Bogs, Bugs, Borgs, and Bacteriophages: Metagenomic and Biochemical Insights into the Enigmatic World of Extrachromosomal Genetic Elements

by

Basem Al-Shayeb

A dissertation submitted in partial satisfaction of the

requirements for the degree of

Doctor of Philosophy

in

Microbiology

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor Jillian F. Banfield, Co-chair
Professor Jennifer A. Doudna, Co-chair
Professor Kimberley D. Seed
Professor Donald C. Rio

Spring 2022

Abstract

Bogs, Bugs, Borgs, and Bacteriophages: Metagenomic and Biochemical Insights into the Enigmatic World of Extrachromosomal Genetic Elements
by

Basem Al-Shayeb

Doctor of Philosophy in Microbiology

University of California, Berkeley

Professor Jillian F. Banfield, Chair
Professor Jennifer A. Doudna, Chair

As a Ph.D. Candidate and National Science Foundation Predoctoral Fellow at the University of California, Berkeley, working in the labs of Dr. Jillian Banfield and Dr. Jennifer Doudna, I have dedicated my Ph.D. to the discovery and investigation of novel extrachromosomal elements and tools for biotechnological applications through a combination of genomics and biochemistry.

The first chapter of this thesis uncovers 10 new clades of the largest bacteriophages ever found across many ecosystems worldwide, with genome sizes rivaling those of the smallest bacteria. We found that the phages are not only equipped with a wide variety of features typically associated with life and cellular organisms such as ribosomal proteins, tRNA synthetases and initiation and elongation factors, but also some of the viruses intriguingly utilize alternative genetic codes to translate their proteins. Notably, I discovered that the huge phage genomes encode CRISPR-Cas systems that may be used for inter-viral warfare. Some of these are miniature, previously undescribed CRISPR-Cas systems that are about half of the size of Cas9. This work was published in *Nature*.

The second chapter describes the analysis and testing of one of the novel phage CRISPR-Cas systems, CRISPR-CasΦ, that we have shown can indeed exclude mobile elements such as plasmids from infecting the same host cell despite their small size, and can be applicable for programmable genome editing in bacterial, plant, and mammalian cells as the most compact functional CRISPR-Cas systems to date, potentially circumventing cell delivery barriers exhibited with CRISPR-Cas9 gene editing. Intriguingly, the CRISPR-CasΦ system exhibited a previously undescribed consolidation of chemistries in a Cas nuclease as the RuvC active site mediated both double-stranded DNA cleavage and RNA processing in a metal-dependent manner. This work was published in *Science*.

The third chapter examines the discovery of enigmatic giant linear extrachromosomal elements, which we refer to as “Borgs”, inhabiting archaea. These elements that are about 1 Mbp long were recovered from multiple environments and

may play a previously unrecognized role in controlling greenhouse gas emissions. Their genomes are represented in 2 uneven replichores, with inverted repeats >1.5kbp long on either end and dozens of tandem repeats throughout their genomes. They contain no obvious hallmarks of previously reported viruses or plasmids, and ~80% of their genes consist of novel and uncharacterized proteins. Our analysis of horizontal gene transfer suggests that many ribosomal, metabolic, and extracellular electron transfer genes and operons recently transferred from their hosts, including the *nif* operon for Nitrogen fixation and the MCR complex which was recently proposed to be involved in oxidation of methane. Evidence also suggests recent recombination events between different Borgs presumably within the same host cell. This work is currently in review at *Nature*.

The fourth chapter describes an open-science effort for robust viral discovery computational pipelines driven by the COVID-19 pandemic. Working with a truly collaborative global team of bioinformaticians, this work describes the discovery of over 100,000 species of viruses to which I have contributed novel huge phage genomes. This manuscript was published in *Nature.*

The final chapter examines the discovery of thousands of viruses encoding CRISPR-Cas systems, many of which target competing cryptic mobile elements that are predicted to infect the same bacterial hosts. From genome-resolved metagenomics and bioinformatics-enabled phylogenetic insights to biochemistry, structural biology, and eukaryotic genome editing, I describe hundreds of novel hypercompact and divergent CRISPR-Cas systems, with special consideration towards the novel Casλ family. Casλ possesses an aberrant RNA structure reminiscent of a naturally-occurring sgRNA and processes its own crRNA at the 3' end, unlike any previously described single-RNA CRISPR-Cas system. The tertiary structure determined via cryo-EM reveals the machinery for PAM recognition, hybrid assembly, and DNA cleavage. RNA-targeting systems on viruses lack crucial residues or accessory proteins that would, in their bacterial counterparts, result in acute abortive infection, suggesting a potential strategy for phage systems to maintain host viability while preventing superinfection. In addition to their streamlined nature that is advantageous for cellular delivery, hypercompact phage systems can produce efficient genome editing in endogenous genes in mammalian and plant cells on par with, or in some cases, exceeding gold-standard Cas12a editing, demonstrating significant utility for biotechnological applications.

Overall, this dissertation describes the use of a combination of bioinformatics and biochemistry to shed light on gigantic bacterial viruses, the proteins they encode on their genomes, and elements such as Borgs which we are only beginning to understand. Huge phages and Borgs represent little-known biology, the platforms for which are distinct from previously known systems, and significantly broaden our overall understanding of "non-living" selfish genetic entities. The metagenomic discovery and biochemical and structural characterization of hypercompact CRISPR-Cas systems in addition to analyses of their genome editing utility in eukaryotic cells pave the road for efficacious delivery of treatments to human cells in the near future.

Table of contents

Introduction

Bacteriophages (phages), viruses of bacteria, are considered distinct from cellular life due to their inability to conduct most biological processes required for reproduction. They are agents of ecosystem change because they prey upon specific bacterial populations, mediate lateral gene transfer, alter host metabolism, and redistribute bacterially-derived compounds via cell lysis (Breitbart et al., 2018; Emerson et al., 2018; Rascovan et al., 2016). They spread antibiotic resistance (Balcazar, 2014) and disperse pathogenicity factors that cause disease in humans and animals (Brown-Jaque et al., 2018; Penadés et al., 2015). Some phages can even impact Earth's climate as they decrease methane oxidation rates by infection and lysis of methane-oxidizing bacteria (Lee et al., 2021), and others with the critical subunit of MMO (Chen et al., 2020) likely increase the ability of their host bacteria to conserve energy during phage replication. Most knowledge about phages is based on laboratory-studied examples, the vast majority of which have genomes a few 10s of kbp in length. This motivated a more comprehensive analysis of microbial communities to evaluate the prevalence, diversity, and ecosystem distribution of phage with large genomes. This research expands our understanding of phage biodiversity and reveals the wide variety of ecosystems in which phage have genomes with sizes that rival those of small celled bacteria (Castelle et al., 2018; Nakabachi et al., 2006; Pérez-Brocal et al., 2006). We postulate that these phages have evolved a distinct 'life' strategy that involves extensive interception and augmentation of host biology while they replicate their huge genomes.

Akin to phages, other extra-chromosomal elements also have the capacity to impact global biogeochemical cycles. Methane (CH_4) is a greenhouse gas roughly 30 times more potent than carbon dioxide (CO_2), and approximately 1 gigaton is produced annually by methanogenic (methane-producing) archaea that inhabit anoxic environments(Thauer et al., 2008). The efflux of methane into the atmosphere is mitigated by methane-oxidizing microorganisms (methanotrophs). In oxic environments, CH_4 is consumed by aerobic bacteria that use a methane monooxygenase (MMO) and O_2 as terminal electron acceptor(Hanson and Hanson, 1996), whereas in anoxic environments anaerobic methanotrophic archaea (ANME) use a reverse methanogenesis pathway to oxidize CH_4, the key enzyme of which is methyl-CoM reductase (MCR)(Boetius et al., 2000; Hallam et al., 2003). Some ANMEs rely on a syntrophic partner to couple CH_4 oxidation to the reduction of terminal electron acceptors, yet *Methanoperedens* (ANME-2d, phylum *Euryarchaeota*) can directly couple CH_4 oxidation to the reduction of iron, nitrate or manganese(Ettwig et al., 2016; Leu et al., 2020). In this thesis, we report the discovery of novel extrachromosomal elements (ECEs) that are inferred to replicate within *Methanoperedens* spp. Their numerous and diverse metabolism-relevant genes, huge size, and distinctive genome architecture distinguish these archaeal ECEs from all previously reported elements associated with archaea(Ausiannikava et al., 2018; Ng et al., 1998; Wang et al., 2015) and from bacteriophages, which typically have one or a few biogeochemically relevant genes(Anantharaman et al., 2014; Lindell et al., 2004). We hypothesize that these novel ECEs may substantially impact the capacity of *Methanoperedens* spp to oxidize methane.

CRISPR-Cas systems are adaptive immune systems present in 40% of bacteria and 85% of archaea to confer resistance against invading extrachromosomal elements such as plasmids or the aforementioned viruses (Makarova et al., 2019). While CRISPR-Cas9 is undoubtedly the most well-known and utilized CRISPR-associated RNA-guided nuclease to date, there is an exceptionally high diversity of CRISPR-Cas systems that have been discovered in recent years. Early in my Ph.D., I attended a short introductory seminar that was, in some ways, the most pivotal twenty minutes in my graduate career. Dr Kimberley Seed described a type I-F system that was inserted into the genomes of a group of ICP1-related phages infecting *Vibrio cholerae.* She had reported these phages in 2013 (Seed et al., 2013). This motivated me to survey the abundance, as well as the sequence and biochemical diversity of CRISPR-Cas systems throughout the virosphere, which remained poorly understood. As a result, we reported the presence of CRISPR-Cas systems that were hijacked by several clades of huge as well as run-of-the-mill phages, where we posited their role in inter-viral warfare to abrogate superinfection, and validated systems found, sometimes exclusively, in bacteriophages as functional CRISPR-Cas systems within bacteria. Moreover, I studied several novel systems biochemically and structurally and established them as programmable gene editing tools in human and plant cells. With advantages in both vector-based delivery into cells and a wider range of targetable genomic sequences, these hypercompact systems provide a powerful addition to the CRISPR-Cas toolbox.

Acknowledgments

This body of work would not have been possible without the support of multitudes of people. Through historic wildfires and a global pandemic, lab shifts and power outages, we've made it this far together.

First and foremost, to my family whose hard work and tremendous sacrifices enabled me to be here today. Your perseverance despite all odds and your belief in my abilities continue to inspire me to do the best that I possibly can. Thank you to my school teachers, whose faith in me far exceeded my own as a child, always pushing me to dream bigger and aim higher. I never could have imagined, as a child in Cairo, that I would one day have my work cited in a Nobel Lecture.

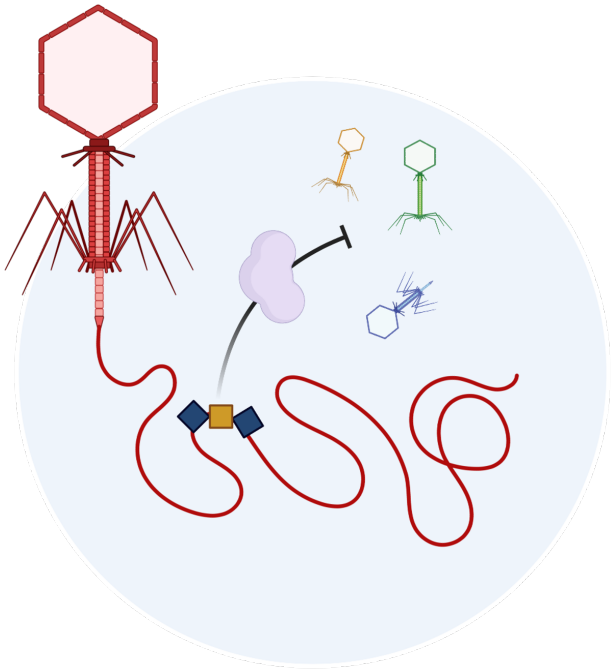
Thank you to Dr. Jill Banfield, for providing much guidance, feedback, and green tea in our meetings. Thank you for teaching me the joys of discovering fascinating new phenomena from the most unsuspecting of sources: from muddy fields to a pipe by the highway to groundwater wells and everything in between. Thank you to Dr. Jennifer Doudna, who has been very encouraging of my progress and development from day one and treated me as a colleague rather than a research subordinate. Thank you for providing an invigorating lab environment to work in, and for giving me the freedom to explore what I found to be most interesting and trusting that I would find fascinating biological questions to work on for myself and for others, often including me in meetings to provide my input. Your infectious excitement describing Cas13 snipping RNA at our first meeting sold me instantly. Through both labs, I was exposed to the scientific delights of working on challenging questions, mentally and physically, in very competitive fields.

I also would be remiss not to thank Dr. Kimberley Seed. Early in my Ph.D., I attended her introductory seminar that was, in several ways, one of the most pivotal moments in my graduate career and contributed to my research directions via my quals and thesis committees. Thank you to David Colognori and Enrique Lin Shiao, who quickly grew into close friends. Thank you to my rotation mentor Gavin Knott, for his enthusiasm and validation of my often-crazy ideas that encouraged future pursuits. Patrick Pausch, Kasia Soczek, and Petr Skopintsev, for being driven and reliable collaborators. Rohan Sachdeva, my first desk mate and partner-in-crime early on. Joy Wang, who was always willing to hear about interesting new Cas1-related findings and pursue mechanistic questions of mutual interest. Alex Crits-Christoph, Alex Jaffe, Victor Reyes-Umana, and Hector Trujillo for their helpful advice navigating my way through grad school. Thank you to Brady Cress, Departmental advisor Dr. Arash Komeili, IGI Director Dr. Brad Ringeisen, and Thesis Committee member Dr. Donald Rio, who were generous with their guidance and support throughout the years.

1 Chapter 1: Clades of huge phages from across Earth's ecosystems

Basem Al-Shayeb # , Rohan Sachdeva # , Lin-Xing Chen, Fred Ward, Patrick Munk, Audra Devoto, Cindy J. Castelle, Matthew R. Olm, Keith Bouma-Gregson, Yuki Amano, Christine He, Raphaël Méheust, Brandon Brooks, Alex Thomas, Adi Lavy, Paula Matheus-Carnevali, Christine Sun, Daniela S. A. Goltsman, Mikayla A. Borton, Allison Sharrar, Alexander L. Jaffe, Tara C. Nelson, Rose Kantor, Ray Keren, Katherine R. Lane, Ibrahim F. Farag, Shufei Lei, Kari Finstad, Ronald Amundson, Karthik Anantharaman, Jinglie Zhou, Alexander J. Probst, Mary E. Power, Susannah G. Tringe, Wen-Jun Li, Kelly Wrighton, Sue Harrison, Michael Morowitz, David A. Relman, Jennifer A. Doudna, Anne-Catherine Lehours, Lesley Warren, Jamie H. D. Cate, Joanne M. Santini & Jillian F. Banfield.

Published in *Nature*, 2020.



1.1 **Abstract**

Phages typically have small genomes(Yuan and Gao, 2017) and depend on their bacterial hosts for replication(Breitbart et al., 2018). DNA sequenced from many diverse ecosystems revealed hundreds of phage genomes of >200 kbp, including a genome of 735 kbp, the largest phage genome to date. Thirty-five genomes were manually curated to completion (circular and no gaps). Expanded genetic repertoires include diverse and new CRISPR-Cas systems, tRNAs, tRNA synthetases, tRNA modification enzymes, translation initiation, and elongation factors, and ribosomal proteins. Phage CRISPR-Cas systems have the capacity to silence host transcription factors and translational genes, potentially as part of a larger interaction network that intercepts translation to redirect biosynthesis to phage-encoded functions. In addition, some phages may repurpose bacterial CRISPR-Cas systems to eliminate competing phages. We phylogenetically define major clades of huge phage from human and other animal microbiomes, oceans, lakes, sediments, soils, and the built environment. We conclude that their large gene inventories reflect a conserved biological strategy, observed over a broad bacterial host range and across Earth's ecosystems.

N.B. All main figures for this manuscript can be found below in their dedicated section. All supplementary files (including figures and tables) can be found online with the published manuscript.

1.2 Introduction

Bacteriophages (phages), viruses of bacteria, are considered distinct from cellular life due to their inability to conduct most biological processes required for reproduction. They are agents of ecosystem change because they prey upon specific bacterial populations, mediate lateral gene transfer, alter host metabolism, and redistribute bacterially-derived compounds via cell lysis(Breitbart et al., 2018; Emerson et al., 2018; Rascovan et al., 2016). They spread antibiotic resistance(Balcazar, 2014) and disperse pathogenicity factors that cause disease in humans and animals(Brown-Jaque et al., 2018; Penadés et al., 2015). Most knowledge about phage is based on laboratory-studied examples, the vast majority of which have genomes a few 10s of kbp in length. Widely used isolation-based methods select against large phage particles, and they can be excluded from phage concentrates obtained by passage through 100 nm or 200 nm filters(Yuan and Gao, 2017). In 2017, only 93 isolated phages with genomes of >200 kbp in length were published(Yuan and Gao, 2017). Sequencing of whole community DNA can uncover phage-derived fragments, yet large genomes can still escape detection due to fragmentation(Shkoporov and Hill, 2019). However, a new clade of human and animal-associated megaphages was recently described based on genomes manually curated to completion from metagenomic datasets(Devoto et al., 2019). This finding motivated a more comprehensive analysis of microbial communities to evaluate the prevalence, diversity, and ecosystem distribution of phage with large genomes. Previously, phages with genomes >200 kbp have been referred to as “jumbo”(Yuan and Gao, 2017) or, in the case of >500 kbp, “megaphage”(Devoto et al., 2019). As the set reconstructed here span both size ranges we simply refer to them as “huge phage”. A graphical abstract provides an overview of our approach and main findings (**Extended Data Figure 1**). The research expands our understanding of phage biodiversity and reveals the wide variety of ecosystems in which phage have genomes with sizes that rival those of small celled bacteria(Castelle et al., 2018; Nakabachi et al., 2006; Pérez-Brocal et al., 2006). We postulate that these phages have evolved a distinct ‘life’ strategy that involves extensive interception and augmentation of host biology while they replicate their huge genomes.

1.3 Results and Discussion

Genome sizes and basic features

We reconstructed 351 phage, 6 plasmid-like, and 4 sequences of unknown classification (**Extended Data Figure 2**). We excluded additional sequences inferred to be plasmids (see Methods), retaining only those encoding CRISPR-Cas loci. We included 3 phage sequences of ≤200 kbp in length due to the presence of CRISPR-Cas loci. Consistent with classification as phage, we identified a wide variety of phage-relevant genes, including those involved in lysis and encoding structural proteins, and documented other expected phage genomic features (SI). Some predicted proteins are large, up to 7,694 amino acids in length; some were tentatively annotated as structural proteins. 175 phage sequences were circularized and 35 were manually curated to completion, in some cases by resolving complex repeat regions, revealing their encoded proteins (see Methods and

Table S1). The remaining genomes are likely incomplete, but some may be linear. Approximately 30% of genomes show clear GC skew indicative of bi-directional replication and 30% have patterns indicative of unidirectional replication (SI and **Extended Data Figure 3**)(Lobry, 1996).

Our four largest complete, manually curated, and circularized phage genomes are 634, 636, 642, and 735 kbp in length and represent the largest phage genomes reported to date. Previously, the largest circularized phage genome was 596 kbp in length(Paez-Espino et al., 2016). The same prior study reported a circularized genome of 630 kbp in length, but this is an assembly artifact (SI). The problem of concatenation artifacts was sufficiently prominent in IMG/VR(Paez-Espino et al., 2017) that we did not include these data in further analyses. We used both complete and circularized genomes from our study and published phage genomes to depict a current view of the distribution of phage genome sizes (Methods). Without the huge phage reported here, the median genome size for complete phage is ~52 kbp (**Figure 1A**). Thus, sequences reported here substantially expand the inventory of phage with unusually large genomes (**Figure 1B**).

Some of our reported genomes have very low coding density (nine <78%, see SI), probably due to the use of a genetic code different from the standard code (Methods). This phenomenon has been rarely noted in phages, but was reported for Lak phages(Devoto et al., 2019), and by Ivanova *et al(Ivanova et al., 2014)*. In the current study, some genomes (mostly human/animal associated) appear to have reassigned the UAG (amber) stop codon to code for an amino acid (SI and **Extended Data Figure 4**).

In only one case, we identified a sequence of >200 kbp that was classified as a prophage based on a transition into flanking bacterial genome sequence. However, around half of the genomes were not circularized, so their potential integration as prophage cannot be ruled out. The presence of integrases in some genomes is suggestive of a temperate lifestyle under some conditions.

Hosts, diversity, and distribution

An intriguing question relates to the evolutionary history of phages with huge genomes. Are they the result of recent genome expansion within clades of normal-sized phage or is a large inventory of genes an established, persistent strategy? To investigate this, we constructed phylogenetic trees for large terminase subunit (**Figure 2**) and major capsid (**Extended Data Figure 5A)** proteins using sequences in public databases as context (Methods). Many of the sequences from our phage genomes cluster together with high bootstrap support, defining clades. Analysis of the genome size information for database sequences shows that the public sequences that fall into these clades are from phages with genomes of at least 120 kbp in length. The largest clade, referred to here as Mahaphage (Maha being Sanskrit for huge), includes all of our biggest genomes as well as the 540 - 552 kbp Lak genomes from human and animal microbiomes(Devoto et al., 2019). We identified nine other clusters of large phages, and refer to them using the words for “huge” in the languages of some authors of this publication. We acknowledge that the detailed tree topologies for different genes and datasets vary somewhat, but the

clustering is broadly supported by protein family and capsid analyses (**Extended Data Figure 5A, B**). The consistent grouping together of large phages into clades establishes that large genome size is a relatively stable trait. Within each clade, phages were sampled from a wide variety of environmental types (**Figure 2**), indicating diversification of these huge phages and their hosts across ecosystems. We also examined the environmental distribution of phages that are so closely related that their genomes can be aligned and found 20 cases where they occur in at least two distinct cohorts or habitat types (**Table S2**).

