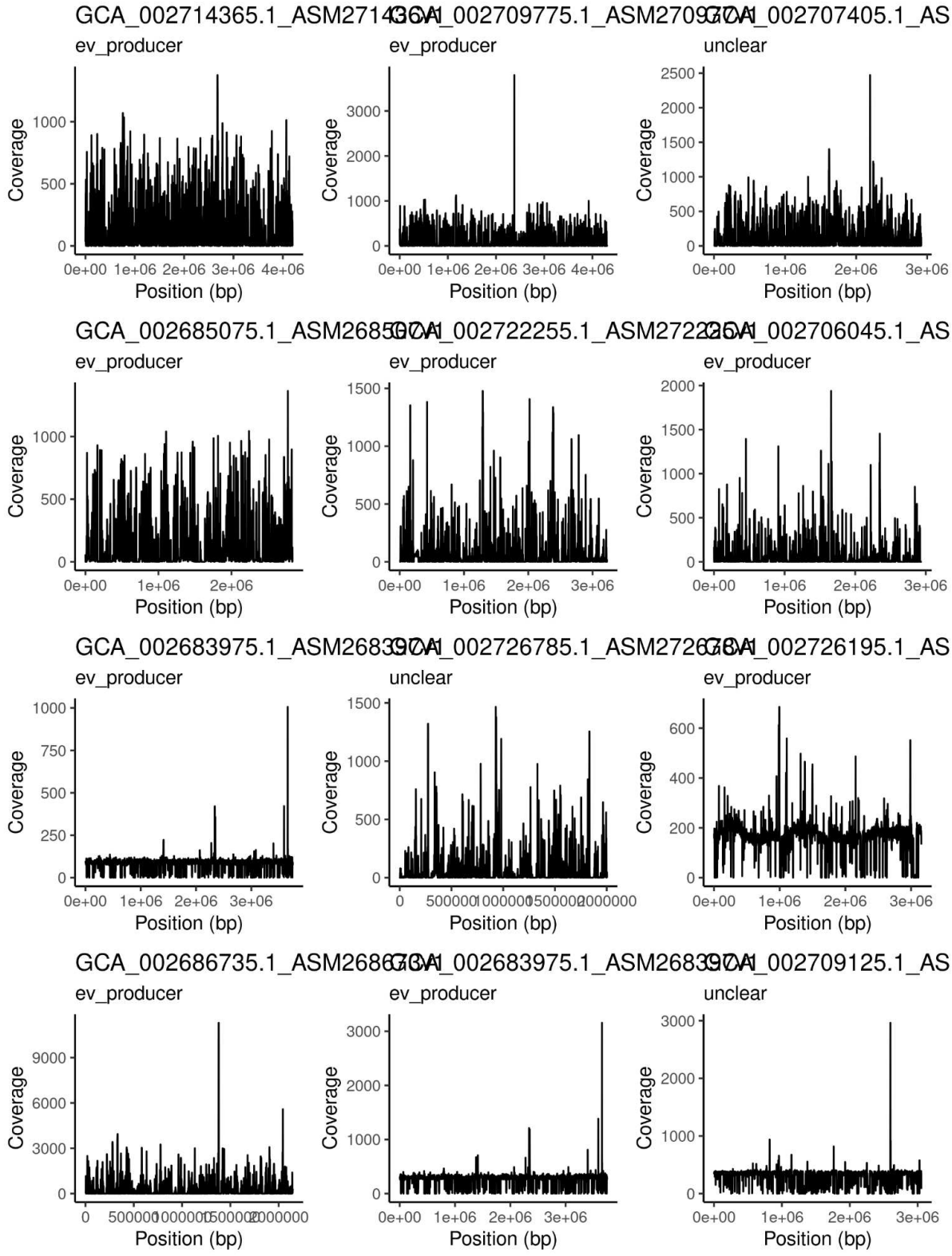

GCA_002683475.1_ASM268347v1
GCA_002710605.1_ASM271060v1
GCA_002696125.1_ASM
GCA_002732055.1_ASM273205v1
GCA_002685015.1_ASM268501v1
GCA_002696385.1_ASM
GCA_002693205.1_ASM269320v1
GCA_002695155.1_ASM269515v1
GCA_002729395.1_ASM
GCA_002683915.1_ASM268391v1
GCA_002714765.1_ASM271476v1
GCA_002695155.1_ASM
ev_producer
Coverage
Position (bp)

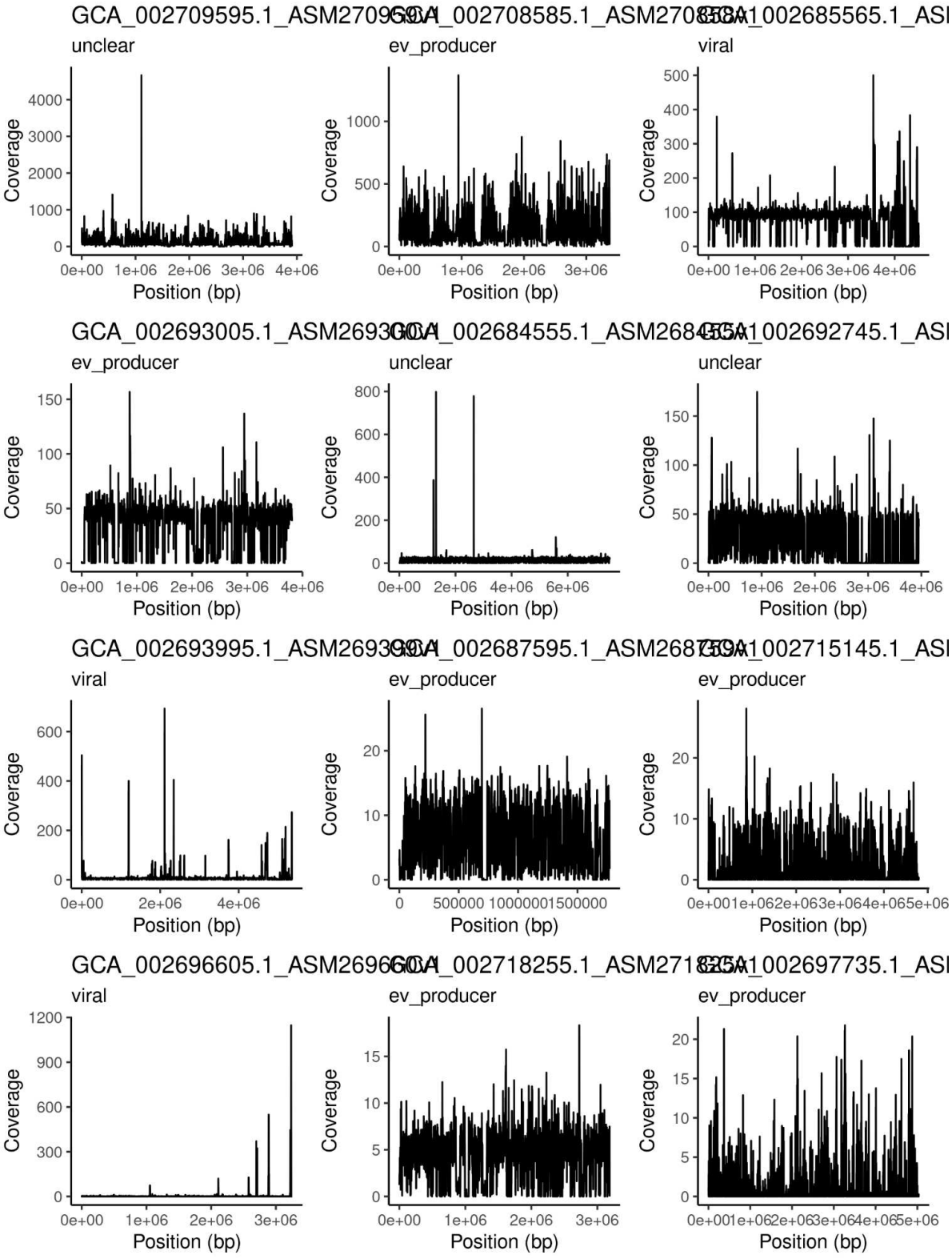

GCA_002696105.1_ASM269610v1_002729395.1_ASM272939v1_002685555.1_AS
ev_producer
Coverage
Position (bp)
GCA_002714765.1_ASM271476v1_002703665.1_ASM270366v1_002715685.1_AS
GCA_002693845.1_ASM269384v1_002696305.1_ASM269630v1_002693505.1_AS
GCA_002683615.1_ASM268361v1_002710445.1_ASM271044v1_002716205.1_AS

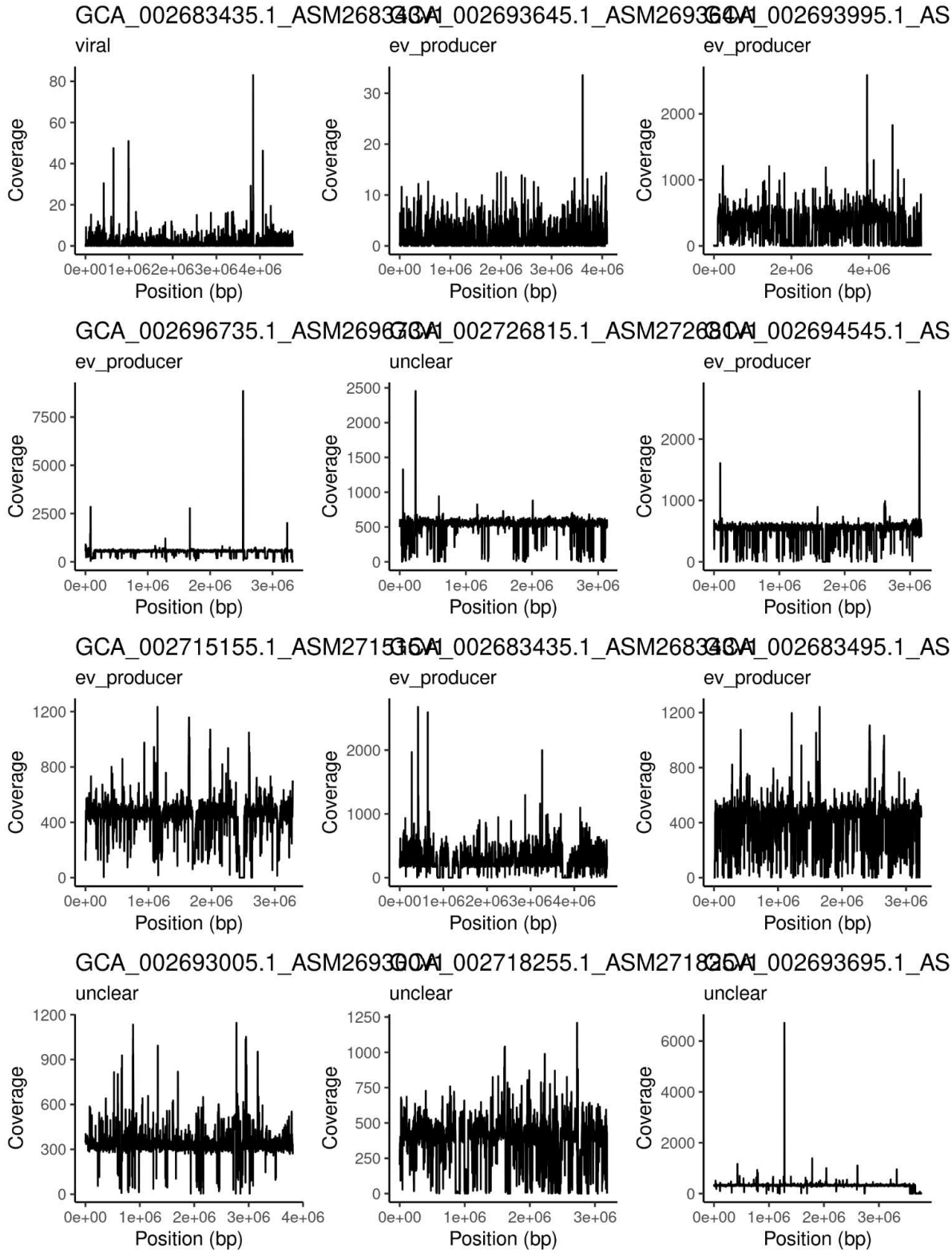

GCA_002685555.1_ASM268555v1_002703665.1_ASM270366v1_002715685.1_ASM
ev_producer
ev_producer
ev_producer
Coverage
Position (bp)
GCA_002693845.1_ASM269384v1_002729845.1_ASM272984v1_002695465.1_ASM
ev_producer
ev_producer
ev_producer
GCA_002689395.1_ASM268939v1_002714095.1_ASM271409v1_002723335.1_ASM
ev_producer
ev_producer
viral
GCA_002723335.1_ASM272333v1_002723335.1_ASM272333v1_002723335.1_ASM
viral
ev_producer
ev_producer


GCA_002685565.1_ASM268556v1_002684005.1_ASM268400v1_002696075.1_AS
ev_producer
ev_producer
ev_producer
GCA_002684275.1_ASM268427v1_002696605.1_ASM269660v1_002684095.1_AS
unclear
viral
unclear
GCA_002699145.1_ASM269914v1_002704665.1_ASM270466v1_002696005.1_AS
unclear
viral
ev_producer
GCA_002683435.1_ASM268343v1_002726345.1_ASM272634v1_002726615.1_AS
ev_producer
ev_producer
ev_producer
Coverage
Position (bp)


GCA_002694205.1_ASM269420v1
ev_producer
Coverage
Position (bp)
GCA_002693995.1_ASM269399v1
ev_producer
Coverage
Position (bp)
GCA_002685395.1_ASM268539v1
ev_producer
Coverage
Position (bp)
GCA_002723205.1_ASM272320v1
unclear
Coverage
Position (bp)
GCA_002685565.1_ASM268556v1
ev_producer
Coverage
Position (bp)
GCA_002683435.1_ASM268343v1
viral
Coverage
Position (bp)
GCA_002723235.1_ASM272323v1
viral
Coverage
Position (bp)
GCA_002685395.1_ASM268539v1
ev_producer
Coverage
Position (bp)
GCA_002726615.1_ASM272661v1
ev_producer
Coverage
Position (bp)
GCA_002684075.1_ASM268407v1
ev_producer
Coverage
Position (bp)
GCA_002707445.1_ASM270744v1
ev_producer
Coverage
Position (bp)
GCA_002694405.1_ASM269440v1
ev_producer
Coverage
Position (bp)


GCA_002714365.1_ASM271436v1_002709775.1_ASM270977v1_002707405.1_AS
ev_producer
ev_producer
unclear
GCA_002685075.1_ASM268507v1_002722255.1_ASM272225v1_002706045.1_AS
ev_producer
ev_producer
ev_producer
GCA_002683975.1_ASM268397v1_002726785.1_ASM272678v1_002726195.1_AS
ev_producer
unclear
ev_producer
GCA_002686735.1_ASM268673v1_002683975.1_ASM268397v1_002709125.1_AS
ev_producer
ev_producer
unclear
Coverage
Position (bp)


GCA_002709595.1_ASM270959v1
unclear
GCA_002708585.1_ASM270858v1
ev_producer
GCA_002685565.1_ASM268556v1
viral
GCA_002693005.1_ASM269300v1
ev_producer
GCA_002684555.1_ASM268455v1
unclear
GCA_002692745.1_ASM269274v1
unclear
GCA_002693995.1_ASM269399v1
viral
GCA_002687595.1_ASM268759v1
ev_producer
GCA_002715145.1_ASM271514v1
ev_producer
GCA_002696605.1_ASM269660v1
viral
GCA_002718255.1_ASM271825v1
ev_producer
GCA_002697735.1_ASM269773v1
ev_producer
Coverage
Position (bp)


GCA_002683435.1_ASM268343v1
viral
GCA_002693645.1_ASM269364v1
ev_producer
GCA_002693995.1_AS
ev_producer
GCA_002696735.1_ASM269673v1
ev_producer
GCA_002726815.1_ASM272681v1
unclear
GCA_002694545.1_AS
ev_producer
GCA_002715155.1_ASM271515v1
ev_producer
GCA_002683435.1_ASM268343v1
ev_producer
GCA_002683495.1_AS
ev_producer
GCA_002693005.1_ASM269300v1
unclear
GCA_002718255.1_ASM271825v1
unclear
GCA_002693695.1_AS
unclear
Coverage
Position (bp)

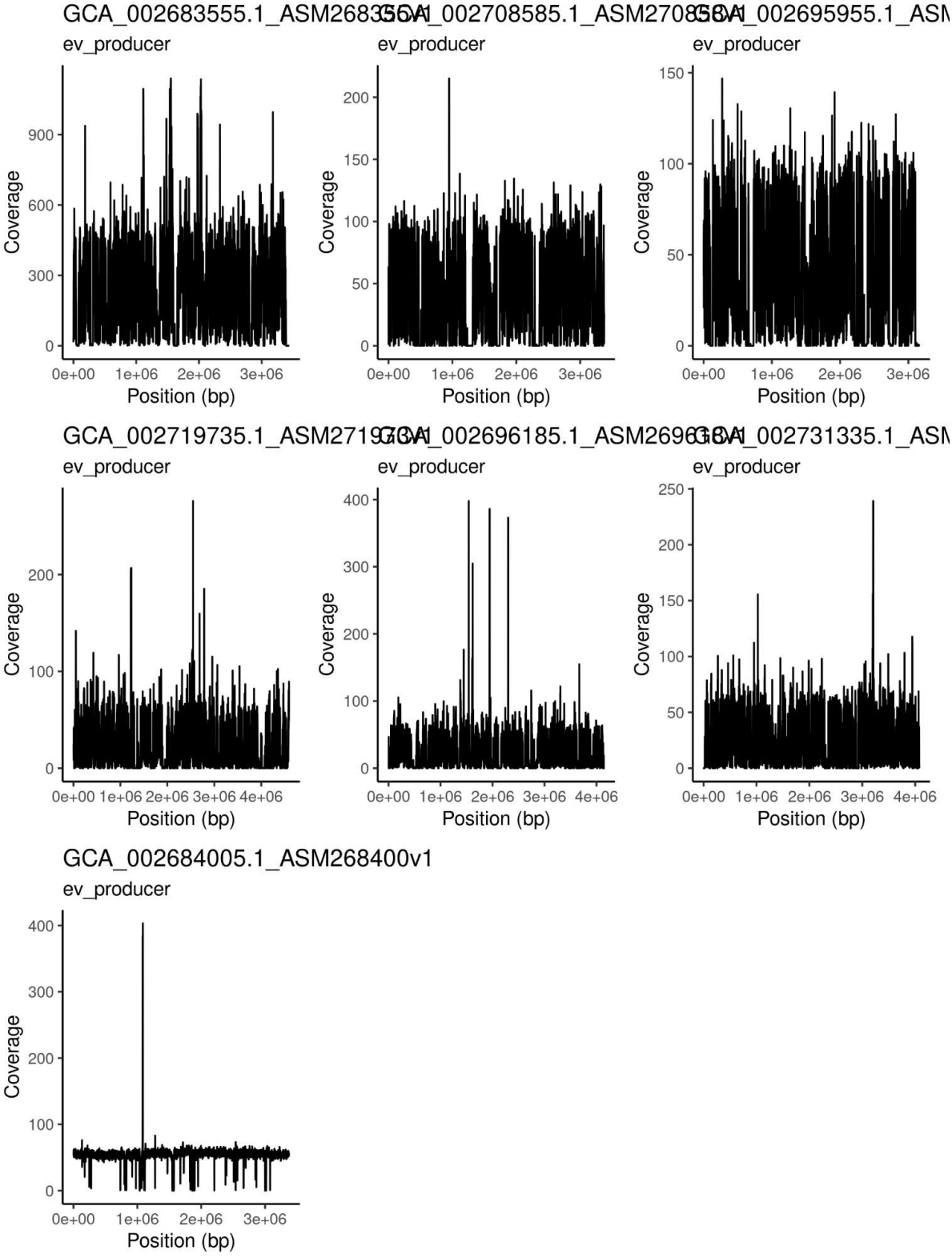

GCA_002683555.1_ASM268355v1
GCA_002708585.1_ASM270858v1
GCA_002695955.1_ASM
ev_producer
Coverage
Position (bp)
GCA_002719735.1_ASM271973v1
GCA_002696185.1_ASM269618v1
GCA_002731335.1_ASM
GCA_002684005.1_ASM268400v1

Table S1 - Sample Overview

Original Name	Sample Accessi	BioProject Acce	Sequencing Lat	Sample Descrip	Sampling Date	DNase treatment	Bands visible	Volume (ml)	DNA Concentra	Sequencing	No of Reads
H-4	ERS14766242	PRJEB60526	4976_D	sample H, CsCl f	10.06.23	after gradient	no	1	6,5	50 µl	10467617
H-7	ERS14766243	PRJEB60526	4976_G	sample H, CsCl f	10.06.23	after gradient	no	1	66	40 µl	9308451
G-3	ERS14766244	PRJEB60526	4976_P	sample G, CsCl f	09.06.2023	after gradient	no	1	5,3	pooled with GI-4,	9899585
G-5	ERS14766245	PRJEB60526	4976_Q	sample G, CsCl f	09.06.2023	after gradient	no	1	66,7	40 µl	9283264
G-7	ERS14766246	PRJEB60526	4976_S	sample G, CsCl f	09.06.2023	after gradient	no	1	2,68	pooled with GI-8,	97366701
G-9	ERS14766318	PRJEB60526	4976_T	sample G, CsCl f	09.06.2023	after gradient	no	1	34,679	40 µl	10522933
G+A	ERS14766238	PRJEB60526	4976_AA	sample G, CsCl b	09.06.2023	before gradient	yes	0,5	1,8	pooled with GI+E	9420774
G+D	ERS14766239	PRJEB60526	4976_AB	sample G, CsCl b	09.06.2023	before gradient	yes	0,5	3,6	pooled with GI+E	9696592
I+A	ERS14766240	PRJEB60526	4976_AD	sample I, CsCl ba	11.06.2023	before gradient	yes	1	2	pooled with II+B,	9007671
I+C	ERS14766241	PRJEB60526	4976_AE	sample I, CsCl ba	11.06.2023	before gradient	yes	1	4,5	50 µl	9622684
H_022	ERS14766237	PRJEB60526	4903_L	sample H, fractio	10.06.23	no	not applicable	not applicable	50,5	30 µl	10555905
H_045	ERS14766236	PRJEB60526	4903_K	sample H, fractio	10.06.23	no	not applicable	not applicable	211,7	15 µl	11175788
H_080	ERS14766235	PRJEB60526	4903_J	sample H, fractio	10.06.23	no	not applicable	not applicable	67,8	30 µl	11193342
H_300	ERS14766184	PRJEB60526	4903_I	sample H, fractio	10.06.23	no	not applicable	not applicable	43,2	30 µl	11138263