To determine the extent to which bacterial host phylogeny correlates with phage clades, we identified some phage hosts using CRISPR spacer targeting from bacteria in the same or related samples and phylogenies of normally host-associated phage genes (see below; **Table S3**). We also tested the predictive value of bacterial taxonomic affiliations of the phage gene inventories (Methods) and found that in every case, CRISPR spacer targeting and phylogeny agreed with phylum-level taxonomic profiles. Consequently, we used taxonomic profiles to predict the bacterial host phylum for many phages (**Table S4**). The results establish the importance of Firmicutes and Proteobacteria as hosts (**Extended Data Figure 2**) (P = 2.5 x 10^{-5}; n = 74; W = 606; one-sided Wilcoxon signed-rank test). The higher prevalence of Firmicutes huge phage in the human and animal gut compared to other environments reflects the potential host compositions of the microbiomes (P = 9.3 x 10^{-7}; n = 37; U = 238; one-sided Mann-Whitney U-test). Notably, the five genomes >634 kbp in length are all for phage predicted to replicate in Bacteroidetes, as do Lak phage(Devoto et al., 2019), and all cluster within Mahaphage. Overall, phages grouped together phylogenetically are predicted to replicate in bacteria of the same phylum (**Figure 2**).

Metabolism, transcription, translation

The phage genomes encode proteins predicted to localize to the bacterial membrane or cell surface. These may impact host susceptibility to infection by other phages (**Table S5** and SI). We identified almost all previously reported categories of genes suggested to augment host metabolism (SI). Many phages have genes involved in *de novo* biosynthesis of purines and pyrimidines, and the interconversion of nucleic and ribonucleic acids and nucleotide phosphorylation states. These gene sets are intriguingly similar to those of bacteria with very small cells and putative symbiotic lifestyles(Castelle et al., 2018) (**Table S5**).

Notably, many phages have genes whose predicted functions are in transcription and translation (**Table S6)**. Complete phage genomes encode up to 67 tRNAs, with sequences distinct from those of their hosts (**Table S7**). Generally, the number of tRNAs per genome increases with genome length (**Figure 1**) (Spearman's ρ = 0.61; P = 4.5 x 10^{-22}; n = 201) . They have up to 15 tRNA synthetases per genome (**Table S7**), also distinct from but related to those of their hosts (**Extended Data Figure 7A** and SI). Phage may use these proteins to charge their own tRNA variants with host-derived amino acids. A subset of genomes have genes for tRNA modification and ligation of tRNAs cleaved by host defenses.

Many phages carry genes implicated in interception and redirection of host translation. These genes include initiation factors IF1 and IF3, as well as ribosomal proteins S4, S1, S21, and L7/L12 (ribosomal proteins were only recently reported in phage(Mizuno et al., 2019) (**Figure 3**)). Both rpS1 and rpS21 are important for translation initiation in bacteria(Farwell et al., 1992; Sørensen et al., 1998; Van Duin and Wijnands, 1981), making them likely useful for the hijacking of host ribosomes. Further analysis of rpS21 proteins revealed N-terminal extensions rich in basic and aromatic residues important for RNA binding. We predict that these phage ribosomal proteins substitute for host proteins(Mizuno et al., 2019), and their extensions assist in competitive ribosome binding or preferential initiation of phage mRNAs.

Because rpS1 is often studied in the context of Shine Dalgarno (SD) sequence recognition by the ribosome(Farwell et al., 1992; Sørensen et al., 1998), we predicted the ribosomal binding sites for each phage genome (Methods). While most phages have canonical SD sequences, huge phages from this study that carry possible rpS1s rarely have identifiable SD sequences (**SI** and **Table S8**). It is difficult to confirm "true" rpS1 proteins due to the ubiquity of the S1 domain, but this correlation with non-canonical SD sequences suggests a role in translation initiation, either on or off the ribosome.

While assuming control of initiation may be the most logical step for phage redirection of host translation, efficiency of elongation and termination is necessary for robust infection and replication. Accordingly, we found many genes associated with the latter steps of translation in phage genomes. These include elongation factors G, Tu, and Ts, rpL7/12, and the processing enzyme peptide deformylase (PDF) (**Figure 3**), previously reported in phage genomes(Frank et al., 2013). We hypothesize that phage-encoded elongation factors maintain overall translation efficiency during infection, much like PDF's prior predicted role in sustaining translation of necessary host photosynthetic proteins(Frank et al., 2013). Translation termination factors are also represented in our huge phage genomes, including release factor 1 and 2, ribosome recycling factor, as well as tmRNAs and small protein B (SmpB), which rescue ribosomes stalled on damaged transcripts and trigger the degradation of aberrant proteins. These tmRNAs are also used by phages to sense the physiological state of host cells and can induce lysis when the number of stalled ribosomes in the host is high(Janssen and Hayes, 2012). Interestingly, some large putative plasmids have analogous suites of translation-relevant genes (**Table S5**).

CRISPR-Cas mediated interactions

We identified most major types of CRISPR-Cas systems on phage, including Cas9-based Type II, the recently described Type V-I(Yan et al., 2019), new variants of Type V-U systems(Shmakov et al., 2017), and new subtypes of Type V-F (Harrington et al., 2018) (**Extended Data Figure 8**). The Class II systems (types II and V) are reported in phage for the first time. Most phage effector nucleases (for interference) have conserved catalytic residues, implying that they are functional (**Supplementary Data File**).

Unlike the well-described case of a phage with a CRISPR system(Seed et al., 2013), almost all phage CRISPR systems lack spacer acquisition machinery (Cas1, Cas2, and Cas4) and many lack recognizable genes for interference (**Table S1 and Extended Data Figure 9**). For example, two related phages have a Type I-C variant system lacking Cas1 and Cas2 and have a helicase protein in lieu of Cas3. They also harbor a second system containing a new candidate ~750 aa Type V effector protein, Cas12J (**Figure 4 and Table S1**), that occurs proximal to CRISPR arrays.

In some cases, phage lacking genes for interference and spacer integration have similar CRISPR repeats as their hosts (**Figure 4C**), and thus may utilize host Cas proteins. Alternatively, systems lacking an effector nuclease may repress transcription of the target sequences without cleavage(Luo et al., 2015; Stachler and Marchfelder, 2016). Alternatively, spacer-repeat guide RNAs may act in an RNAi-like mechanism to silence host CRISPR systems or nucleic acids to which they can hybridize. The phage-encoded CRISPR arrays are often compact (median 6 repeats per array; **Extended Data Figure 10**). This range is substantially smaller than typically found in prokaryotic genomes (mean of 41 for Class I systems)(Toms and Barrangou, 2017). Some phage spacers target core structural and regulatory genes of other phages (**Figure 4C, Table S10**). Thus, phages apparently augment their hosts' immune arsenal to prevent infection by competing phages.

Some phage-encoded CRISPR loci have spacers that target bacteria in the same sample or in a sample from the same study. We suppose that the targeted bacteria are the hosts for these phages, an inference supported by other host prediction analyses (**Table S4**). Some loci with bacterial chromosome-targeting spacers encode Cas proteins that could cleave the host chromosome, whereas others do not. Targeting host genes could disable or alter their regulation, which may be advantageous during the phage infection cycle. Some phage CRISPR spacers target bacterial intergenic regions, possibly interfering with genome regulation by blocking promoters or silencing non-coding RNAs.

Interesting examples of CRISPR targeting of bacterial chromosomes involve transcription and translation genes. For instance, one phage targets a σ^{70} in its host's genome and encodes its own σ^{70} transcription factor (SI). Some huge phage genomes encode anti-sigma factor-like proteins (AsiA), consistent with prior reports of σ^{70} hijacking by phage with AsiA(Brown and Hughes, 1995). In another example, a phage spacer targets the host glycyl tRNA synthetase, but the Cas14 effector lacks one of the required catalytic residues for cleavage, suggesting a role in repression (as a "dCas14"), rather than in cleavage (SI).

Interestingly, we found no evidence of host-encoded spacers targeting any CRISPR-bearing phage. However, phage CRISPR targeting of other phages that are also targeted by bacterial CRISPR (**Figure 4C**) suggested phage-host associations that were broadly confirmed by the phage taxonomic profile (**Table S4**).

Some large *Pseudomonas* phages encode Anti-CRISPRs(Bondy-Denomy et al., 2015; Pawluk et al., 2016) (Acr) and proteins that assemble a nucleus-like compartment segregating their replicating genomes from host defense and other bacterial systems(Chaikeeratisak et al., 2017a). We identified proteins encoded in huge phage genomes that cluster with AcrVA5, AcrVA2, AcrIIA7, and AcrIIA11 and may function as Acrs. Also identified were tubulin-homologs (PhuZ) and proteins (**SI**) that create a proteinaceous phage "nucleus"(Chaikeeratisak et al., 2017b). The phage nucleus was recently shown to protect the phage genome against host defense by physically blocking CRISPR-Cas degradation(Mendoza et al., 2018).

1.4 **Conclusions**

We show that phages with huge genomes are widespread across Earth's ecosystems. We manually completed 35 genomes, distinguishing them from prophage, providing accurate genome lengths and complete inventories of genes, including those encoded in complex repeat regions that break automated assemblies. Even closely related phages have diversified across habitats. Host and phage migration could transfer genes relevant in medicine and agriculture (e.g., pathogenicity factors and antibiotic resistance, SI). Additional medical significance could involve direct or indirect activation of immune responses. For example, some phages directly stimulate IFN-g via a TLR9-dependent pathway and exacerbate colitis(Gogokhia et al., 2019). Huge phage may represent a reservoir of novel nucleic acid manipulation tools with applications in genome editing and might be harnessed to improve human and animal health. For instance, huge phage equipped with CRISPR-Cas systems might be tamed and used to modulate bacterial microbiome function or eliminate unwanted bacteria.

The huge phages define massive clades, suggesting that a gene inventory comparable in size to those of many symbiotic bacteria is a conserved strategy for phage survival. Overall, their genes appear to redirect the host's protein production capacity to favor phage genes by first intercepting the earliest steps of translation and then ensuring efficient protein production thereafter. These inferences are aligned with findings for some eukaryotic viruses, which control every phase of protein synthesis(Jaafar and Kieft, 2019). Some acquired CRISPR-Cas systems with unusual compositions that may function to control host genes and eliminate competing phages.

More broadly, huge phages represent little-known biology, the platforms for which are distinct from those of small phages and partially analogous to those of symbiotic bacteria, somewhat blurring the distinctions between life and non-life. Given phylogenetic evidence for large radiations of huge phages, we wonder if they are ancient and arose simultaneously with free-living cells, their symbionts, and other phages from a pre-life (protogenote) state(Woese, 1998) rather than appearing more recently via episodes of genome expansion.

Figures

Figure 1: Distribution of phage genome sizes and tRNAs. A. Size distribution of circularized bacteriophage genomes from this study, Lak megaphage genomes reported recently from a subset of the same samples(Devoto et al., 2019), and reference sources. Reference genomes were collected from all complete RefSeq r92 dsDNA genomes and non-artifactual assemblies >200 kb from Páez-Espino(Paez-Espino et al., 2016). **B.** Histogram of the genome size distribution of phage with genomes >200 kb from this study, Lak, and reference genomes. Box and whisker plot of tRNA counts per genome from this study and Lak phage as a function of genome size (n = 201 individual phage genomes). The middle line for each box marks the median tRNA count for each size bin, the box marks the interquartile range, and the whiskers represent the maxima and minima.

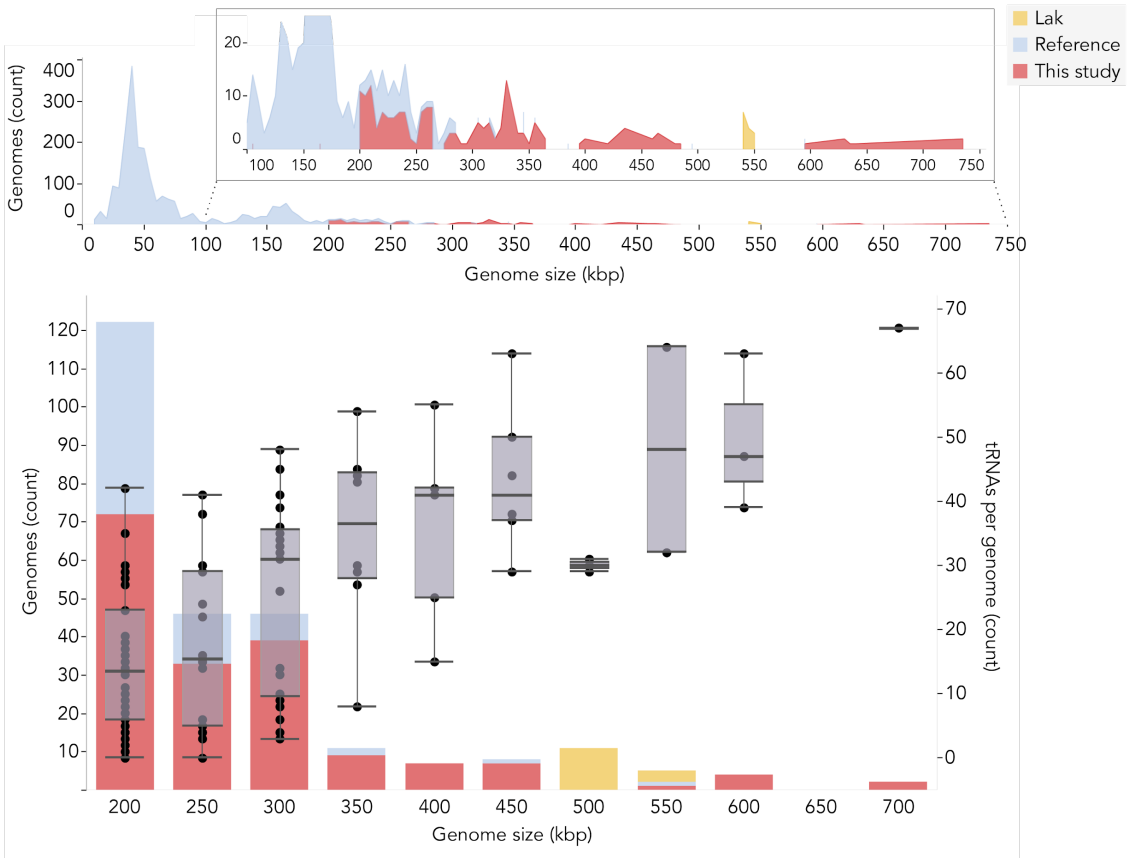


Figure 2: Phylogenetic reconstruction of huge phage evolutionary history. Phage phylogeny was reconstructed using large terminase sequences from this study and similar matches from all RefSeq r92 proteins. The tree also includes large terminase sequences from complete RefSeq phage, the Lak megaphage clade(Devoto et al., 2019) and non-artifactual phage genomes that are >200 kbp from(Paez-Espino et al., 2016). Huge phage clades identified in this study were independently corroborated with a

phylogenetic reconstruction of major capsid genes (**Extended Data Figure 5A**) and protein clustering (**Extended Data Figure 5B**). The tree was rooted using 13 eukaryotic Herpesvirus terminases. The inner to outer rings display the presence of CRISPR-Cas from this study, host phylum, environmental sampling type, and genome size. Host phylum and genome size were not included for RefSeq protein database matches where the sequence may be integrated prophage or part of organismal genome projects.

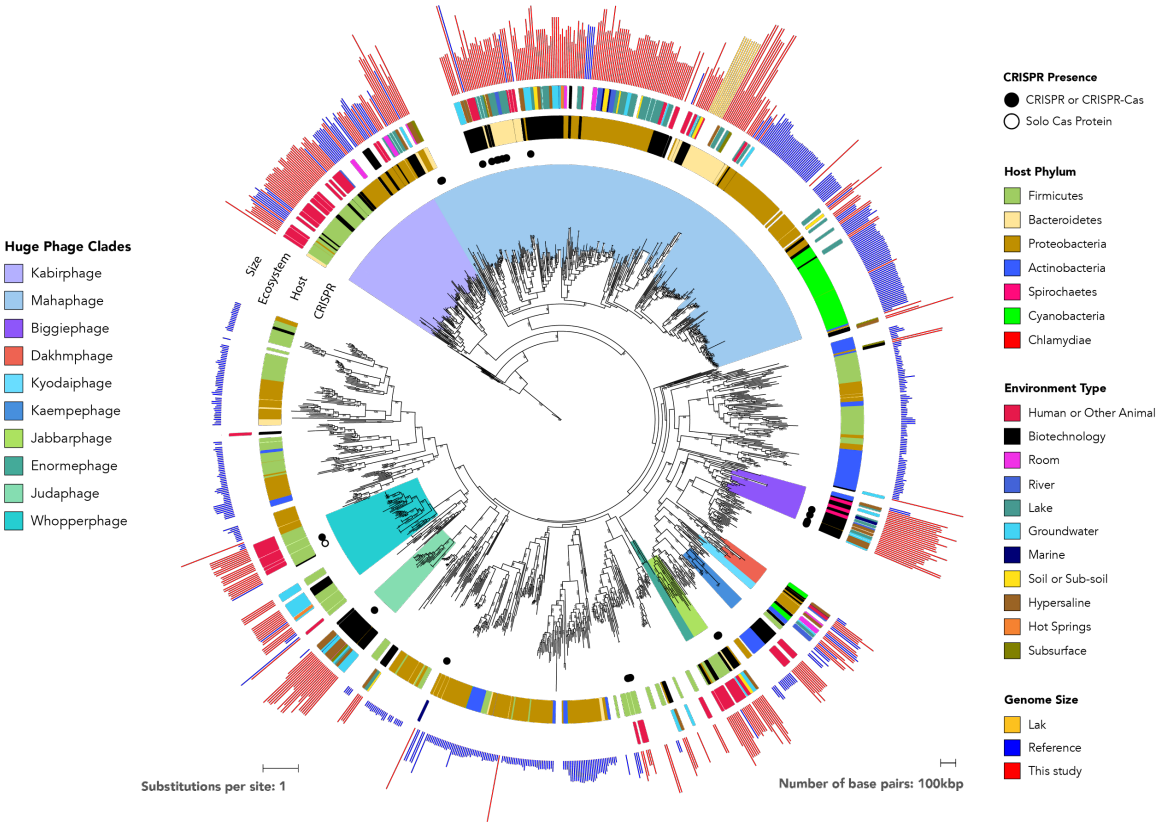


Figure 3: A model for phage interception and redirection of host translational systems. Potential mechanisms for how phage-encoded capacities could function to redirect the host's translational system to produce phage proteins (bacterial components in blue, phage in red). No huge phage has all translation related genes, but many have tRNAs and tRNA synthetases (see **Table S6**). Phage proteins with up to 6 ribosomal protein S1 domains occur in a few genomes. The S1 binds mRNA to bring it into the site on the ribosome where it is decoded(Subramanian, 1983). Phage ribosomal protein S21 might promote translation initiation of phage mRNAs, and many sequences have N-terminal extensions that may be involved in binding RNA (dashed blue line in ribosome insert, PDB: 6BU8(Loveland and Korostelev, 2018), analyzed with UCSF Chimera(Pettersen et al., 2004). Many other proteins of the translational apparatus are encoded by huge phage, belonging to all steps of the translation cycle.

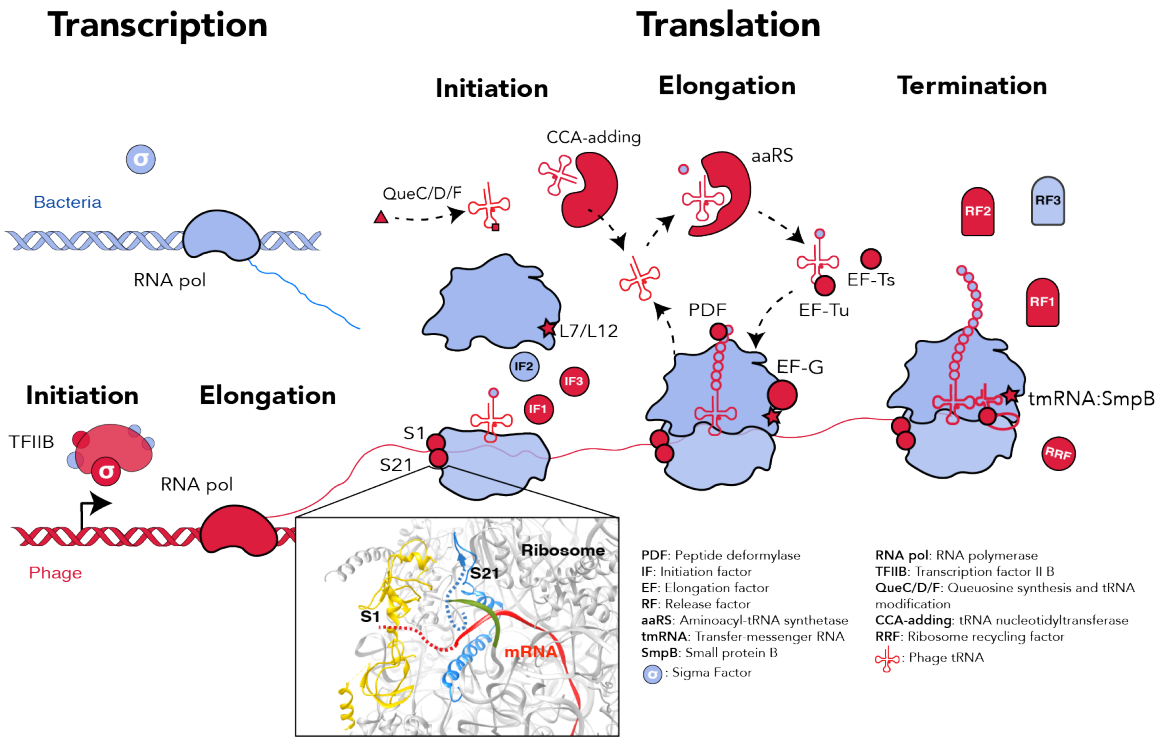
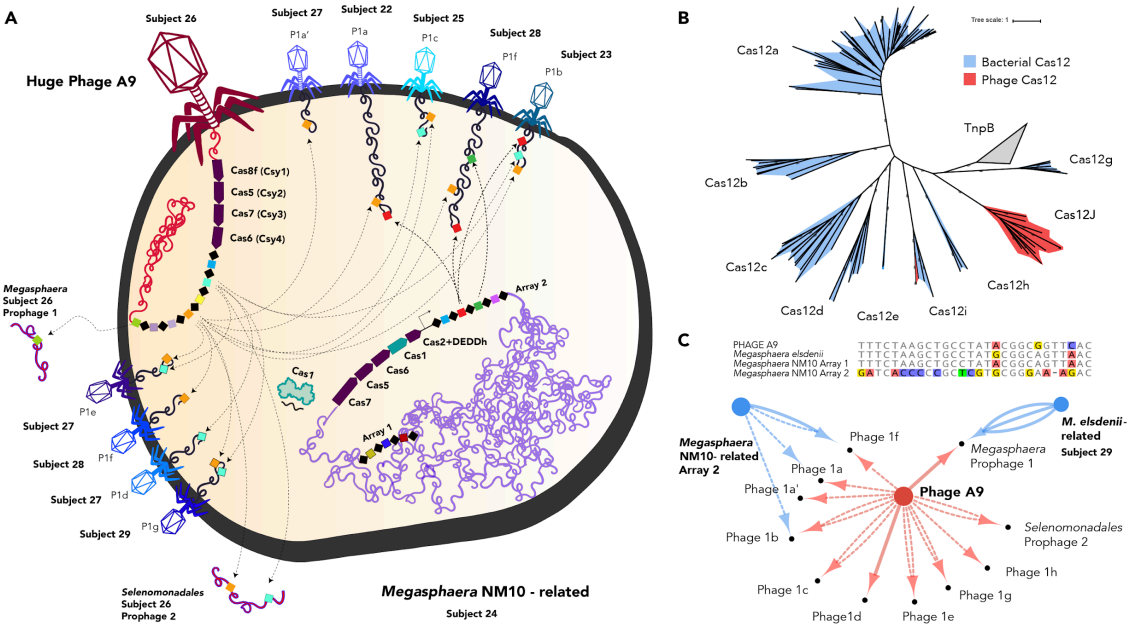
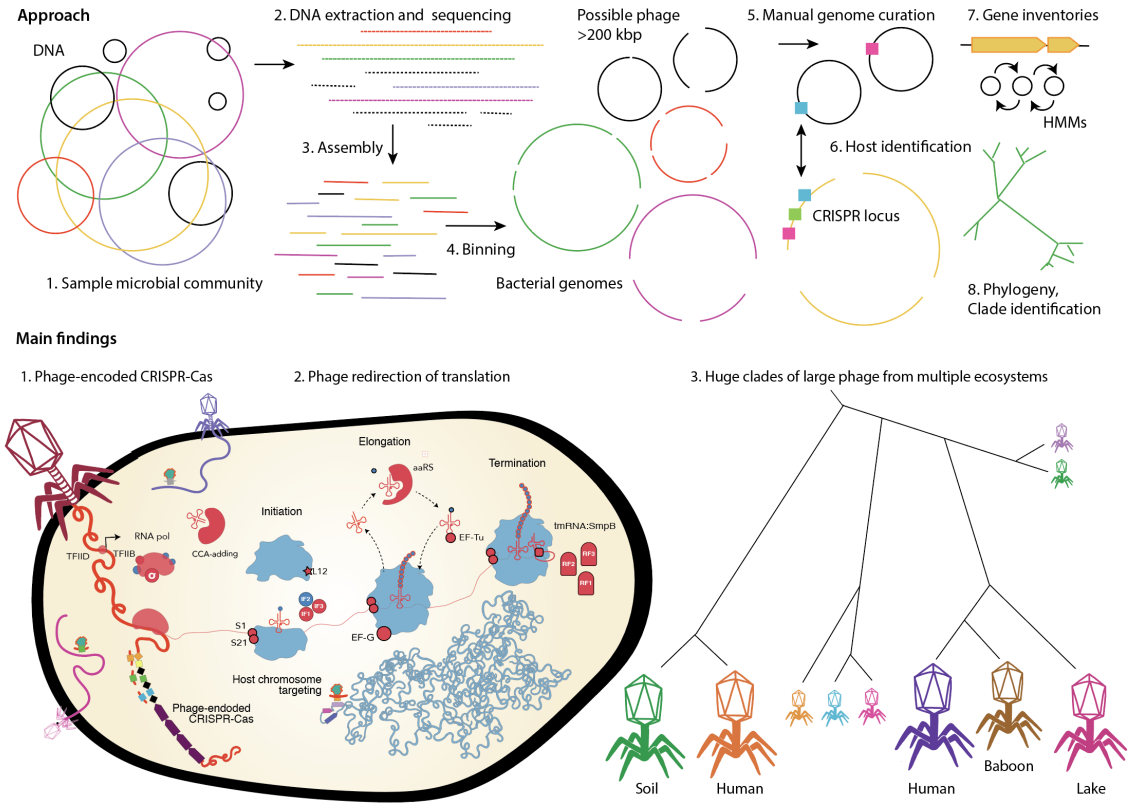


Figure 4: Phage and bacterial CRISPR interaction dynamics. A. Cell diagram of bacterium-phage and phage-phage interactions involving CRISPR targeting during superinfection. Arrows indicate CRISPR-Cas targeting of the prophage and phage genomes. Phage names indicate related groups delineated via whole genome alignment. We only included CRISPR interactions from samples of subjects of the same human cohort. **B.** Maximum likelihood phylogenetic tree of Cas12 subtypes a-i. Phage-encoded Cas12i and Cas12J, the new effector, are outlined in red, with bacterial-encoded proteins in blue. Bootstrap values >90 are shown on the branches (circles). Cas14 and Type V-U trees are provided separately (**Figure S11**). **C.** Top panel shows the alignment of the consensus repeats from the A9 phage array and predicted host bacterial arrays. Bottom panel is an interaction network showing targeting of bacterial- (blue) and phage- (red) encoded CRISPR spacers. Number of edges indicate number of spacers from the array with targets to the smaller node. Solid edges denote spacer targets with no or 1 mismatch, and dashed edges denote 2-3 mismatches (to account for degeneration in old-end phage spacers, diversity in different subjects, or phage mutation to avoid targeting).

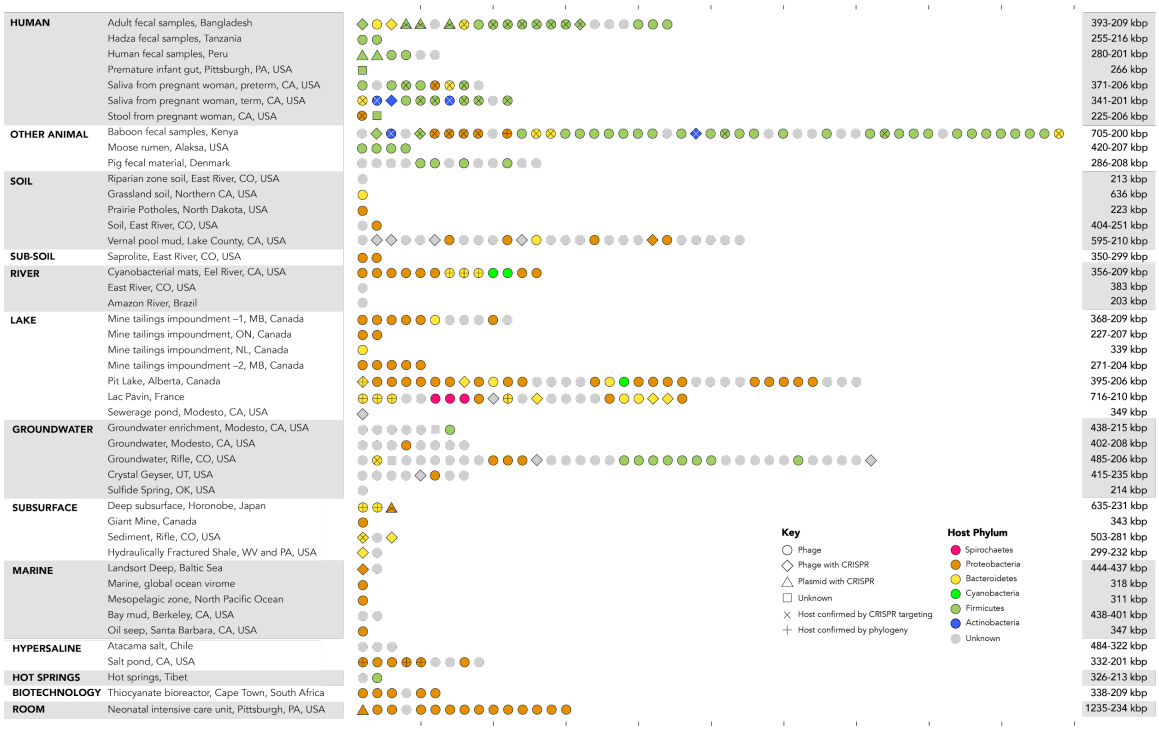


Extended Data

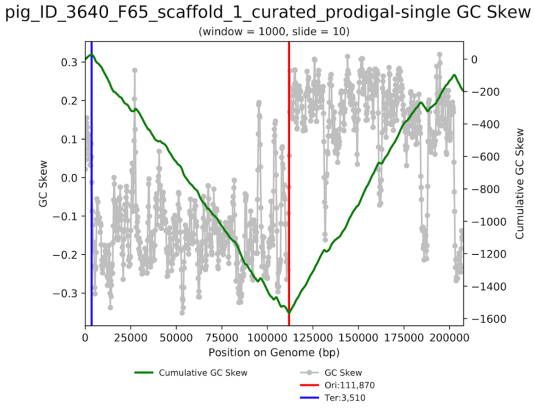
Extended Data Figure 1: Graphical abstract describing the approach and main findings of this study.

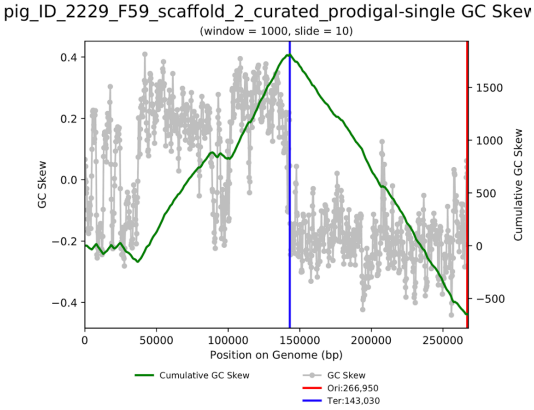


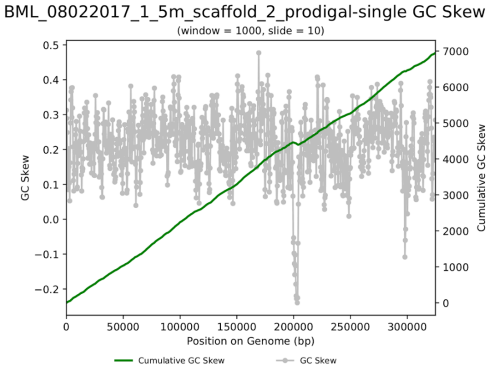
Extended Data Figure 2: Ecosystems with phage genomes and plasmid-like sequences >200 kb. Genomes grouped by sampling site type. Each box represents a phage genome or plasmid-like sequence , and boxes are horizontally arranged in order of decreasing genome size. Size range for each site type is listed to the right. Colors indicate putative host phylum based on genome taxonomic profile, with confirmation by CRISPR spacer targeting (X) or information system gene phylogenetic analyses (+).

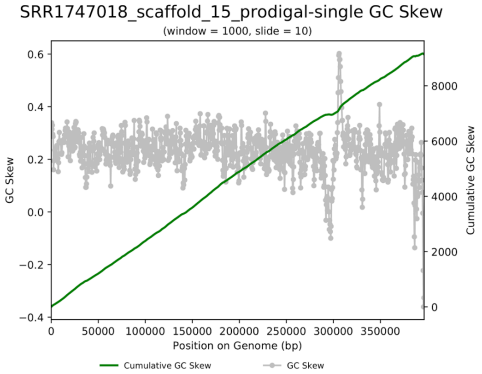


Extended Data Figure 3: Examples of phage genomes that display GC skew indicative of bidirectional replication. Some have a pattern strongly indicative of bi-directional replication (origin-to-terminus) typically found in bacteria, as shown in the first two panels (however, the origin may not correspond to the start of the genome). Others have skew suggestive of unidirectional replication (bottom panel).

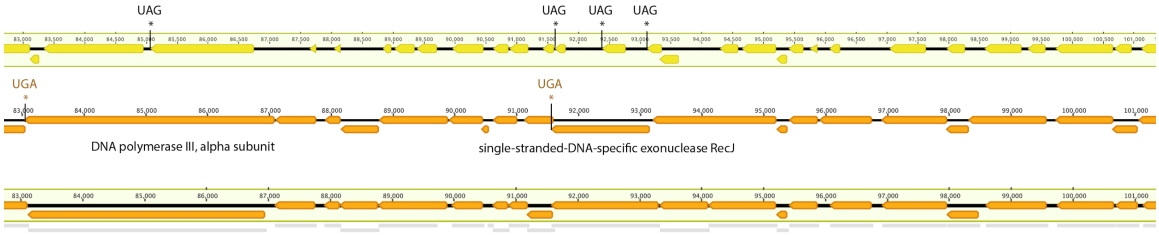




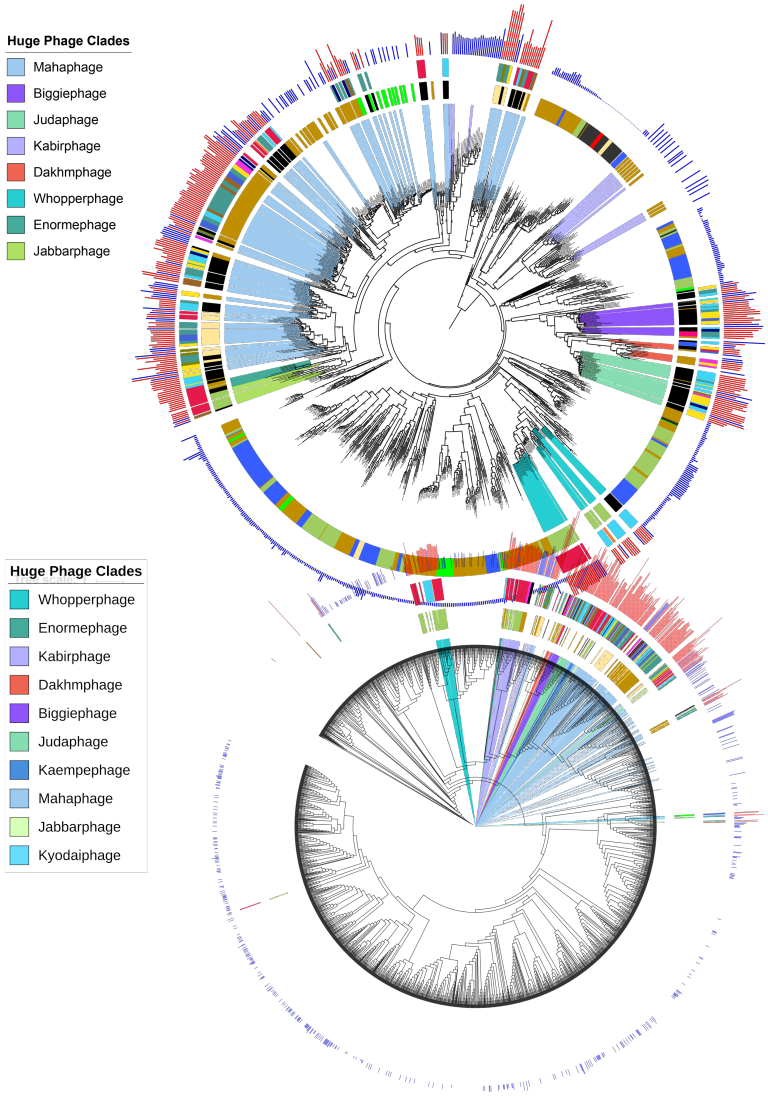




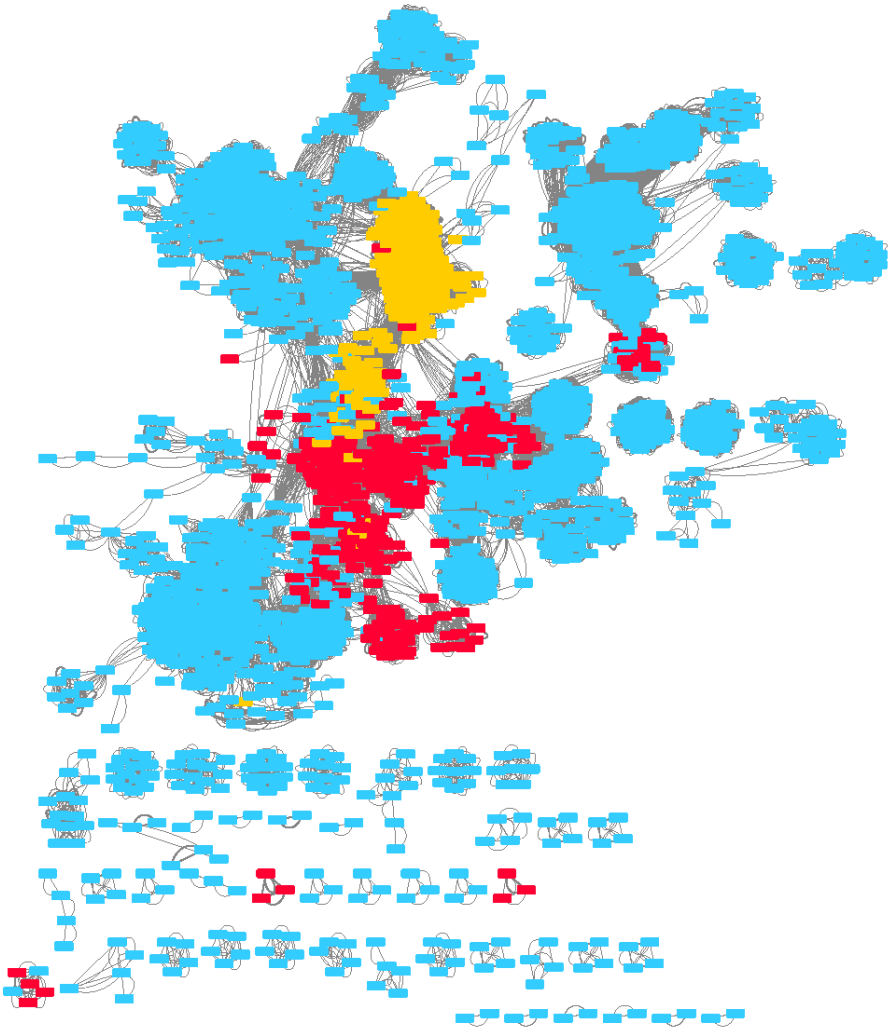
Extended Data Figure 4: **Example of phage alternative coding.** Comparisons of gene predictions for a region with genes of clearly predicted function in M05_PHAGE_COMPLETE_32_3. Top: the standard (code 11) genetic code. Middle: both TAG and TAA repurposed (code 6). Bottom: with just TAG repurposed (code 16). Overall, analysis of well annotated genes supported code 16 as the best choice (TAG -> X, as X could not be clearly resolved based on sequence alignments with related proteins).