Table S1 - top200 MAGs

shortname	final_label	sample	viral	no_of_GTA_hits	gta	virus_active	length	genus	Bases	Coverage	Reads	Dataset
20110404_Bin_29_1	gta	heligoland	TRUE	17	TRUE	TRUE	4.628.818	g__Sulfitobacter	252.163.885	544.769	1717774	Heligoland
20160419_Bin_12_2	ev_producer	heligoland	FALSE	0	FALSE	FALSE	3.954.532	g__Balneola	152.198.153	38.487	886257	Heligoland
20110523_Bin_120_1	ev_producer	heligoland	FALSE	0	FALSE	FALSE	2.022.917	g__MAG-120531	53.953.240	26.671	330502	Heligoland
20120412_Bin_67	gta	heligoland	TRUE	15	TRUE	TRUE	3377690	g__Sulfitobacter	49.121.781	14.543	332479	Heligoland
20160316_Bin_33_4	ev_producer	heligoland	FALSE	1	FALSE	FALSE	3595865	g__Alcanivorax	52217634	145.216	273300	Heligoland
20160419_Bin_37_1	ev_producer	heligoland	TRUE	4	FALSE	TRUE	3504045	g__Sulfitobacter	36386099	10.384	246991	Heligoland
20160419_Bin_12_1	ev_producer	heligoland	TRUE	0	FALSE	TRUE	3527269	g__Balneola	29322804	8.3132	173910	Heligoland
20110523_Bin_85_1	unclear	heligoland	FALSE	6	TRUE	FALSE	3850895	g__Sulfitobacter	29247339	7.5949	204278	Heligoland
20160512_Bin_61_3	ev_producer	heligoland	FALSE	0	FALSE	FALSE	2336703	g__Winogradskyella	11117258	4.7577	64549	Heligoland
20120412_Bin_112	ev_producer	heligoland	TRUE	0	FALSE	TRUE	3693838	g__Polaribacter	16289127	4.4098	95695	Heligoland
20100413_Bin_59	ev_producer	heligoland	FALSE	0	FALSE	FALSE	2963208	g__Polaribacter	12298406	4.1504	74795	Heligoland
20160419_Bin_28_1	viral	heligoland	TRUE	0	FALSE	TRUE	1981290	g__SCGC-AAA160-P02	7858251	3.9662	46598	Heligoland
20120412_Bin_42_1	gta	heligoland	FALSE	13	TRUE	FALSE	3196934	g__Roseovarius	11549896	3.6128	82913	Heligoland
20110516_Bin_58_1	gta	heligoland	TRUE	16	TRUE	TRUE	3950844	g__Tateyamaria	13640860	3.4526	96430	Heligoland
20120503_Bin_84_1	ev_producer	heligoland	FALSE	0	FALSE	FALSE	3029998	g__Algibacter_B	9704007	3.2026	55999	Heligoland
20120412_Bin_12	gta	heligoland	FALSE	13	TRUE	FALSE	4442588	g__Sulfitobacter	14173169	3.1903	102532	Heligoland
20160419_Bin_8_3	ev_producer	heligoland	FALSE	0	FALSE	FALSE	2654212	g__Polaribacter	8253099	3.1094	48382	Heligoland
20100511_Bin_80	ev_producer	heligoland	TRUE	0	FALSE	TRUE	2614996	g__BACL24	8130579	3.1092	51002	Heligoland
20110526_Bin_113_1	remove	heligoland	FALSE	0	FALSE	FALSE	1931754	g__UBA3031	5152208	2.6671	30586	Heligoland
20120412_Bin_49	gta	heligoland	TRUE	15	TRUE	TRUE	3066456	g__Flavimaricola	6981749	2.2768	51910	Heligoland
GCA_002703405.1_ASM270340v1	gta	122_DCM	TRUE	15	TRUE	TRUE	3142791	Sulfitobacter sp.	3586162	7,7623	40100	Tara Oceans
GCA_002703405.1_ASM270340v1	gta	122_MES	TRUE	15	TRUE	TRUE	3142791	Sulfitobacter sp.	9943040	75,2274	105604	Tara Oceans
GCA_002691425.1_ASM269142v1	gta	158_MES	TRUE	14	TRUE	TRUE	3587593	Roseobacter sp.	8793107	316,541	90496	Tara Oceans
GCA_002706785.1_ASM270678v1	gta	158_MES	TRUE	15	TRUE	TRUE	3666882	Roseobacter sp.	8770061	301,4187	90278	Tara Oceans
GCA_002726465.1_ASM272646v1	gta	158_MES	FALSE	15	TRUE	TRUE	3487762	Roseobacter sp.	8308982	300,2197	85478	Tara Oceans
GCA_002716265.1_ASM271626v1	gta	158_MES	FALSE	14	TRUE	TRUE	3023545	Roseobacter sp.	7784054	324,2048	80096	Tara Oceans
GCA_002693125.1_ASM269312v1	gta	158_MES	FALSE	15	TRUE	TRUE	3106942	Roseobacter sp.	7389115	297,722	76058	Tara Oceans
GCA_002731875.1_ASM273187v1	gta	158_MES	TRUE	14	TRUE	TRUE	4169734	Roseobacter sp.	6467656	187,9895	66412	Tara Oceans
GCA_002708965.1_ASM270896v1	gta	158_MES	TRUE	12	TRUE	TRUE	3704350	Roseobacter sp.	5808913	188,1465	59644	Tara Oceans
GCA_002684135.1_ASM268413v1	gta	158_MES	FALSE	14	TRUE	TRUE	2514760	Roseobacter sp.	5740673	285,9807	59078	Tara Oceans
GCA_002703405.1_ASM270340v1	gta	158_MES	TRUE	15	TRUE	TRUE	3142791	Sulfitobacter sp.	4074316	112,3462	41944	Tara Oceans
GCA_002706905.1_ASM270690v1	gta	158_MES	TRUE	14	TRUE	TRUE	4842610	Pseudooceanicola sp.	3945640	63,3259	40644	Tara Oceans
GCA_002685095.1_ASM268509v1	gta	158_MES	TRUE	15	TRUE	TRUE	6660740	Rhodobacteraceae bacterium	3833448	43,7798	39642	Tara Oceans
GCA_002731875.1_ASM273187v1	gta	158_SRF	TRUE	14	TRUE	TRUE	4169734	Roseobacter sp.	8728021	51,308	89110	Tara Oceans
GCA_002691425.1_ASM269142v1	gta	158_SRF	TRUE	14	TRUE	TRUE	3587593	Roseobacter sp.	8476340	53,3775	86656	Tara Oceans
GCA_002706785.1_ASM270678v1	gta	158_SRF	TRUE	15	TRUE	TRUE	3666882	Roseobacter sp.	8447750	51,02	86358	Tara Oceans
GCA_002726465.1_ASM272646v1	gta	158_SRF	FALSE	15	TRUE	TRUE	3487762	Roseobacter sp.	7966882	50,6651	81414	Tara Oceans
GCA_002708965.1_ASM270896v1	gta	158_SRF	TRUE	12	TRUE	TRUE	3704350	Roseobacter sp.	7945463	52,1533	81122	Tara Oceans
GCA_002716265.1_ASM271626v1	gta	158_SRF	FALSE	14	TRUE	TRUE	3023545	Roseobacter sp.	7421936	54,1589	75874	Tara Oceans
GCA_002693125.1_ASM269312v1	gta	158_SRF	FALSE	15	TRUE	TRUE	3106942	Roseobacter sp.	7080020	49,8634	72386	Tara Oceans
GCA_002684135.1_ASM268413v1	gta	158_SRF	FALSE	14	TRUE	TRUE	2514760	Roseobacter sp.	5569298	48,2923	56902	Tara Oceans
GCA_002703405.1_ASM270340v1	gta	158_SRF	TRUE	15	TRUE	TRUE	3142791	Sulfitobacter sp.	5180568	32,6911	53024	Tara Oceans
GCA_002722375.1_ASM272237v1	gta	158_SRF	TRUE	16	TRUE	TRUE	3436664	Sulfitobacter sp.	4167337	22,2621	42626	Tara Oceans
GCA_002695005.1_ASM269500v1	gta	64_SRF	TRUE	15	TRUE	TRUE	4294341	Maritimibacter sp.	16284457	44,3197	177480	Tara Oceans

GCA_002692855.1_ASM269285v1	unclear	122_DCM	TRUE	7	FALSE	TRUE	4244573	Maricaulis sp.	6301803	7,6048	66124	Tara Oceans
GCA_002692855.1_ASM269285v1	unclear	64_DCM	TRUE	7	FALSE	TRUE	4244573	Maricaulis sp.	8889915	8,0229	91142	Tara Oceans
GCA_002707095.1_ASM270709v1	ev_producer	122_DCM	FALSE	0	FALSE	FALSE	4099214	Haliea sp.	20340491	11,5146	225628	Tara Oceans
GCA_002726835.1_ASM272683v1	ev_producer	122_DCM	FALSE	0	FALSE	FALSE	4147862	Haliea sp.	19782451	11,1043	219486	Tara Oceans
GCA_002719515.1_ASM271951v1	ev_producer	122_DCM	FALSE	0	FALSE	FALSE	3009494	Haliea sp.	15421536	11,6862	171184	Tara Oceans
GCA_002715105.1_ASM271510v1	ev_producer	122_DCM	TRUE	0	FALSE	FALSE	3591368	Balneola sp.	13375301	8,4488	139826	Tara Oceans
GCA_002683475.1_ASM268347v1	ev_producer	122_DCM	FALSE	0	FALSE	TRUE	2872426	Haliea sp.	12427068	9,7752	138070	Tara Oceans
GCA_002692925.1_ASM269292v1	remove	122_DCM	FALSE	1	FALSE	TRUE	6086193	Hyphomonas sp.	10744341	20,2181	114402	Tara Oceans
GCA_002692965.1_ASM269296v1	ev_producer	122_DCM	FALSE	0	FALSE	FALSE	3830706	Haliea sp.	5920431	2,4733	65432	Tara Oceans
GCA_002691705.1_ASM269170v1	remove	122_DCM	FALSE	0	FALSE	TRUE	1938045	Candidatus Pelagibacter sp.	4736414	20,6005	49458	Tara Oceans
GCA_002683615.1_ASM268361v1	ev_producer	122_DCM	FALSE	1	FALSE	FALSE	3675011	Hyphomonas sp.	2790794	6,2194	31866	Tara Oceans
GCA_002716025.1_ASM271602v1	ev_producer	122_MES	FALSE	0	FALSE	FALSE	2758896	Idiomarina sp.	25285981	233,93	260606	Tara Oceans
GCA_002714765.1_ASM271476v1	ev_producer	122_MES	FALSE	0	FALSE	FALSE	4643030	Marinobacter sp.	22844750	118,7958	240138	Tara Oceans
GCA_002718115.1_ASM271811v1	ev_producer	122_MES	FALSE	0	FALSE	FALSE	2661514	Idiomarina sp.	22160590	209,8885	228508	Tara Oceans
GCA_002714925.1_ASM271492v1	ev_producer	122_MES	FALSE	0	FALSE	FALSE	2370866	Idiomarina sp.	19367147	205,1823	199670	Tara Oceans
GCA_002683675.1_ASM268367v1	ev_producer	122_MES	FALSE	0	FALSE	FALSE	4225935	Rheinheimera sp.	17661793	105,3408	182436	Tara Oceans
GCA_002721435.1_ASM272143v1	ev_producer	122_MES	FALSE	0	FALSE	FALSE	3202986	Marinobacter sp.	16613350	125,0769	174788	Tara Oceans
GCA_002684295.1_ASM268429v1	ev_producer	122_MES	FALSE	0	FALSE	FALSE	2210035	Idiomarina sp.	16925637	190,0814	174480	Tara Oceans
GCA_002721625.1_ASM272162v1	ev_producer	122_MES	FALSE	0	FALSE	FALSE	3198669	Rheinheimera sp.	14182148	111,9521	146500	Tara Oceans
GCA_002683235.1_ASM268323v1	ev_producer	122_MES	FALSE	0	FALSE	FALSE	2253521	Idiomarina sp.	11489224	96,9979	118096	Tara Oceans
GCA_002729395.1_ASM272939v1	ev_producer	122_MES	FALSE	0	FALSE	FALSE	3017982	Marinobacter sp.	9727122	62,4581	102114	Tara Oceans
GCA_002707095.1_ASM270709v1	ev_producer	122_SRF	FALSE	0	FALSE	FALSE	4099214	Haliea sp.	28106858	39,257	311418	Tara Oceans
GCA_002726835.1_ASM272683v1	ev_producer	122_SRF	FALSE	0	FALSE	FALSE	4147862	Haliea sp.	27319691	37,7321	302706	Tara Oceans
GCA_002719515.1_ASM271951v1	ev_producer	122_SRF	FALSE	0	FALSE	FALSE	3009494	Haliea sp.	20777305	39,4241	230232	Tara Oceans
GCA_002683675.1_ASM268367v1	ev_producer	122_SRF	FALSE	0	FALSE	FALSE	4225935	Rheinheimera sp.	19065644	26,5411	206506	Tara Oceans
GCA_002683475.1_ASM268347v1	ev_producer	122_SRF	FALSE	0	FALSE	TRUE	2872426	Haliea sp.	16028499	32,1533	177930	Tara Oceans
GCA_002721625.1_ASM272162v1	ev_producer	122_SRF	FALSE	0	FALSE	FALSE	3198669	Rheinheimera sp.	15151689	27,8743	164218	Tara Oceans
GCA_002714765.1_ASM271476v1	ev_producer	122_SRF	FALSE	0	FALSE	FALSE	4643030	Marinobacter sp.	14729884	19,4126	162984	Tara Oceans
GCA_002692925.1_ASM269292v1	remove	122_SRF	FALSE	1	FALSE	TRUE	6086193	Hyphomonas sp.	13243915	45,6602	138134	Tara Oceans
GCA_002721435.1_ASM272143v1	ev_producer	122_SRF	FALSE	0	FALSE	FALSE	3202986	Marinobacter sp.	11832270	21,468	131064	Tara Oceans
GCA_002685555.1_ASM268555v1	ev_producer	122_SRF	FALSE	0	FALSE	FALSE	3721159	Alcanivorax sp.	11519864	24,0679	130080	Tara Oceans
GCA_002715685.1_ASM271568v1	ev_producer	122_SRF	FALSE	0	FALSE	FALSE	3609167	Alcanivorax sp.	11393782	24,3561	128584	Tara Oceans
GCA_002703665.1_ASM270366v1	ev_producer	122_SRF	FALSE	1	FALSE	FALSE	3528367	Alcanivorax sp.	11335304	25,0058	128054	Tara Oceans
GCA_002693845.1_ASM269384v1	ev_producer	122_SRF	FALSE	1	FALSE	FALSE	3189430	Alcanivorax sp.	10194678	24,1513	115106	Tara Oceans
GCA_002692955.1_ASM269295v1	ev_producer	158_DCM	FALSE	0	FALSE	FALSE	3372341	Thalassospira sp.	6560012	136,7779	66920	Tara Oceans
GCA_002731595.1_ASM273159v1	ev_producer	158_DCM	FALSE	0	FALSE	FALSE	2670106	Thalassospira sp.	4792235	153,1601	48884	Tara Oceans
GCA_002684455.1_ASM268445v1	ev_producer	158_DCM	FALSE	0	FALSE	FALSE	3075968	Arcobacter sp.	4889537	103,1731	48808	Tara Oceans
GCA_002694765.1_ASM269476v1	ev_producer	158_DCM	FALSE	0	FALSE	FALSE	3010433	Xanthomarina sp.	4398397	74,1019	44024	Tara Oceans
GCA_002714865.1_ASM271486v1	ev_producer	158_DCM	FALSE	0	FALSE	FALSE	2912717	Xanthomarina sp.	4347609	75,0631	43518	Tara Oceans
GCA_002711885.1_ASM271188v1	remove	158_DCM	FALSE	0	FALSE	FALSE	2388162	Thalassospira sp.	4258921	122,5861	43396	Tara Oceans
GCA_002682935.1_ASM268293v1	ev_producer	158_DCM	FALSE	0	FALSE	FALSE	3568067	Flavobacteriaceae bacterium	3936550	50,0704	39378	Tara Oceans
GCA_002713425.1_ASM271342v1	ev_producer	158_DCM	FALSE	0	FALSE	FALSE	2315894	Flavobacteriaceae bacterium	2274474	42,3205	22752	Tara Oceans
GCA_002692925.1_ASM269292v1	remove	158_MES	FALSE	1	FALSE	TRUE	6086193	Hyphomonas sp.	4615252	97,7043	47892	Tara Oceans
GCA_002683615.1_ASM268361v1	ev_producer	158_MES	FALSE	1	FALSE	FALSE	3675011	Hyphomonas sp.	4134406	139,5938	42932	Tara Oceans
GCA_002693505.1_ASM269350v1	remove	158_MES	FALSE	1	FALSE	FALSE	3569926	Hyphomonadaceae bacterium	4091453	141,6307	42448	Tara Oceans
GCA_002710445.1_ASM271044v1	remove	158_MES	FALSE	1	FALSE	FALSE	2980843	Hyphomonas sp.	4035425	168,0794	41844	Tara Oceans
GCA_002707095.1_ASM270709v1	ev_producer	158_SRF	FALSE	0	FALSE	FALSE	4099214	Haliea sp.	10388284	85,0236	107134	Tara Oceans
GCA_002726835.1_ASM272683v1	remove	158_SRF	FALSE	0	FALSE	FALSE	4147862	Haliea sp.	10184870	83,0977	105030	Tara Oceans

GCA_002719515.1_ASM271951v1	ev_producer	158_SRF	FALSE	0	FALSE	FALSE	3009494	Haliea sp.	7971980	83,6534	82244	Tara Oceans
GCA_002683475.1_ASM268347v1	ev_producer	158_SRF	FALSE	0	FALSE	TRUE	2872426	Haliea sp.	7011278	77,1612	72406	Tara Oceans
GCA_002710605.1_ASM271060v1	ev_producer	158_SRF	FALSE	0	FALSE	TRUE	2937613	Flavobacteriaceae bacterium	4405709	42,0741	44482	Tara Oceans
GCA_002696125.1_ASM269612v1	ev_producer	64_DCM	FALSE	0	FALSE	FALSE	2280097	Halomonas sp.	15112403	10,6045	162654	Tara Oceans
GCA_002732055.1_ASM273205v1	ev_producer	64_DCM	FALSE	0	FALSE	FALSE	2427037	Halomonas sp.	11472398	7,835	122808	Tara Oceans
GCA_002692925.1_ASM269292v1	remove	64_DCM	FALSE	1	FALSE	TRUE	6086193	Hyphomonas sp.	10810576	14,6409	110850	Tara Oceans
GCA_002685015.1_ASM268501v1	ev_producer	64_DCM	FALSE	0	FALSE	FALSE	1975268	Halomonas sp.	9585134	7,7006	102494	Tara Oceans
GCA_002696385.1_ASM269638v1	ev_producer	64_DCM	FALSE	0	FALSE	FALSE	2154320	Halomonas sp.	8496190	6,7079	90792	Tara Oceans
GCA_002693205.1_ASM269320v1	ev_producer	64_DCM	FALSE	0	FALSE	FALSE	1533498	Halomonas sp.	8206630	8,7215	87856	Tara Oceans
GCA_002695155.1_ASM269515v1	ev_producer	64_DCM	FALSE	0	FALSE	FALSE	3097567	Marinobacter sp.	6849101	5,0245	73486	Tara Oceans
GCA_002729395.1_ASM272939v1	ev_producer	64_DCM	FALSE	0	FALSE	FALSE	3017982	Marinobacter sp.	6682163	5,0892	71690	Tara Oceans
GCA_002691705.1_ASM269170v1	remove	64_DCM	FALSE	0	FALSE	TRUE	1938045	Candidatus Pelagibacter sp.	6946362	19,4561	71084	Tara Oceans
GCA_002683915.1_ASM268391v1	ev_producer	64_DCM	FALSE	0	FALSE	FALSE	4501600	Marinobacter sp.	4434794	1,9806	47252	Tara Oceans
GCA_002714765.1_ASM271476v1	ev_producer	64_DCM	FALSE	0	FALSE	FALSE	4643030	Marinobacter sp.	3410885	1,3483	36298	Tara Oceans
GCA_002695155.1_ASM269515v1	ev_producer	64_MES	FALSE	0	FALSE	FALSE	3097567	Marinobacter sp.	13874049	412,2258	139408	Tara Oceans
GCA_002696105.1_ASM269610v1	ev_producer	64_MES	FALSE	0	FALSE	FALSE	2694643	Flavobacteriaceae bacterium	13608915	466,2021	135622	Tara Oceans
GCA_002729395.1_ASM272939v1	ev_producer	64_MES	FALSE	0	FALSE	FALSE	3017982	Marinobacter sp.	13407496	406,5838	134750	Tara Oceans
GCA_002685555.1_ASM268555v1	ev_producer	64_MES	FALSE	0	FALSE	FALSE	3721159	Alcanivorax sp.	13088273	335,942	132178	Tara Oceans
GCA_002714765.1_ASM271476v1	ev_producer	64_MES	FALSE	0	FALSE	FALSE	4643030	Marinobacter sp.	12996708	226,3591	130416	Tara Oceans
GCA_002703665.1_ASM270366v1	ev_producer	64_MES	FALSE	1	FALSE	FALSE	3528367	Alcanivorax sp.	12832139	349,525	129610	Tara Oceans
GCA_002715685.1_ASM271568v1	ev_producer	64_MES	FALSE	0	FALSE	FALSE	3609167	Alcanivorax sp.	12540388	334,5163	126602	Tara Oceans
GCA_002693845.1_ASM269384v1	ev_producer	64_MES	FALSE	1	FALSE	FALSE	3189430	Alcanivorax sp.	11611246	345,6535	117260	Tara Oceans
GCA_002696305.1_ASM269630v1	ev_producer	64_MES	FALSE	0	FALSE	FALSE	2412656	Aequorivita sp.	11355493	438,1439	113222	Tara Oceans
GCA_002692925.1_ASM269292v1	remove	64_SRF	FALSE	1	FALSE	TRUE	6086193	Hyphomonas sp.	32979492	64,0688	352640	Tara Oceans
GCA_002693505.1_ASM269350v1	ev_producer	64_SRF	FALSE	1	FALSE	FALSE	3569926	Hyphomonadaceae bacterium	30607158	85,515	328024	Tara Oceans
GCA_002683615.1_ASM268361v1	ev_producer	64_SRF	FALSE	1	FALSE	FALSE	3675011	Hyphomonas sp.	30444968	82,3688	326320	Tara Oceans
GCA_002710445.1_ASM271044v1	ev_producer	64_SRF	FALSE	1	FALSE	FALSE	2980843	Hyphomonas sp.	26763364	88,8192	286876	Tara Oceans
GCA_002716205.1_ASM271620v1	ev_producer	64_SRF	FALSE	1	FALSE	FALSE	2348295	Hyphomonas sp.	21516188	90,9534	230694	Tara Oceans
GCA_002685555.1_ASM268555v1	ev_producer	64_SRF	FALSE	0	FALSE	FALSE	3721159	Alcanivorax sp.	19187288	53,0463	208966	Tara Oceans
GCA_002703665.1_ASM270366v1	ev_producer	64_SRF	FALSE	1	FALSE	FALSE	3528367	Alcanivorax sp.	18870137	55,3965	205560	Tara Oceans
GCA_002715685.1_ASM271568v1	ev_producer	64_SRF	FALSE	0	FALSE	FALSE	3609167	Alcanivorax sp.	18477048	52,9434	201338	Tara Oceans
GCA_002693845.1_ASM269384v1	ev_producer	64_SRF	FALSE	1	FALSE	FALSE	3189430	Alcanivorax sp.	17040822	53,8909	185586	Tara Oceans
GCA_002729845.1_ASM272984v1	ev_producer	64_SRF	FALSE	1	FALSE	FALSE	2880502	Alcanivorax sp.	15338212	55,0699	167136	Tara Oceans
GCA_002695465.1_ASM269546v1	ev_producer	64_SRF	FALSE	1	FALSE	FALSE	3214342	Hyphomonas sp.	14837106	33,5036	157984	Tara Oceans
GCA_002689395.1_ASM268939v1	ev_producer	64_SRF	FALSE	1	FALSE	FALSE	3254579	Hyphomonas sp.	14699306	32,6303	156506	Tara Oceans
GCA_002714095.1_ASM271409v1	ev_producer	64_SRF	FALSE	0	FALSE	FALSE	2143045	Alcanivorax sp.	11505549	55,4109	125506	Tara Oceans
GCA_002723335.1_ASM272333v1	viral	122_DCM	TRUE	1	FALSE	TRUE	3754379	Haliea sp.	20623266	12,869	230488	Tara Oceans
GCA_002723335.1_ASM272333v1	viral	122_MES	TRUE	1	FALSE	TRUE	3754379	Haliea sp.	11237941	87,5536	119102	Tara Oceans
GCA_002723335.1_ASM272333v1	ev_producer	122_SRF	TRUE	1	FALSE	TRUE	3754379	Haliea sp.	14294595	17,7749	156794	Tara Oceans
GCA_002723335.1_ASM272333v1	ev_producer	158_SRF	TRUE	1	FALSE	TRUE	3754379	Haliea sp.	6062259	40,0236	62304	Tara Oceans
GCA_002685565.1_ASM268556v1	ev_producer	122_DCM	TRUE	0	FALSE	TRUE	4517979	Haliea sp.	20528998	10,5841	227666	Tara Oceans
GCA_002684005.1_ASM268400v1	ev_producer	122_DCM	TRUE	0	FALSE	TRUE	3376779	Balneola sp.	13431709	8,9799	140492	Tara Oceans
GCA_002696075.1_ASM269607v1	ev_producer	122_DCM	TRUE	0	FALSE	TRUE	3275288	Balneola sp.	13300417	9,2017	139008	Tara Oceans
GCA_002684275.1_ASM268427v1	unclear	122_DCM	TRUE	0	FALSE	TRUE	3202815	Bacteroidetes bacterium	12866875	9,0929	134364	Tara Oceans
GCA_002696605.1_ASM269660v1	viral	122_DCM	TRUE	0	FALSE	TRUE	3248730	Pseudomonadales bacterium	7151078	3,9792	75142	Tara Oceans
GCA_002684095.1_ASM268409v1	unclear	122_DCM	TRUE	0	FALSE	TRUE	4627664	Haliea sp.	5373479	2,1712	59346	Tara Oceans
GCA_002699145.1_ASM269914v1	unclear	122_DCM	TRUE	0	FALSE	TRUE	5261423	Halioglobus sp.	4674604	1,5795	51880	Tara Oceans
GCA_002704665.1_ASM270466v1	viral	122_DCM	TRUE	0	FALSE	TRUE	1784933	Proteobacteria bacterium	4333358	4,8171	45010	Tara Oceans