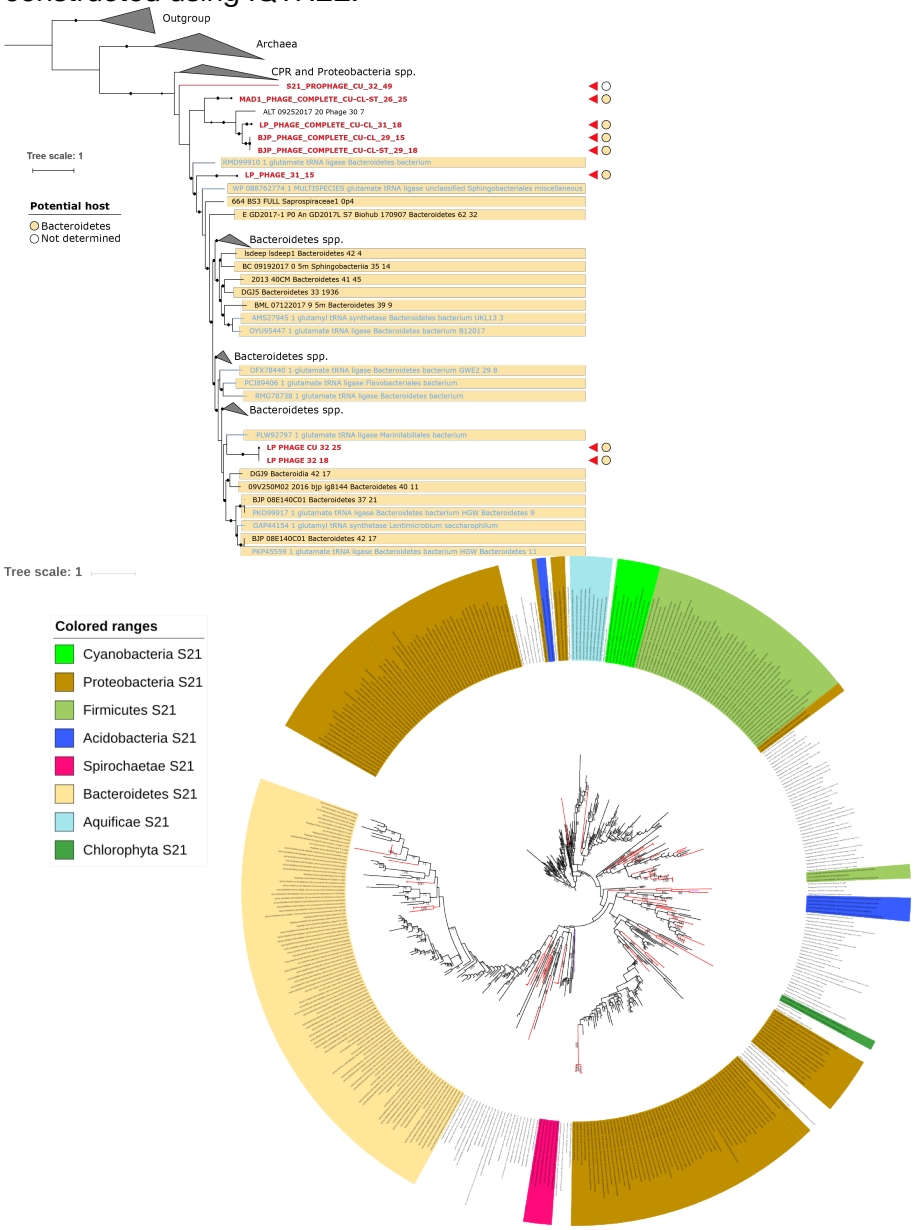
Extended Data Figure 5: Phage phylogenetic and protein-cluster relationships. A. Phage phylogenetic tree based on the major capsid protein. Outer ring shows genome length; bars in red are for genomes reconstructed and reported in this study and bars in blue are for database genomes. The next ring indicates environment of origin, see **Figure 2** for key. The inner ring indicates phylum of host (black indicates unknown), see **Figure 2** for key. Superimposed colors indicate named clades comprised of huge phage that were identified in the terminase tree. **B.** Hierarchical clustering dendrogram of phage genomes based on jaccard distance between the presence or absence profiles of protein families, performed using an average linkage method. Outermost ring shows phage genome length, next ring shows environment of origin, then predicted phylum affiliation of bacterial hosts. For color key see **Figure 2**. Superimposed colors indicate named clades comprised of huge phage that were identified in the terminase tree. The clustering supports the phylogenetic analyses shown in **Figure 2** and **Extended Data Figure 5A**.



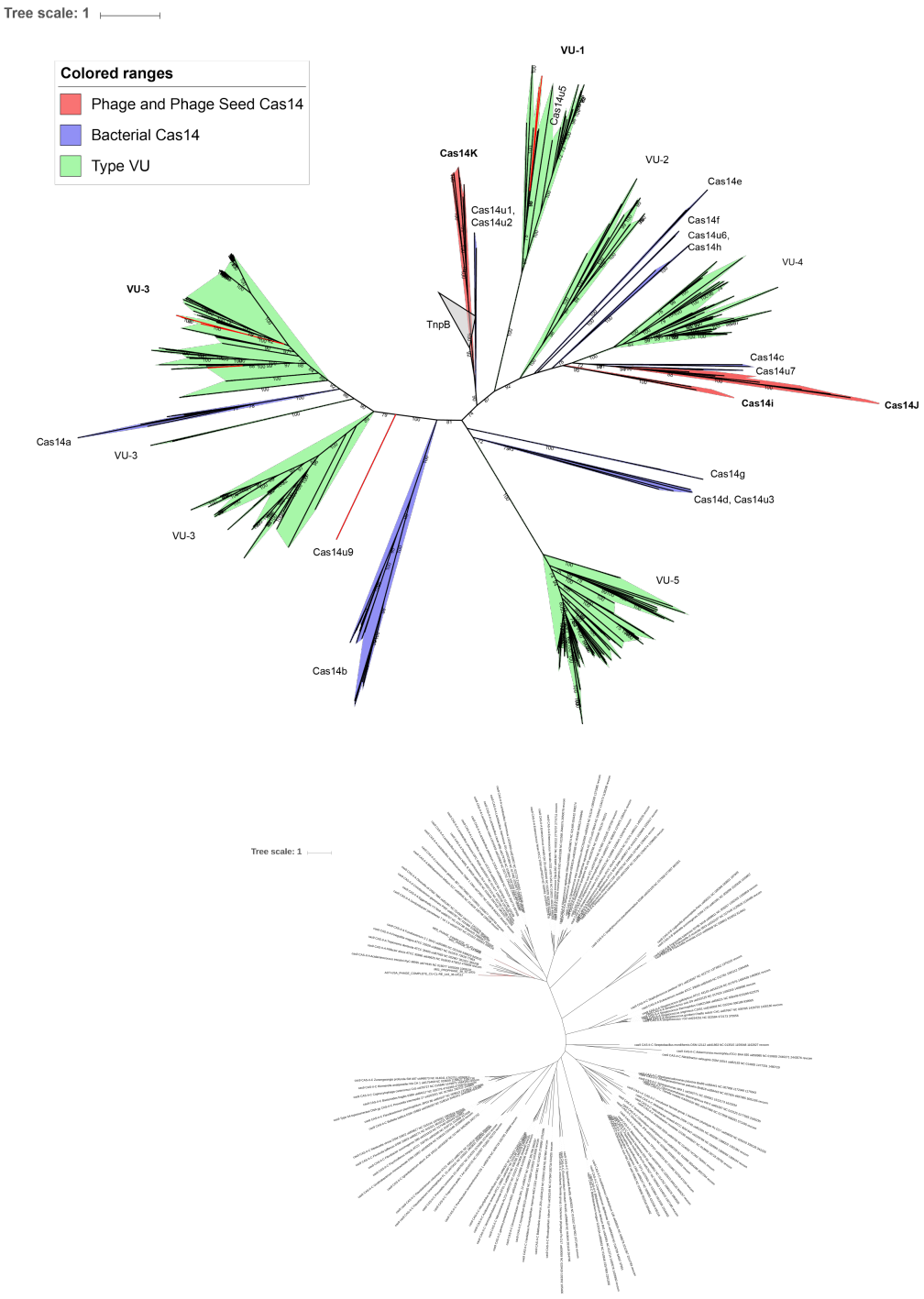
Extended Data Figure 6: Phage and plasmid protein clustering network. Network analysis using vContact2 and Cytoscape(Smoot et al., 2011) based on the number of shared protein clusters between the genomes in this study (red), RefSeq prokaryotic virus (blue) genomes, and 400 randomly sampled plasmid sequences (yellow) from RefSeq. Each node represents a genome and each edge is the hypergeometric similarity (>30) between genomes based on shared protein clusters. This analysis was used to help distinguish between the classification of genomes as phage, plasmid, or unknown.



Extended Data Figure 7: tRNA synthetase phylogenetic analysis. A. Aminoacyl tRNA synthetases were detected in many huge phage reported in this study (**Table S6**). This figure shows the phylogenetic sub-tree for glutamate-tRNA synthetase sequences from phage (red text and indicated by small triangles) that place within or close to those from Bacteroidetes hosts is shown as an example. Bacterial sequences from public databases are indicated by black text and those from metagenomes from which huge phage genomes were reconstructed are indicated by blue text. Colored circles indicate the predicted phylum of the bacterial host for each phage. **B.** Phylogenetic tree of phage encoded ribosomal protein S21 and the top refseq hits for each protein, constructed using IQTREE.

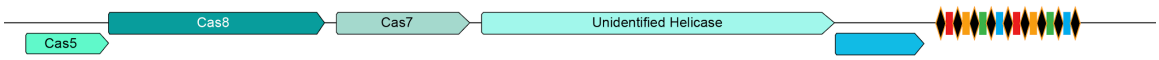


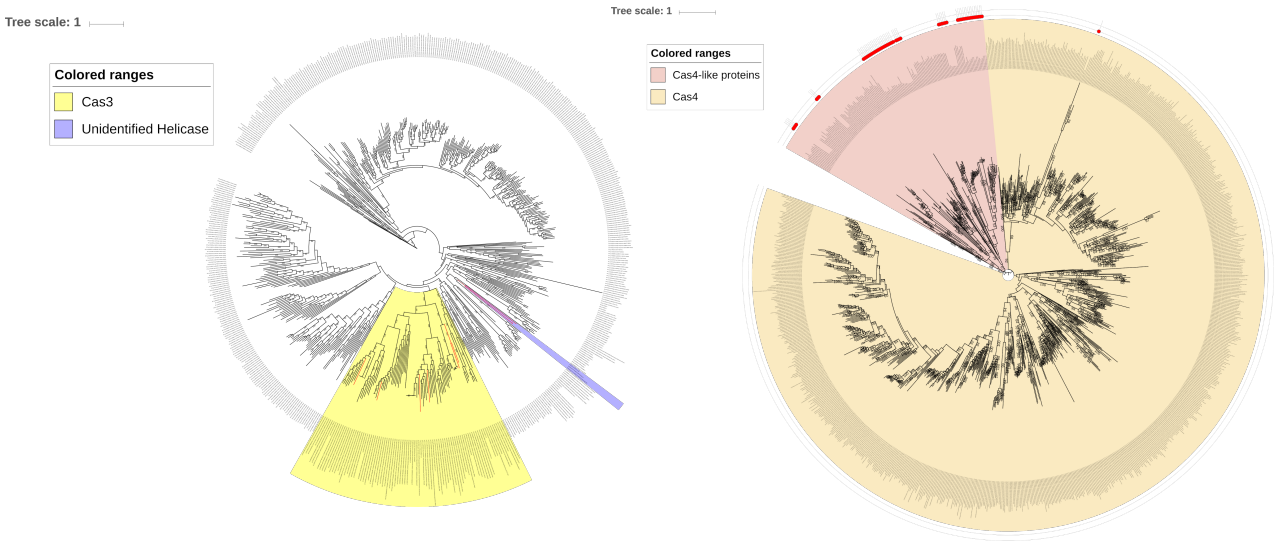
Extended Data Figure 8: Phylogenetic trees of Cas14, Type V-U, and Cas9. A. Phylogenetic tree for Cas14 and Type V-U. **B**. Phylogenetic tree for Cas9.



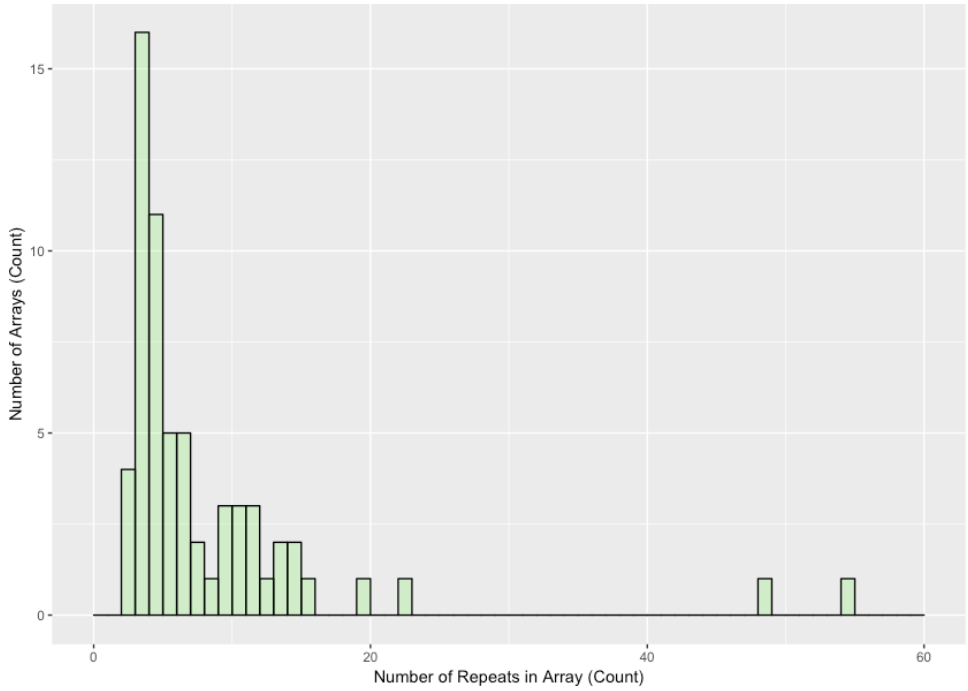
Extended Data Figure 9: Variant Type I CRISPR-Cas system and Cas4-like proteins found in huge phage genomes. A. Locus architecture for Type-I-var CRISPR phage. An interesting type I system identified in huge phage lacks Cas6 but harbors

Cas5, most similar to the Cas5d protein from type I-C, which acts as the pre-crRNA endonuclease (a role commonly reserved for Cas6). The proposed active site residues of Cas5d are to some extent different in the Cas5 of this system, though this may still confer processing activity, since this change is also observed for other Cas6 homologs. **B.** Phylogenetic tree of Superfamily 1-6 helicases, including Cas3 and the unidentified helicase in the Type I-C variant system. **C.** Phylogenetic tree of Cas4, Cas4-like proteins from the phage and plasmid genomes reported here, and the top 50 RefSeq hits to the Cas4-like proteins. Cas4-like genes from this study are denoted with red circles.





Extended Data Figure 10: Distribution of phage and plasmid-encoded CRISPR array sizes. The indicated count is the number of recovered repeats.



1.5 **Methods**

Ecosystem sampling

Metagenomic datasets were acquired from human fecal and oral samples, fecal samples from other animals, freshwater lakes and rivers, marine ecosystems, sediments, hot springs, soils, deep subsurface habitats, and the built environment (**Extended Data Figure 2**). Genome sequences that were clearly not bacterial, archaeal, archaeal virus, eukaryotic or eukaryotic virus were classified as phage, plasmid-like or mobile genetic elements of uncertain nature based on their gene inventories (see **Supplementary Information, SI**). *De novo* assembled fragments close to or >200 kbp in length were tested for circularization and a subset selected for manual verification and curation to completion.

Phage and plasmid genome identification

Datasets generated in the current study, those from prior research conducted by our team, the Tara Oceans microbiomes(Karsenti et al., 2011), and the Global Oceans Virome (GOV)(Roux et al., 2016) were searched for sequence assemblies that could have derived from phage with genomes of >200 kbp in length. Read assembly, gene prediction, and

initial gene annotation followed standard methods reported previously(Bushnell, 2016; Edgar, 2015; Joshi and Fass, 2011; Nurk et al., 2017; Peng et al., 2012).

Phage candidates were initially found by retrieving sequences that were not assigned to a genome and had no clear taxonomic profile at the domain level. Taxonomic profiles were determined through a voting scheme, where there had to be a winner taxonomy >50% votes at each taxonomic rank based on UniProt and ggKbase (ggkbase.berkeley.edu) database protein annotations(Raveh-Sadka et al., 2015). Phage were further narrowed down by identifying sequences with a high number of hypothetical protein annotations and/or the presence of phage specific genes, *e.g.*, capsid, tail, terminase, spike, holin, portal, and baseplate. All candidate phage sequences were checked throughout to distinguish putative prophage from phage. Prophage were identified based on a clear transition into genome with a high fraction of confident functional predictions, often associated with core metabolic functions, and much higher similarity to bacterial genomes. Plasmids were distinguished from phage based on matches to plasmid partitioning and conjugative transfer genes. Those that did not have phage specific genes were assigned using phylogenetic tree placement using *recA*, *polA*, *polB*, *dnaE*, and the DNA sliding clamp loader gene. Phage and placement assignments were further verified using a network of protein clustering with proteins from RefSeq prokaryotic viruses and 400 randomly sampled plasmids >200 kb using vContact2(Bolduc et al., 2017) (**Extended Data Figure 6**).

Phage and plasmid genome manual curation

All scaffolds classified were tested for end overlaps indicative of circularization. Assembled sequences that could be perfectly circularized were considered potentially "complete". Erroneous concatenated sequence assemblies were initially flagged by searching for direct repeats >5 kb using Vmatch(Kurtz, 2003). Potentially concatenated sequence assemblies were manually checked for multiple large repeating sequences using the dotplot and RepeatFinder features in Geneious v9. Sequences were corrected and removed from further analysis if the corrected length was <200 kbp.

A subset of the phage sequences were selected for manual curation, with the goal of finishing (replacing all Ns at scaffolding gaps or local misassemblies by the correct nucleotide sequences and circularization). Curation generally followed methods described previously(Devoto et al., 2019). In brief, reads from the appropriate dataset were mapped using Bowtie2 v2.3.4.1(Langmead and Salzberg, 2012) to the *de novo* assembled sequences. Unplaced mate pairs of mapped reads were retained with shrinksam (github.com/bcthomas/shrinksam). Mappings were manually checked throughout to identify local misassemblies using Geneious v9. N-filled gaps or misassembly corrections made use of unplaced paired reads, in some cases using reads relocated from sites where they were mis-mapped. In such cases, mis-mappings were identified based on much larger than expected paired read distances, high polymorphism densities, backwards mapping of one read pair, or any combination of these. Similarly, ends were extended using unplaced or incorrectly placed paired reads until circularization could be established. In some cases, extended ends were used to recruit new scaffolds that were then added to the assembly. The accuracy of all extensions and local assembly changes were verified in a subsequent phase of read mapping. In many

cases, assemblies were terminated or internally corrupted by the presence of repeated sequences. In these cases, blocks of repeated sequence as well as unique flanking sequence were identified. Reads were then manually relocated, respecting paired read placement rules and unique flanking sequences. After gap closure, circularization, and verification of accuracy throughout, end overlap was eliminated, genes were predicted, and the start moved to an intergenic region, in some cases suspected to be origin based on a combination of coverage trends and GC skew(Brown et al., 2016). Finally, the sequences were checked to identify any repeated sequences that could have led to an incorrect path choice because the repeated regions were larger than the distance spanned by paired reads. This step also ruled out artifactual long phage sequences generated by end to end repeats of smaller phage, which occur in previously described datasets(Devoto et al., 2019).

Structural and functional annotation

Following identification and curation of phage genomes, coding sequences (CDS) and Shine-Dalgarno ribosomal binding site (RBS) motifs were predicted with prodigal using genetic code 11 (-m -g 11 -p single). The resulting CDS were annotated as previously described by searching against UniProt, UniRef100, and KEGG(Wrighton et al., 2014). Functional annotations were further assigned by searching proteins against PFAM r32(Finn et al., 2014), TIGRFAMS r15(Haft et al., 2013), Virus Orthologous Groups r90 (VOG) (vogdb.org), and Prokaryotic Virus Orthologous Groups(Grazziotin et al., 2017) (pVOG). tRNAs were identified with tRNAscan-SE 2.0(Lowe and Eddy, 1997) using the bacterial model. tmRNAs were assigned using ARAGORN v1.2.38(Laslett and Canback, 2004) with the bacterial/plant genetic code.

Clustering of the CDS into families was achieved using a two-step procedure. A first protein clustering was done using the fast and sensitive protein sequence searching software MMseqs(Hauser et al., 2016). An all-vs.-all sequences search was performed using e-value: 1 x 10^{-3}, sensitivity: 7.5 and coverage: 0.5. A sequence similarity network was built based on the pairwise similarities and the greedy set cover algorithm from MMseqs was performed to define protein subclusters. The resulting subclusters were defined as subfamilies. In order to test for distant homology, we grouped subfamilies into protein families using an HMM-HMM comparison. The proteins of each subfamily with at least two protein members were aligned using the result2msa parameter of MMseqs, and from the multiple sequence alignments HMM profiles were built using the HHpred(Remmert et al., 2011) suite. The subfamilies were then compared to each other using HHblits from the HHpred suite (with parameters -v 0 -p 50 -z 4 -Z 32000 -B 0 -b 0). For subfamilies with probability scores of ≥ 95% and coverage ≥ 0.50, a similarity score (probability × coverage) was used as weights of the input network in the final clustering using the Markov clustering algorithm(Enright et al., 2002), with 2.0 as the inflation parameter. These clusters were defined as the protein families. Protein sequences were functionally annotated based on their best hmmsearch match (version 3.1) (E-value cut-off 1 x 10^{-3}) against an HMM database constructed based on orthologous groups defined by the KEGG database(Kanehisa et al., 2016) (downloaded on June 10, 2015). Domains were predicted using the same hmmsearch procedure against the PFAM r31 database(Finn et al., 2014). The domain architecture of each protein sequence was predicted using the DAMA software(Bernardes et al., 2016) (default parameters). SIGNALP(Petersen et al., 2011) (version 4.1) (parameters: -f short -t gram+) and

PSORT(Peabody et al., 2016) v3 (parameters: --long --positive) were used to predict the putative cellular localization of the proteins. Prediction of transmembrane helices in proteins was performed using TMHMM(Krogh et al., 2001) (version 2.0) (default parameters). Hairpins (palindromes, based on identical overlapping repeats in the forward and reverse directions) were identified using the Geneious Repeat Finder and located dataset-wide using Vmatch(Kurtz, 2003). Repeats >25 bp with 100% similarity were tabulated.

Reference genomes for size comparisons

RefSeq r92 genomes were recovered by using the NCBI Virus portal and selecting only complete dsDNA genomes with bacterial hosts. Genomes from(Paez-Espino et al., 2016) were downloaded from IMG/VR and only sequence assemblies labeled “circular” with predicted bacterial hosts were retained. Given the presence of sequences in IMG/VR that are based on erroneous concatenations, we only considered sequences from this source that are >200 kb, but a subset of these were removed as artifactual sequences.

Alternative genetic codes

In cases where gene prediction using the standard bacterial code (code 11) resulted in seemingly anomalously low coding densities, potential alternative genetic codes were investigated. In addition to making a prediction using the Fast and Accurate genetic Code Inference and Logo(Dutilh et al., 2011) (FACIL) web server, we identified genes with well defined functions (*e.g.*, polymerase, nuclease) and determined the stop codons terminating genes that were shorter than expected. We then re-predicted genes using GLIMMER3 v1.5(Delcher et al., 1999) and prodigal with TAG not interpreted as a stop codon. Other combinations of repurposed stop codons were evaluated, and candidate codes (*e.g.*, code 6, with only one stop codon) were ruled out due to unlikely gene fusion predictions.

Large terminase subunit and major capsid phylogenetic analysis

The large terminase subunit phylogenetic tree was constructed by recovering large terminases from the aforementioned protein clustering and annotation pipeline. CDS that matched with > 30 bitscore against PFAM, TIGRFAMS, VOG, and pVOG were retained. Any CDS that had a hit to large terminase, regardless of bitscore, was searched using HHblits(Steinegger et al.) against the uniclust30_2018_08 database. The resulting alignment was then further searched against the PDB70 database. Remaining CDS that clustered in protein families with a large terminase HMM were also included after manual verification. Detected large terminases were manually verified using the HHPred(Steinegger et al.) and jPred(Cole et al., 2008) webservers. Large terminases from the > 200 kbp(Paez-Espino et al., 2016) phage genomes and all >200 kbp complete dsDNA phage genomes from RefSeq r92 were also included by protein family clustering with the phage CDS from this study. The resulting terminases were clustered at 95% amino acid identity (AAI) to reduce redundancy using CD-HIT (Huang et al., 2010). Smaller phage genomes were included by searching the resulting CDS set against the full RefSeq protein database and retaining the top 10 best hits. Those hits that had no large terminase match against PFAM, TIGRFAMS, VOG, or pVOG were removed from further consideration and the remaining set was clustered at 90% AAI. The final set of large terminase CDS that were >100 aa were aligned using MAFFT(Katoh and Standley, 2013) v7.407 (--localpair --maxiterate 1000), and poorly aligned sequences were removed and

the resulting set was realigned. The phylogenetic tree was inferred using IQTREE v1.6.6 using automatic model selection(Nguyen et al., 2015). The phylogenetic tree of major capsid protein (MCP) genes was constructed by retrieving all MCPs annotated by combining the PFAM annotations of protein families and direct annotations by PFAM, TIGRFAMS, VOG, and pVOG. Reference MCP gene sequences were collected using the same strategy and sources as for the large terminase subunit tree. The resulting set were further screened by searching against PFAM, TIGRFAMS, VOG, and pVOG and removing matches that had no large terminase match regardless of bitscore. The final set of major capsid sequences were aligned with MAFFT(--localpair --maxiterate 1000) and the phylogenetic tree was constructed using IQTREE with automatic model selection and 1000 bootstrap replicates.

Whole genome scale clustering

To identify phage genomes that were closely related at the whole genome level we compared sequences using whole genome alignments. The goal of this analysis was to further corroborate the identified phylogenetic clades and test for the presence of very similar phages in different habitats and environments. Genomes grouped together in the primary clusters from dRep v2(Olm et al., 2017) were evaluated for genome alignment using Mauve(Darling et al., 2004) within Geneious v9.

CRISPR-Cas Locus and target detection

Phage and host encoded CRISPR loci (repeats and spacers) were identified using a combination of MinCED (github.com/ctSkennerton/minced) and CRISPRDetect(Biswas et al., 2016). A custom database of Cas genes was built by collecting Cas gene sequences from(Burstein et al., 2017; Harrington et al., 2018; Makarova et al., 2015; Shmakov et al., 2015; Smargon et al., 2017; Yan et al., 2018, 2019) and built with MAFFT (--localpair --maxiterate 1000) and hmmbuild. CDS from this study were searched against the HMM database using hmmsearch with e-value < 1 x 10^{-5}. Matches were checked using a combination of hmmscan and BLAST searches against the NCBI nr database and manually verified by identifying co-located CRISPR arrays and Cas genes. Spacers extracted from between repeats of the CRISPR locus were compared to sequences assemblies from the same site using BLASTN-short(Altschul et al., 1990) Matches with alignment length >24 bp and ≤1 mismatch were retained and targets were classified as bacterial, phage, or other. CRISPR arrays that had at least one ≤1 mismatch, were further searched for more spacer matches in the target sequence by finding more hits with ≤3 mismatches.

Host identification

The phylum affiliations of bacterial hosts for phage and plasmid-like sequences were predicted by considering the UniProt taxonomic profiles of every CDS for each phage genome. The phylum level matches for each phage genome were summed and the phylum with the most hits was considered as the potential host phylum. However, only cases where this phylum that had 3x as many counts as the next most counted phylum were assigned as the tentative phage host phylum. Phage hosts were further assigned and verified using the aforementioned CRISPR targeting strategy with the phage and plasmid-like genomes as targets. CRISPR arrays were predicted on all sequence assemblies from the same site that each phage genome was reconstructed. Sequence assemblies containing spacers with a match of length >24 bp and ≤1 mismatch. In the

case of phage, the match was used to infer a phage-host relationship. In all cases, the predicted host phylum based on taxonomic profiling and CRISPR targeting were in complete agreement. Similarly, the phyla of hosts were predicted based on phylogenetic analysis of phage genes also found in host genomes (*e.g.*, involved in translation and nucleotide reactions). Inferences based on computed taxonomic profiles and phylogenetic trees were also in complete agreement.

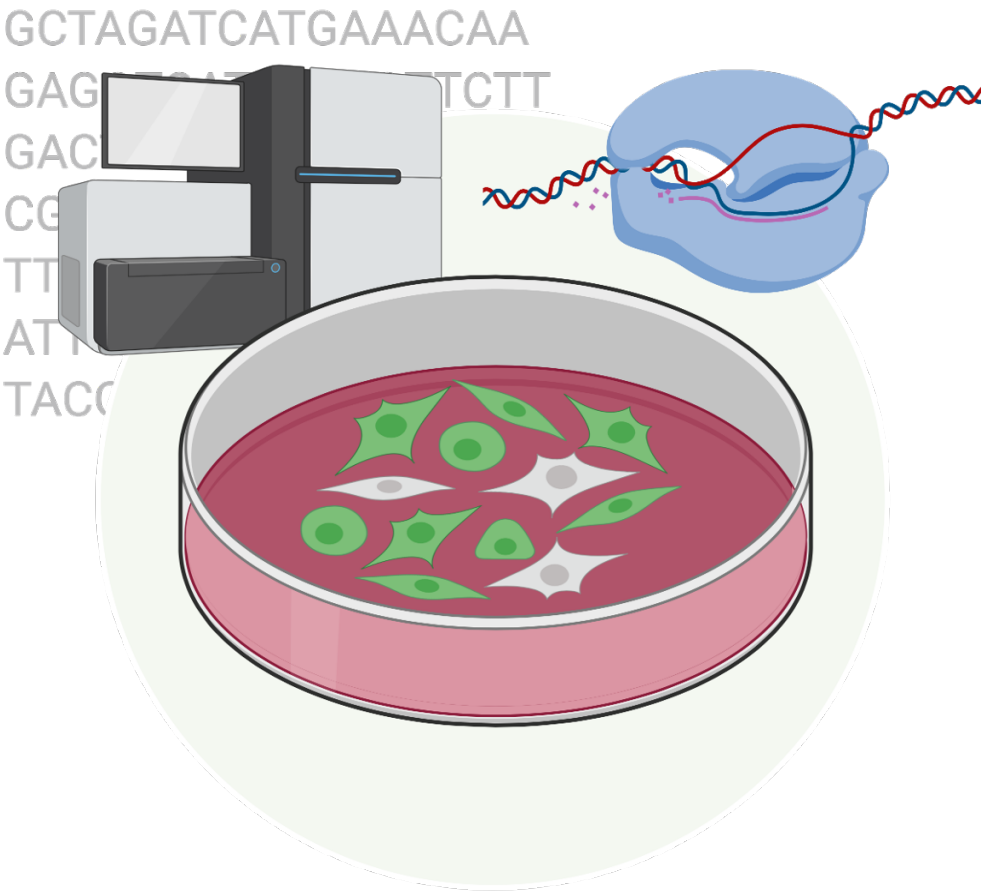
Phage encoded tRNA synthetase trees

Phylogenetic trees were constructed for phage encoded tRNA synthetase, ribosomal, and initiation factor protein sequences using a set of the closest reference sequences from NCBI and bacterial genomes from the current study. The tRNA synthetases were identified based on annotation of genes via the standard ggKbase pipeline (see above), and confirmed by HMMs with datasets from TIGRFAMs. For each type of tRNA synthetase, references were selected by comparing all the corresponding genes of this type against NCBI nr using DIAMOND v0.9.24(Buchfink et al., 2015), their top 100 hits were clustered by CD-HIT with 90% similarity threshold(Huang et al., 2010). The phylogenetic tree of each tRNA synthetase was constructed using RAxML v8.0.26(Stamatakis, 2014) with the PROTGAMMALG model.

2 Chapter 2: CRISPR-CasΦ from huge phages is a hypercompact genome editor

Patrick Pausch #, Basem Al-Shayeb #, Ezra Bisom-Rapp, Connor A Tsuchida, Zheng Li, Brady F Cress, Gavin J Knott, Steven E Jacobsen, Jillian F Banfield, Jennifer A Doudna

Published in *Science*, 2020.



2.1 Abstract

CRISPR-Cas systems are found widely in prokaryotes where they provide adaptive immunity against virus infection and plasmid transformation. We describe a minimal functional CRISPR-Cas system, comprising a single ~70 kilodalton protein, CasΦ, and a CRISPR array, encoded exclusively in the genomes of huge bacteriophages. CasΦ employs a single active site for both CRISPR RNA (crRNA) processing and crRNA-guided DNA cutting to target foreign nucleic acids. This hypercompact system is active *in vitro* and in human and plant cells with expanded target recognition capabilities relative to other CRISPR-Cas proteins. Useful for genome editing and DNA detection but with a molecular weight half that of Cas9 and Cas12a genome-editing enzymes, CasΦ offers advantages for cellular delivery that expand the genome editing toolbox.

N.B. All main figures for this manuscript can be found below in their dedicated section. All supplementary files (including figures and tables) can be found online with the published manuscript.

2.2 Introduction

Competition between viruses and their host microbes fostered the evolution of CRISPR-Cas systems that employ nucleases and non-coding CRISPR RNAs (crRNAs) to target foreign nucleic acids by complementary base pairing (Barrangou et al., 2007). Processing of CRISPR array transcripts, consisting of repeats and spacer sequences acquired from viruses or other mobile genetic elements (MGEs) (McGinn and Marraffini, 2019), generates mature crRNAs that guide Cas proteins (Hille et al., 2018) to detect and destroy previously encountered viruses. Although found almost exclusively in microbial genomes, the recent discovery of ubiquitous huge bacteriophages (viruses of bacteria) revealed the surprising prevalence of CRISPR-Cas systems encoded in their genomes (Al-Shayeb et al., 2020). These systems notably lack CRISPR spacer acquisition machinery (Cas1, Cas2 and Cas4 proteins) and generally harbor compact CRISPR arrays (median: 5 spacers per array), some of which target the genes of competing phages or phage hosts. CasΦ (Cas12j) is a family of Cas proteins encoded in the Biggiephage clade (Al-Shayeb et al., 2020). CasΦ contains a C-terminal RuvC domain with remote homology to that of the TnpB nuclease superfamily from which type V CRISPR-Cas proteins are thought to have evolved (Al-Shayeb et al., 2020; Shmakov et al., 2017) (fig. S1). However, CasΦ shares <7% amino acid identity with other type V CRISPR-Cas proteins and is most closely related to a TnpB group distinct from miniature type V (Cas14) proteins (Fig. 1A).

2.3 **Results and Discussion**

CasΦ's unusually small size of ~70-80 kDa, about half the size of the Cas9 and Cas12a (Fig. 1B), and its lack of co-occurring genes raised the question of whether CasΦ functions as a *bona fide* CRISPR-Cas system. We investigated three divergent CasΦ orthologs from metagenomic assemblies (fig. S2), hereafter referred to as CasΦ-1, CasΦ-2 and CasΦ-3. To examine CasΦ's ability to recognize and target DNA in bacterial cells, we tested whether CasΦ could protect *Escherichia coli* from plasmid transformation. CRISPR–Cas systems target DNA sequences following or preceding a 2–5 base pair (bp) Protospacer Adjacent Motif (PAM) for self-versus-non-self discrimination (Gleditzsch et al., 2019). To determine whether CasΦ uses a PAM, we transformed a library of plasmids containing randomized regions adjacent to crRNA-complementary target sites, thereby depleting plasmids harboring functional PAMs. This revealed the crRNA-guided double-strand DNA (dsDNA) targeting capability of CasΦ and minimal T-rich PAM sequences, including 5′-TBN-3′ PAMs (where B is G, T, or C) depleted for CasΦ-2 (Fig. 1C).

We next used the *E. coli* expression system and plasmid interference assay to determine the components required for CRISPR-CasΦ system function. RNA-sequencing analysis revealed transcription of the *casΦ* gene and the reduced CRISPR array but no evidence of other non-coding RNA such as a trans-activating CRISPR RNA (tracrRNA) within the locus (Fig. 1D). In addition, CasΦ activity could be readily reprogrammed to target other plasmid sequences by altering the guide RNA (fig. S3).

These findings suggest that in its native environment, CasΦ is a functional phage protein and *bona fide* CRISPR-Cas effector capable of cleaving crRNA-complementary DNA such as other phage (Fig. 1E). Furthermore, these results demonstrate that this single-RNA system is much more compact than other active CRISPR-Cas systems (Fig. 1F).

We next investigated the DNA recognition and cleavage requirements of CasΦ *in vitro*. RNA-seq revealed that the crRNA spacer, which is complementary to DNA targets, is 14-20 nucleotides (nt) long (Fig. 1D). Incubation of purified CasΦ (fig. S4) with crRNAs of different spacer sizes along with supercoiled plasmid or linear dsDNA revealed that DNA cleavage requires the presence of a cognate PAM and a spacer of ≥ 14 nt (Fig. 2A; fig. S5A). Analysis of the cleavage products showed that CasΦ generated staggered 5′-overhangs of 8-12 nt (Fig. 2B, C; fig. S5B, C), similar to the staggered DNA cuts observed for other type V CRISPR-Cas enzymes including Cas12a and CasX (Liu et al., 2019a; Zetsche et al., 2015). We also observed that CasΦ-2 and CasΦ-3 were more active *in vitro* than CasΦ-1, and the non-target strand (NTS) was cleaved faster than the target-strand (TS) within the RuvC active site (Fig. 2D; figs. S6A, S7; Supplementary Text). Furthermore, CasΦ was found to cleave ssDNA but not ssRNA *in cis* and *in trans* (fig. S6B, S8), suggesting that CasΦ may also target ssDNA MGEs or ssDNA intermediates. The trans-cleavage activity of CasΦ, observed only upon DNA recognition *in cis* (fig. S8), coupled with a minimal PAM requirement (Fig. 1C), may be useful for broader nucleic acid detection as previously demonstrated for type V and type VI Cas proteins (Chen et al., 2018; East-Seletsky et al., 2016; Gootenberg et al., 2017).

CRISPR-CasΦ systems must produce mature crRNA to guide foreign DNA cleavage. Other type V CRISPR-Cas proteins process pre-crRNAs using an internal active site distinct from the RuvC domain (Fonfara et al., 2016) or by recruiting Ribonuclease III to cleave a pre-crRNA:tracrRNA duplex (Burstein et al., 2017; Harrington et al., 2018; Shmakov et al., 2015; Yan et al., 2019). The absence of a detectable tracrRNA for CasΦ hinted that CasΦ may catalyze crRNA maturation on its own. To test this possibility, we incubated purified CasΦ with substrates designed to mimic the pre-crRNA structure (Fig. 3A). Reaction products corresponding to a 26-29 nt-long repeat and 20 nt spacer sequence of the crRNA were observed only in the presence of wild type CasΦ, corroborated by RNA-seq analysis of native loci (Figs. 1D; 3A, C; fig. S9). In control experiments, we found that pre-crRNA processing is strictly magnesium-dependent (Fig. 3B; fig. S9), which is different from other CRISPR-Cas RNA processing reactions and suggested a distinct cleavage mechanism. Notably, the RuvC domain requires magnesium to cleave DNA (Nowotny, 2009), and some RuvC domains have been reported to have endoribonucleolytic activity (Yan et al., 2019). Based on these observations, we tested CasΦ containing a RuvC-inactivating mutation and found it to be incapable of processing pre-crRNAs (Fig. 3B; fig. S9A, B). Both wild-type and catalytically inactivated CasΦ proteins bind crRNA, and their reconstituted complexes with pre-crRNA have similar elution profiles from a size exclusion column, suggesting no pre-crRNA binding or protein stability defect resulting from the RuvC mutation (fig. S10).

We hypothesized that if the RuvC domain is responsible for pre-crRNA processing, the products should contain 5′-phosphate and 2′- and 3′-hydroxyl moieties as observed in RNAs generated by the RuvC-related RNase HI enzymes (Nowotny, 2009). In contrast, other type V CRISPR-Cas enzymes process pre-crRNA by metal-

independent acid-base catalysis in an active site distinct from the RuvC, generating 2′-3′-cyclic phosphate crRNA termini, as observed for Cas12a (Swarts et al., 2017). Phosphatase treatment of CasΦ-generated crRNA followed by denaturing acrylamide gel analysis showed no change in the crRNA migration, distinct from the change in mobility detected for crRNA generated by Cas12a (Fig. 3C; fig. S9C). This result implies that no 2′-3′-cyclic phosphate was formed during the reaction catalyzed by CasΦ, in contrast to the acid-base catalyzed processing reaction by Cas12a (Fig. 3C, D). Together, these data demonstrate that CasΦ uses a single RuvC active site for both pre-crRNA processing and DNA cleavage.

The versatility and programmability of CRISPR-Cas systems for genome editing in virtually any organism have sparked a revolution in biotechnology and fundamental research (Knott and Doudna, 2018). To investigate whether CasΦ can be harnessed for human genome editing, we performed a gene disruption assay (Liu et al., 2019a) using CasΦ co-expressed with a crRNA in HEK293 cells (Fig. 4A). We found that CasΦ-2 and CasΦ-3, can induce targeted disruption of a genomically integrated EGFP gene (Fig. 4A; fig. S11). In one case, CasΦ-2 with an individual guide RNA was able to edit up to 33% of cells (Fig. 4A), comparable to levels initially reported for CRISPR–Cas9, CRISPR–Cas12a, and CRISPR–CasX (Liu et al., 2019a; Mali et al., 2013; Zetsche et al., 2015). We next tested if CasΦ-2 can be delivered as RNPs into plant protoplasts to edit the endogenous *Arabidopsis thaliana PDS3* gene (Fig. 4B; fig. S12). Next generation sequencing revealed that CasΦ-2 introduces primarily 8-10 bp deletions (Fig. 4B), consistent with the cleavage pattern observed *in vitro* (Fig. 2C). The small size of CasΦ in combination with its minimal PAM requirement will be particularly advantageous for both vector-based delivery into cells and a wider range of targetable genomic sequences, providing a powerful addition to the CRISPR-Cas toolbox.

Supplementary Text

To assess the role of the RuvC domain in DNA cleavage, the active site was mutated (D371A, D394A, or D413A) to produce a deactivated CasΦ variant (dCasΦ) that did not cleave dsDNA, ssDNA or ssRNA *in vitro* (fig. S6A, B). When expressed in *E. coli* along with crRNA, dCasΦ could not prevent transformation of a crRNA-complementary plasmid, consistent with a requirement for RuvC-catalyzed DNA cutting (fig. S3). This observation, together with the delayed cleavage of the TS after NTS cleavage (Fig. 2D; fig. S7), suggests that CasΦ cleaves each strand sequentially within the RuvC active site. Sequential strand cleavage is consistent with the dsDNA cutting mechanism of the type V CRISPR-Cas proteins (Cofsky et al., 2020; Swarts and Jinek, 2019) that share closest evolutionary origin with CasΦ.