GCA_002696005.1_ASM269600v1	ev_producer	122_MES	TRUE	0	FALSE	TRUE	2923195	Idiomarina sp.	25801446	225,1086	265966	Tara Oceans
GCA_002683435.1_ASM268343v1	ev_producer	122_MES	TRUE	0	FALSE	TRUE	4750670	Marinobacter sp.	22518907	113,9505	236878	Tara Oceans
GCA_002726345.1_ASM272634v1	ev_producer	122_MES	TRUE	0	FALSE	TRUE	2676639	Idiomarina sp.	21816597	208,5298	224916	Tara Oceans
GCA_002726615.1_ASM272661v1	ev_producer	122_MES	TRUE	0	FALSE	TRUE	3581496	Marinobacter sp.	16781068	111,0482	176530	Tara Oceans
GCA_002694205.1_ASM269420v1	ev_producer	122_MES	TRUE	0	FALSE	TRUE	2269208	Idiomarina sp.	15040387	163,8289	155046	Tara Oceans
GCA_002693995.1_ASM269399v1	ev_producer	122_MES	TRUE	1	FALSE	TRUE	5357492	Marinobacter sp.	13247767	49,0889	139036	Tara Oceans
GCA_002685395.1_ASM268539v1	ev_producer	122_MES	TRUE	0	FALSE	TRUE	5142000	Rheinheimera sp.	12268876	49,7085	126536	Tara Oceans
GCA_002723205.1_ASM272320v1	unclear	122_MES	TRUE	0	FALSE	TRUE	5730820	Alteromonadaceae bacterium	11727178	53,9464	120150	Tara Oceans
GCA_002685565.1_ASM268556v1	ev_producer	122_SRF	TRUE	0	FALSE	TRUE	4517979	Haliea sp.	28527489	36,1759	315996	Tara Oceans
GCA_002683435.1_ASM268343v1	viral	122_SRF	TRUE	0	FALSE	TRUE	4750670	Marinobacter sp.	15933440	20,3921	176508	Tara Oceans
GCA_002723235.1_ASM272323v1	viral	122_SRF	TRUE	2	FALSE	TRUE	4493577	Spongiibacteraceae bacterium	12399910	15,4357	138988	Tara Oceans
GCA_002685395.1_ASM268539v1	ev_producer	122_SRF	TRUE	0	FALSE	TRUE	5142000	Rheinheimera sp.	12596553	11,9849	135482	Tara Oceans
GCA_002726615.1_ASM272661v1	ev_producer	122_SRF	TRUE	0	FALSE	TRUE	3581496	Marinobacter sp.	11355491	18,6664	125688	Tara Oceans
GCA_002696165.1_ASM269616v1	remove	122_SRF	TRUE	0	FALSE	TRUE	7185140	Phycisphaerae bacterium	9640088	11,8453	107456	Tara Oceans
GCA_002684075.1_ASM268407v1	ev_producer	158_DCM	TRUE	0	FALSE	TRUE	4756359	Thalassospira sp.	8928523	132,7951	91058	Tara Oceans
GCA_002707445.1_ASM270744v1	ev_producer	158_DCM	TRUE	0	FALSE	TRUE	4697496	Thalassospira sp.	8339605	123,1472	85006	Tara Oceans
GCA_002694405.1_ASM269440v1	ev_producer	158_DCM	TRUE	0	FALSE	TRUE	4659979	Thalassospira sp.	7936309	116,7512	80972	Tara Oceans
GCA_002714365.1_ASM271436v1	ev_producer	158_DCM	TRUE	0	FALSE	TRUE	4201089	Thalassospira sp.	7052931	112,5395	71896	Tara Oceans
GCA_002709775.1_ASM270977v1	ev_producer	158_DCM	TRUE	0	FALSE	TRUE	4304835	Thalassospira sp.	6725282	103,1114	68622	Tara Oceans
GCA_002707405.1_ASM270740v1	unclear	158_DCM	TRUE	0	FALSE	TRUE	2917325	Thalassospira sp.	5300838	125,3435	54034	Tara Oceans
GCA_002685075.1_ASM268507v1	ev_producer	158_DCM	TRUE	0	FALSE	TRUE	2839798	Arcobacter sp.	4586269	104,7924	45796	Tara Oceans
GCA_002722255.1_ASM272225v1	ev_producer	158_DCM	TRUE	0	FALSE	TRUE	3224965	Winogradskyella sp.	3947320	61,1856	39518	Tara Oceans
GCA_002706045.1_ASM270604v1	ev_producer	158_DCM	TRUE	0	FALSE	TRUE	2936064	Flavobacteriaceae bacterium	3170749	51,9005	31714	Tara Oceans
GCA_002683975.1_ASM268397v1	ev_producer	158_DCM	TRUE	0	FALSE	TRUE	3749540	Pseudomonadales bacterium	2524243	83,9178	25834	Tara Oceans
GCA_002726785.1_ASM272678v1	unclear	158_DCM	TRUE	0	FALSE	TRUE	2003842	Flavobacteriaceae bacterium	2545427	63,0842	25472	Tara Oceans
GCA_002726195.1_ASM272619v1	ev_producer	158_DCM	TRUE	0	FALSE	TRUE	3156483	Flavobacteriales bacterium	2315996	164,526	23214	Tara Oceans
GCA_002686735.1_ASM268673v1	ev_producer	158_MES	TRUE	0	FALSE	TRUE	2145676	Psychrobacter sp.	7810816	311,1968	78652	Tara Oceans
GCA_002683975.1_ASM268397v1	ev_producer	158_MES	TRUE	0	FALSE	TRUE	3749540	Pseudomonadales bacterium	5851730	272,8051	60270	Tara Oceans
GCA_002709125.1_ASM270912v1	unclear	158_MES	TRUE	0	FALSE	TRUE	3053049	Rickettsiales bacterium	5273294	299,4767	53524	Tara Oceans
GCA_002709595.1_ASM270959v1	unclear	158_MES	TRUE	1	FALSE	TRUE	3908704	Hyphomonas sp.	4045488	111,1665	42090	Tara Oceans
GCA_002708585.1_ASM270858v1	ev_producer	158_MES	TRUE	3	FALSE	TRUE	3375859	Hyphomonadaceae bacterium	3777678	129,4546	39286	Tara Oceans
GCA_002685565.1_ASM268556v1	viral	158_SRF	TRUE	0	FALSE	TRUE	4517979	Haliea sp.	10674022	79,0136	110016	Tara Oceans
GCA_002693005.1_ASM269300v1	ev_producer	158_SRF	TRUE	0	FALSE	TRUE	3810386	Flavobacteriaceae bacterium	4992103	36,5708	50408	Tara Oceans
GCA_002684555.1_ASM268455v1	unclear	158_SRF	TRUE	0	FALSE	TRUE	7492206	Gimesia sp.	4303237	17,866	43816	Tara Oceans
GCA_002692745.1_ASM269274v1	unclear	158_SRF	TRUE	1	FALSE	TRUE	3951760	Flavobacteriaceae bacterium	4194715	29,0341	42330	Tara Oceans
GCA_002693995.1_ASM269399v1	viral	64_DCM	TRUE	1	FALSE	TRUE	5357492	Marinobacter sp.	11569087	7,921	122654	Tara Oceans
GCA_002687595.1_ASM268759v1	ev_producer	64_DCM	TRUE	0	FALSE	TRUE	1771232	Halomonas sp.	8105990	6,9367	86710	Tara Oceans
GCA_002715145.1_ASM271514v1	ev_producer	64_DCM	TRUE	0	FALSE	TRUE	4794930	Halomonas sp.	7965333	1,9876	84786	Tara Oceans
GCA_002696605.1_ASM269660v1	viral	64_DCM	TRUE	0	FALSE	TRUE	3248730	Pseudomonadales bacterium	7218204	3,4689	74422	Tara Oceans
GCA_002718255.1_ASM271825v1	ev_producer	64_DCM	TRUE	0	FALSE	TRUE	3186133	Marinobacter sp.	5793511	4,2778	62118	Tara Oceans
GCA_002697735.1_ASM269773v1	ev_producer	64_DCM	TRUE	0	FALSE	TRUE	5033673	Halomonas sp.	4216001	0,8615	44916	Tara Oceans
GCA_002683435.1_ASM268343v1	viral	64_DCM	TRUE	0	FALSE	TRUE	4750670	Marinobacter sp.	3798858	1,5787	40338	Tara Oceans
GCA_002693645.1_ASM269364v1	ev_producer	64_DCM	TRUE	0	FALSE	TRUE	4095594	Marinobacter sp.	3724636	1,7061	39672	Tara Oceans
GCA_002693995.1_ASM269399v1	ev_producer	64_MES	TRUE	1	FALSE	TRUE	5357492	Marinobacter sp.	18653864	323,0831	187480	Tara Oceans
GCA_002696735.1_ASM269673v1	ev_producer	64_MES	TRUE	0	FALSE	TRUE	3324714	Methylophaga sp.	17497779	532,2323	174756	Tara Oceans
GCA_002726815.1_ASM272681v1	unclear	64_MES	TRUE	0	FALSE	TRUE	3134966	Methylophaga sp.	16515130	533,3054	164958	Tara Oceans
GCA_002694545.1_ASM269454v1	ev_producer	64_MES	TRUE	0	FALSE	TRUE	3172870	Methylophaga sp.	16380284	523,52	163614	Tara Oceans
GCA_002715155.1_ASM271515v1	ev_producer	64_MES	TRUE	0	FALSE	TRUE	3285734	Aequorivita sp.	15207766	436,8369	151634	Tara Oceans

GCA_002683435.1_ASM268343v1	ev_producer	64_MES	TRUE	0	FALSE	TRUE	4750670	Marinobacter sp.	14383248	253,6732	144408	Tara Oceans
GCA_002683495.1_ASM268349v1	ev_producer	64_MES	TRUE	0	FALSE	TRUE	3238790	Aequorivita sp.	14016663	405,9823	139766	Tara Oceans
GCA_002693005.1_ASM269300v1	unclear	64_MES	TRUE	0	FALSE	TRUE	3810386	Flavobacteriaceae bacterium	13656499	341,0468	136234	Tara Oceans
GCA_002718255.1_ASM271825v1	unclear	64_MES	TRUE	0	FALSE	TRUE	3186133	Marinobacter sp.	12559244	363,694	126212	Tara Oceans
GCA_002693695.1_ASM269369v1	unclear	64_MES	TRUE	0	FALSE	TRUE	3771012	Pseudohongiella sp.	11736976	321,384	117912	Tara Oceans
GCA_002683555.1_ASM268355v1	ev_producer	64_MES	TRUE	0	FALSE	TRUE	3436895	Aequorivita sp.	11500489	247,2382	114506	Tara Oceans
GCA_002708585.1_ASM270858v1	ev_producer	64_SRF	TRUE	3	FALSE	TRUE	3375859	Hyphomonadaceae bacterium	18804384	44,9981	200660	Tara Oceans
GCA_002695955.1_ASM269595v1	ev_producer	64_SRF	TRUE	3	FALSE	TRUE	3159258	Hyphomonadaceae bacterium	17087492	42,8997	182356	Tara Oceans
GCA_002719735.1_ASM271973v1	ev_producer	64_SRF	TRUE	0	FALSE	TRUE	4593791	Alcanivorax sp.	13817544	23,9139	149608	Tara Oceans
GCA_002696185.1_ASM269618v1	ev_producer	64_SRF	TRUE	0	FALSE	TRUE	4147124	Alcanivorax sp.	12366011	23,665	133868	Tara Oceans
GCA_002731335.1_ASM273133v1	ev_producer	64_SRF	TRUE	0	FALSE	TRUE	4077693	Alcanivorax sp.	12336210	23,6536	133788	Tara Oceans
GCA_002684005.1_ASM268400v1	ev_producer	64_SRF	TRUE	0	FALSE	TRUE	3376779	Balneola sp.	12086109	53,7234	124106	Tara Oceans

Table S1 - Tara Ocean MAGs (Tully 2018)

Selection! Only top rows are shown, full table available online

# Assembly	Level	WGS	BioSample	Isolate	Taxonomy
GCA_002727995.1	Contig	PBUV00000000	SAMN07618132	RS446	Acidiferrobacter sp.
GCA_002694065.1	Contig	NZKZ00000000	SAMN07618131	IN47	Acidiferrobacter sp.
GCA_002693245.1	Contig	NZPA00000000	SAMN07618135	EAC671	Acidiferrobacteraceae bacterium
GCA_002687565.1	Contig	NZAN00000000	SAMN07618133	ARS29	Acidiferrobacteraceae bacterium
GCA_002686015.1	Contig	NYZW00000000	SAMN07618134	ARS46	Acidiferrobacteraceae bacterium
GCA_002714255.1	Contig	PAWW00000000	SAMN07618144	SAT1331	Acidiferrobacteraceae bacterium
GCA_002713825.1	Contig	PAUU00000000	SAMN07618146	SAT1475	Acidiferrobacteraceae bacterium
GCA_002713125.1	Contig	PAVN00000000	SAMN07618145	SAT1416	Acidiferrobacteraceae bacterium
GCA_002705025.1	Contig	PAIB00000000	SAMN07618136	NAT194	Acidiferrobacteraceae bacterium
GCA_002726415.1	Contig	PBSU00000000	SAMN07618140	NP79	Acidiferrobacteraceae bacterium
GCA_002724185.1	Contig	PBOB00000000	SAMN07618143	RS826	Acidiferrobacteraceae bacterium
GCA_002725835.1	Contig	PBRH00000000	SAMN07618141	NP959	Acidiferrobacteraceae bacterium
GCA_002725185.1	Contig	PBQY00000000	SAMN07618142	NP97	Acidiferrobacteraceae bacterium
GCA_002721105.1	Contig	PBJL00000000	SAMN07618147	SP295	Acidiferrobacteraceae bacterium
GCA_002731295.1	Contig	PCBE00000000	SAMN07618137	NP122	Acidiferrobacteraceae bacterium
GCA_002729715.1	Contig	PBYR00000000	SAMN07618138	NP24	Acidiferrobacteraceae bacterium
GCA_002729195.1	Contig	PBTT00000000	SAMN07618139	NP54	Acidiferrobacteraceae bacterium
GCA_002729125.1	Contig	PBTQ00000000	SAMN07618178	NP57	Acidimicrobiaceae bacterium
GCA_002729765.1	Contig	PBYV00000000	SAMN07618177	NP20	Acidimicrobiaceae bacterium
GCA_002731135.1	Contig	PCAA00000000	SAMN07618176	NP1395	Acidimicrobiaceae bacterium
GCA_002730955.1	Contig	PCBN00000000	SAMN07618174	NP1171	Acidimicrobiaceae bacterium
GCA_002730515.1	Contig	PCAJ00000000	SAMN07618175	NP134	Acidimicrobiaceae bacterium
GCA_002721075.1	Contig	PBJH00000000	SAMN07618206	SP2971	Acidimicrobiaceae bacterium
GCA_002720565.1	Contig	PBJA00000000	SAMN07618207	SP2979	Acidimicrobiaceae bacterium
GCA_002720485.1	Contig	PBIQ00000000	SAMN07618208	SP2997	Acidimicrobiaceae bacterium
GCA_002720045.1	Contig	PBGE00000000	SAMN07618211	SP3070	Acidimicrobiaceae bacterium

GCA_002719335.1	Contig	PBHE00000000	SAMN07618209	SP304	Acidimicrobiaceae bacterium
GCA_002719235.1	Contig	PBGN00000000	SAMN07618210	SP3061	Acidimicrobiaceae bacterium
GCA_002722565.1	Contig	PBMG00000000	SAMN07618203	SP219	Acidimicrobiaceae bacterium
GCA_002722525.1	Contig	PBMJ00000000	SAMN07618202	SP215	Acidimicrobiaceae bacterium
GCA_002722465.1	Contig	PBMC00000000	SAMN07618204	SP223	Acidimicrobiaceae bacterium
GCA_002721305.1	Contig	PBKA00000000	SAMN07618205	SP281	Acidimicrobiaceae bacterium
GCA_002725235.1	Contig	PBRA00000000	SAMN07618182	NP966	Acidimicrobiaceae bacterium
GCA_002724825.1	Contig	PBPU00000000	SAMN07618183	RS349	Acidimicrobiaceae bacterium
GCA_002726075.1	Contig	PBRY00000000	SAMN07618181	NP937	Acidimicrobiaceae bacterium
GCA_002725355.1	Contig	PBOQ00000000	SAMN07618185	RS387	Acidimicrobiaceae bacterium
GCA_002724495.1	Contig	PBOT00000000	SAMN07618184	RS384	Acidimicrobiaceae bacterium
GCA_002726315.1	Contig	PBSS00000000	SAMN07618180	NP80	Acidimicrobiaceae bacterium
GCA_002726525.1	Contig	PBTF00000000	SAMN07618179	NP68	Acidimicrobiaceae bacterium
GCA_002727235.1	Contig	PBWL00000000	SAMN07618190	RS817	Acidimicrobiaceae bacterium
GCA_002727955.1	Contig	PBUU00000000	SAMN07618186	RS447	Acidimicrobiaceae bacterium
GCA_002727895.1	Contig	PBUO00000000	SAMN07618187	RS454	Acidimicrobiaceae bacterium
GCA_002727855.1	Contig	PBUJ00000000	SAMN07618189	RS460	Acidimicrobiaceae bacterium
GCA_002728515.1	Contig	PBUN00000000	SAMN07618188	RS455	Acidimicrobiaceae bacterium
GCA_002704945.1	Contig	PAIN00000000	SAMN07618168	NAT207	Acidimicrobiaceae bacterium
GCA_002705305.1	Contig	PAHW00000000	SAMN07618167	NAT188	Acidimicrobiaceae bacterium
GCA_002703225.1	Contig	NZTT00000000	SAMN07618153	EAC105	Acidimicrobiaceae bacterium
GCA_002703045.1	Contig	PAEG00000000	SAMN07618162	MED867	Acidimicrobiaceae bacterium
GCA_002702085.1	Contig	NZYA00000000	SAMN07618171	NAT284	Acidimicrobiaceae bacterium
GCA_002699725.1	Contig	PACE00000000	SAMN07618163	NAT100	Acidimicrobiaceae bacterium
GCA_002699625.1	Contig	PABW00000000	SAMN07618164	NAT108	Acidimicrobiaceae bacterium
GCA_002699505.1	Contig	PABC00000000	SAMN07618165	NAT13	Acidimicrobiaceae bacterium
GCA_002699455.1	Contig	PAAP00000000	SAMN07618166	NAT143	Acidimicrobiaceae bacterium
GCA_002698945.1	Contig	NZZV00000000	SAMN07618169	NAT239	Acidimicrobiaceae bacterium
GCA_002698445.1	Contig	NZXV00000000	SAMN07618172	NAT292	Acidimicrobiaceae bacterium
GCA_002698125.1	Contig	NZVA00000000	SAMN07618173	NAT71	Acidimicrobiaceae bacterium

GCA_002697965.1	Contig	NZZE00000000	SAMN07618170	NAT255	Acidimicrobiaceae bacterium
GCA_002696225.1	Contig	NZRM00000000	SAMN07618154	EAC35	Acidimicrobiaceae bacterium
GCA_002713765.1	Scaffold	PAUL00000000	SAMN07618195	SAT1505	Acidimicrobiaceae bacterium
GCA_002713545.1	Contig	PATJ00000000	SAMN07618197	SAT1608	Acidimicrobiaceae bacterium
GCA_002713525.1	Contig	PATF00000000	SAMN07618199	SAT1613	Acidimicrobiaceae bacterium
GCA_002713485.1	Contig	PAWI00000000	SAMN07618193	SAT136	Acidimicrobiaceae bacterium
GCA_002714225.1	Contig	PAWU00000000	SAMN07618192	SAT1333	Acidimicrobiaceae bacterium
GCA_002713975.1	Contig	PAVP00000000	SAMN07618194	SAT1405	Acidimicrobiaceae bacterium
GCA_002714485.1	Contig	PAXM00000000	SAMN07618191	SAT127	Acidimicrobiaceae bacterium
GCA_002717325.1	Contig	PBDM00000000	SAMN07618212	SP349	Acidimicrobiaceae bacterium
GCA_002717365.1	Contig	PBDJ00000000	SAMN07618213	SP351	Acidimicrobiaceae bacterium
GCA_002716285.1	Contig	PBAM00000000	SAMN07618214	SP49	Acidimicrobiaceae bacterium
GCA_002715405.1	Contig	PAZD00000000	SAMN07618215	SP83	Acidimicrobiaceae bacterium
GCA_002712445.1	Contig	PATP00000000	SAMN07618196	SAT1581	Acidimicrobiaceae bacterium
GCA_002712325.1	Contig	PATH00000000	SAMN07618198	SAT161	Acidimicrobiaceae bacterium
GCA_002709645.1	Contig	PAQG00000000	SAMN07618200	SP142	Acidimicrobiaceae bacterium
GCA_002708935.1	Contig	PAPO00000000	SAMN07618201	SP160	Acidimicrobiaceae bacterium
GCA_002687745.1	Contig	NZBE00000000	SAMN07618151	ARS1426	Acidimicrobiaceae bacterium
GCA_002687365.1	Contig	NZAA00000000	SAMN07618152	ARS42	Acidimicrobiaceae bacterium
GCA_002690205.1	Contig	NZDT00000000	SAMN07618148	ARS1030	Acidimicrobiaceae bacterium
GCA_002690175.1	Contig	NZDQ00000000	SAMN07618149	ARS1034	Acidimicrobiaceae bacterium
GCA_002688245.1	Contig	NZCI00000000	SAMN07618150	ARS1215	Acidimicrobiaceae bacterium
GCA_002693375.1	Contig	NZPP00000000	SAMN07618155	EAC65	Acidimicrobiaceae bacterium
GCA_002693885.1	Contig	NZKM00000000	SAMN07618160	IN924	Acidimicrobiaceae bacterium
GCA_002692945.1	Contig	NZMP00000000	SAMN07618157	IN1213	Acidimicrobiaceae bacterium
GCA_002692545.1	Contig	NZKP00000000	SAMN07618159	IN919	Acidimicrobiaceae bacterium
GCA_002694825.1	Contig	NZNH00000000	SAMN07618156	EAC9	Acidimicrobiaceae bacterium
GCA_002694625.1	Contig	NZMO00000000	SAMN07618158	IN1214	Acidimicrobiaceae bacterium
GCA_002691865.1	Contig	NZJD00000000	SAMN07618161	MED569	Acidimicrobiaceae bacterium
GCA_002730525.1	Contig	PCAK00000000	SAMN07618221	NP133	Acidobacteria bacterium