2.4 Conclusions

Three other well-characterized Cas enzymes Cas9, Cas12a, and CasX, use one (Cas12a and CasX) or two active sites (Cas9) for DNA cutting and rely on a separate active site (Cas12a) or additional factors (CasX and Cas9) for crRNA processing (Fig. 4C). The finding that a single RuvC active site in CasΦ is capable of crRNA processing and DNA cutting suggests that size limitations of phage genomes, possibly in combination with large population sizes and higher mutation rates in phages compared to prokaryotes (Duffy et al., 2008; Lee and Marx, 2012; Lynch, 2006), led to a

consolidation of chemistries within one catalytic center. Such compact proteins may be particularly amenable to engineering and laboratory evolution to create new functionalities for genome manipulation, and highlight huge phages as an exciting forefront for discovery and biotechnological applications for human health.

2.5 **Figures**

Fig. 1. CasΦ is a *bona fide* CRISPR-Cas system from huge phages. (**A**) Maximum Likelihood phylogenetic tree of type V effector proteins and respective predicted ancestral TnpB nucleases. Bootstrap and approximate likelihood-ratio test values ≥ 90 are denoted on the branches with black circles. (**B**) Illustrations of genomic CRISPR-Cas loci of CasΦ, Cas14, and systems previously employed in genome editing applications. (**C**) Graphical representation of the PAM depletion assay and the resulting PAMs for three CasΦ orthologs. (**D**) RNA-sequencing results (left) mapped onto the native genomic loci of CasΦ orthologs and their upstream and downstream non-coding regions as cloned with reduced CRISPR-arrays into expression plasmids. Enlarged view of RNA mapped onto the first repeat-spacer pair (right). (**E**) Schematic of the hypothesized function of Biggiephage-encoded CasΦ in an instance of superinfection of its host. CasΦ may be used by the huge phage to eliminate competing mobile genetic elements. (**F**) Predicted molecular weights of the ribonucleoprotein (RNP) complexes of small CRISPR-Cas effectors and those functional in editing of mammalian cells.

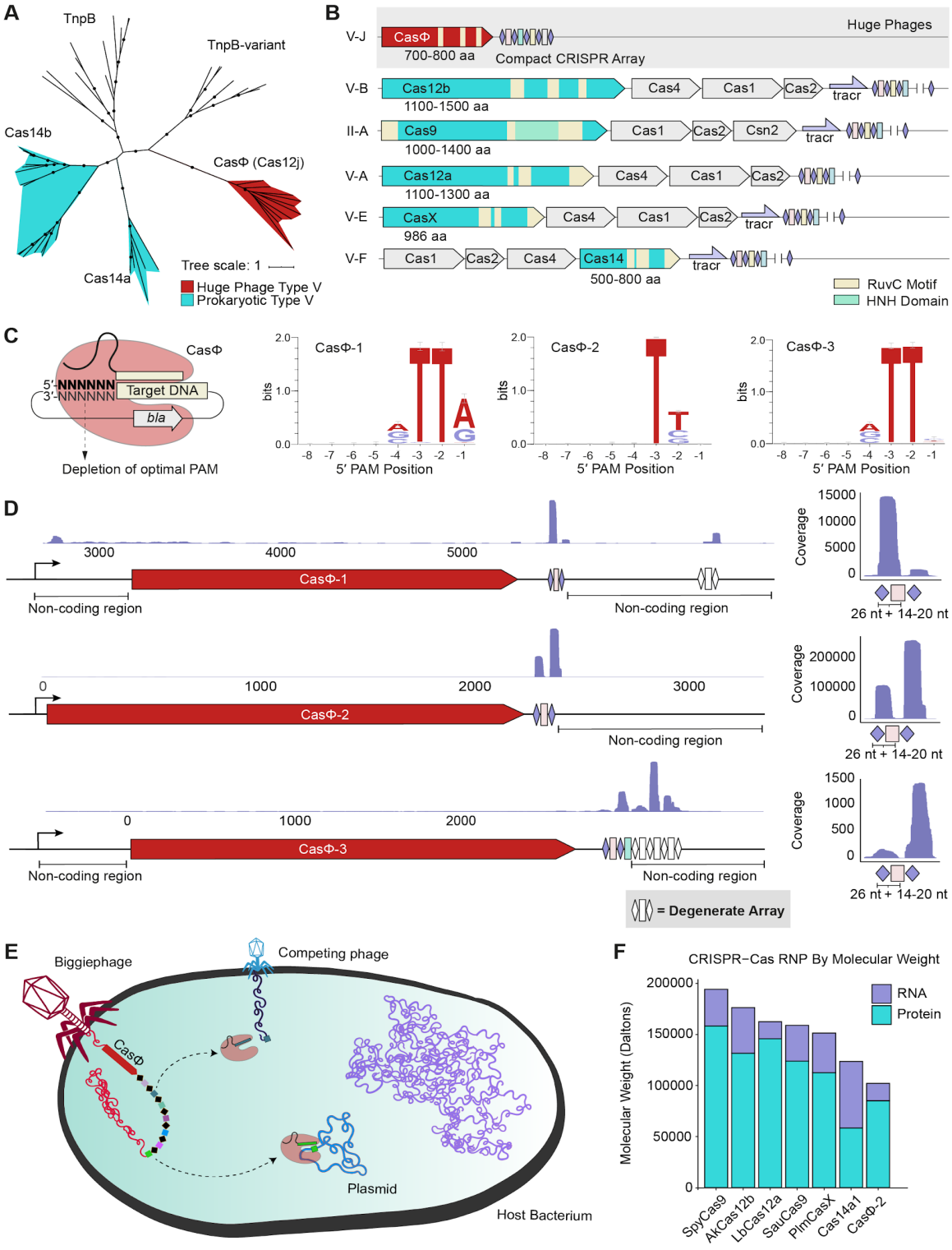

A
TnpB
TnpB-variant
Cas14b
CasΦ (Cas12j)
Cas14a
Tree scale: 1
Huge Phage Type V
Prokaryotic Type V
B
Huge Phages
V-J
CasΦ
700-800 aa
Compact CRISPR Array
V-B
Cas12b
1100-1500 aa
Cas4
Cas1
Cas2
tracr
II-A
Cas9
1000-1400 aa
Csn2
V-A
Cas12a
1100-1300 aa
V-E
CasX
986 aa
V-F
Cas14
500-800 aa
RuvC Motif
HNH Domain
C
CasΦ
5′-NNNNNN
3′-NNNNNN
Target DNA
bla
Depletion of optimal PAM
CasΦ-1
CasΦ-2
CasΦ-3
bits
5′ PAM Position
D
CasΦ-1
CasΦ-2
CasΦ-3
Non-coding region
Coverage
26 nt + 14-20 nt
= Degenerate Array
E
Biggiephage
Competing phage
CasΦ
Plasmid
Host Bacterium
F
CRISPR–Cas RNP By Molecular Weight
RNA
Protein
Molecular Weight (Daltons)
SpyCas9
AkCas12b
LbCas12a
SauCas9
PlmCasX
Cas14a1
CasΦ-2

Fig. 2. CasΦ cleaves DNA. (**A**) Supercoiled plasmid cleavage assay testing CasΦ RNPs reconstituted with crRNAs of different spacer lengths. (**B**) Cleavage assay targeting dsDNA oligo-duplices for mapping of the cleavage structure. (**C**) Scheme illustrating the cleavage pattern. (**D**) NTS and TS DNA cleavage efficiency (n = 3 each, mean ± s.d.). Data is shown in fig. S7B.

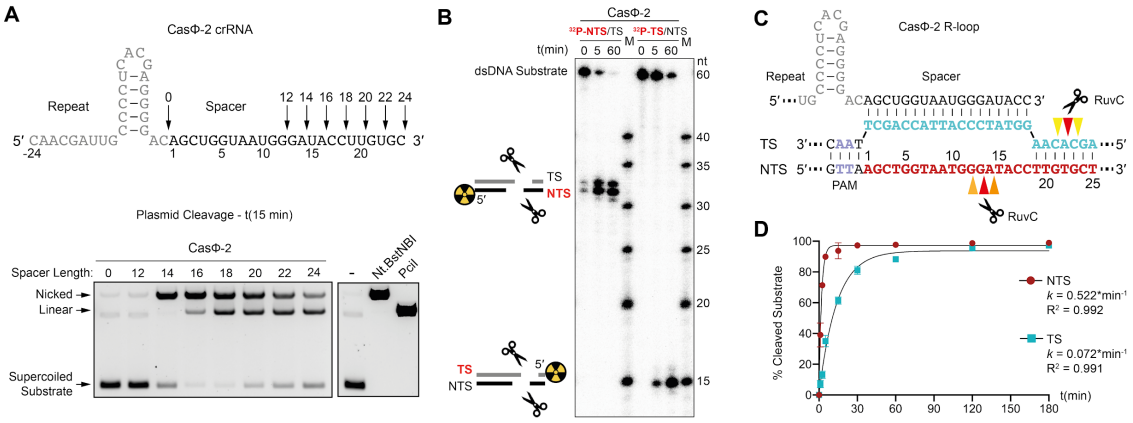


Fig. 3. CasΦ processes pre-crRNA within the RuvC active site. (**A**) pre-crRNA substrates and processing sites (red triangles) as derived from the OH-ladder in panel C. (**B**) Pre-crRNA processing assay for CasΦ-1 and CasΦ-2 in dependence of Mg^{2+} and RuvC active site residue variation (D371A and D394A) (n = 3 each, mean ± s.d.; t = 60 min). Data is shown in fig. S9B. (**C**) Left and middle: Alkaline hydrolysis ladder (OH) of the pre-crRNA substrate. Right: PNK-phosphatase treatment of the CasΦ and *Acidaminococcus sp.* Cas12a cleavage products. (**D**) Graphical representation of the mature crRNA termini chemistry of CasΦ and Cas12a and PNK-phosphorylase treatment outcomes.

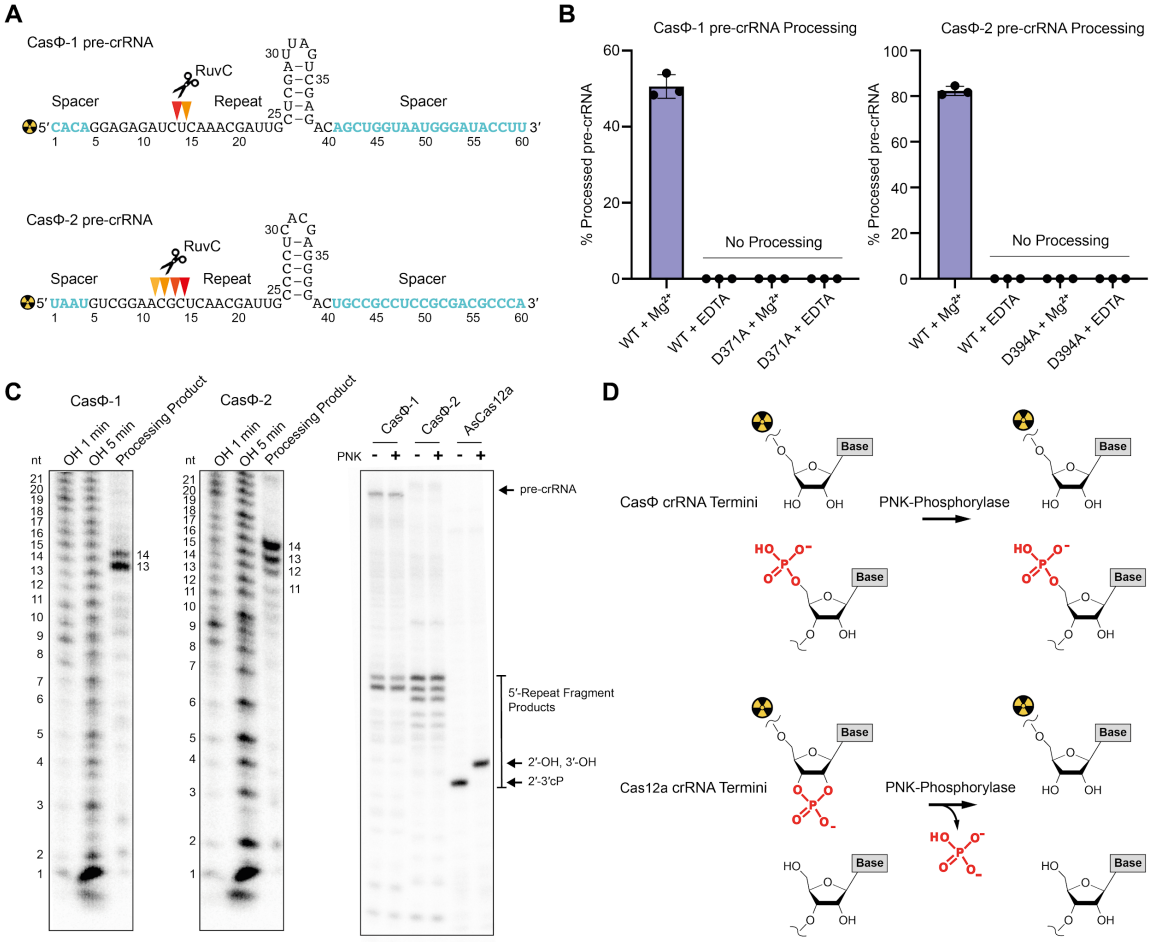
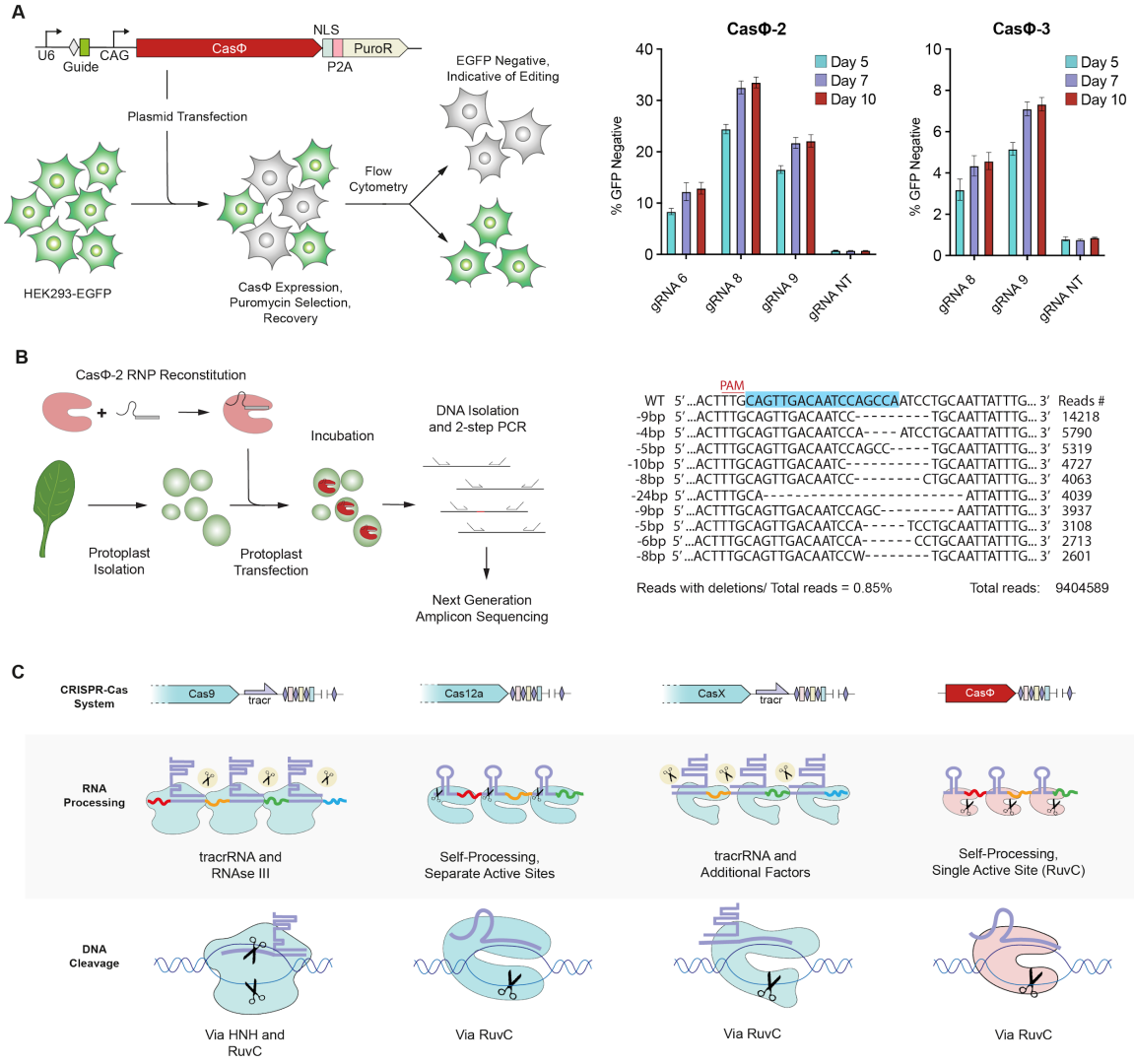


Fig. 4. CasΦ is functional for genome editing. (**A**) Experimental workflow of the GFP disruption assay (left) and GFP disruption using CasΦ-2 and CasΦ-3 and a non-targeting (NT) guide as a negative control (n = 3 each, mean ± s.d.). (**B**) Experimental workflow of CasΦ-2 RNP-mediated genome-editing in *A. thaliana* mesophyll protoplasts (left) and amplicon sequencing data (right) showing the most frequent deletions for gRNA33 in the targeted region (blue) within the *AtPDS3* gene. (**C**) Scheme illustrating the differences in RNA processing and DNA cutting for Cas9, Cas12a, CasX, and CasΦ.



2.6 **Methods**

Metagenomic assemblies, genome curation, and CRISPR-CasΦ detection

Metagenomic sequencing data was assembled using previously described methods (Al-Shayeb et al., 2020). Coding sequences (CDS) were predicted from sequence assemblies using prodigal with genetic code 11 (-m -g 11 -p single) and (-m -g 11 -p meta) and preliminary annotations and phage genome curations were performed as previously described (Al-Shayeb et al., 2020). Bowtie2 v2.3.4.1 was used to map reads to the de novo assembled sequences, and we retained unplaced mate pairs of mapped reads with shrinksam (github.com/bcthomas/shrinksam). N-filled gaps and local misassemblies were identified and corrected, and unplaced or incorrectly placed paired reads allowed extension of contig ends. Local assembly changes and extensions were verified with further read mapping. A database of CasΦ sequences from (Al-Shayeb et al., 2020) was generated using MAFFT v7.407 and hmmbuild. CDS from new assemblies were searched against the HMM database using hmmsearch with e-value < 1×10^{-5} and added to the database upon verification.

Phylogenetic analysis of type V systems

Cas protein sequences were collected from (Al-Shayeb et al., 2020; Burstein et al., 2017; Harrington et al., 2018; Makarova et al., 2019; Shmakov et al., 2015; Yan et al., 2019) and representatives from the TnpB superfamily were collected from (Makarova et al., 2019) and top BLAST hits from RefSeq. The resulting set was clustered at 90% amino acid identity to reduce redundancy. A new alignment of CasΦ with the resulting sequence set was generated using MAFFT LINSI with 1000 iterations and filtered to remove columns composed of gaps in 95% of sequences. Poorly aligned sequences were removed and the resulting set was realigned. The phylogenetic tree was inferred using IQTREE v1.6.6 using automatic model selection and 1000 bootstraps.

crRNA sequence analysis

CRISPR-RNA (crRNA) repeats from Phage-encoded CRISPR loci were identified using MinCED (github.com/ctSkennerton/minced) and CRISPRDetect. The repeats were compared by generating pairwise similarity scores using the Needleman-Wunsch algorithm followed by EMBOSS Needle. A heatmap was built using the similarity score matrix and hierarchical clustering produced dendrograms that were overlaid onto the heatmap to delineate different clusters of repeats.

Generation of plasmids

CasΦ loci, including an additional *E. coli* RBS upstream of *casΦ,* were ordered as G-blocks from Integrated DNA Technologies (IDT) and cloned using Golden Gate assembly (GG) under the control of a tetracycline-inducible promoter for RNA seq and PAM depletion plasmid interference experiments. Perfect repeat-spacer units of the by metagenomics identified CRISPR-arrays were reduced to a single repeat-spacer-repeat unit, amenable to stuffer-spacer exchange by GG-assembly (AarI-restriction sites). Subsequently, *CasΦ* gene sequences were subcloned by GG-assembly into pRSFDuet-1 (Novagen) within MCSI without tags for efficiency of transformation plasmid

interference assays, or fused to a C-terminal hexa-histidine tag for protein purification. For plasmid interference assays, mini-CRISPR arrays (repeat-spacer-repeat, or repeat-spacer-HDV ribozyme) amenable to stuffer-spacer exchange by GG-assembly (AarI-restriction sites) were cloned into MCS II of pRSFDuet. For genome editing experiments in human cells, *casΦ* genes were ordered as G-blocks from IDT encoding codon optimized genes for expression in human cells. G-blocks were cloned via GG-assembly into the vector backbone of pBLO62.5, downstream fused to two SV40 NLSs via a GSG linker encoding sequence. The guide encoding sequence of pBLO62.5 was exchanged to encode for a single CRISPR-repeat of the respective homologue, followed by a 20 bp stuffer spacer sequence amenable to GG-assembly exchange using the restriction enzyme SapI. For production of NLS tagged CasΦ for *in planta* genome editing, *E. coli* codon optimized *casΦ* was cloned using GG assembly into MCSI of pRSFDuet-1 (Novagen) downstream fused to two SV40 NLS sequences and a hexa-histidine tag. A list of plasmids and a brief description is given in table S1. Plasmid sequences and maps will be made available on addgene. To reprogram the CasΦ vectors to target different loci, stuffer-spacer were exchanged via GG-assembly to encode the guide for the selected target site (guide spacer sequences are listed in table S2). Mutations in the *casΦ* genes were introduced by GG-assembly to create *dcasΦ* genes.

PAM depletion DNA interference assay

PAM depletion assays were performed with both, CasΦ plasmids that either carried the whole CasΦ locus as derived from metagenomics (pPP049, pPP056 and pPP062), or with plasmids that contained only the *casΦ* gene and a mini CRISPR (pPP097, pPP102 and pPP107). Assays were performed as three individual biological replicates. Plasmids containing *casΦ* and mini CRISPRs were transformed into *E. coli* BL21(DE3) (NEB) and constructs containing CasΦ genomic loci were transformed into *E. coli* DH5α (QB3-Macrolab, UC Berkeley). Subsequently, electrocompetent cells were prepared by ice cold H_20 and 10 % glycerol washing. A plasmid library was constructed with 8 randomized nucleotides upstream (5') end of the target sequence (kind gift of Hannah Spinner). Competent cells were transformed in triplicate by electroporation with 200 ng library plasmids (0.1 mm electroporation cuvettes (Bio-Rad) on a Micropulser electroporator (Bio-Rad)). After a two-hour recovery period, cells were plated on selective media and colony forming units were determined to ensure appropriate coverage of all possible combinations of the randomized 5' PAM region. Strains were grown at 25 °C for 48 hours on media containing appropriate antibiotics (either 100 µg/mL carbenicillin and 34 µg/mL chloramphenicol, or 100 µg/mL carbenicillin and 50 µg/mL kanamycin) and 0.05 mM isopropyl-β-D-thiogalactopyranoside (IPTG), or 200 nM anhydrotetracycline (aTc), depending on the vector to ensure propagation of plasmids and CasΦ effector production. Subsequently, propagated plasmids were isolated using a QIAprep Spin Miniprep Kit (Qiagen).

PAM depletion sequencing analysis

Amplicon sequencing of the targeted plasmid was used to identify PAM motifs that are preferentially depleted. Sequencing reads were mapped to the respective plasmids and PAM randomized regions were extracted. The abundance of each possible 8 nucleotide combination was counted from the aligned reads and normalized to the total reads for each sample. Enriched PAMs were computed by calculating the

log ratio compared to the abundance in the control plasmids, and were used to produce sequence logos.

RNA preparation for RNAseq

Plasmids containing CasΦ loci were transformed into chemically competent *E. coli* DH5α (QB3-Macrolab, UC Berkeley). Preparations were performed as three individual biological replicates. Single colonies were picked to inoculate 5 mL starter cultures (LB, 34 µg/mL chloramphenicol) which were incubated at 37 °C shaking vigorously overnight. The next morning, main cultures were inoculated 1:100 (LB, 34 µg/mL chloramphenicol) and locus expression was induced with 200 nM aTc for 24 h at 16 °C. Cells were harvested by centrifugation, resuspended in lysis buffer (20 mM Hepes-Na pH 7.5 RT, 200 mM NaCl) and lysed using glass beads (0.1 mm glass beads, 4x 30 s vortex at 4 °C, interspaced by 30 s cool-down on ice). 200 µL cell lysis supernatant were transferred into Trizol for RNA extraction according to the manufacturer's protocol (Ambion). 10 µg RNA were treated with 20 units of T4-PNK (NEB) for 6 h at 37 °C for 2′-3′-dephosphorylation. Subsequently, 1 mM ATP was added and the sample was incubated for 1 h at 37 °C for *5′*-phosphorylation before heat inactivation at 65 °C for 20 min and subsequent Trizol purification.

RNA analysis by RNAseq

cDNA libraries were prepared using the RealSeq-AC miRNA library kit illumina sequencing (somagenics). cDNA libraries were subjected to Illumina MiSeq sequencing, and raw sequencing data was processed to remove adapters and sequencing artifacts, and high-quality reads were maintained. The resulting reads were mapped to their respective plasmids to determine the CRISPR locus expression and crRNA processing, and coverage was calculated at each region.

Efficiency of transformation plasmid interference assay

CasΦ vectors were transformed into chemically competent *E. coli* BL21(DE3) (NEB). Individual colonies for biological replicates were picked to inoculate three 5 mL (LB, Kanamycin 50 µg/mL) starter cultures to prepare electrocompetent cells the following day. 50 mL (LB, Kanamycin 50 µg/mL) main cultures were inoculated 1:100 and grown vigorously shaking at 37 °C to an OD_{600} of 0.3. Subsequently, the cultures were cooled to room temperature and *casΦ* expression was induced with 0.2 mM IPTG. Cultures were grown to an OD_{600} of 0.6-0.7 at 25 °C, before preparation of electrocompetent cells by repeated ice-cold H_20 and 10% glycerol washes. Cells were resuspended in 250 µL 10% glycerol. 90 µL aliquots were flash frozen in liquid nitrogen and stored at -80 °C. The next day, 80 µL competent cells were combined with 3.2 µL plasmid (20 ng/µL pUC19 target plasmid, or 20 ng/µL pYTK001 control plasmid), incubated for 30 min on ice and split into three individual 25 µL transformation reactions. After electroporation in 0.1 mm electroporation cuvettes (Bio-Rad) on a Micropulser electroporator (Bio-Rad), cells were recovered in 1 mL recovery medium (Lucigen) supplemented with 0.2 mM IPTG, shaking at 37 °C for one hour. Subsequently, 10-fold dilution series were prepared and 5 µL of the respective dilution steps were spot-plated on LB-Agar containing the appropriate antibiotics. Plates were incubated overnight at 37 °C and colonies were counted the following day to determine the transformation efficiency. To assess the transformation efficiency, the mean and

standard deviations were calculated from the cell forming units per ng transformed plasmids for the electroporation triplicates.

Protein production and purification

CasΦ overexpression vectors were transformed into chemically competent *E. coli* BL21(DE3)-Star (QB3-Macrolab, UC Berkeley) and incubated overnight at 37 °C on LB-Kan agar plates (50 µg/mL Kanamycin). Single colonies were picked to inoculate 80 mL (LB, Kanamycin 50 µg/mL) starter cultures which were incubated at 37 °C shaking vigorously overnight. The next day, 1.5 L TB-Kan medium (50 µg/mL Kanamycin) were inoculated with 40 mL starter culture and grown at 37 °C to an OD_{600} of 0.6, cooled down on ice for 15 min and gene expression was subsequently induced with 0.5 mM IPTG followed by incubation overnight at 16 °C. Cells were harvested by centrifugation and resuspended in wash buffer (50 mM HEPES-Na pH 7.5 RT, 1 M NaCl, 20 mM imidazole, 5 % glycerol and 0.5 mM TCEP), subsequently lysed by sonication, followed by lysate clarification by centrifugation. The soluble fraction was loaded on a 5 mL Ni-NTA Superflow Cartridge (Qiagen) pre-equilibrated in wash buffer. Bound proteins were washed with 20 column volumes (CV) wash buffer and subsequently eluted in 5 CV elution buffer (50 mM HEPES-Na pH 7.5 RT, 500 mM NaCl, 500 mM imidazole, 5 % glycerol and 0.5 mM TCEP). The eluted proteins were concentrated to 1 mL before injection into a HiLoad 16/600 Superdex 200pg column (GE Healthcare) pre-equilibrated in size-exclusion chromatography buffer (20 mM HEPES-Na pH 7.5 RT, 500 mM NaCl, 5 % glycerol and 0.5 mM TCEP). Peak fractions were concentrated to 1 mL and concentrations were determined using a NanoDrop 8000 Spectrophotometer (Thermo Scientific). Proteins were purified at a constant temperature of 4 °C and concentrated proteins were kept on ice to prevent aggregation, snap frozen in liquid nitrogen and stored at -80 °C. AsCas12a was purified as previously described (Knott et al., 2019).

In vitro cleavage assays - spacer tiling

Plasmid targets were cloned by GG-assembly of spacer 2, found in the CRISPR-array of CasΦ-1, downstream to a cognate 5'-TTA PAM, or non-cognate 5'-CCA PAM into pYTK095 (Target sequences are given in table S3). Supercoiled plasmids were prepared by propagation of the plasmid overnight at 37 °C in *E. coli* Mach1 (QB3-Macrolab, UC Berkeley) in LB and Carbenicillin (100 µg/mL) and subsequent preparation using a Qiagen Miniprep kit (Qiagen). Linear DNA targets were prepared by PCR from the plasmid target. crRNA guides were ordered as synthetic RNA oligos from IDT (table S4), dissolved in DEPC H_20 and heated for 3 min at 95 °C before cool down at RT. Active RNP complexes were assembled at a concentration of 1.25 µM by mixing protein and crRNA (IDT) in a 1:1 molar ratio in cleavage buffer (10 mM Hepes-K pH 7.5 RT, 150 mM KCl, 5 mM $MgCl_2$, 0.5 mM TCEP) and incubation at RT for 30 min. Cleavage reactions were initiated by addition of DNA (10 nM) to preformed RNP (1 µM) in reaction buffer (10 mM Hepes-K pH 7.5 RT, 150 mM KCl, 5 mM $MgCl_2$, 0.5 mM TCEP). The reactions were incubated at 37 °C, quenched with 50 mM EDTA and stored in liquid nitrogen. Samples were thawed and treated with 0.8 units proteinase K (NEB) for 20 min at 37 °C. Loading dye was added (Gel Loading Dye Purple 6X, NEB) and samples were analyzed by electrophoresis on a 1% agarose gel and stained with SYBR Safe (Thermo Fisher Scientific). For comparison to cleavage products, supercoiled plasmids were digested with PciI (NEB) for linearization and Nt.BstNBI (NEB) for

plasmid nicking and open circle formation. Comparable cleavage assays under varied conditions (n ≥ 3) showed consistent results.

In vitro cleavage assays - radiolabeled nucleic acids

Active CasΦ RNP complexes were assembled in a 1:1.2 molar ratio by diluting CasΦ protein to 4 µM and crRNA (IDT) to 5 µM in RNP assembly buffer (20 mM HEPES-Na pH 7.5 RT, 300 mM KCl, 10 mM $MgCl_2$, 20 % glycerol, 1 mM TCEP) and incubation for 30 min at RT. Substrates were 5′-end-labelled using T4-PNK (NEB) in the presence of ^{32}P-γ-ATP (Substrate sequences are given in table S3). Oligo-duplex targets were generated by combining ^{32}P-labelled and unlabelled complementary oligonucleotides in a 1:1.5 molar ratio. Oligos were hybridized to a DNA-duplex concentration of 50 nM in hybridization buffer (10 mM Tris-Cl pH 7.5 RT, 150 mM KCl), by heating for 5 min to 95 °C and a slow cool down to RT in a heating block. Cleavage reactions were initiated by combining 200 nM RNP with 2 nM substrate in reaction buffer (10 mM HEPES-Na pH 7.5 RT, 150 mM KCl, 5 mM $MgCl_2$, 10 % glycerol, 0.5 mM TCEP) and subsequently incubated at 37 °C. For trans-cleavage assays, guide complementary activator substrates were diluted in oligonucleotide hybridization buffer (10 mM Tris pH 7.8 RT, 150 mM KCl) to a concentration of 4 µM, heated to 95 °C for 5 min, and subsequently cooled down at RT to allow duplex formation for double stranded activator substrates. Cleavage reactions were set up by combining 200 nM RNP with 100 nM activator substrate and incubation for 10 min at RT before addition of 2 nM ssDNA, or ssRNA, trans cleavage substrates. Reactions were stopped by addition of two volumes formamide loading buffer (96 % formamide, 100 µg/mL bromophenol blue, 50 µg/mL xylene cyanol, 10 mM EDTA, 50 µg/mL heparin), heated to 95 °C for 5 min, and cooled down on ice before separation on a 12.5 % denaturing urea-PAGE. Gels were dried for 4 h at 80 °C before phosphor-imaging visualization using an Amersham Typhoon scanner (GE Healthcare). Technical replicates (n ≥ 2) and comparable cleavage assays under varied conditions (n ≥ 3) of biological replicates (n ≥ 2) showed consistent results. Bands were quantified using ImageQuant TL (GE) and cleaved substrate was calculated from the intensity relative to the intensity observed at t = 0 min. Curves were fit to a One-Phase-Decay model in Prism 8 (graphpad) to derive the rate of cleavage.

In vitro pre-crRNA processing assay

Pre-crRNA substrates were 5′-end-labelled using T4-PNK (NEB) in the presence of ^{32}P-γ-ATP (Substrate sequences are given in table S3). Processing reactions were initiated by combining 50 nM CasΦ with 1 nM substrate in pre-crRNA processing buffer (10 mM Tris pH 8 RT, 200 mM KCl, 5 mM $MgCl_2$ or 25 mM EDTA, 10 % glycerol, 1 mM DTT) and subsequently incubated at 37 °C. Substrate hydrolysis ladders were prepared using the alkaline hydrolysis buffer according to the manufacturer's protocol (Ambion). 10 µL of the processing reaction products were treated with 10 units T4-PNK (NEB) for 1 h at 37 °C in the absence of ATP for termini chemistry analysis. Reactions were stopped by addition of two volumes formamide loading buffer (96 % formamide, 100 µg/mL bromophenol blue, 50 µg/mL xylene cyanol, 10 mM EDTA, 50 µg/mL heparin), heated to 95 °C for 3 min, and cooled down on ice before separation on a 12.5 %, or 20 %, denaturing urea-PAGE. Gels were dried for 4 h at 80 °C before phosphor-imaging visualization using an Amersham Typhoon scanner (GE Healthcare). Technical replicates (n ≥ 3) and comparable cleavage assays under varied conditions (n ≥ 3) of biological replicates (n ≥ 2) showed consistent results. Bands were quantified using

ImageQuant TL (GE) and processed RNA was calculated from the intensity at t = 60 min relative to the intensity observed at t = 0 min.

Analytical size exclusion chromatography

500 µL sample (5-10 µM protein, RNA, or reconstituted RNPs) were injected onto a S200 XK10/300 size exclusion chromatography (SEC) column (GE Healthcare) pre-equilibrated in SEC buffer (20 mM HEPES-Cl pH 7.5 RT, 250 mM KCl, 5 mM $MgCl_2$, 5 % glycerol and 0.5 mM TCEP). Prior to SEC, CasΦ RNP complexes were assembled by incubating CasΦ protein and pre-crRNA for 1 h in 2X pre-crRNA processing buffer (20 mM Tris pH 8 RT, 400 mM KCl, 10 mM $MgCl_2$, 20 % glycerol, 2 mM DTT).

Genome editing in human cells

The GFP HEK293 reporter cells were generated via lentiviral integration as previously described (Richardson et al., 2016). Cells were routinely tested for absence of mycoplasma using the MycoAlert Mycoplasma Detection Kit (Lonza), according to the manufacturer's protocol. GFP HEK293 reporter cells were seeded into 96-well plates and transfected at 60-70% confluency the next day according to the manufacturer's protocol with lipofectamine 3000 (Life Technologies) and 200 ng of plasmid DNA encoding the CasΦ gRNA and CasΦ–P2A–PAC fusion. As a comparison control, 200 ng of plasmid DNA encoding the SpyCas9 sgRNA and SpyCas9–P2A–PAC fusion was transfected identically, with target sequences adjusted for PAM differences. 24 hours post-transfection, successfully transfected cells were selected for by adding 1.5 µg/mL puromycin to the cell culture media for 72 hours. Cells were passaged regularly to maintain sub-confluent conditions and then analyzed on an Attune NxT Flow Cytometer with an autosampler. Cells were analyzed on the flow cytometer after 10 days to allow for clearance of GFP from cells.

Protoplast isolation and transfection

A. thaliana plants (Col-0 ecotype) were grown with 12 h light/12 h dark photoperiod under low light (75 µE m^{-2} s^{-1}) and mesophyll protoplasts were isolated from leaves of 4-week-old plants as described previously (Yoo et al., 2007) . In brief, *A. thaliana* leaves were cut into 0.5-1 mm stripes with sharp razor blades and submerged in enzyme solution (20 mM MES pH 5.7, 0.4 M mannitol, 20 mM KCl, 1.5% cellulase R10, 0.4% macerozyme R10, enzymes from Yakult Pharmaceutical Ind. Co., Ltd., Japan). The leaf stripes in enzyme solution were vacuum infiltrated for 30 min in dark and then incubated in dark for 3 h at room temperature. The protoplasts were released during this incubation. After the incubation, the enzyme/protoplast solution was diluted with equal volume of W5 solution (2 mM MES pH 5.7, 154 mM NaCl, 125 mM $CaCl_2$, 5 mM KCl), and filtered through 70-µm nylon mesh (Carolina Biological Supplies, cat 65222N) into round bottom tubes. Protoplasts were collected by centrifuging the flow-through at 100 g for 2 min at 4 °C. Supernatant was removed and protoplasts (pellet) were resuspended in W5 solution at 2 x 10^5 cells/ml. Resuspended protoplasts were kept on ice for 30 min for resting. During the resting, the protoplasts were re-collected at the bottom of tubes by gravity. Then the supernatant was removed as much as possible and the protoplasts were resuspended with MMG solution (4 mM MES PH 5.7, 0.4 M mannitol, 15 mM $MgCl_2$) to the same volume (2 x 10^5 cells/ml). Sterile and RNase-

free reagents were used for protoplast isolation. Active CasΦ-2 RNP complexes were reconstituted by diluting CasΦ-2-NLS protein, purified as described above, to 4 µM and gRNA to 5 µM in RNP assembly buffer as described above and incubated for 30 min at RT. 26 µL of 4 µM RNP were first added to a round-bottom 2 mL tube. Then 200 µL of protoplasts (at 2 x 10^5 cells/mL) were added to the tube. 2 µL of 5 µg/µL salmon sperm DNA was added and mixed gently by tapping the tube 3-4 times. Then, 228 µL of fresh, sterile and RNase free PEG-$CaCl_2$ solution (40% PEG4000, 0.2 M mannitol, 100 mM $CaCl_2$) was added to the protoplast-RNP mixture and mixed well by gently tapping tubes. The protoplasts with PEG solution were incubated at room temperature for 10 min, then 880 µL of W5 solution was added and mixed with the protoplasts by inverting the tube 2-3 times to stop the transfection. Protoplasts were harvested by centrifugation at 100 rcf for 2 min, resuspended in 1 mL WI solution (4 mM MES pH 5.7, 0.5 M mannitol, 20 mM KCl) and plated into 6-well plates pre-coated with 5% calf serum. The lids of the 6-well plates were closed to begin the incubation of the protoplasts. For control samples, 10 µg of HBT-sGFP plasmid (ABRC stock CD3-911) were added to 200 µl protoplasts and followed the same transfection and plating procedure as stated above. For the initial RNP screening experiment, the protoplasts were incubated at RT for 12 h, then moved to 37 °C for 2.5 h. Then, the protoplasts were moved back to room temperature and incubated for a total duration of 36 h. For the independent experiment where gRNA28, gRNA31 and gRNA33 were tested, the protoplasts were incubated at RT for 12 hours, then moved to 37 °C for 2.5 h. Then, the protoplasts were moved back to room temperature and incubated for a total duration of 48 h. At the end of the incubations, the protoplasts were collected by a first centrifugation at 100 rcf for 2-3 min. Keeping the pellet, the supernatant was moved to another tube and went through another centrifugation at 3000 rcf for 3 min to collect any residue protoplasts. Pellets from these two centrifugations were combined and flash frozen for further analysis.

Amplicon sequencing

DNAs of protoplast samples were extracted using the Qiagen DNeasy plant mini kit. Amplicons were obtained by two rounds of PCR (2-step PCR). Amplification primers for the first round of PCR were designed to have the 3′ part of primer with sequences flanking a 200-300 bp fragment of the *AtPDS3* gene around the guide RNA of interest. The 5′ part of the primer contained sequences to be bound by common sequencing primers (for reading paired-end reads, read 1 and read 2). The primers were designed so that the gRNA sequence started from within 100 bp from the beginning of read 1. The first round of PCR was done with Phusion High-Fidelity Polymerase (ThermoFisher cat F530N). Half of all DNA from a protoplast transfection sample was used as the template, and 25 cycles of amplification were done for the first round. Then the reaction was cleaned by 1x Ampure XP beads (Beckman Coulter A63881). The elution from the cleanup was used as the template for the second round of PCR by Phusion High-Fidelity Polymerase with 12 cycles. The second round of PCR was designed so that indices were added to each sample. The samples were then purified by 0.8-1 X Ampure beads for 1-2 rounds until no primer dimers were seen, with fragments below 200 bp considered primer dimers. Then amplicons were sent for paired-end 150 bp next generation sequencing.