Table S1 - Heligoland MAGs (Orellana 2019)

accession	realm	phylum	class	order	family	genus	species	Reference
20120607_Bin_39_1-contigs	d__Bacteria	p__Actinobacteriota	c__Acidimicrobiia	o__Acidimicrobiales	f__Ilumatobacteraceae	g__Ilumatobacter_A	s__	
20120607_Bin_77-contigs	d__Bacteria	p__Actinobacteriota	c__Acidimicrobiia	o__Acidimicrobiales	f__Ilumatobacteraceae	g__Ilumatobacter_A	s__	
20160316_Bin_20_5-contigs	d__Bacteria	p__Actinobacteriota	c__Acidimicrobiia	o__Acidimicrobiales	f__UBA10347	g__UBA10347	s__UBA10347	
20160426_Bin_31_1-contigs	d__Bacteria	p__Actinobacteriota	c__Acidimicrobiia	o__Acidimicrobiales	f__UBA10347	g__UBA10347	s__UBA10347	
20110331_Bin_50_1-contigs	d__Bacteria	p__Actinobacteriota	c__Acidimicrobiia	o__Actinomarinales	f__Actinomarinaceae	g__Actinomarina	s__Actinomarina	
20160517_Bin_10_1-contigs	d__Bacteria	p__Actinobacteriota	c__Actinomycetia	o__Actinomycetales	f__Microbacteriaceae	g__Pontimonas	s__	
20110523_Bin_64_1-contigs	d__Bacteria	p__Actinobacteriota	c__Actinomycetia	o__Nanopelagicales	f__S36-B12	g__	s__	
20110523_Bin_63_1-contigs	d__Bacteria	p__Actinobacteriota	c__Actinomycetia	o__Nanopelagicales	f__S36-B12	g__UBA4592	s__UBA4592	
20110523_Bin_63_3-contigs	d__Bacteria	p__Actinobacteriota	c__Actinomycetia	o__Nanopelagicales	f__S36-B12	g__UBA4592	s__UBA4592	
20100430_Bin_38-contigs	d__Bacteria	p__Actinobacteriota	c__Actinomycetia	o__Propionibacteriales	f__Nocardioidaceae	g__Nocardioides	s__	
20160512_Bin_69_4-contigs	d__Bacteria	p__Bacteroidota	c__Bacteroidia	o__Chitinophagales	f__Saprospiraceae	g__AM-2324	s__	
20160517_Bin_20_2-contigs	d__Bacteria	p__Bacteroidota	c__Bacteroidia	o__Chitinophagales	f__Saprospiraceae	g__RFSX01	s__	
20110428_Bin_119-contigs	d__Bacteria	p__Bacteroidota	c__Bacteroidia	o__Chitinophagales	f__Saprospiraceae	g__UBA1994	s__UBA1994	
20100504_Bin_77-contigs	d__Bacteria	p__Bacteroidota	c__Bacteroidia	o__Cytophagales	f__Cyclobacteriaceae	g__Fabibacter	s__Fabibacter	
20110428_Bin_104_1-contigs	d__Bacteria	p__Bacteroidota	c__Bacteroidia	o__Cytophagales	f__Cyclobacteriaceae	g__	s__	
20160316_Bin_26_2_5-contigs	d__Bacteria	p__Bacteroidota	c__Bacteroidia	o__Cytophagales	f__Cyclobacteriaceae	g__UBA4465	s__	
20160412_Bin_34_4-contigs	d__Bacteria	p__Bacteroidota	c__Bacteroidia	o__Cytophagales	f__Cyclobacteriaceae	g__UBA4465	s__	
20160502_Bin_26_3-contigs	d__Bacteria	p__Bacteroidota	c__Bacteroidia	o__Cytophagales	f__Cyclobacteriaceae	g__UBA4465	s__	
20100303_Bin_85_1-contigs	d__Bacteria	p__Bacteroidota	c__Bacteroidia	o__Cytophagales	f__Cyclobacteriaceae	g__UBA4465	s__UBA4465	
20100413_Bin_46_2-contigs	d__Bacteria	p__Bacteroidota	c__Bacteroidia	o__Cytophagales	f__Cyclobacteriaceae	g__UBA4465	s__UBA4465	
20160512_Bin_21_3-contigs	d__Bacteria	p__Bacteroidota	c__Bacteroidia	o__Cytophagales	f__WTJX01	g__	s__	
20110526_Bin_151_1-contigs	d__Bacteria	p__Bacteroidota	c__Bacteroidia	o__Flavobacteriales	f__BACL11	g__	s__	
20160331_Bin_33_5-contigs	d__Bacteria	p__Bacteroidota	c__Bacteroidia	o__Flavobacteriales	f__BACL11	g__	s__	
20160426_Bin_8_6-contigs	d__Bacteria	p__Bacteroidota	c__Bacteroidia	o__Flavobacteriales	f__BACL11	g__	s__	
20160419_Bin_44_1-contigs	d__Bacteria	p__Bacteroidota	c__Bacteroidia	o__Flavobacteriales	f__BACL11	g__UWMA-0277	s__	
20110428_Bin_15-contigs	d__Bacteria	p__Bacteroidota	c__Bacteroidia	o__Flavobacteriales	f__Crocinitomicaceae	g__UBA4466	s__	
20120426_Bin_71_1-contigs	d__Bacteria	p__Bacteroidota	c__Bacteroidia	o__Flavobacteriales	f__Crocinitomicaceae	g__UBA4466	s__UBA4466	
20160502_Bin_56_4-contigs	d__Bacteria	p__Bacteroidota	c__Bacteroidia	o__Flavobacteriales	f__Crocinitomicaceae	g__UBA952	s__	
20160512_Bin_69_2-contigs	d__Bacteria	p__Bacteroidota	c__Bacteroidia	o__Flavobacteriales	f__Crocinitomicaceae	g__UBA952	s__	
20120503_Bin_84_1-contigs	d__Bacteria	p__Bacteroidota	c__Bacteroidia	o__Flavobacteriales	f__Flavobacteriaceae	g__Algibacter_B	s__	
20160517_Bin_62_2-contigs	d__Bacteria	p__Bacteroidota	c__Bacteroidia	o__Flavobacteriales	f__Flavobacteriaceae	g__BACL21	s__BACL21	
20160412_Bin_12_4-contigs	d__Bacteria	p__Bacteroidota	c__Bacteroidia	o__Flavobacteriales	f__Flavobacteriaceae	g__CAU-1491	s__	
20110404_Bin_35-contigs	d__Bacteria	p__Bacteroidota	c__Bacteroidia	o__Flavobacteriales	f__Flavobacteriaceae	g__Dokdonia	s__	
20110428_Bin_94_1-contigs	d__Bacteria	p__Bacteroidota	c__Bacteroidia	o__Flavobacteriales	f__Flavobacteriaceae	g__Flavicella	s__	
20160331_Bin_72_1-contigs	d__Bacteria	p__Bacteroidota	c__Bacteroidia	o__Flavobacteriales	f__Flavobacteriaceae	g__Flavicella	s__	

20100504_Bin_76_2-contigs	d__Bacteria	p__Bacteroidota	c__Bacteroidia	o__Flavobacteriales	f__Flavobacteriaceae	g__Flavicella	s__Flavicella
20100511_Bin_72_1-contigs	d__Bacteria	p__Bacteroidota	c__Bacteroidia	o__Flavobacteriales	f__Flavobacteriaceae	g__Flavicella	s__Flavicella
20100518_Bin_64-contigs	d__Bacteria	p__Bacteroidota	c__Bacteroidia	o__Flavobacteriales	f__Flavobacteriaceae	g__Flavicella	s__Flavicella
20120607_Bin_74-contigs	d__Bacteria	p__Bacteroidota	c__Bacteroidia	o__Flavobacteriales	f__Flavobacteriaceae	g__Flavicella	s__Flavicella
20160331_Bin_62-contigs	d__Bacteria	p__Bacteroidota	c__Bacteroidia	o__Flavobacteriales	f__Flavobacteriaceae	g__Flavobacterium	s__
20100511_Bin_40-contigs	d__Bacteria	p__Bacteroidota	c__Bacteroidia	o__Flavobacteriales	f__Flavobacteriaceae	g__Flavobacterium	s__Flavobacterium
20160502_Bin_23_4-contigs	d__Bacteria	p__Bacteroidota	c__Bacteroidia	o__Flavobacteriales	f__Flavobacteriaceae	g__GCA-002733185	s__
20160331_Bin_37_1-contigs	d__Bacteria	p__Bacteroidota	c__Bacteroidia	o__Flavobacteriales	f__Flavobacteriaceae	g__GCA-002733185	s__GCA-002733185
20160512_Bin_59_1-contigs	d__Bacteria	p__Bacteroidota	c__Bacteroidia	o__Flavobacteriales	f__Flavobacteriaceae	g__GCA-002733185	s__GCA-002733185
20120412_Bin_51_1-contigs	d__Bacteria	p__Bacteroidota	c__Bacteroidia	o__Flavobacteriales	f__Flavobacteriaceae	g__Gillisia	s__
20120412_Bin_64_1-contigs	d__Bacteria	p__Bacteroidota	c__Bacteroidia	o__Flavobacteriales	f__Flavobacteriaceae	g__Gillisia	s__
20120412_Bin_64_2-contigs	d__Bacteria	p__Bacteroidota	c__Bacteroidia	o__Flavobacteriales	f__Flavobacteriaceae	g__Gillisia	s__
20100504_Bin_97_1-contigs	d__Bacteria	p__Bacteroidota	c__Bacteroidia	o__Flavobacteriales	f__Flavobacteriaceae	g__Hel1-33-131	s__
20100518_Bin_106_1-contigs	d__Bacteria	p__Bacteroidota	c__Bacteroidia	o__Flavobacteriales	f__Flavobacteriaceae	g__Hel1-33-131	s__
20100518_Bin_124_1-contigs	d__Bacteria	p__Bacteroidota	c__Bacteroidia	o__Flavobacteriales	f__Flavobacteriaceae	g__Hel1-33-131	s__
20110516_Bin_81_1-contigs	d__Bacteria	p__Bacteroidota	c__Bacteroidia	o__Flavobacteriales	f__Flavobacteriaceae	g__Hel1-33-131	s__
20110516_Bin_85-contigs	d__Bacteria	p__Bacteroidota	c__Bacteroidia	o__Flavobacteriales	f__Flavobacteriaceae	g__Hel1-33-131	s__
20110519_Bin_85_1-contigs	d__Bacteria	p__Bacteroidota	c__Bacteroidia	o__Flavobacteriales	f__Flavobacteriaceae	g__Hel1-33-131	s__
20110523_Bin_125_1-contigs	d__Bacteria	p__Bacteroidota	c__Bacteroidia	o__Flavobacteriales	f__Flavobacteriaceae	g__Hel1-33-131	s__
20110523_Bin_38_1-contigs	d__Bacteria	p__Bacteroidota	c__Bacteroidia	o__Flavobacteriales	f__Flavobacteriaceae	g__Hel1-33-131	s__
20110523_Bin_61_1-contigs	d__Bacteria	p__Bacteroidota	c__Bacteroidia	o__Flavobacteriales	f__Flavobacteriaceae	g__Hel1-33-131	s__
20110526_Bin_11_1-contigs	d__Bacteria	p__Bacteroidota	c__Bacteroidia	o__Flavobacteriales	f__Flavobacteriaceae	g__Hel1-33-131	s__
20120510_Bin_76_1-contigs	d__Bacteria	p__Bacteroidota	c__Bacteroidia	o__Flavobacteriales	f__Flavobacteriaceae	g__Hel1-33-131	s__
20120607_Bin_49_1-contigs	d__Bacteria	p__Bacteroidota	c__Bacteroidia	o__Flavobacteriales	f__Flavobacteriaceae	g__Hel1-33-131	s__
20120607_Bin_49_2-contigs	d__Bacteria	p__Bacteroidota	c__Bacteroidia	o__Flavobacteriales	f__Flavobacteriaceae	g__Hel1-33-131	s__
20120607_Bin_49_3-contigs	d__Bacteria	p__Bacteroidota	c__Bacteroidia	o__Flavobacteriales	f__Flavobacteriaceae	g__Hel1-33-131	s__
20160512_Bin_35_3-contigs	d__Bacteria	p__Bacteroidota	c__Bacteroidia	o__Flavobacteriales	f__Flavobacteriaceae	g__Hel1-33-131	s__
20100423_Bin_16_1-contigs	d__Bacteria	p__Bacteroidota	c__Bacteroidia	o__Flavobacteriales	f__Flavobacteriaceae	g__Hel1-33-131	s__Hel1-33-131
20110523_Bin_120_1-contigs	d__Bacteria	p__Bacteroidota	c__Bacteroidia	o__Flavobacteriales	f__Flavobacteriaceae	g__MAG-120531	s__MAG-120531
20110428_Bin_3_3-contigs	d__Bacteria	p__Bacteroidota	c__Bacteroidia	o__Flavobacteriales	f__Flavobacteriaceae	g__MAG-121220-bin8	s__
20110523_Bin_47_1-contigs	d__Bacteria	p__Bacteroidota	c__Bacteroidia	o__Flavobacteriales	f__Flavobacteriaceae	g__MAG-121220-bin8	s__
20120426_Bin_77_1-contigs	d__Bacteria	p__Bacteroidota	c__Bacteroidia	o__Flavobacteriales	f__Flavobacteriaceae	g__MAG-121220-bin8	s__
20160512_Bin_46_2-contigs	d__Bacteria	p__Bacteroidota	c__Bacteroidia	o__Flavobacteriales	f__Flavobacteriaceae	g__MAG-121220-bin8	s__
20100303_Bin_98_1-contigs	d__Bacteria	p__Bacteroidota	c__Bacteroidia	o__Flavobacteriales	f__Flavobacteriaceae	g__MAG-121220-bin8	s__MAG-121220-bin8
20110530_Bin_58-contigs	d__Bacteria	p__Bacteroidota	c__Bacteroidia	o__Flavobacteriales	f__Flavobacteriaceae	g__Maribacter	s__
20120412_Bin_55-contigs	d__Bacteria	p__Bacteroidota	c__Bacteroidia	o__Flavobacteriales	f__Flavobacteriaceae	g__Maribacter	s__
20120412_Bin_71_1-contigs	d__Bacteria	p__Bacteroidota	c__Bacteroidia	o__Flavobacteriales	f__Flavobacteriaceae	g__Marixanthomonas	s__
20160517_Bin_30_2-contigs	d__Bacteria	p__Bacteroidota	c__Bacteroidia	o__Flavobacteriales	f__Flavobacteriaceae	g__MED-G11	s__
20110324_Bin_41_4-contigs	d__Bacteria	p__Bacteroidota	c__Bacteroidia	o__Flavobacteriales	f__Flavobacteriaceae	g__MED-G11	s__MED-G11

20110523_Bin_69_2-contigs	d__Bacteria	p__Bacteroidota	c__Bacteroidia	o__Flavobacteriales	f__Flavobacteriaceae	g__MS024-2A	s__
20110526_Bin_123_1-contigs	d__Bacteria	p__Bacteroidota	c__Bacteroidia	o__Flavobacteriales	f__Flavobacteriaceae	g__MS024-2A	s__
20120503_Bin_93_1-contigs	d__Bacteria	p__Bacteroidota	c__Bacteroidia	o__Flavobacteriales	f__Flavobacteriaceae	g__MS024-2A	s__MS024-2A
20120607_Bin_121_1-contigs	d__Bacteria	p__Bacteroidota	c__Bacteroidia	o__Flavobacteriales	f__Flavobacteriaceae	g__MS024-2A	s__MS024-2A
20160512_Bin_10_8-contigs	d__Bacteria	p__Bacteroidota	c__Bacteroidia	o__Flavobacteriales	f__Flavobacteriaceae	g__MS024-2A	s__MS024-2A
20120412_Bin_69_1-contigs	d__Bacteria	p__Bacteroidota	c__Bacteroidia	o__Flavobacteriales	f__Flavobacteriaceae	g__Nonlabens	s__
20160321_Bin_55_1-contigs	d__Bacteria	p__Bacteroidota	c__Bacteroidia	o__Flavobacteriales	f__Flavobacteriaceae	g__Polaribacter_A	s__
20100408_Bin_85_1-contigs	d__Bacteria	p__Bacteroidota	c__Bacteroidia	o__Flavobacteriales	f__Flavobacteriaceae	g__Polaribacter	s__
20100430_Bin_65_1-contigs	d__Bacteria	p__Bacteroidota	c__Bacteroidia	o__Flavobacteriales	f__Flavobacteriaceae	g__Polaribacter	s__
20100504_Bin_100-contigs	d__Bacteria	p__Bacteroidota	c__Bacteroidia	o__Flavobacteriales	f__Flavobacteriaceae	g__Polaribacter	s__
20100518_Bin_42_1-contigs	d__Bacteria	p__Bacteroidota	c__Bacteroidia	o__Flavobacteriales	f__Flavobacteriaceae	g__Polaribacter	s__
20110426_Bin_22-contigs	d__Bacteria	p__Bacteroidota	c__Bacteroidia	o__Flavobacteriales	f__Flavobacteriaceae	g__Polaribacter	s__
20110512_Bin_60_1-contigs	d__Bacteria	p__Bacteroidota	c__Bacteroidia	o__Flavobacteriales	f__Flavobacteriaceae	g__Polaribacter	s__
20110512_Bin_84_1-contigs	d__Bacteria	p__Bacteroidota	c__Bacteroidia	o__Flavobacteriales	f__Flavobacteriaceae	g__Polaribacter	s__
20110523_Bin_132_1-contigs	d__Bacteria	p__Bacteroidota	c__Bacteroidia	o__Flavobacteriales	f__Flavobacteriaceae	g__Polaribacter	s__
20110530_Bin_26-contigs	d__Bacteria	p__Bacteroidota	c__Bacteroidia	o__Flavobacteriales	f__Flavobacteriaceae	g__Polaribacter	s__
20120412_Bin_112-contigs	d__Bacteria	p__Bacteroidota	c__Bacteroidia	o__Flavobacteriales	f__Flavobacteriaceae	g__Polaribacter	s__
20120426_Bin_74_1-contigs	d__Bacteria	p__Bacteroidota	c__Bacteroidia	o__Flavobacteriales	f__Flavobacteriaceae	g__Polaribacter	s__
20120503_Bin_22_2-contigs	d__Bacteria	p__Bacteroidota	c__Bacteroidia	o__Flavobacteriales	f__Flavobacteriaceae	g__Polaribacter	s__
20120607_Bin_111_1-contigs	d__Bacteria	p__Bacteroidota	c__Bacteroidia	o__Flavobacteriales	f__Flavobacteriaceae	g__Polaribacter	s__
20160412_Bin_14_1-contigs	d__Bacteria	p__Bacteroidota	c__Bacteroidia	o__Flavobacteriales	f__Flavobacteriaceae	g__Polaribacter	s__
20160419_Bin_29_2-contigs	d__Bacteria	p__Bacteroidota	c__Bacteroidia	o__Flavobacteriales	f__Flavobacteriaceae	g__Polaribacter	s__
20160419_Bin_8_3-contigs	d__Bacteria	p__Bacteroidota	c__Bacteroidia	o__Flavobacteriales	f__Flavobacteriaceae	g__Polaribacter	s__
20160512_Bin_59_4-contigs	d__Bacteria	p__Bacteroidota	c__Bacteroidia	o__Flavobacteriales	f__Flavobacteriaceae	g__Polaribacter	s__
20100413_Bin_59-contigs	d__Bacteria	p__Bacteroidota	c__Bacteroidia	o__Flavobacteriales	f__Flavobacteriaceae	g__Polaribacter	s__Polaribacter
20110526_Bin_92-contigs	d__Bacteria	p__Bacteroidota	c__Bacteroidia	o__Flavobacteriales	f__Flavobacteriaceae	g__	s__
20110530_Bin_12_1-contigs	d__Bacteria	p__Bacteroidota	c__Bacteroidia	o__Flavobacteriales	f__Flavobacteriaceae	g__	s__
20160517_Bin_69_2-contigs	d__Bacteria	p__Bacteroidota	c__Bacteroidia	o__Flavobacteriales	f__Flavobacteriaceae	g__	s__
20100330_Bin_54_1-contigs	d__Bacteria	p__Bacteroidota	c__Bacteroidia	o__Flavobacteriales	f__Flavobacteriaceae	g__SCGC-AAA160-P02	s__
20100511_Bin_71_1-contigs	d__Bacteria	p__Bacteroidota	c__Bacteroidia	o__Flavobacteriales	f__Flavobacteriaceae	g__SCGC-AAA160-P02	s__
20110428_Bin_65_1-contigs	d__Bacteria	p__Bacteroidota	c__Bacteroidia	o__Flavobacteriales	f__Flavobacteriaceae	g__SCGC-AAA160-P02	s__
20110506_Bin_62_1-contigs	d__Bacteria	p__Bacteroidota	c__Bacteroidia	o__Flavobacteriales	f__Flavobacteriaceae	g__SCGC-AAA160-P02	s__
20110519_Bin_89-contigs	d__Bacteria	p__Bacteroidota	c__Bacteroidia	o__Flavobacteriales	f__Flavobacteriaceae	g__SCGC-AAA160-P02	s__
20160419_Bin_11_4-contigs	d__Bacteria	p__Bacteroidota	c__Bacteroidia	o__Flavobacteriales	f__Flavobacteriaceae	g__SCGC-AAA160-P02	s__
20160419_Bin_28_1-contigs	d__Bacteria	p__Bacteroidota	c__Bacteroidia	o__Flavobacteriales	f__Flavobacteriaceae	g__SCGC-AAA160-P02	s__
20100511_Bin_81_1-contigs	d__Bacteria	p__Bacteroidota	c__Bacteroidia	o__Flavobacteriales	f__Flavobacteriaceae	g__SCGC-AAA160-P02	s__SCGC-AAA160-P02
20110526_Bin_69_2-contigs	d__Bacteria	p__Bacteroidota	c__Bacteroidia	o__Flavobacteriales	f__Flavobacteriaceae	g__SGZJ01	s__
20160426_Bin_5_5-contigs	d__Bacteria	p__Bacteroidota	c__Bacteroidia	o__Flavobacteriales	f__Flavobacteriaceae	g__SGZJ01	s__SGZJ01
20120412_Bin_69_2-contigs	d__Bacteria	p__Bacteroidota	c__Bacteroidia	o__Flavobacteriales	f__Flavobacteriaceae	g__Subsaximicrobium	s__

20120607_Bin_128_1-contigs	d__Bacteria	p__Bacteroidota	c__Bacteroidia	o__Flavobacteriales	f__Flavobacteriaceae	g__UBA11891	s__UBA11891
20120607_Bin_109-contigs	d__Bacteria	p__Bacteroidota	c__Bacteroidia	o__Flavobacteriales	f__Flavobacteriaceae	g__UBA724	s__
20110523_Bin_35_1-contigs	d__Bacteria	p__Bacteroidota	c__Bacteroidia	o__Flavobacteriales	f__Flavobacteriaceae	g__UBA7428	s__
20110523_Bin_35_2-contigs	d__Bacteria	p__Bacteroidota	c__Bacteroidia	o__Flavobacteriales	f__Flavobacteriaceae	g__UBA7428	s__
20110523_Bin_59_1-contigs	d__Bacteria	p__Bacteroidota	c__Bacteroidia	o__Flavobacteriales	f__Flavobacteriaceae	g__UBA7428	s__
20120531_Bin_89_1-contigs	d__Bacteria	p__Bacteroidota	c__Bacteroidia	o__Flavobacteriales	f__Flavobacteriaceae	g__UBA7428	s__
20120607_Bin_103-contigs	d__Bacteria	p__Bacteroidota	c__Bacteroidia	o__Flavobacteriales	f__Flavobacteriaceae	g__UBA7428	s__
20160331_Bin_37_3-contigs	d__Bacteria	p__Bacteroidota	c__Bacteroidia	o__Flavobacteriales	f__Flavobacteriaceae	g__UBA7428	s__
20160502_Bin_17_4-contigs	d__Bacteria	p__Bacteroidota	c__Bacteroidia	o__Flavobacteriales	f__Flavobacteriaceae	g__UBA7428	s__
20160512_Bin_35_6-contigs	d__Bacteria	p__Bacteroidota	c__Bacteroidia	o__Flavobacteriales	f__Flavobacteriaceae	g__UBA7428	s__
20100408_Bin_88-contigs	d__Bacteria	p__Bacteroidota	c__Bacteroidia	o__Flavobacteriales	f__Flavobacteriaceae	g__UBA7428	s__UBA7428
20100504_Bin_107_1-contigs	d__Bacteria	p__Bacteroidota	c__Bacteroidia	o__Flavobacteriales	f__Flavobacteriaceae	g__UBA7428	s__UBA7428
20120607_Bin_104_1-contigs	d__Bacteria	p__Bacteroidota	c__Bacteroidia	o__Flavobacteriales	f__Flavobacteriaceae	g__UBA8316	s__
20120607_Bin_76_1-contigs	d__Bacteria	p__Bacteroidota	c__Bacteroidia	o__Flavobacteriales	f__Flavobacteriaceae	g__UBA8316	s__UBA8316
20120607_Bin_76_2-contigs	d__Bacteria	p__Bacteroidota	c__Bacteroidia	o__Flavobacteriales	f__Flavobacteriaceae	g__UBA8316	s__UBA8316
20120412_Bin_40-contigs	d__Bacteria	p__Bacteroidota	c__Bacteroidia	o__Flavobacteriales	f__Flavobacteriaceae	g__Ulvibacter_A	s__
20100330_Bin_61_1-contigs	d__Bacteria	p__Bacteroidota	c__Bacteroidia	o__Flavobacteriales	f__Flavobacteriaceae	g__Ulvibacter_B	s__
20100330_Bin_62_1-contigs	d__Bacteria	p__Bacteroidota	c__Bacteroidia	o__Flavobacteriales	f__Flavobacteriaceae	g__Ulvibacter_B	s__
20110428_Bin_109_1-contigs	d__Bacteria	p__Bacteroidota	c__Bacteroidia	o__Flavobacteriales	f__Flavobacteriaceae	g__Ulvibacter_B	s__
20110530_Bin_37-contigs	d__Bacteria	p__Bacteroidota	c__Bacteroidia	o__Flavobacteriales	f__Flavobacteriaceae	g__Ulvibacter_B	s__
20120524_Bin_77_1-contigs	d__Bacteria	p__Bacteroidota	c__Bacteroidia	o__Flavobacteriales	f__Flavobacteriaceae	g__Ulvibacter_B	s__
20120531_Bin_42-contigs	d__Bacteria	p__Bacteroidota	c__Bacteroidia	o__Flavobacteriales	f__Flavobacteriaceae	g__Ulvibacter_B	s__
20160316_Bin_63_4-contigs	d__Bacteria	p__Bacteroidota	c__Bacteroidia	o__Flavobacteriales	f__Flavobacteriaceae	g__Ulvibacter_B	s__
20160412_Bin_47_1-contigs	d__Bacteria	p__Bacteroidota	c__Bacteroidia	o__Flavobacteriales	f__Flavobacteriaceae	g__Ulvibacter_B	s__
20160419_Bin_47_2-contigs	d__Bacteria	p__Bacteroidota	c__Bacteroidia	o__Flavobacteriales	f__Flavobacteriaceae	g__Ulvibacter_B	s__
20120412_Bin_36_1-contigs	d__Bacteria	p__Bacteroidota	c__Bacteroidia	o__Flavobacteriales	f__Flavobacteriaceae	g__Ulvibacter	s__
20160512_Bin_61_3-contigs	d__Bacteria	p__Bacteroidota	c__Bacteroidia	o__Flavobacteriales	f__Flavobacteriaceae	g__Winogradskyella	s__
20120607_Bin_21_1-contigs	d__Bacteria	p__Bacteroidota	c__Bacteroidia	o__Flavobacteriales	f__GCA-002722245	g__	s__
20120426_Bin_4_1-contigs	d__Bacteria	p__Bacteroidota	c__Bacteroidia	o__Flavobacteriales	f__	g__	s__
20160517_Bin_55_15-contigs	d__Bacteria	p__Bacteroidota	c__Bacteroidia	o__Flavobacteriales	f__Salibacteraceae	g__	s__
20120405_Bin_61_1-contigs	d__Bacteria	p__Bacteroidota	c__Bacteroidia	o__Flavobacteriales	f__Salibacteraceae	g__SHAN690	s__
20120510_Bin_26_1-contigs	d__Bacteria	p__Bacteroidota	c__Bacteroidia	o__Flavobacteriales	f__Salibacteraceae	g__SHAN690	s__
20110523_Bin_111_1-contigs	d__Bacteria	p__Bacteroidota	c__Bacteroidia	o__Flavobacteriales	f__Salibacteraceae	g__SHAN690	s__SHAN690
20120510_Bin_116-contigs	d__Bacteria	p__Bacteroidota	c__Bacteroidia	o__Flavobacteriales	f__Salibacteraceae	g__SHAN690	s__SHAN690
20160426_Bin_35_2-contigs	d__Bacteria	p__Bacteroidota	c__Bacteroidia	o__Flavobacteriales	f__Salibacteraceae	g__SHAN690	s__SHAN690
20120405_Bin_31-contigs	d__Bacteria	p__Bacteroidota	c__Bacteroidia	o__Flavobacteriales	f__Salibacteraceae	g__UBA2108	s__
20120416_Bin_76_1-contigs	d__Bacteria	p__Bacteroidota	c__Bacteroidia	o__Flavobacteriales	f__Salibacteraceae	g__UBA4419	s__
20110421_Bin_62_1-contigs	d__Bacteria	p__Bacteroidota	c__Bacteroidia	o__Flavobacteriales	f__Salibacteraceae	g__UBA4419	s__UBA4419
20110530_Bin_91_1-contigs	d__Bacteria	p__Bacteroidota	c__Bacteroidia	o__Flavobacteriales	f__Salibacteraceae	g__UBA6770	s__UBA6770

20110428_Bin_106_1-contigs	d__Bacteria	p__Bacteroidota	c__Bacteroidia	o__Flavobacteriales	f__Schleiferiaceae	g__TMED14	s__
20120524_Bin_100_1-contigs	d__Bacteria	p__Bacteroidota	c__Bacteroidia	o__Flavobacteriales	f__Schleiferiaceae	g__TMED14	s__TMED14
20160316_Bin_8_7-contigs	d__Bacteria	p__Bacteroidota	c__Bacteroidia	o__Flavobacteriales	f__UA16	g__UA16	s__
20160412_Bin_32_2-contigs	d__Bacteria	p__Bacteroidota	c__Bacteroidia	o__Flavobacteriales	f__UA16	g__UA16	s__
20100303_Bin_111_1-contigs	d__Bacteria	p__Bacteroidota	c__Bacteroidia	o__Flavobacteriales	f__UA16	g__UA16	s__UA16
20120503_Bin_85_1-contigs	d__Bacteria	p__Bacteroidota	c__Bacteroidia	o__Flavobacteriales	f__UA16	g__UA16	s__UA16
20160426_Bin_40_5-contigs	d__Bacteria	p__Bacteroidota	c__Bacteroidia	o__Flavobacteriales	f__UA16	g__UA16	s__UA16
20160412_Bin_32_3-contigs	d__Bacteria	p__Bacteroidota	c__Bacteroidia	o__Flavobacteriales	f__UA16	g__UBA8752	s__
20160512_Bin_24_7-contigs	d__Bacteria	p__Bacteroidota	c__Bacteroidia	o__Flavobacteriales	f__UA16	g__UBA8752	s__
20120426_Bin_67_1-contigs	d__Bacteria	p__Bacteroidota	c__Bacteroidia	o__Flavobacteriales	f__UA16	g__UBA8752	s__UBA8752
20160331_Bin_10_3-contigs	d__Bacteria	p__Bacteroidota	c__Bacteroidia	o__Flavobacteriales	f__UA16	g__UBA8752	s__UBA8752
20100303_Bin_107_1-contigs	d__Bacteria	p__Bacteroidota	c__Bacteroidia	o__Flavobacteriales	f__UA16	g__UBA974	s__UBA974
20120531_Bin_86_1-contigs	d__Bacteria	p__Bacteroidota	c__Bacteroidia	o__Flavobacteriales	f__UBA10066	g__MED-G20	s__
20120607_Bin_116-contigs	d__Bacteria	p__Bacteroidota	c__Bacteroidia	o__Flavobacteriales	f__UBA10066	g__MED-G20	s__
20120607_Bin_122_1-contigs	d__Bacteria	p__Bacteroidota	c__Bacteroidia	o__Flavobacteriales	f__UBA10066	g__MED-G20	s__
20160426_Bin_19_1-contigs	d__Bacteria	p__Bacteroidota	c__Bacteroidia	o__Flavobacteriales	f__UBA10066	g__MED-G20	s__MED-G20
20100518_Bin_100-contigs	d__Bacteria	p__Bacteroidota	c__Bacteroidia	o__Flavobacteriales	f__UBA10066	g__	s__
20120416_Bin_46_2-contigs	d__Bacteria	p__Bacteroidota	c__Bacteroidia	o__Flavobacteriales	f__UBA10066	g__	s__
20160512_Bin_26_10-contigs	d__Bacteria	p__Bacteroidota	c__Bacteroidia	o__Flavobacteriales	f__UBA10066	g__	s__
20120510_Bin_29_1-contigs	d__Bacteria	p__Bacteroidota	c__Bacteroidia	o__Flavobacteriales	f__UBA10329	g__UBA10329	s__UBA10329
20100518_Bin_123-contigs	d__Bacteria	p__Bacteroidota	c__Bacteroidia	o__Flavobacteriales	f__UBA7430	g__UBA6772	s__
20100518_Bin_73_1-contigs	d__Bacteria	p__Bacteroidota	c__Bacteroidia	o__Flavobacteriales	f__UBA7430	g__UBA6772	s__UBA6772
20120607_Bin_21_2-contigs	d__Bacteria	p__Bacteroidota	c__Bacteroidia	o__Flavobacteriales	f__UBA7430	g__UBA6772	s__UBA6772
20160426_Bin_73_1-contigs	d__Bacteria	p__Bacteroidota	c__Bacteroidia	o__Flavobacteriales	f__UBA7430	g__UBA7430	s__UBA7430
20160502_Bin_30_5-contigs	d__Bacteria	p__Bacteroidota	c__Bacteroidia	o__Flavobacteriales	f__Vicingaceae	g__BCD18	s__
20160316_Bin_37_3-contigs	d__Bacteria	p__Bacteroidota	c__Bacteroidia	o__Flavobacteriales	f__Vicingaceae	g__GCA-002793235	s__
20100518_Bin_34-contigs	d__Bacteria	p__Bacteroidota	c__Bacteroidia	o__Flavobacteriales	f__Vicingaceae	g__UBA1494_A	s__
20160512_Bin_48_3-contigs	d__Bacteria	p__Bacteroidota	c__Bacteroidia	o__NS11-12g	f__	g__	s__
20160321_Bin_59-contigs	d__Bacteria	p__Bacteroidota	c__Bacteroidia	o__NS11-12g	f__UBA9320	g__UBA9320	s__
20160426_Bin_77_3-contigs	d__Bacteria	p__Bacteroidota	c__Bacteroidia	o__NS11-12g	f__UBA9320	g__UBA9320	s__
20160512_Bin_2_11-contigs	d__Bacteria	p__Bacteroidota	c__Bacteroidia	o__NS11-12g	f__UBA9320	g__UBA9320	s__
20160512_Bin_35_4-contigs	d__Bacteria	p__Bacteroidota	c__Bacteroidia	o__NS11-12g	f__UBA9320	g__UBA9320	s__
20100504_Bin_78-contigs	d__Bacteria	p__Bacteroidota	c__Bacteroidia	o__NS11-12g	f__UBA9320	g__UBA9320	s__UBA9320
20110426_Bin_16_2-contigs	d__Bacteria	p__Bacteroidota	c__Bacteroidia	o__NS11-12g	f__UBA9320	g__UBA9320	s__UBA9320
20110526_Bin_64-contigs	d__Bacteria	p__Bacteroidota	c__Bacteroidia	o__NS11-12g	f__UBA9320	g__UBA9320	s__UBA9320
20160321_Bin_74_5-contigs	d__Bacteria	p__Bacteroidota	c__Bacteroidia	o__NS11-12g	f__UBA9320	g__UBA9320	s__UBA9320
20160502_Bin_66_1-contigs	d__Bacteria	p__Bacteroidota	c__Bacteroidia	o__NS11-12g	f__UBA9320	g__UBA987	s__
20100430_Bin_69_1-contigs	d__Bacteria	p__Bacteroidota	c__Bacteroidia	o__NS11-12g	f__UBA955	g__	s__
20160512_Bin_35_1-contigs	d__Bacteria	p__Bacteroidota	c__Bacteroidia	o__NS11-12g	f__UBA955	g__	s__

20160419_Bin_12_1-contigs	d__Bacteria	p__Bacteroidota	c__Rhodothermia	o__Balneolales	f__Balneolaceae	g__Balneola	s__
20160419_Bin_12_2-contigs	d__Bacteria	p__Bacteroidota	c__Rhodothermia	o__Balneolales	f__Balneolaceae	g__Balneola	s__
20120405_Bin_52_2-contigs	d__Bacteria	p__Bacteroidota	c__Rhodothermia	o__Rhodothermales	f__UBA10348	g__	s__
20120405_Bin_52_1-contigs	d__Bacteria	p__Bacteroidota	c__Rhodothermia	o__Rhodothermales	f__UBA10348	g__UBA10348	s__UBA10348
20120412_Bin_15_2-contigs	d__Bacteria	p__Bdellovibrionota	c__Bacteriovoracia	o__Bacteriovoracales	f__Bacteriovoracaceae	g__	s__
20100430_Bin_56-contigs	d__Bacteria	p__Bdellovibrionota	c__Bacteriovoracia	o__Bacteriovoracales	f__Bacteriovoracaceae	g__UBA6144	s__
20120412_Bin_129_1-contigs	d__Bacteria	p__Bdellovibrionota	c__Bdellovibrionia	o__Bdellovibrionales	f__UBA1609	g__	s__
20120531_Bin_93_1-contigs	d__Bacteria	p__Campylobacterota	c__Campylobacteria	o__Campylobacterales	f__Arcobacteraceae	g__Poseidonibacter	s__
20100413_Bin_17-contigs	d__Bacteria	p__Cyanobacteria	c__Cyanobacteriia	o__Phormidesmiales	f__Phormidesmiaceae	g__Phormidesmis	s__
20100413_Bin_21_1-contigs	d__Bacteria	p__Cyanobacteria	c__Cyanobacteriia	o__Phormidesmiales	f__Phormidesmiaceae	g__Phormidesmis	s__
20120308_Bin_132_1-contigs	d__Bacteria	p__Marinisomatota	c__Marinisomatia	o__Marinisomatales	f__Marinisomataceae	g__Marinisoma	s__
20120405_Bin_65_1-contigs	d__Bacteria	p__Marinisomatota	c__Marinisomatia	o__Marinisomatales	f__Marinisomataceae	g__Marinisoma	s__Marinisoma
20120416_Bin_84_1-contigs	d__Bacteria	p__Marinisomatota	c__Marinisomatia	o__Marinisomatales	f__TCS55	g__TCS55	s__
20120510_Bin_97_2-contigs	d__Bacteria	p__Marinisomatota	c__Marinisomatia	o__Marinisomatales	f__UBA1611	g__	s__
20120531_Bin_27_3-contigs	d__Bacteria	p__Marinisomatota	c__Marinisomatia	o__Marinisomatales	f__UBA1611	g__	s__
20120607_Bin_68_2-contigs	d__Bacteria	p__Marinisomatota	c__Marinisomatia	o__Marinisomatales	f__UBA1611	g__TMED80	s__
20160512_Bin_60_7-contigs	d__Bacteria	p__Myxococcota	c__UBA796	o__UBA796	f__UBA796	g__	s__
20100504_Bin_90-contigs	d__Archaea	p__Nanoarchaeota	c__Nanoarchaeia	o__Woesearchaeales	f__BM511	g__	s__
20120416_Bin_85_1-contigs	d__Archaea	p__Nanoarchaeota	c__Nanoarchaeia	o__Woesearchaeales	f__	g__	s__
20120416_Bin_112_1-contigs	d__Bacteria	p__Patescibacteria	c__Gracilibacteria	o__UBA4473	f__UBA4473	g__UBA4473	s__UBA4473
20120531_Bin_98-contigs	d__Bacteria	p__Patescibacteria	c__Paceibacteria	o__UBA6257	f__2-01-FULL-56-20	g__	s__
20120412_Bin_156-contigs	d__Bacteria	p__Patescibacteria	c__Paceibacteria	o__UBA9983_A	f__GCA-2747955	g__	s__
20100423_Bin_77-contigs	d__Bacteria	p__Patescibacteria	c__Paceibacteria	o__UBA9983_A	f__UBA2163	g__OLB19	s__
20120412_Bin_137_1-contigs	d__Bacteria	p__Patescibacteria	c__Paceibacteria	o__UBA9983_A	f__UBA2163	g__OLB19	s__
20100430_Bin_91-contigs	d__Bacteria	p__Patescibacteria	c__Paceibacteria	o__UBA9983_A	f__UBA2163	g__OLB19	s__OLB19
20120412_Bin_65_1-contigs	d__Bacteria	p__Patescibacteria	c__Saccharimonadia	o__Saccharimonadales	f__Saccharimonadacea	g__	s__
20100330_Bin_74_1-contigs	d__Bacteria	p__Patescibacteria	c__Saccharimonadia	o__Saccharimonadales	f__Saccharimonadacea	g__UBA1020	s__
20100420_Bin_75_2-contigs	d__Bacteria	p__Patescibacteria	c__Saccharimonadia	o__Saccharimonadales	f__Saccharimonadacea	g__UBA6022	s__
20120405_Bin_9-contigs	d__Bacteria	p__Planctomycetota	c__Planctomycetes	o__Pirellulales	f__Pirellulaceae	g__JABAAE01	s__
20100504_Bin_5-contigs	d__Bacteria	p__Planctomycetota	c__UBA11346	o__	f__	g__	s__
20160517_Bin_10_6-contigs	d__Bacteria	p__Proteobacteria	c__Alphaproteobacteri	o__Bin95	f__Bin95	g__VMCJ01	s__
20160502_Bin_20_1-contigs	d__Bacteria	p__Proteobacteria	c__Alphaproteobacteri	o__Bin95	f__	g__	s__
20100511_Bin_69_1-contigs	d__Bacteria	p__Proteobacteria	c__Alphaproteobacteri	o__Caulobacterales	f__Caulobacteraceae	g__Brevundimonas	s__
20160412_Bin_55-contigs	d__Bacteria	p__Proteobacteria	c__Alphaproteobacteri	o__Caulobacterales	f__Caulobacteraceae	g__Brevundimonas	s__
20120412_Bin_9_1-contigs	d__Bacteria	p__Proteobacteria	c__Alphaproteobacteri	o__Caulobacterales	f__Maricaulaceae	g__Algimonas	s__
20160331_Bin_7_2-contigs	d__Bacteria	p__Proteobacteria	c__Alphaproteobacteri	o__Caulobacterales	f__Maricaulaceae	g__Hellea	s__
20160426_Bin_23_2-contigs	d__Bacteria	p__Proteobacteria	c__Alphaproteobacteri	o__Caulobacterales	f__Maricaulaceae	g__Hellea	s__
20120412_Bin_66_2-contigs	d__Bacteria	p__Proteobacteria	c__Alphaproteobacteri	o__Micavibrionales	f__Micavibrionaceae	g__JAACQG01	s__
20160512_Bin_57_1-contigs	d__Bacteria	p__Proteobacteria	c__Alphaproteobacteri	o__Micavibrionales	f__Micavibrionaceae	g__UBA1664	s__

20160321_Bin_13_2-contigs	d__Bacteria	p__Proteobacteria	c__Alphaproteobacteri	o__MPNO01	f__MPNO01	g__GCA-2689465	s__
20120607_Bin_133-contigs	d__Bacteria	p__Proteobacteria	c__Alphaproteobacteri	o__Parvibaculales	f__RS24	g__UBA8337	s__
20120607_Bin_69_1-contigs	d__Bacteria	p__Proteobacteria	c__Alphaproteobacteri	o__Parvibaculales	f__RS24	g__UBA8337	s__
20110426_Bin_1_22-contigs	d__Bacteria	p__Proteobacteria	c__Alphaproteobacteri	o__Pelagibacterales	f__Pelagibacteraceae	g__IMCC9063	s__IMCC9063
20160321_Bin_57_4-contigs	d__Bacteria	p__Proteobacteria	c__Alphaproteobacteri	o__Pelagibacterales	f__Pelagibacteraceae	g__IMCC9063	s__IMCC9063
20120416_Bin_7_11-contigs	d__Bacteria	p__Proteobacteria	c__Alphaproteobacteri	o__Pelagibacterales	f__Pelagibacteraceae	g__Pelagibacter_A	s__
20100303_Bin_49_2-contigs	d__Bacteria	p__Proteobacteria	c__Alphaproteobacteri	o__Pelagibacterales	f__Pelagibacteraceae	g__Pelagibacter	s__
20110428_Bin_1_24-contigs	d__Bacteria	p__Proteobacteria	c__Alphaproteobacteri	o__Pelagibacterales	f__Pelagibacteraceae	g__Pelagibacter	s__
20120405_Bin_1_10-contigs	d__Bacteria	p__Proteobacteria	c__Alphaproteobacteri	o__Pelagibacterales	f__Pelagibacteraceae	g__Pelagibacter	s__
20120416_Bin_7_16-contigs	d__Bacteria	p__Proteobacteria	c__Alphaproteobacteri	o__Pelagibacterales	f__Pelagibacteraceae	g__Pelagibacter	s__
20120416_Bin_7_6-contigs	d__Bacteria	p__Proteobacteria	c__Alphaproteobacteri	o__Pelagibacterales	f__Pelagibacteraceae	g__Pelagibacter	s__
20120510_Bin_2_13-contigs	d__Bacteria	p__Proteobacteria	c__Alphaproteobacteri	o__Pelagibacterales	f__Pelagibacteraceae	g__Pelagibacter	s__
20160502_Bin_2_4-contigs	d__Bacteria	p__Proteobacteria	c__Alphaproteobacteri	o__Pelagibacterales	f__Pelagibacteraceae	g__Pelagibacter	s__
20120531_Bin_19_1-contigs	d__Bacteria	p__Proteobacteria	c__Alphaproteobacteri	o__Pelagibacterales	f__Pelagibacteraceae	g__Pelagibacter	s__Pelagibacter
20110526_Bin_21_1-contigs	d__Bacteria	p__Proteobacteria	c__Alphaproteobacteri	o__Puniceispirillales	f__AAA536-G10	g__AAA536-G10	s__
20160321_Bin_5_3_2-contigs	d__Bacteria	p__Proteobacteria	c__Alphaproteobacteri	o__Puniceispirillales	f__AAA536-G10	g__AAA536-G10	s__
20110407_Bin_15-contigs	d__Bacteria	p__Proteobacteria	c__Alphaproteobacteri	o__Puniceispirillales	f__Puniceispirillaceae	g__UBA3439	s__
20160321_Bin_11_2-contigs	d__Bacteria	p__Proteobacteria	c__Alphaproteobacteri	o__Puniceispirillales	f__Puniceispirillaceae	g__UBA3439	s__
20160321_Bin_11_7-contigs	d__Bacteria	p__Proteobacteria	c__Alphaproteobacteri	o__Puniceispirillales	f__Puniceispirillaceae	g__UBA3439	s__
20160412_Bin_22_2-contigs	d__Bacteria	p__Proteobacteria	c__Alphaproteobacteri	o__Puniceispirillales	f__Puniceispirillaceae	g__UBA3439	s__
20120405_Bin_14_5-contigs	d__Bacteria	p__Proteobacteria	c__Alphaproteobacteri	o__Puniceispirillales	f__Puniceispirillaceae	g__UBA3439	s__UBA3439
20160517_Bin_40_2-contigs	d__Bacteria	p__Proteobacteria	c__Alphaproteobacteri	o__Puniceispirillales	f__Puniceispirillaceae	g__UBA4588	s__UBA4588
20120308_Bin_24_2-contigs	d__Bacteria	p__Proteobacteria	c__Alphaproteobacteri	o__Rhizobiales_B	f__TMED25	g__MED-G09	s__
20110526_Bin_130_1-contigs	d__Bacteria	p__Proteobacteria	c__Alphaproteobacteri	o__Rhodobacterales	f__Rhodobacteraceae	g__Amylibacter	s__
20100303_Bin_42_2-contigs	d__Bacteria	p__Proteobacteria	c__Alphaproteobacteri	o__Rhodobacterales	f__Rhodobacteraceae	g__Amylibacter	s__Amylibacter
20100408_Bin_75_1-contigs	d__Bacteria	p__Proteobacteria	c__Alphaproteobacteri	o__Rhodobacterales	f__Rhodobacteraceae	g__Amylibacter	s__Amylibacter
20120607_Bin_14_1-contigs	d__Bacteria	p__Proteobacteria	c__Alphaproteobacteri	o__Rhodobacterales	f__Rhodobacteraceae	g__Amylibacter	s__Amylibacter
20160512_Bin_25_4-contigs	d__Bacteria	p__Proteobacteria	c__Alphaproteobacteri	o__Rhodobacterales	f__Rhodobacteraceae	g__Amylibacter	s__Amylibacter
20160512_Bin_36_1-contigs	d__Bacteria	p__Proteobacteria	c__Alphaproteobacteri	o__Rhodobacterales	f__Rhodobacteraceae	g__Amylibacter	s__Amylibacter
20160419_Bin_13_3-contigs	d__Bacteria	p__Proteobacteria	c__Alphaproteobacteri	o__Rhodobacterales	f__Rhodobacteraceae	g__Ascidiaceihabitans	s__
20100408_Bin_33_1-contigs	d__Bacteria	p__Proteobacteria	c__Alphaproteobacteri	o__Rhodobacterales	f__Rhodobacteraceae	g__Ascidiaceihabitans	s__Ascidiaceihabitans
20110523_Bin_42_1-contigs	d__Bacteria	p__Proteobacteria	c__Alphaproteobacteri	o__Rhodobacterales	f__Rhodobacteraceae	g__CPC320	s__CPC320
20160517_Bin_7_6-contigs	d__Bacteria	p__Proteobacteria	c__Alphaproteobacteri	o__Rhodobacterales	f__Rhodobacteraceae	g__CPC320	s__CPC320
20120412_Bin_49-contigs	d__Bacteria	p__Proteobacteria	c__Alphaproteobacteri	o__Rhodobacterales	f__Rhodobacteraceae	g__Flavimaricola	s__
20110428_Bin_29_1-contigs	d__Bacteria	p__Proteobacteria	c__Alphaproteobacteri	o__Rhodobacterales	f__Rhodobacteraceae	g__GCA-002705045	s__
20160321_Bin_30_5-contigs	d__Bacteria	p__Proteobacteria	c__Alphaproteobacteri	o__Rhodobacterales	f__Rhodobacteraceae	g__GCA-2697345	s__GCA-2697345
20120405_Bin_34_1-contigs	d__Bacteria	p__Proteobacteria	c__Alphaproteobacteri	o__Rhodobacterales	f__Rhodobacteraceae	g__Lentibacter	s__Lentibacter
20120607_Bin_93_1-contigs	d__Bacteria	p__Proteobacteria	c__Alphaproteobacteri	o__Rhodobacterales	f__Rhodobacteraceae	g__LGRT01	s__LGRT01
20100408_Bin_26_1-contigs	d__Bacteria	p__Proteobacteria	c__Alphaproteobacteri	o__Rhodobacterales	f__Rhodobacteraceae	g__Planktomarina	s__

20120503_Bin_5_4-contigs	d__Bacteria	p__Proteobacteria	c__Alphaproteobacteri	o__Rhodobacterales	f__Rhodobacteraceae	g__Planktomarina	s__
20160512_Bin_7_10-contigs	d__Bacteria	p__Proteobacteria	c__Alphaproteobacteri	o__Rhodobacterales	f__Rhodobacteraceae	g__Planktomarina	s__
20110421_Bin_58_1-contigs	d__Bacteria	p__Proteobacteria	c__Alphaproteobacteri	o__Rhodobacterales	f__Rhodobacteraceae	g__Planktomarina	s__Planktomarina
20110523_Bin_97_1-contigs	d__Bacteria	p__Proteobacteria	c__Alphaproteobacteri	o__Rhodobacterales	f__Rhodobacteraceae	g__Planktotalea	s__
20160502_Bin_8_5-contigs	d__Bacteria	p__Proteobacteria	c__Alphaproteobacteri	o__Rhodobacterales	f__Rhodobacteraceae	g__Planktotalea	s__
20160426_Bin_44_1-contigs	d__Bacteria	p__Proteobacteria	c__Alphaproteobacteri	o__Rhodobacterales	f__Rhodobacteraceae	g__Planktotalea	s__Planktotalea
20120412_Bin_42_1-contigs	d__Bacteria	p__Proteobacteria	c__Alphaproteobacteri	o__Rhodobacterales	f__Rhodobacteraceae	g__Roseovarius	s__
20160517_Bin_25_4-contigs	d__Bacteria	p__Proteobacteria	c__Alphaproteobacteri	o__Rhodobacterales	f__Rhodobacteraceae	g__RR4-56	s__
20110404_Bin_29_1-contigs	d__Bacteria	p__Proteobacteria	c__Alphaproteobacteri	o__Rhodobacterales	f__Rhodobacteraceae	g__Sulfitobacter	s__
20110523_Bin_85_1-contigs	d__Bacteria	p__Proteobacteria	c__Alphaproteobacteri	o__Rhodobacterales	f__Rhodobacteraceae	g__Sulfitobacter	s__
20120412_Bin_12-contigs	d__Bacteria	p__Proteobacteria	c__Alphaproteobacteri	o__Rhodobacterales	f__Rhodobacteraceae	g__Sulfitobacter	s__
20120412_Bin_67-contigs	d__Bacteria	p__Proteobacteria	c__Alphaproteobacteri	o__Rhodobacterales	f__Rhodobacteraceae	g__Sulfitobacter	s__Sulfitobacter
20160419_Bin_37_1-contigs	d__Bacteria	p__Proteobacteria	c__Alphaproteobacteri	o__Rhodobacterales	f__Rhodobacteraceae	g__Sulfitobacter	s__Sulfitobacter
20110516_Bin_58_1-contigs	d__Bacteria	p__Proteobacteria	c__Alphaproteobacteri	o__Rhodobacterales	f__Rhodobacteraceae	g__Tateyamaria	s__
20160412_Bin_32_4-contigs	d__Bacteria	p__Proteobacteria	c__Alphaproteobacteri	o__Rhodobacterales	f__Rhodobacteraceae	g__Tateyamaria	s__
20160331_Bin_45_1-contigs	d__Bacteria	p__Proteobacteria	c__Alphaproteobacteri	o__Rhodobacterales	f__Rhodobacteraceae	g__Tateyamaria	s__Tateyamaria
20160321_Bin_44_2-contigs	d__Bacteria	p__Proteobacteria	c__Alphaproteobacteri	o__Rhodobacterales	f__Rhodobacteraceae	g__UBA10365	s__UBA10365
20120405_Bin_28-contigs	d__Bacteria	p__Proteobacteria	c__Alphaproteobacteri	o__Rhodobacterales	f__Rhodobacteraceae	g__WLWX01	s__
20110516_Bin_8_1-contigs	d__Bacteria	p__Proteobacteria	c__Alphaproteobacteri	o__Rhodobacterales	f__Rhodobacteraceae	g__Yoonia	s__
20110523_Bin_9_1-contigs	d__Bacteria	p__Proteobacteria	c__Alphaproteobacteri	o__Rhodobacterales	f__Rhodobacteraceae	g__Yoonia	s__
20100330_Bin_4_2-contigs	d__Bacteria	p__Proteobacteria	c__Alphaproteobacteri	o__Rhodobacterales	f__Rhodobacteraceae	g__Yoonia	s__Yoonia
20100303_Bin_41_1-contigs	d__Bacteria	p__Proteobacteria	c__Alphaproteobacteri	o__Rhodospirillales	f__2-02-FULL-58-16_A	g__GCA-2717285	s__
20120416_Bin_28_1-contigs	d__Bacteria	p__Proteobacteria	c__Alphaproteobacteri	o__Rhodospirillales	f__Casp-alpha2	g__UBA4479	s__UBA4479
20160426_Bin_22_2-contigs	d__Bacteria	p__Proteobacteria	c__Alphaproteobacteri	o__Rhodospirillales	f__Casp-alpha2	g__UBA4479	s__UBA4479
20160321_Bin_51_4-contigs	d__Bacteria	p__Proteobacteria	c__Alphaproteobacteri	o__Rickettsiales	f__JABSSG01	g__	s__
20100408_Bin_13_1-contigs	d__Bacteria	p__Proteobacteria	c__Alphaproteobacteri	o__Rickettsiales	f__Rickettsiaceae	g__GCA-2402195	s__
20160412_Bin_29_2-contigs	d__Bacteria	p__Proteobacteria	c__Alphaproteobacteri	o__Rickettsiales	f__Rickettsiaceae	g__GCA-2402195	s__
20110428_Bin_98_1-contigs	d__Bacteria	p__Proteobacteria	c__Alphaproteobacteri	o__Rickettsiales	f__UBA1997	g__UBA1997	s__UBA1997
20110523_Bin_92_1-contigs	d__Bacteria	p__Proteobacteria	c__Alphaproteobacteri	o__Sphingomonadales	f__Emcibacteraceae	g__UBA4441	s__
20110526_Bin_132_1-contigs	d__Bacteria	p__Proteobacteria	c__Alphaproteobacteri	o__Sphingomonadales	f__Emcibacteraceae	g__UBA4441	s__UBA4441
20160316_Bin_33_5-contigs	d__Bacteria	p__Proteobacteria	c__Alphaproteobacteri	o__Sphingomonadales	f__Sphingomonadaceae	g__Erythrobacter_A	s__Erythrobacter_A
20160512_Bin_76_1-contigs	d__Bacteria	p__Proteobacteria	c__Alphaproteobacteri	o__Sphingomonadales	f__Sphingomonadaceae	g__Sphingomonas	s__
20100303_Bin_27-contigs	d__Bacteria	p__Proteobacteria	c__Alphaproteobacteri	o__Sphingomonadales	f__Sphingomonadaceae	g__Sphingomonas	s__Sphingomonas
20160512_Bin_50_2-contigs	d__Bacteria	p__Proteobacteria	c__Alphaproteobacteri	o__Sphingomonadales	f__Sphingomonadaceae	g__Sphingomonas	s__Sphingomonas
20110324_Bin_70_1-contigs	d__Bacteria	p__Proteobacteria	c__Alphaproteobacteri	o__TMED109	f__TMED109	g__GCA-2684605	s__GCA-2684605
20120405_Bin_99_1-contigs	d__Bacteria	p__Proteobacteria	c__Alphaproteobacteri	o__UBA6615	f__UBA6615	g__	s__
20110519_Bin_80_1-contigs	d__Bacteria	p__Proteobacteria	c__Alphaproteobacteri	o__UBA7985	f__UBA7985	g__UBA7985	s__
20160512_Bin_39_1-contigs	d__Bacteria	p__Proteobacteria	c__Alphaproteobacteri	o__UBA7985	f__UBA7985	g__UBA7985	s__
20120510_Bin_61_1-contigs	d__Bacteria	p__Proteobacteria	c__Alphaproteobacteri	o__UBA7985	f__UBA7985	g__UBA7985	s__UBA7985

20160331_Bin_15_2-contigs	d__Bacteria	p__Proteobacteria	c__Gammaproteobact	o__Arenicellales	f__Arenicellaceae	g__Arenicella	s__Arenicella
20110512_Bin_4-contigs	d__Bacteria	p__Proteobacteria	c__Gammaproteobact	o__Burkholderiales	f__Burkholderiaceae	g__Acidovorax	s__Acidovorax
20110526_Bin_17_1-contigs	d__Bacteria	p__Proteobacteria	c__Gammaproteobact	o__Burkholderiales	f__Burkholderiaceae	g__RS62	s__RS62
20100303_Bin_45_1-contigs	d__Bacteria	p__Proteobacteria	c__Gammaproteobact	o__Burkholderiales	f__Methylophilaceae	g__BACL14	s__
20110523_Bin_2_1-contigs	d__Bacteria	p__Proteobacteria	c__Gammaproteobact	o__Burkholderiales	f__Methylophilaceae	g__BACL14	s__
20120412_Bin_30_2-contigs	d__Bacteria	p__Proteobacteria	c__Gammaproteobact	o__Burkholderiales	f__Methylophilaceae	g__BACL14	s__
20110324_Bin_35_2-contigs	d__Bacteria	p__Proteobacteria	c__Gammaproteobact	o__Burkholderiales	f__Methylophilaceae	g__BACL14	s__BACL14
20110526_Bin_113_1-contigs	d__Bacteria	p__Proteobacteria	c__Gammaproteobact	o__Burkholderiales	f__UBA3031	g__UBA3031	s__UBA3031
20100511_Bin_42-contigs	d__Bacteria	p__Proteobacteria	c__Gammaproteobact	o__Enterobacterales	f__Aeromonadaceae	g__Aeromonas	s__Aeromonas
20160331_Bin_18_1-contigs	d__Bacteria	p__Proteobacteria	c__Gammaproteobact	o__Enterobacterales	f__Alteromonadaceae	g__Alteromonas_E	s__
20110530_Bin_39_1-contigs	d__Bacteria	p__Proteobacteria	c__Gammaproteobact	o__Enterobacterales	f__Alteromonadaceae	g__Glaciecola	s__
20120412_Bin_29_2-contigs	d__Bacteria	p__Proteobacteria	c__Gammaproteobact	o__Enterobacterales	f__Alteromonadaceae	g__Glaciecola	s__
20120412_Bin_87_1-contigs	d__Bacteria	p__Proteobacteria	c__Gammaproteobact	o__Enterobacterales	f__Alteromonadaceae	g__Glaciecola	s__
20110328_Bin_54-contigs	d__Bacteria	p__Proteobacteria	c__Gammaproteobact	o__Enterobacterales	f__Alteromonadaceae	g__Glaciecola	s__Glaciecola
20160412_Bin_76_1-contigs	d__Bacteria	p__Proteobacteria	c__Gammaproteobact	o__Enterobacterales	f__Alteromonadaceae	g__Glaciecola	s__Glaciecola
20160316_Bin_57-contigs	d__Bacteria	p__Proteobacteria	c__Gammaproteobact	o__Enterobacterales	f__Alteromonadaceae	g__Idiomarina	s__
20110512_Bin_39-contigs	d__Bacteria	p__Proteobacteria	c__Gammaproteobact	o__Enterobacterales	f__Alteromonadaceae	g__Idiomarina	s__Idiomarina
20110523_Bin_59_2-contigs	d__Bacteria	p__Proteobacteria	c__Gammaproteobact	o__Enterobacterales	f__Alteromonadaceae	g__Paraglaciecola	s__
20120412_Bin_75-contigs	d__Bacteria	p__Proteobacteria	c__Gammaproteobact	o__Enterobacterales	f__Alteromonadaceae	g__Paraglaciecola	s__
20160419_Bin_63_3-contigs	d__Bacteria	p__Proteobacteria	c__Gammaproteobact	o__Enterobacterales	f__Alteromonadaceae	g__Paraglaciecola	s__
20100511_Bin_53-contigs	d__Bacteria	p__Proteobacteria	c__Gammaproteobact	o__Enterobacterales	f__Alteromonadaceae	g__Pararheinheimera	s__
20120607_Bin_2_1-contigs	d__Bacteria	p__Proteobacteria	c__Gammaproteobact	o__Enterobacterales	f__Alteromonadaceae	g__Psychrosphaera	s__Psychrosphaera
20110321_Bin_75_1-contigs	d__Bacteria	p__Proteobacteria	c__Gammaproteobact	o__Ga0077536	f__Ga0077536	g__UBA4457	s__UBA4457
20100303_Bin_58_1-contigs	d__Bacteria	p__Proteobacteria	c__Gammaproteobact	o__Ga0077536	f__Ga0077536	g__UBA981	s__UBA981
20120510_Bin_82_2-contigs	d__Bacteria	p__Proteobacteria	c__Gammaproteobact	o__GCA-002705445	f__GCA-002705445	g__GCA-2705445	s__
20110428_Bin_3_1-contigs	d__Bacteria	p__Proteobacteria	c__Gammaproteobact	o__GCA-002705445	f__GCA-002716945	g__GCA-2698665	s__
20100303_Bin_108_1-contigs	d__Bacteria	p__Proteobacteria	c__Gammaproteobact	o__Methylococcales	f__Cycloclasticaceae	g__Cycloclasticus	s__Cycloclasticus
20160316_Bin_33_2-contigs	d__Bacteria	p__Proteobacteria	c__Gammaproteobact	o__Nevskiales	f__Salinisphaeraceae	g__Salinisphaera	s__
20120524_Bin_116-contigs	d__Bacteria	p__Proteobacteria	c__Gammaproteobact	o__Nitrosococcales	f__Methylophagaceae	g__	s__
20160512_Bin_52_1-contigs	d__Bacteria	p__Proteobacteria	c__Gammaproteobact	o__Nitrosococcales	f__Methylophagaceae	g__	s__
20160512_Bin_52_3-contigs	d__Bacteria	p__Proteobacteria	c__Gammaproteobact	o__Nitrosococcales	f__Methylophagaceae	g__	s__
20160517_Bin_69_5-contigs	d__Bacteria	p__Proteobacteria	c__Gammaproteobact	o__Nitrosococcales	f__Methylophagaceae	g__	s__
20120308_Bin_133-contigs	d__Bacteria	p__Proteobacteria	c__Gammaproteobact	o__PS1	f__Thioglobaceae	g__Thioglobus_A	s__
20110324_Bin_7_5-contigs	d__Bacteria	p__Proteobacteria	c__Gammaproteobact	o__PS1	f__Thioglobaceae	g__Thioglobus	s__
20120426_Bin_6_3-contigs	d__Bacteria	p__Proteobacteria	c__Gammaproteobact	o__PS1	f__Thioglobaceae	g__Thioglobus	s__Thioglobus
20100430_Bin_20_1-contigs	d__Bacteria	p__Proteobacteria	c__Gammaproteobact	o__Pseudomonadales	f__Alcanivoracaceae	g__Alcanivorax	s__
20160316_Bin_33_4-contigs	d__Bacteria	p__Proteobacteria	c__Gammaproteobact	o__Pseudomonadales	f__Alcanivoracaceae	g__Alcanivorax	s__
20120412_Bin_122_1-contigs	d__Bacteria	p__Proteobacteria	c__Gammaproteobact	o__Pseudomonadales	f__Halieaceae	g__Luminiphilus	s__Luminiphilus
20120524_Bin_98_1-contigs	d__Bacteria	p__Proteobacteria	c__Gammaproteobact	o__Pseudomonadales	f__Halieaceae	g__Luminiphilus	s__Luminiphilus

20160502_Bin_1_3-contigs	d__Bacteria	p__Proteobacteria	c__Gammaproteobact	o__Pseudomonadales	f__Halieaceae	g__Luminiphilus	s__Luminiphilus
20120412_Bin_16_1-contigs	d__Bacteria	p__Proteobacteria	c__Gammaproteobact	o__Pseudomonadales	f__Halieaceae	g__Parahaliea	s__
20110519_Bin_83_1-contigs	d__Bacteria	p__Proteobacteria	c__Gammaproteobact	o__Pseudomonadales	f__HTCC2089	g__UBA4421	s__UBA4421
20110530_Bin_15_1-contigs	d__Bacteria	p__Proteobacteria	c__Gammaproteobact	o__Pseudomonadales	f__HTCC2089	g__UBA4421	s__UBA4421
20160512_Bin_19_2-contigs	d__Bacteria	p__Proteobacteria	c__Gammaproteobact	o__Pseudomonadales	f__HTCC2089	g__UBA4421	s__UBA4421
20160512_Bin_29_5-contigs	d__Bacteria	p__Proteobacteria	c__Gammaproteobact	o__Pseudomonadales	f__HTCC2089	g__UBA4421	s__UBA4421
20160316_Bin_51-contigs	d__Bacteria	p__Proteobacteria	c__Gammaproteobact	o__Pseudomonadales	f__HTCC2089	g__UBA4582	s__
20100518_Bin_10_2-contigs	d__Bacteria	p__Proteobacteria	c__Gammaproteobact	o__Pseudomonadales	f__HTCC2089	g__UBA4582	s__UBA4582
20160517_Bin_37_5-contigs	d__Bacteria	p__Proteobacteria	c__Gammaproteobact	o__Pseudomonadales	f__HTCC2089	g__UBA4582	s__UBA4582
20160517_Bin_40_1-contigs	d__Bacteria	p__Proteobacteria	c__Gammaproteobact	o__Pseudomonadales	f__HTCC2089	g__UBA9659	s__
20100303_Bin_46-contigs	d__Bacteria	p__Proteobacteria	c__Gammaproteobact	o__Pseudomonadales	f__HTCC2089	g__UBA9926	s__
20160331_Bin_23_7-contigs	d__Bacteria	p__Proteobacteria	c__Gammaproteobact	o__Pseudomonadales	f__HTCC2089	g__UBA9926	s__UBA9926
20100408_Bin_69_1-contigs	d__Bacteria	p__Proteobacteria	c__Gammaproteobact	o__Pseudomonadales	f__Nitrincolaceae	g__ASP10-02a	s__ASP10-02a
20100420_Bin_40_1-contigs	d__Bacteria	p__Proteobacteria	c__Gammaproteobact	o__Pseudomonadales	f__Porticoccaceae	g__HTCC2207	s__
20100423_Bin_61_1-contigs	d__Bacteria	p__Proteobacteria	c__Gammaproteobact	o__Pseudomonadales	f__Porticoccaceae	g__HTCC2207	s__
20100518_Bin_122-contigs	d__Bacteria	p__Proteobacteria	c__Gammaproteobact	o__Pseudomonadales	f__Porticoccaceae	g__HTCC2207	s__
20100518_Bin_32_1-contigs	d__Bacteria	p__Proteobacteria	c__Gammaproteobact	o__Pseudomonadales	f__Porticoccaceae	g__HTCC2207	s__
20110506_Bin_45_1-contigs	d__Bacteria	p__Proteobacteria	c__Gammaproteobact	o__Pseudomonadales	f__Porticoccaceae	g__HTCC2207	s__
20110509_Bin_66_1-contigs	d__Bacteria	p__Proteobacteria	c__Gammaproteobact	o__Pseudomonadales	f__Porticoccaceae	g__HTCC2207	s__
20110509_Bin_84_1-contigs	d__Bacteria	p__Proteobacteria	c__Gammaproteobact	o__Pseudomonadales	f__Porticoccaceae	g__HTCC2207	s__
20110523_Bin_106_1-contigs	d__Bacteria	p__Proteobacteria	c__Gammaproteobact	o__Pseudomonadales	f__Porticoccaceae	g__HTCC2207	s__
20120607_Bin_31_1-contigs	d__Bacteria	p__Proteobacteria	c__Gammaproteobact	o__Pseudomonadales	f__Porticoccaceae	g__HTCC2207	s__
20160331_Bin_23_1-contigs	d__Bacteria	p__Proteobacteria	c__Gammaproteobact	o__Pseudomonadales	f__Porticoccaceae	g__HTCC2207	s__
20160412_Bin_46_3-contigs	d__Bacteria	p__Proteobacteria	c__Gammaproteobact	o__Pseudomonadales	f__Porticoccaceae	g__HTCC2207	s__
20160412_Bin_52_3-contigs	d__Bacteria	p__Proteobacteria	c__Gammaproteobact	o__Pseudomonadales	f__Porticoccaceae	g__HTCC2207	s__
20160419_Bin_51_3-contigs	d__Bacteria	p__Proteobacteria	c__Gammaproteobact	o__Pseudomonadales	f__Porticoccaceae	g__HTCC2207	s__
20160426_Bin_12_1-contigs	d__Bacteria	p__Proteobacteria	c__Gammaproteobact	o__Pseudomonadales	f__Porticoccaceae	g__HTCC2207	s__
20160426_Bin_42_2-contigs	d__Bacteria	p__Proteobacteria	c__Gammaproteobact	o__Pseudomonadales	f__Porticoccaceae	g__HTCC2207	s__
20160502_Bin_10_2-contigs	d__Bacteria	p__Proteobacteria	c__Gammaproteobact	o__Pseudomonadales	f__Porticoccaceae	g__HTCC2207	s__
20160512_Bin_34_5-contigs	d__Bacteria	p__Proteobacteria	c__Gammaproteobact	o__Pseudomonadales	f__Porticoccaceae	g__HTCC2207	s__
20100408_Bin_49_1-contigs	d__Bacteria	p__Proteobacteria	c__Gammaproteobact	o__Pseudomonadales	f__Porticoccaceae	g__HTCC2207	s__HTCC2207
20100423_Bin_57_1-contigs	d__Bacteria	p__Proteobacteria	c__Gammaproteobact	o__Pseudomonadales	f__Porticoccaceae	g__HTCC2207	s__HTCC2207
20160502_Bin_33_3-contigs	d__Bacteria	p__Proteobacteria	c__Gammaproteobact	o__Pseudomonadales	f__Porticoccaceae	g__HTCC2207	s__HTCC2207
20160512_Bin_33_2-contigs	d__Bacteria	p__Proteobacteria	c__Gammaproteobact	o__Pseudomonadales	f__Porticoccaceae	g__HTCC2207	s__HTCC2207
20110526_Bin_155_1-contigs	d__Bacteria	p__Proteobacteria	c__Gammaproteobact	o__Pseudomonadales	f__Porticoccaceae	g__Porticoccus	s__Porticoccus
20160316_Bin_43_3-contigs	d__Bacteria	p__Proteobacteria	c__Gammaproteobact	o__Pseudomonadales	f__Porticoccaceae	g__Porticoccus	s__Porticoccus
20160321_Bin_19_1-contigs	d__Bacteria	p__Proteobacteria	c__Gammaproteobact	o__Pseudomonadales	f__Porticoccaceae	g__Porticoccus	s__Porticoccus
20110519_Bin_47_1-contigs	d__Bacteria	p__Proteobacteria	c__Gammaproteobact	o__Pseudomonadales	f__Pseudohongiellacea	g__OM182	s__OM182
20160331_Bin_1_7-contigs	d__Bacteria	p__Proteobacteria	c__Gammaproteobact	o__Pseudomonadales	f__Pseudohongiellacea	g__OM182	s__OM182

20120405_Bin_39_1-contigs	d__Bacteria	p__Proteobacteria	c__Gammaproteobact	o__Pseudomonadales	f__Pseudohongiellacea	g__RFVC01	s__
20110324_Bin_31_1-contigs	d__Bacteria	p__Proteobacteria	c__Gammaproteobact	o__Pseudomonadales	f__Pseudohongiellacea	g__UBA5109	s__
20120510_Bin_79_1-contigs	d__Bacteria	p__Proteobacteria	c__Gammaproteobact	o__Pseudomonadales	f__Pseudohongiellacea	g__UBA9145	s__
20100303_Bin_67_1-contigs	d__Bacteria	p__Proteobacteria	c__Gammaproteobact	o__Pseudomonadales	f__Pseudohongiellacea	g__UBA9145	s__UBA9145
20100511_Bin_24_1-contigs	d__Bacteria	p__Proteobacteria	c__Gammaproteobact	o__Pseudomonadales	f__Pseudomonadaceae	g__Pseudomonas_E	s__Pseudomonas_E
20100511_Bin_59-contigs	d__Bacteria	p__Proteobacteria	c__Gammaproteobact	o__Pseudomonadales	f__Pseudomonadaceae	g__Pseudomonas_E	s__Pseudomonas_E
20160419_Bin_15-contigs	d__Bacteria	p__Proteobacteria	c__Gammaproteobact	o__Pseudomonadales	f__Pseudomonadaceae	g__Pseudomonas_E	s__Pseudomonas_E
20160502_Bin_20_4-contigs	d__Bacteria	p__Proteobacteria	c__Gammaproteobact	o__Pseudomonadales	f__Pseudomonadaceae	g__Pseudomonas_E	s__Pseudomonas_E
20110321_Bin_56_1-contigs	d__Bacteria	p__Proteobacteria	c__Gammaproteobact	o__Pseudomonadales	f__Saccharospirillaceae	g__Oleispira	s__Oleispira
20120607_Bin_29_1-contigs	d__Bacteria	p__Proteobacteria	c__Gammaproteobact	o__Pseudomonadales	f__Saccharospirillaceae	g__Oleispira	s__Oleispira
20100423_Bin_43_1-contigs	d__Bacteria	p__Proteobacteria	c__Gammaproteobact	o__Pseudomonadales	f__Saccharospirillaceae	g__Reinekea	s__Reinekea
20160512_Bin_14_8-contigs	d__Bacteria	p__Proteobacteria	c__Gammaproteobact	o__Pseudomonadales	f__Spongiibacteraceae_	g__Oceanicoccus	s__Oceanicoccus
20100413_Bin_15_1-contigs	d__Bacteria	p__Proteobacteria	c__Gammaproteobact	o__Pseudomonadales	f__UBA6940	g__UBA6940	s__
20100430_Bin_2_1-contigs	d__Bacteria	p__Proteobacteria	c__Gammaproteobact	o__Pseudomonadales	f__UBA6940	g__UBA6940	s__
20120412_Bin_100_1-contigs	d__Bacteria	p__Proteobacteria	c__Gammaproteobact	o__Pseudomonadales	f__UBA6940	g__UBA6940	s__
20100423_Bin_12_1-contigs	d__Bacteria	p__Proteobacteria	c__Gammaproteobact	o__Pseudomonadales	f__UBA7434	g__UBA7434	s__
20110506_Bin_49_1-contigs	d__Bacteria	p__Proteobacteria	c__Gammaproteobact	o__Pseudomonadales	f__UBA7434	g__UBA7434	s__UBA7434
20120607_Bin_72_1-contigs	d__Bacteria	p__Proteobacteria	c__Gammaproteobact	o__SAR86	f__D2472	g__CACEJU01	s__
20160316_Bin_35_5-contigs	d__Bacteria	p__Proteobacteria	c__Gammaproteobact	o__SAR86	f__D2472	g__D2472	s__
20160426_Bin_18_7-contigs	d__Bacteria	p__Proteobacteria	c__Gammaproteobact	o__SAR86	f__D2472	g__D2472	s__
20110526_Bin_7_1-contigs	d__Bacteria	p__Proteobacteria	c__Gammaproteobact	o__SAR86	f__D2472	g__D2472	s__D2472
20100303_Bin_10_1-contigs	d__Bacteria	p__Proteobacteria	c__Gammaproteobact	o__SAR86	f__D2472	g__SCGC-AAA076-P13	s__
20100511_Bin_79_1-contigs	d__Bacteria	p__Proteobacteria	c__Gammaproteobact	o__SAR86	f__D2472	g__SCGC-AAA076-P13	s__
20160426_Bin_18_6-contigs	d__Bacteria	p__Proteobacteria	c__Gammaproteobact	o__SAR86	f__D2472	g__SCGC-AAA076-P13	s__
20120308_Bin_24_1-contigs	d__Bacteria	p__Proteobacteria	c__Gammaproteobact	o__SAR86	f__SAR86	g__GCA-2707915	s__
20120405_Bin_65_3-contigs	d__Bacteria	p__Proteobacteria	c__Gammaproteobact	o__SAR86	f__SAR86	g__	s__
20100303_Bin_68_1-contigs	d__Bacteria	p__Proteobacteria	c__Gammaproteobact	o__UBA4486	f__UBA4486	g__UBA7359	s__
20100303_Bin_105-contigs	d__Bacteria	p__Proteobacteria	c__Gammaproteobact	o__UBA4486	f__UBA4486	g__UBA7359	s__UBA7359
20110324_Bin_71_1-contigs	d__Bacteria	p__Proteobacteria	c__Gammaproteobact	o__UBA4486	f__UBA4486	g__UBA7359	s__UBA7359
20160426_Bin_28_5-contigs	d__Bacteria	p__Proteobacteria	c__Gammaproteobact	o__UBA4575	f__UBA4575	g__UBA4575	s__UBA4575
20110324_Bin_93_1-contigs	d__Bacteria	p__Proteobacteria	c__Gammaproteobact	o__UBA7366	f__UBA7366	g__UBA7366	s__UBA7366
20160502_Bin_19_1-contigs	d__Bacteria	p__Proteobacteria	c__Zetaproteobacteria	o__Mariprofundales	f__Mariprofundaceae	g__GCA-2401635	s__
20120308_Bin_19-contigs	d__Bacteria	p__SAR324	c__SAR324	o__SAR324	f__NAC60-12	g__JCVI-SCAAA005	s__JCVI-SCAAA005
20120531_Bin_91_1-contigs	d__Archaea	p__Thermoplasmatota	c__Poseidoniia_A	o__Poseidoniales	f__Poseidoniaceae	g__MGIIa-L1	s__MGIIa-L1
20160316_Bin_38_1-contigs	d__Archaea	p__Thermoplasmatota	c__Poseidoniia_A	o__Poseidoniales	f__Poseidoniaceae	g__MGIIa-L1	s__MGIIa-L1
20160316_Bin_75_1-contigs	d__Archaea	p__Thermoplasmatota	c__Poseidoniia_A	o__Poseidoniales	f__Poseidoniaceae	g__MGIIa-L2	s__MGIIa-L2
20110526_Bin_122_1-contigs	d__Archaea	p__Thermoplasmatota	c__Poseidoniia_A	o__Poseidoniales	f__Poseidoniaceae	g__Poseidonia	s__Poseidonia
20120308_Bin_120_1-contigs	d__Archaea	p__Thermoplasmatota	c__Poseidoniia_A	o__Poseidoniales	f__Poseidoniaceae	g__Poseidonia	s__Poseidonia
20110324_Bin_45_4-contigs	d__Archaea	p__Thermoplasmatota	c__Poseidoniia_A	o__Poseidoniales	f__Thalassarchaeaceae	g__MGIIb-O2	s__

20110321_Bin_88_1-contigs	d__Archaea	p__Thermoplasmatota	c__Poseidoniia_A	o__Poseidoniales	f__Thalassarchaeaceae	g__MGIIb-O2	s__MGIIb-O2
20120308_Bin_122_2-contigs	d__Archaea	p__Thermoplasmatota	c__Poseidoniia_A	o__Poseidoniales	f__Thalassarchaeaceae	g__MGIIb-O2	s__MGIIb-O2
20160316_Bin_11_1-contigs	d__Archaea	p__Thermoplasmatota	c__Poseidoniia_A	o__Poseidoniales	f__Thalassarchaeaceae	g__MGIIb-O3	s__
20110428_Bin_141_1-contigs	d__Archaea	p__Thermoplasmatota	c__Poseidoniia_A	o__Poseidoniales	f__Thalassarchaeaceae	g__MGIIb-O3	s__MGIIb-O3
20110321_Bin_55_1-contigs	d__Archaea	p__Thermoproteota	c__Nitrososphaeria	o__Nitrososphaerales	f__Nitrosopumilaceae	g__Nitrosopumilus	s__
20120405_Bin_51-contigs	d__Bacteria	p__Verrucomicrobiota	c__Lentisphaeria	o__Lentisphaerales	f__UBA12247	g__UBA12247	s__UBA12247
20100511_Bin_23_1-contigs	d__Bacteria	p__Verrucomicrobiota	c__Verrucomicrobiae	o__Opitutales	f__DSM-45221	g__BACL24	s__
20100511_Bin_80-contigs	d__Bacteria	p__Verrucomicrobiota	c__Verrucomicrobiae	o__Opitutales	f__DSM-45221	g__BACL24	s__
20160502_Bin_28_2-contigs	d__Bacteria	p__Verrucomicrobiota	c__Verrucomicrobiae	o__Opitutales	f__DSM-45221	g__BACL24	s__
20160512_Bin_25_7-contigs	d__Bacteria	p__Verrucomicrobiota	c__Verrucomicrobiae	o__Opitutales	f__DSM-45221	g__BACL24	s__
20120412_Bin_56_2-contigs	d__Bacteria	p__Verrucomicrobiota	c__Verrucomicrobiae	o__Opitutales	f__DSM-45221	g__BACL24	s__BACL24
20120503_Bin_68_1-contigs	d__Bacteria	p__Verrucomicrobiota	c__Verrucomicrobiae	o__Opitutales	f__DSM-45221	g__BACL24	s__BACL24
20100511_Bin_22_1-contigs	d__Bacteria	p__Verrucomicrobiota	c__Verrucomicrobiae	o__Opitutales	f__DSM-45221	g__UBA7441	s__UBA7441
20110421_Bin_23_1-contigs	d__Bacteria	p__Verrucomicrobiota	c__Verrucomicrobiae	o__Opitutales	f__MB11C04	g__MB11C04	s__
20110428_Bin_19_2-contigs	d__Bacteria	p__Verrucomicrobiota	c__Verrucomicrobiae	o__Opitutales	f__MB11C04	g__MB11C04	s__
20110526_Bin_2_13-contigs	d__Bacteria	p__Verrucomicrobiota	c__Verrucomicrobiae	o__Opitutales	f__MB11C04	g__MB11C04	s__
20110526_Bin_54_1-contigs	d__Bacteria	p__Verrucomicrobiota	c__Verrucomicrobiae	o__Opitutales	f__MB11C04	g__MB11C04	s__
20120416_Bin_11_2-contigs	d__Bacteria	p__Verrucomicrobiota	c__Verrucomicrobiae	o__Opitutales	f__MB11C04	g__MB11C04	s__
20120510_Bin_9_1-contigs	d__Bacteria	p__Verrucomicrobiota	c__Verrucomicrobiae	o__Opitutales	f__MB11C04	g__MB11C04	s__
20160502_Bin_38_2-contigs	d__Bacteria	p__Verrucomicrobiota	c__Verrucomicrobiae	o__Opitutales	f__MB11C04	g__MB11C04	s__
20160512_Bin_63_1-contigs	d__Bacteria	p__Verrucomicrobiota	c__Verrucomicrobiae	o__Opitutales	f__MB11C04	g__MB11C04	s__
20160331_Bin_58_1-contigs	d__Bacteria	p__Verrucomicrobiota	c__Verrucomicrobiae	o__Opitutales	f__MB11C04	g__MB11C04	s__MB11C04
20110506_Bin_6-contigs	d__Bacteria	p__Verrucomicrobiota	c__Verrucomicrobiae	o__Opitutales	f__Verruco-01	g__	s__
20100518_Bin_135_1-contigs	d__Bacteria	p__Verrucomicrobiota	c__Verrucomicrobiae	o__Verrucomicrobiales	f__Akkermansiaceae	g__UBA4581	s__UBA4581
20100518_Bin_140-contigs	d__Bacteria	p__Verrucomicrobiota	c__Verrucomicrobiae	o__Verrucomicrobiales	f__Akkermansiaceae	g__UBA4581	s__UBA4581
20100420_Bin_8_1-contigs	d__Bacteria	p__Verrucomicrobiota	c__Verrucomicrobiae	o__Verrucomicrobiales	f__Akkermansiaceae	g__UBA985	s__
20160512_Bin_18_4-contigs	d__Bacteria	p__Verrucomicrobiota	c__Verrucomicrobiae	o__Verrucomicrobiales	f__Akkermansiaceae	g__UBA985	s__
20100330_Bin_30_1-contigs	d__Bacteria	p__Verrucomicrobiota	c__Verrucomicrobiae	o__Verrucomicrobiales	f__Akkermansiaceae	g__UBA985	s__UBA985
20160321_Bin_63-contigs	d__Bacteria	p__Verrucomicrobiota	c__Verrucomicrobiae	o__Verrucomicrobiales	f__Akkermansiaceae	g__UBA985	s__UBA985
20160502_Bin_33_1-contigs	d__Bacteria	p__Verrucomicrobiota	c__Verrucomicrobiae	o__Verrucomicrobiales	f__Akkermansiaceae	g__UBA985	s__UBA985
20160502_Bin_33_8-contigs	d__Bacteria	p__Verrucomicrobiota	c__Verrucomicrobiae	o__Verrucomicrobiales	f__Akkermansiaceae	g__UBA985	s__UBA985
20160517_Bin_39_3-contigs	d__Bacteria	p__Verrucomicrobiota	c__Verrucomicrobiae	o__Verrucomicrobiales	f__Akkermansiaceae	g__UBA985	s__UBA985

Table S1 - peDNA/virome datasets

Sample	DRR accession	DRX accession	Sample Name	Environment	Reference
DRR173063	DRR173063	DRX163674	Wsample_nameSample_1_P2	Global Deep Ocean Conveyor	De Corte et al 2019
DRR173064	DRR173064	DRX163675	Wsample_nameSample_1_P3	Global Deep Ocean Conveyor	De Corte et al 2019
DRR173065	DRR173065	DRX163676	Wsample_nameSample_2_P2	Global Deep Ocean Conveyor	De Corte et al 2019
DRR173066	DRR173066	DRX163677	Wsample_nameSample_2_P3	Global Deep Ocean Conveyor	De Corte et al 2019
DRR173067	DRR173067	DRX163678	Wsample_nameSample_3_P2	Global Deep Ocean Conveyor	De Corte et al 2019
DRR173068	DRR173068	DRX163679	Wsample_nameSample_3_P3	Global Deep Ocean Conveyor	De Corte et al 2019
DRR173069	DRR173069	DRX163680	Wsample_nameSample_4_P2	Global Deep Ocean Conveyor	De Corte et al 2019
DRR173070	DRR173070	DRX163681	Wsample_nameSample_4_P3	Global Deep Ocean Conveyor	De Corte et al 2019
DRR173071	DRR173071	DRX163682	Wsample_nameSample_5_P2	Global Deep Ocean Conveyor	De Corte et al 2019
DRR173072	DRR173072	DRX163683	Wsample_nameSample_5_P3	Global Deep Ocean Conveyor	De Corte et al 2019
DRR173073	DRR173073	DRX163684	Wsample_nameSample_6_P2	Global Deep Ocean Conveyor	De Corte et al 2019
DRR173074	DRR173074	DRX163685	Wsample_nameSample_6_P3	Global Deep Ocean Conveyor	De Corte et al 2019
DRR173075	DRR173075	DRX163686	Wsample_nameSample_7_P2	Global Deep Ocean Conveyor	De Corte et al 2019
DRR173076	DRR173076	DRX163687	Wsample_nameSample_7_P3	Global Deep Ocean Conveyor	De Corte et al 2019
DRR173077	DRR173077	DRX163688	Wsample_nameSample_8_P2	Global Deep Ocean Conveyor	De Corte et al 2019
DRR173078	DRR173078	DRX163689	Wsample_nameSample_8_P3	Global Deep Ocean Conveyor	De Corte et al 2019
DRR173079	DRR173079	DRX163690	Wsample_nameSample_9_P2	Global Deep Ocean Conveyor	De Corte et al 2019
DRR173080	DRR173080	DRX163691	Wsample_nameSample_9_P3	Global Deep Ocean Conveyor	De Corte et al 2019
DRR173081	DRR173081	DRX163692	Wsample_nameSample_10_P2	Global Deep Ocean Conveyor	De Corte et al 2019
DRR173082	DRR173082	DRX163693	Wsample_nameSample_10_P3	Global Deep Ocean Conveyor	De Corte et al 2019
DRR173083	DRR173083	DRX163694	Wsample_nameSample_11_P2	Global Deep Ocean Conveyor	De Corte et al 2019
DRR173084	DRR173084	DRX163695	Wsample_nameSample_11_P3	Global Deep Ocean Conveyor	De Corte et al 2019
DRR173085	DRR173085	DRX163696	Wsample_nameSample_12_P2	Global Deep Ocean Conveyor	De Corte et al 2019
DRR173086	DRR173086	DRX163697	Wsample_nameSample_12_P3	Global Deep Ocean Conveyor	De Corte et al 2019
DRR173087	DRR173087	DRX163698	Wsample_nameSample_13_P2	Global Deep Ocean Conveyor	De Corte et al 2019
DRR173088	DRR173088	DRX163699	Wsample_nameSample_13_P3	Global Deep Ocean Conveyor	De Corte et al 2019
DRR173089	DRR173089	DRX163700	Wsample_nameSample_14_P2	Global Deep Ocean Conveyor	De Corte et al 2019
DRR173090	DRR173090	DRX163701	Wsample_nameSample_14_P3	Global Deep Ocean Conveyor	De Corte et al 2019
DRR173091	DRR173091	DRX163702	Wsample_nameSample_15_P2	Global Deep Ocean Conveyor	De Corte et al 2019
DRR173092	DRR173092	DRX163703	Wsample_nameSample_15_P3	Global Deep Ocean Conveyor	De Corte et al 2019

DRR173093	DRR173093	DRX163704	Wsample_nameSample_16_P2	Global Deep Ocean Conveyor	De Corte et al 2019
DRR173094	DRR173094	DRX163705	Wsample_nameSample_16_P3	Global Deep Ocean Conveyor	De Corte et al 2019

Sample	Comment	SRA Accession	Reference
med4	Culture	SRR1013844, SRR	Biller 2014
SRR1015709	Environmental	SRR1015709,	Biller 2015
SRR1015707	Environmental	SRR1015707	Biller 2016

Sample label	Station	Depth	ERR Accession	Sample Accession	Run Accession	Corresponding	Reference
100_DCM	TARA_100	50	ERR599359	ERS491897	ERR599359	Station100_DCM	Gregory & Zayed et al 2019
100_SRF	TARA_100	5	ERR599342	ERS491851	ERR599342; SM	Station100_SUR_A	Gregory & Zayed et al 2019
102_DCM	TARA_102	40	ERR599349	ERS492035	ERR599349	Station102_DCM_	Gregory & Zayed et al 2019
102_MES	TARA_102	480	ERR599346	ERS491995	ERR599346	Station102_MES_	Gregory & Zayed et al 2019
102_SRF	TARA_102	5	ERR599353	ERS491955	ERR599353; SM	Station102_SUR_A	Gregory & Zayed et al 2019
109_DCM	TARA_109	30	ERR594387	ERS492198	ERR594387, ERR	Station109_DCM_	Gregory & Zayed et al 2019
109_SRF	TARA_109	5	ERR594412	ERS492160	ERR594412	Station109_SUR_A	Gregory & Zayed et al 2019
11_SRF	TARA_011	9	ERR599383	ERS488253	ERR599383	Station11_SUR_A	Gregory & Zayed et al 2019
111_DCM	TARA_111	90	ERR599369	ERS492376	ERR599369	Station111_DCM_	Gregory & Zayed et al 2019
111_MES	TARA_111	350	ERR2752161	ERS492399	ERR599368; Res	Station111_MES_(	Gregory & Zayed et al 2019
111_SRF	TARA_111	5	ERR599357	ERS492342	ERR599357	Station111_SUR_A	Gregory & Zayed et al 2019
122_DCM	TARA_122	115	ERR599377	ERS492714	ERR599377	Station122_DCM_	Gregory & Zayed et al 2019
122_MES	TARA_122	600	ERR599341	ERS492692	ERR599341; Res	Station122_MES_(	Gregory & Zayed et al 2019
122_SRF	TARA_122	5	ERR599380	ERS492657	ERR599380	Station122_SUR_A	Gregory & Zayed et al 2019
123_MIX	TARA_123	150	ERR599348	ERS492797	ERR599348, ERR	Station123_MXL_	Gregory & Zayed et al 2019
123_SRF	TARA_123	5	ERR599378	ERS492748	ERR599381, ERR	Station123_SUR_A	Gregory & Zayed et al 2019
124_MIX	TARA_124	120	ERR599367	ERS492877	ERR599367	Station124_MXL_	Gregory & Zayed et al 2019
124_SRF	TARA_124	5	ERR599354	ERS492843	ERR599354	Station124_SUR_A	Gregory & Zayed et al 2019
125_MIX	TARA_125	140	ERR599338	ERS492941	ERR599338	Station125_MXL_	Gregory & Zayed et al 2019
125_SRF	TARA_125	5	ERR599337	ERS492903	ERR599337	Station125_SUR_A	Gregory & Zayed et al 2019
137_DCM	TARA_137	40	ERR599340	ERS493684	ERR599340	Station137_DCM_	Gregory & Zayed et al 2019
137_MES	TARA_137	375	ERR599366	ERS493718	ERR599366; Res	Station137_MES_(	Gregory & Zayed et al 2019
137_SRF	TARA_137	5	ERR599363	ERS493651	ERR599363	Station137_SUR_A	Gregory & Zayed et al 2019
138_DCM	TARA_138	60	ERR599345	ERS493803	ERR599345	Station138_DCM_	Gregory & Zayed et al 2019
138_MES	TARA_138	450	ERR599365	ERS493835	ERR599365; Res	Station138_MES_(	Gregory & Zayed et al 2019

138_SRF	TARA_138		5	ERR599355	ERS493767	ERR599355	Station138_SUR_A	Gregory & Zayed et al 2019
155_DCM	TARA_155		40	ERR2762100	ERS1307964	ERR2762100, ER	Station155_DCM_	Gregory & Zayed et al 2019
155_MES	TARA_155		751	ERR2762102	ERS1307938	ERR2762102, ER	Station155_MES_	Gregory & Zayed et al 2019
155_SRF	TARA_155		5	ERR2762104	ERS1307890	ERR2762104, ER	Station155_SUR_A	Gregory & Zayed et al 2019
158_DCM	TARA_158		25	ERR2762106	ERS1308437	ERR2762106, ER	Station158_DCM_	Gregory & Zayed et al 2019
158_MES	TARA_158		651	ERR2762108	ERS1308472	ERR2762108, ER	Station158_MES_	Gregory & Zayed et al 2019
158_SRF	TARA_158		5	ERR2762110	ERS1308390	ERR2762110, ER	Station158_SUR_A	Gregory & Zayed et al 2019
163_MES	TARA_163		499	ERR2762112	ERS1308540	ERR2762112, ER	Station163_MES_	Gregory & Zayed et al 2019
163_SRF	TARA_163		5	ERR2762114	ERS1308500	ERR2762114, ER	Station163_SUR_A	Gregory & Zayed et al 2019
168_DCM	TARA_168		40	ERR2762116	ERS1308623	ERR2762116, ER	Station168_DCM_	Gregory & Zayed et al 2019
168_MIX	TARA_168	10-100		ERR2762118	ERS1308664	ERR2762118, ER	Station168_IZZ_A	Gregory & Zayed et al 2019
168_SRF	TARA_168		5	ERR2762120	ERS1308567	ERR2762120, ER	Station168_SUR_A	Gregory & Zayed et al 2019
173_DCM	TARA_173		35	ERR2762122	ERS1308744	ERR2762122, ER	Station173_DCM_	Gregory & Zayed et al 2019
173_SRF	TARA_173		5	ERR2762124	ERS1308704	ERR2762124, ER	Station173_SUR_A	Gregory & Zayed et al 2019
175_MIX	TARA_175	10-100		ERR2762126	ERS1308858	ERR2762126, ER	Station175_IZZ_A	Gregory & Zayed et al 2019
175_SRF	TARA_175		5	ERR2762128	ERS1308779	ERR2762128, ER	Station175_SUR_A	Gregory & Zayed et al 2019
175_ZZZ	TARA_175		199	ERR2762130	ERS1308812	ERR2762130, ER	Station175_ZZZ_A	Gregory & Zayed et al 2019
178_SRF	TARA_178		5	ERR2762132	ERS1308894	ERR2762132, ER	Station178_SUR_A	Gregory & Zayed et al 2019
18_DCM	TARA_018		60	ERR594352	ERS488354	ERR594352	Station18_DCM_A	Gregory & Zayed et al 2019
18_SRF	TARA_018		5	ERR594358	ERS488340	ERR594358	Station18_SUR_A	Gregory & Zayed et al 2019
180_DCM	TARA_180		30	ERR2762136	ERS1308985	ERR2762136, ER	Station180_SUR_A	Gregory & Zayed et al 2019
180_SRF	TARA_180		5	ERR2762134	ERS1308938	ERR2762134, ER	Station180_ZZZ_A	Gregory & Zayed et al 2019
188_DCM	TARA_188		17	ERR2762138	ERS1309076	ERR2762138, ER	Station188_DCM_	Gregory & Zayed et al 2019
188_SRF	TARA_188		5	ERR2762140	ERS1309035	ERR2762140, ER	Station188_SUR_A	Gregory & Zayed et al 2019
189_DCM	TARA_189		20	ERR2762142	ERS1309171	ERR2762142, ER	Station189_DCM_	Gregory & Zayed et al 2019
189_MES	TARA_189		299	ERR2762144	ERS1309196	ERR2762144, ER	Station189_MES_	Gregory & Zayed et al 2019
191_SRF	TARA_191		5	ERR2762148	ERS1309221	ERR2762148, ER	Station191_SUR_A	Gregory & Zayed et al 2019
193_SRF	TARA_193		5	ERR2762150	ERS1309264	ERR2762150, ER	Station193_SUR_A	Gregory & Zayed et al 2019
194_DCM	TARA_194		35	ERR2762153	ERS1309369	ERR2762153, ER	Station194_DCM_	Gregory & Zayed et al 2019
194_SRF	TARA_194		5	ERR2762155	ERS1309308	ERR2762155, ER	Station194_SUR_A	Gregory & Zayed et al 2019
196_SRF	TARA_196		5	ERR2762157	ERS1309397	ERR2762157, ER	Station196_SUR_A	Gregory & Zayed et al 2019
201_DCM	TARA_201		36	ERR2762159	ERS1309532	ERR2762159, ER	Station201_DCM_	Gregory & Zayed et al 2019
201_MES	TARA_201		441	ERR2762162	ERS1309450	ERR2762162, ER	Station201_MES_	Gregory & Zayed et al 2019
201_SRF	TARA_201		5	ERR2762164	ERS1309476	ERR2762164, ER	Station201_SUR_A	Gregory & Zayed et al 2019

205_MES	TARA_205	491	ERR2762167	ERS1309617	ERR2762167, EF	Station205_MES_	Gregory & Zayed et al 2019
205_SRF	TARA_205	5	ERR2762169	ERS1309560	ERR2762169, EF	Station205_SUR_A	Gregory & Zayed et al 2019
206_MES	TARA_206	411	ERR2762171	ERS1309687	ERR2762171, EF	Station206_MES_	Gregory & Zayed et al 2019
206_SRF	TARA_206	5	ERR2762173	ERS1309642	ERR2762173, EF	Station206_SUR_A	Gregory & Zayed et al 2019
208_SRF	TARA_208	5	ERR2762176	ERS1309715	ERR2762176, EF	Station208_SUR_A	Gregory & Zayed et al 2019
209_MES	TARA_209	351	ERR2762178	ERS1309809	ERR2762178, EF	Station209_MES_	Gregory & Zayed et al 2019
209_SRF	TARA_209	5	ERR2762180	ERR2762180	ERR2762180, EF	Station209_SUR_A	Gregory & Zayed et al 2019
210_MES	TARA_210	391	ERR2762183	ERS1309881	ERR2762183, EF	Station210_MES_	Gregory & Zayed et al 2019
210_SRF	TARA_210	5	ERR2762186	ERS1309836	ERR2762186, EF	Station210_SUR_A	Gregory & Zayed et al 2019
22_SRF	TARA_022	5	ERR594406	ERS488448	ERR594406,ERF	Station22_SUR_A	Gregory & Zayed et al 2019
23_DCM	TARA_023	55	ERR594408	ERS478007	ERR594408	Station23_DCM_A	Gregory & Zayed et al 2019
25_DCM	TARA_025	50	ERR594375	ERS488518	ERR594375	Station25_DCM_A	Gregory & Zayed et al 2019
25_SRF	TARA_025	5	ERR594396	ERS488499	ERR594396	Station25_SUR_A	Gregory & Zayed et al 2019
30_DCM	TARA_030	70	ERR594405	ERS478052	ERR594405	Station30_DCM_A	Gregory & Zayed et al 2019
31_SRF	TARA_031	5	ERR594401	ERS488558	ERR594401, ERI	Station31_SUR_A	Gregory & Zayed et al 2019
32_DCM	TARA_032	80	ERR594360	ERS488613	ERR594360	Station32_DCM_A	Gregory & Zayed et al 2019
32_SRF	TARA_032	5	ERR594393	ERS488589,ERS488590	ERR594393	Station32_SUR_A	Gregory & Zayed et al 2019
33_SRF	TARA_033	5	ERR599350	ERS488636	ERR599350	Station33_SUR_A	Gregory & Zayed et al 2019
34_DCM	TARA_034	60	ERR594390	ERS488701	ERR594390	Station34_DCM_A	Gregory & Zayed et al 2019
34_SRF	TARA_034	5	ERR594370	ERS488673	ERR594370, ERI	Station34_SUR_A	Gregory & Zayed et al 2019
36_DCM	TARA_036	17	ERR594402	ERS488757	ERR594402	Station36_DCM_A	Gregory & Zayed et al 2019
36_SRF	TARA_036	5	ERR594369	ERS488737	ERR594369	Station36_SUR_A	Gregory & Zayed et al 2019
37_MES	TARA_037	600	ERR599372	ERS488791	ERR599372; Res	Station37_MES_C	Gregory & Zayed et al 2019
38_DCM	TARA_038	25	ERR594386	ERS488836,ERS488837	ERR594386, ERI	Station38_DCM_A	Gregory & Zayed et al 2019
38_MES	TARA_038	340	ERR599379	ERS488867	ERR599379	Station38_MES_A	Gregory & Zayed et al 2019
38_SRF	TARA_038	5	ERR594400	ERS488813,ERS488814	ERR594400, ERI	Station38_SUR_A	Gregory & Zayed et al 2019
39_DCM	TARA_039	25	ERR594363	ERS488929,ERS488930	ERR594363, ERI	Station39_DCM_A	Gregory & Zayed et al 2019
39_MES	TARA_039	270	ERR599360	ERS488959	ERR599360; Res	Station39_MES_C	Gregory & Zayed et al 2019
39_SRF	TARA_039	5	ERR594356	ERS488892,ERS488893	ERR594356, ERI	Station39_SUR_A	Gregory & Zayed et al 2019
41_DCM	TARA_041	60	ERR594371	ERS489084	ERR594371, ERI	Station41_DCM_A	Gregory & Zayed et al 2019
41_SRF	TARA_041	5	ERR594384	ERS489059	ERR594384	Station41_SUR_A	Gregory & Zayed et al 2019
42_DCM	TARA_042	80	ERR594413	ERS489148	ERR594413	Station42_DCM_A	Gregory & Zayed et al 2019
42_SRF	TARA_042	5	ERR594398	ERS489113	ERR594398, ERI	Station42_SUR_A	Gregory & Zayed et al 2019
45_SRF	TARA_045	5	ERR599356	ERS489257	ERR599356	Station45_SUR_A	Gregory & Zayed et al 2019

46_SRF	TARA_046	5	ERR594376	ERS489285	ERR594376	Station46_SUR_A	Gregory & Zayed et al 2019
48_SRF	TARA_048	5	ERR599344	ERS489332	ERR599344	Station48_SUR_A	Gregory & Zayed et al 2019
52_DCM	TARA_052	75	ERR594394	ERS489603	ERR594394	Station52_DCM_A	Gregory & Zayed et al 2019
52_SRF	TARA_052	5	ERR599358	ERS489552	ERR599358	Station52_SUR_A	Gregory & Zayed et al 2019
56_MES	TARA_056	1000	ERR599371	ERS489729	ERR599371; Res	Station56_MES_C	Gregory & Zayed et al 2019
56_SRF	TARA_056	5	ERR599376	ERS489724	ERR599376	Station56_SUR_A	Gregory & Zayed et al 2019
58_DCM	TARA_058	66	ERR599362	ERS489867	ERR599362	Station58_DCM_A	Gregory & Zayed et al 2019
62_SRF	TARA_062	5	ERR599339	ERS489904	ERR599339	Station62_SUR_A	Gregory & Zayed et al 2019
64_DCM	TARA_064	65	ERR594385	ERS490026	ERR594385	Station64_DCM_A	Gregory & Zayed et al 2019
64_MES	TARA_064	1000	ERR599351	ERS489996	ERR599351; Res	Station64_MES_C	Gregory & Zayed et al 2019
64_SRF	TARA_064	5	ERR594392	ERS489943	ERR594392	Station64_SUR_A	Gregory & Zayed et al 2019
65_DCM	TARA_065	30	ERR594414	ERS490120,ERS490121	ERR594414, ERI	Station65_DCM_A	Gregory & Zayed et al 2019
65_SRF	TARA_065	5	ERR594361	ERS490053,ERS490054	ERR594361, ERI	Station65_SUR_A	Gregory & Zayed et al 2019
66_DCM	TARA_066	30	ERR594389	ERS490180	ERR594389	Station66_DCM_A	Gregory & Zayed et al 2019
66_SRF	TARA_066	5	ERR594362	ERS490142	ERR594362	Station66_SUR_A	Gregory & Zayed et al 2019
67_SRF	TARA_067	5	ERR594395	ERS490204,ERS490205	ERR594395, ERI	Station67_SUR_A	Gregory & Zayed et al 2019
68_DCM	TARA_068	50	ERR594415	ERS490320	ERR594415	Station68_DCM_A	Gregory & Zayed et al 2019
68_MES	TARA_068	700	ERR599382	ERS490245	ERR599382; Res	Station68_MES_C	Gregory & Zayed et al 2019
68_SRF	TARA_068	5	ERR594391	ERS490285	ERR594391	Station68_SUR_A	Gregory & Zayed et al 2019
70_MES	TARA_070	800	ERR594407	ERS490388	ERR594407; Res	Station70_MES_C	Gregory & Zayed et al 2019
70_SRF	TARA_070	5	ERR594353	ERS490346	ERR594353	Station70_SUR_A	Gregory & Zayed et al 2019
72_DCM	TARA_072	100	ERR594379	ERS490494	ERR594379	Station72_DCM_A	Gregory & Zayed et al 2019
72_MES	TARA_072	800	ERR594388	ERS490522	ERR594388	Station72_MES_A	Gregory & Zayed et al 2019
72_SRF	TARA_072	5	ERR594364	ERS490452	ERR594364	Station72_SUR_A	Gregory & Zayed et al 2019
76_DCM	TARA_076	150	ERR594355	ERS490610	ERR594355	Station76_DCM_A	Gregory & Zayed et al 2019
76_MES	TARA_076	800	ERR599370	ERS490648	ERR599370, ERI	Station76_MES_C	Gregory & Zayed et al 2019
76_SRF	TARA_076	5	ERR594354	ERS490557	ERR594354	Station76_SUR_A	Gregory & Zayed et al 2019
78_DCM	TARA_078	120	ERR599374	ERS490704	ERR599374	Station78_DCM_A	Gregory & Zayed et al 2019
78_MES	TARA_078	800	ERR599361	ERS490727	ERR599361; Res	Station78_MES_C	Gregory & Zayed et al 2019
78_SRF	TARA_078	5	ERR594411	ERS490676	ERR594411	Station78_SUR_A	Gregory & Zayed et al 2019
82_DCM	TARA_082	40	ERR594409	ERS490953	ERR594409; Res	Station82_DCM_C	Gregory & Zayed et al 2019
82_SRF	TARA_082	5	ERR599384	ERS490900	ERR599384; Res	Station82_SUR_C	Gregory & Zayed et al 2019
84_SRF	TARA_084	5	ERR599352	ERS491016	ERR599352; Res	Station84_SUR_C	Gregory & Zayed et al 2019
85_DCM	TARA_085	90	ERR594377	ERS491107	ERR594377; Res	Station85_DCM_C	Gregory & Zayed et al 2019

85_MES	TARA_085	790	ERR599373	ERS491123	ERR599373; Res	Station85_MES_C	Gregory & Zayed et al 2019
85_SRF	TARA_085	5	ERR599364	ERS491063	ERR599364; Res	Station85_SUR_C	Gregory & Zayed et al 2019

Sample ID	Assay Type	Bases	BioSample	Host	strain	Reference
SRR12115929	WGS	880192938	SAMN15366360	Polaribacter sp. HaHaR_3_91	Polaribacter pha	Bartlau et al 2022
SRR12115930	WGS	39747730	SAMN15366359	Polaribacter sp. HaHaR_3_91	Polaribacter pha	Bartlau et al 2023
SRR12115931	WGS	1158470845	SAMN15366358	Polaribacter sp. R2A056_3_3	Polaribacter pha	Bartlau et al 2024
SRR12115932	WGS	44836007	SAMN15366357	Polaribacter sp. R2A056_3_3	Polaribacter pha	Bartlau et al 2025
SRR12115933	WGS	28514404	SAMN15366356	Polaribacter sp. R2A056_3_3	Polaribacter pha	Bartlau et al 2026
SRR12115934	WGS	1005521333	SAMN15366394	Winograskyella sp. HaHa_3_2	Winogradskyella	Bartlau et al 2027
SRR12115938	WGS	43718581	SAMN15366355	Polaribacter sp. R2A056_3_3	Polaribacter pha	Bartlau et al 2028
SRR12115949	WGS	31870241	SAMN15366354	Polaribacter sp. R2A056_3_3	Polaribacter pha	Bartlau et al 2029
SRR12115951	WGS	542201880	SAMN15366379	Maribacter forsetii	Polaribacter pha	Bartlau et al 2030
SRR12115961	WGS	36049615	SAMN15366370	Olleya sp. HaHaR_3_96	Olleya phage Ha	Bartlau et al 2031
SRR12115962	WGS	463276751	SAMN15366369	Tenacibaculum sp. AHE14PA	Tenacibaculum p	Bartlau et al 2032
SRR12115963	WGS	35728282	SAMN15366368	Polaribacter sp. HaHaR_3_91	Polaribacter pha	Bartlau et al 2033
SRR12115964	WGS	35441812	SAMN15366367	Polaribacter sp. HaHaR_3_91	Polaribacter pha	Bartlau et al 2034
SRR12115965	WGS	34918749	SAMN15366366	Polaribacter sp. HaHaR_3_91	Polaribacter pha	Bartlau et al 2035
SRR12115966	WGS	27595684	SAMN15366365	Polaribacter sp. HaHaR_3_91	Polaribacter pha	Bartlau et al 2036
SRR12115967	WGS	34332034	SAMN15366364	Polaribacter sp. HaHaR_3_91	Polaribacter pha	Bartlau et al 2037
SRR12115968	WGS	33112293	SAMN15366363	Polaribacter sp. HaHaR_3_91	Polaribacter pha	Bartlau et al 2038
SRR12115969	WGS	23363874	SAMN15366362	Polaribacter sp. HaHaR_3_91	Polaribacter pha	Bartlau et al 2039
SRR12115970	WGS	30861043	SAMN15366361	Polaribacter sp. HaHaR_3_91	Polaribacter pha	Bartlau et al 2040

Declaration of Contribution

Declaration on the contribution of the candidate to a multi-author article/manuscript which is included as a chapter in the present doctoral thesis

Candidate: Dominik Lücking

Title: Mechanisms of Genetic Transfer Between Archaea via Membrane Vesicles, Plasmid Vesicles and Viruses

Contribution of the candidate in percentage to the total workload (up to 100%):

Task	**Chapter 2**	**Chapter 3**	**Chapter 4**
Experimental concept and design	50 %	NA	50 %
Experimental work / bioinformatic work	85 %	NA	90 %
Data analysis and interpretation	85 %	90 %	80 %
Preparation of figures and tables	90 %	90 %	90 %
Drafting of the manuscript	80 %	90 %	80 %

We declare that all authors in the articles/manuscripts have been informed on and did not object to the listed contributions of the candidate.

City / Date: _____

Signatures

_____ _____

Candidate Supervisor

Versicherung an Eides Statt

Universität Bremen

Universität Bremen | Fachbereich 02 | Postfach 33 04 40, 28334 Bremen

Universität Bremen
Fachbereich 2
Prüfungsamt Chemie
z. Hd. Frau Frauke Ernst
Leobener Straße

28359 Bremen
Deutschland

Prüfungsamt
Chemie

Frauke Ernst
Geschäftsstelle
Fachbereich 02
Leobener Str. / NW2
D-28359 Bremen

Verwaltungspavillon 06
Tel. 0421 218-62802
Fax 0421 218-9862802
frauke.ernst@uni-bremen.de
www.fb2.uni-bremen.de

Versicherung an Eides Statt

Name, Vorname	Lücking, Dominik
Matrikel-Nr.	
Straße	Am Dobben 35
Ort, PLZ	Bremen, 28203

Ich, Dominik Lücking (Vorname, Name)

versichere an Eides Statt durch meine Unterschrift, dass ich die vorstehende Arbeit selbständig und ohne fremde Hilfe angefertigt und alle Stellen, die ich wörtlich dem Sinne nach aus Veröffentlichungen entnommen habe, als solche kenntlich gemacht habe, mich auch keiner anderen als der angegebenen Literatur oder sonstiger Hilfsmittel bedient habe.
Ich versichere an Eides Statt, dass ich die vorgenannten Angaben nach bestem Wissen und Gewissen gemacht habe und dass die Angaben der Wahrheit entsprechen und ich nichts verschwiegen habe.
Die Strafbarkeit einer falschen eidesstattlichen Versicherung ist mir bekannt, namentlich die Strafandrohung gemäß § 156 StGB bis zu drei Jahren Freiheitsstrafe oder Geldstrafe bei vorsätzlicher Begehung der Tat bzw. gemäß § 161 Abs. 1 StGB bis zu einem Jahr Freiheitsstrafe oder Geldstrafe bei fahrlässiger Begehung.

_____
Ort, Datum / Unterschrift