Amplicon sequencing result analysis

Reads were first quality- and adaptor-trimmed with trim-galore (version 0.4.4), then mapped to the *AtPDS3* genomic region. Sorted and indexed bam files were used as input files for further analysis by the CrispRvariants R package. Each mutation pattern with corresponding reads counts were exported by the CrispRvariants R package. After assessing all control samples, a criterion to classify reads containing deletion was established: only reads with ≥ 3 bp deletion of same pattern (deletion of same size starting with same location) with ≥ 100 reads counts from a sample were counted into the reads number with deletion. This criterion was established due to the fact that 1 bp indels and occasionally 2 bp deletions were observed with reads number >100 in control samples. Larger deletions were also observed at very low frequencies in control samples. These observations indicate that occasional PCR inaccuracy and low-quality sequencing in a small fraction of reads can result in the deletion patterns with corresponding read number ranges as stated above in control samples. These stringent criteria were employed so that the counted deletion signals were true signals indicating editing events, though it is possible that CasΦ-2 might be able to create 1-2 bp deletions at lower frequency.

3 Chapter 3: Borgs are giant extrachromosomal elements with the potential to augment methane oxidation

Basem Al-Shayeb, Marie C. Schoelmerich, Jacob West-Roberts, Luis E. Valentin-Alvarado, Rohan Sachdeva, Susan Mullen, Alexander Crits-Christoph, Michael J. Wilkins, Kenneth H. Williams, Jennifer A. Doudna, Jillian F. Banfield

3.1 **Abstract**

Anaerobic methane oxidation exerts a key control on greenhouse gas emissions(Wallenius et al., 2021), yet factors that modulate the activity of microorganisms performing this function remain little explored. In studying groundwater, sediments, and wetland soil where methane production and oxidation occur, we discovered extraordinarily large, diverse DNA sequences that primarily encode hypothetical proteins. Four curated, complete genomes are linear, up to ~1 Mbp in length and share genome organization, including replichore structure, long inverted terminal repeats, and genome-wide unique perfect tandem direct repeats that are intergenic or generate amino acid repeats. We infer that these are highly divergent archaeal extrachromosomal elements with a distinct evolutionary origin. Gene sequence similarity, phylogeny, and local divergence of sequence composition indicate that many of their genes were assimilated from methane-oxidizing *Methanoperedens* archaea. We refer to these elements as "Borgs". We identified at least 19 different Borg types coexisting with *Methanoperedens* spp. in four distinct ecosystems. Borgs provide methane-oxidizing *Methanoperedens* archaea access to genes involved in redox reactions and energy generation (e.g., clusters of multiheme cytochromes, methyl coenzyme M reductase) and response to changing environmental conditions. Thus, Borgs could play previously unrecognized roles in the metabolism of a group of archaea known to modulate greenhouse gas emissions.

N.B. All main figures for this manuscript can be found below in their dedicated section. All supplementary files (including figures and tables) can be found online with the published manuscript.

3.2 Introduction

Of all of Earth's biogeochemical cycles, the methane cycle may be most tightly linked to climate. Methane (CH_4) is a greenhouse gas roughly 30 times more potent than carbon dioxide (CO_2), and approximately 1 gigaton is produced annually by methanogenic (methane-producing) archaea that inhabit anoxic environments(Thauer et al., 2008). The efflux of methane into the atmosphere is mitigated by methane-oxidizing microorganisms (methanotrophs). In oxic environments CH_4 is consumed by aerobic bacteria that use a methane monooxygenase (MMO) and O_2 as terminal electron acceptor(Hanson and Hanson, 1996), whereas in anoxic environments anaerobic methanotrophic archaea (ANME) use a reverse methanogenesis pathway to oxidize CH_4, the key enzyme of which is methyl-CoM reductase (MCR)(Boetius et al., 2000; Hallam et al., 2003). Some ANMEs rely on a syntrophic partner to couple CH_4 oxidation to the reduction of terminal electron acceptors, yet *Methanoperedens* (ANME-2d, phylum *Euryarchaeota*) can directly couple CH_4 oxidation to the reduction of iron, nitrate or manganese(Ettwig et al., 2016; Leu et al., 2020). Some phenomena have been suggested to modulate methane oxidation rates. For example, some phages can decrease methane oxidation rates by infection and lysis of methane-oxidizing bacteria(Lee et al., 2021), and others with the critical subunit of MMO(Chen et al., 2020) likely increase the ability of their host bacteria to conserve energy during phage replication. Here, we report the discovery of novel extrachromosomal elements (ECEs) that are inferred to replicate within *Methanoperedens* spp*.* Their numerous and diverse metabolism-relevant genes, huge size, and distinctive genome architecture distinguish these archaeal ECEs from all previously reported elements associated with archaea(Ausiannikava et al., 2018; Ng et al., 1998; Wang et al., 2015) and from bacteriophages, which typically have one or a few biogeochemically relevant genes(Anantharaman et al., 2014; Lindell et al., 2004). We hypothesize that these novel ECEs may substantially impact the capacity of *Methanoperedens* spp to oxidize methane.

3.3 Results and Discussion

Genome Structure and Features

By analysis of whole-community metagenomic data from wetland soils in CA (**Fig. S1**), we discovered enigmatic genetic elements, the genomes for three of which were carefully manually curated to completion (**methods**). From sediment samples from the Rifle, CO aquifer(Hug et al., 2015), we recovered partial genomes from a single population related to those from the wetland soils; the sequences were combined and manually curated to ultimately yield a fourth complete genome (**methods**). All four curated genomes are linear and terminated by >1 kbp inverted repeats. The genome sizes range from 661,708 to 918,293 kbp (**Fig. 1A; Table 1; Table S1**). Prominent features of all genomes are 25 - 54 regions composed of perfect tandem direct repeats (**Fig. 1B; Table S2)** that are novel (**Fig. S2**) and occur in both intergenic regions and in genes where they usually introduce perfect amino acid repeats (**Table S2**). All genomes have two replichores of unequal lengths and initiate replication at the chromosome ends (**Fig S3**).

Each replichore carries essentially all genes on one strand (**Fig. 1A)**. Although the majority of genes are novel, ~21% of the predicted proteins have best matches to proteins of Archaea (**Fig. S4A**), and the vast majority of these have best matches to proteins of *Methanoperedens spp.* (**Fig. S4B**). Notably, the GC contents of the four genomes are ~10% lower than those of previously reported and coexisting *Methanoperedens* species (**Fig. 2A**). We rule out the possibility that these sequences represent genomes of novel Archaea, as they lack almost all of the single-copy genes found in archaeal genomes and sets of ribosomal proteins that are present even in obligate symbionts **(Figs. S5, S6A, Tables S3-S6**). There are no additional sequences in the datasets that could comprise additional portions of these genomes. Thus, they are clearly neither part of *Methanoperedens* spp. genomes nor parts of the genomes of other archaea.

The sequences are much more abundant in deep, anoxic soil samples (**Fig. S7A, B)**. Abundances of *Methanoperedens* spp. and some ECEs are tightly correlated over a set of 50 different wetland soil samples. This observation supports other indications that these ECEs associate with *Methenoperedens* and suggests that specific ECEs have distinct *Methanoperedens* sp. hosts (**Fig. 2B**). This is true for one ECE whose abundances correlate reasonably well with a specific host group, where ECE : *Methanoperedens spp.* abundance ratios range from 2:1 to 8:1. Given their up to ~1 Mbp length, there may be more ECE DNA in some host cells than host DNA.

A few percent of the genes in the genomes have locally elevated GC contents that approach, and in some cases match, those of coexisting *Methanoperedens spp* (**Fig. 1B**). This, and the very high similarity of some protein sequences to those of *Methanoperedens spp*, indicates that these genes were acquired by lateral gene transfer from *Methanoperedens spp*. Other genes with best matches to *Methanoperedens spp* genes have lower GC contents (closer to those of these ECEs at ~33%), suggesting that their DNA composition has partly or completely ameliorated since acquisition(Lawrence and Ochman, 1997).

Archaeal ECEs include viruses(Hua et al., 2019), plasmids(DasSarma et al., 2009), and mini-chromosomes, sometimes also referred to as megaplasmids(Ausiannikava et al., 2018; Ng et al., 1998; Wang et al., 2015). The genomes reported here are much larger than those of all known archaeal viruses, some of which have small, linear genomes(Wang et al., 2015), and at least three are larger than any known bacteriophage(Al-Shayeb et al., 2020). These linear elements are larger than all of the reported circular plasmids that affiliate with halophiles, methanogens, and archaeal thermophiles. We did not detect genes for plasmid partitioning or conjugative systems, rRNA loci, or encoded viral proteins (Suppl Table S3), and the genomes were markedly different from recently reported *Methanoperedens spp* plasmids(Schoelmerich et al., 2022). The distinctly lower GC content and variable copy number argue against their classification as archaeal minichromosomes(Hall et al., 2022; Wang et al., 2015). Thus, we cannot confidently classify the ECEs as viruses, plasmids, or minichromosomes. Moreover, the protein family profiles are quite distinct from those of archaeal and bacterial ECEs (**Fig. 2D, Fig. S5**). Some bacterial megaplasmids have been reported to be very large and linear, but they typically encode few or no essential genes(Medema et al., 2010), and if they contain repeats, they are interspaced (i.e., not tandem)(Wagenknecht et al., 2010). Each distinctive feature of the ECEs has been reported in microbial genomes, plasmids, or viruses, but the combination of these features in these huge ECEs is unique. Thus, we conclude that the genomes represent novel archaeal ECEs that occur in association with, but not as part of, *Methanoperedens*

spp genomes. We refer to these as Borgs, a name that reflects their propensity to assimilate genes from organisms, most notably *Methanoperedens spp*.
Using criteria based on the features of the four complete Borgs, we searched for additional Borgs in our metagenomic datasets from a wide diversity of environment types. From the wetland soil, we constructed bins for 11 additional Borgs, some of which exceed 1 Mbp in length (**Table 1, Table S1**). Other Borgs were sampled from the Rifle, CO aquifer, discharge from an abandoned Corona mercury mine in Napa County, CA, and from shallow riverbed pore fluids in the East River, CO. In total, we recovered genome bins for 19 different Borgs, each of which was assigned a color-based name. Interestingly, we found no Borgs in some samples, despite the presence of *Methanoperedens spp* at very high abundance levels (**Fig. S7**). Thus, it appears that these ECEs do not associate with all *Methanoperedens spp*.

Pairs of the four complete Borg genomes (Purple, Black, Sky, and Lilac) and three fragments of the Orange Borg are alignable over much of their lengths (**Fig 1A**). The Rose and Sky Borg genomes are also largely syntenous and were reconstructed from different samples that contain these Borgs at very different abundance levels. Intriguingly, despite only sharing <50% average nucleotide identity across most of their genomes, the genomes have multiple regions that share 100% nucleotide identity, one of which is ~11 kbp in length (**Fig S8B, C**). This suggests that these two Borgs recombined, indicating that they recently co-existed within the same host cell.

Borg gene inventories

Many Borg genomes encode mobile element defense systems, including RNA-targeting type III-A CRISPR-Cas systems that lack spacer acquisition machinery, a feature previously noted in huge bacterial viruses(Al-Shayeb et al., 2020). An Orange Borg CRISPR spacer targets a gene in a mobile region in a coexisting *Methanoperedens spp* (**Fig. S8D**), further supporting the conclusion that *Methanoperedens spp* are the Borg hosts.

The four complete genomes and almost all of the near-complete and partial genomes encode ribosomal protein L11 (rpL11), and some have one or two other ribosomal proteins **(Fig. S7A)**. The rpL11 protein sequences form a group that places phylogenetically sibling to those of *Methanoperedens spp* (**Fig. S9**), further reinforcing the link between Borgs and *Methanoperedens spp*. Four additional rpL11 sequences were identified on short contigs from the wetland group with the Borg sequences and likely represent additional Borgs (**Table S1**). The topology of the rpL11 tree, and similar topologies observed for phylogenetic trees constructed using other ribosomal proteins, MCR proteins, electron transfer flavoproteins, and aconitase, may indicate the presence of translation-related genes in the Borg ancestor **(Fig. S7A; Fig. S9)**.

The most highly represented Borg genes are glycosyltransferases, genes involved in DNA and RNA manipulation, transport, energy, and the cell surface (PEGA and S-layer proteins). Also prevalent are many membrane-associated proteins of unknown function that may impact the membrane profile of their host (**Fig. 2C**). At least seven Borgs carry a *nifHDK* operon for nitrogen fixation, also predicted in *Methanoperedens spp* genomes, and may augment their host's influence on nitrogen cycling (**Fig. 1B, Fig. S10, Table S6**). Potentially related to survival under resource limitation are genes in at least 10 Borg genomes for synthesis of the carbon storage compound polyhydroxyalkanoate, a capacity also predicted for *Methanoperedens spp.(Liu et al., 2019b)*. Other stress-related

genes encode tellurium resistance proteins that do not occur in *Methanoperedens spp.* genomes (**Table S5**). Intriguingly, all Borgs carry large FtsZ-tubulin homologs that may be involved in cell division, and proteins with the TEP1-like TROVE domain protein that also do not occur in *Methanoperedens spp.* genomes (**Table S5**). These may form a complex similar to Telomerase, Ro, or Vault ribonucleoproteins, although their function remains unclear(Berger et al., 2009). Several Borgs encode two genes of the TCA cycle (citrate synthase and aconitase, **Fig. S10C)**.

Many Borg genes are predicted to play roles in redox and respiratory reactions. The Black Borg encodes *cfbB* and *cfbC*, genes involved in biosynthesis of F430, the cofactor for methyl-coenzyme M reductase (MCR), the central enzyme involved in methane oxidation by *Methanoperedens spp.* The similarity in GC content of Borg *cfbB* and *cfbC* and protein sequences of coexisting *Methanoperedens spp.* suggests that these genes were acquired from *Methanoperedens spp.* recently. The Blue and Olive Borgs encode *cofE* (coenzyme F420:L-glutamate ligase), which is involved in the biosynthesis of a precursor for F420. The Blue and Pink Borgs have an electron bifurcating complex **(Fig. S10B)** that includes D-Lactate dehydrogenase. Eight Borgs encode genes for biosynthesis of tetrahydromethanopterin, a coenzyme used in methanogenesis, and ferredoxin proteins which may serve as electron carriers. The Green and Sky Borgs also encode 5,6,7,8-tetrahydromethanopterin hydro-lyase (Fae), an enzyme responsible for formaldehyde detoxification and involved in pentose-phosphate synthesis. Also identified were genes encoding carbon monoxide dehydrogenase (CODH), plastocyanin, cupredoxins, and many multiheme cytochromes (MHC). These results indicate substantial Borg potential to augment the energy conservation by *Methanoperedens spp.* This is especially apparent for the Lilac Borg.

Lilac Borgs' potential to augment *Methanoperedens spp.* function

We analyzed the genes of the complete Lilac Borg genome in detail as, unlike the other Borgs, the Lilac Borg co-occurs with a single group of *Methanoperedens spp.* that likely represent the host (**Fig. 3, Table S7**). Remarkably, this Borg genome encodes an MCR complex, which is central to methanogenesis and reverse methanogenesis. The *mcrBDGA* cluster shares high (75-88%) amino acid sequence identity with that of the coexisting *Methanoperedens spp.* genome. This complex is also encoded by a fragment of the Steel Borg. For both the Lilac and Steel Borgs, the GC content of the region encoding this operon is elevated relative to the average Borg values. *Methanoperedens spp.* pass electrons from methane oxidation to terminal electron acceptors (Fe^{3+}, NO_3^- or Mn^{4+}) via MHC(Cai et al., 2018; McGlynn et al., 2015; Scheller et al., 2016). The Lilac Borg genome encodes 16 MHCs with up to 32 heme-binding motifs within one protein. By analogy with experiments showing that cyanophages with a photosystem gene increase host fitness, we suggest that MHC genes may increase the capacity of *Methanoperedens* spp. to oxidize methane(Chen et al., 2020; Lindell et al., 2005). However, this needs to be tested experimentally. Membrane-bound and extracellular MHC may diversify the range of *Methanoperedens spp.* extracellular electron acceptors.

The Lilac Borg encodes a functional NiFe CODH, but this is fragmented in some genomes. Other genes for the acetyl-CoA decarbonylase/synthase complex are present only in *Methanoperedens spp.* The CODH is located in proximity to a cytochrome *b* and cytochrome *c,* so electrons from CO oxidation could be passed to an extracellular terminal acceptor such as Fe^{3+} in an energetically downhill reaction. This would allow the

removal of toxic CO and may contribute to the formation of a proton gradient that can be harnessed for energy conservation.

The Lilac Borg has a gene resembling the gamma subunit of ethylbenzene dehydrogenase (EBDH), which is involved in transferring electrons liberated from the hydroxylation of ethylbenzene and propylbenzene(Heider et al., 2016). This EBDH-like protein is located extracellularly, and given heme binding and cohesin domains, it may be involved in electron transfer and attachment.

Although the Lilac Borg lacks genes for a nitrate reductase, it encodes a probable hydroxylamine reductase (Hcp) that may scavenge toxic NO and hydroxylamine byproducts of *Methanoperedens spp.* nitrate metabolism. As the *hcp* gene was not identified in coexisting *Methanoperedens spp*, the Borg gene may protect *Methanoperedens spp.* from nitrosative stress. Proteins such as H_2O_2-forming NADH oxidase (Nox) and superoxide dismutase (SOD) may protect against reactive oxygen species. An alkylhydroperoxidase, two probable disulfide reductases, and a bacterioferritin all may detoxify the H_2O_2 byproduct of Nox and SOD. The Lilac Borg also encodes genes that likely augment osmotic stress tolerance. This Borg, but not *Methanoperedens spp.*, provides genes to make N^{ε}-acetyl-β-lysine as an osmolyte. An aspartate aminotransferase links the tricarboxylic acid cycle and amino acid synthesis, producing glutamate that can be used for the production of the osmolyte β-glutamate. More importantly, perhaps, it has recently been established that a bacterial homolog of this single enzyme can produce methane from methylamine(Wang et al., 2021), raising the possibility of methane cycling within the Borg - *Methanoperedens spp.* system.

The Lilac Borg has three large clusters of genes. The first may be involved in cell wall modification, as it encodes large membrane-integral proteins with up to 17 transmembrane domains, proteins for polysaccharide synthesis, glycosyltransferases, and likely carbohydrate-active proteins. The second contains key metabolic valves that connect gluconeogenesis with mannose metabolism for the production of glycans. One gene, fructose 1,6-bisphosphatase (FBP), was not identified in the *Methanoperedens spp.* genomes and may regulate carbon flow from gluconeogenesis to mannose metabolism. In between these clusters are 12 genes with PEGA domains with similarity to S-layer proteins. Cell surface proteins, along with these PEGA proteins, account for ~13% of all Lilac Borg genes. We conclude that functionalities related to cell wall architecture and modification are key to the impact of these extrachromosomal elements on their host, perhaps triggering cell wall modification for adaptation to changing environmental conditions (**Fig. 3**).

3.4 **Conclusions**

Borgs are enigmatic extrachromosomal elements that can approach (and likely exceed) 1 Mbp in length (**Table 1**). We can neither prove that they are archaeal viruses or plasmids or mini-chromosomes, nor can we prove that they are not. Although they may ultimately be classified as megaplasmids, they are clearly different from anything that has been reported previously. It is fascinating to ponder their possible evolutionary origins. Borg homologous recombination may indicate movement among hosts, thus their possible roles as gene transfer agents. It has been noted that *Methanoperedens spp.* have been particularly open to gene acquisition from diverse bacteria and archaea(Leu et al., 2020), and Borgs may have contributed to this. The existence of Borgs encoding MCR demonstrates for the first time that MCR and MCR-like proteins for metabolism of

methane and short-chain hydrocarbons can exist on extrachromosomal elements and thus could potentially be dispersed across lineages, as is inferred to have occurred several times over the course of archaeal evolution(Boyd et al., 2019; Hua et al., 2019). Borgs carry numerous metabolic genes, some of which produce variants of *Methanoperedens spp.* proteins that could have distinct biophysical and biochemical properties. Assuming that these genes either augment *Methanoperedens spp.* energy metabolism or extend the conditions under which they can function, Borgs may have far-reaching biogeochemical consequences, with important and unanticipated climate implications.

3.5 **Figures**

Table 1: Manually curated complete and draft genomes for the best sampled Borgs. Length is the genome length. Longest is the size of the largest genome fragment. Status indicates degree of genome completeness: complete genomes have been corrected and fully verified throughout. GC is the genome-wide average GC content. For details for these and less abundant examples, see **Table S1**.

Borg	Site	Depth	Length	Longest	Status	GC
Lilac	Rifle Sediment	600 cm	662 kbp	662 kbp	Complete	32%
Steel	Rifle Aquifer	500 cm	620 kbp	39 kbp	Draft	33%
Orange	Vernal Pool	100 cm	971 kbp	606 kbp	Draft	32%
Green	Vernal Pool	100 cm	1.08 Mbp	244 kbp	Draft	34%
Brown	Vernal Pool	100 cm	913 kbp	400 kbp	Draft	32%
Ochre	Vernal Pool	100 cm	702 kbp	169 kbp	Draft	33%
Sky	Vernal Pool	80 cm	763 kbp	763 kbp	Complete	33%
Purple	Vernal Pool	60 cm	918 kbp	918 kbp	Complete	32%
Pink	Vernal Pool	60 cm	1.14 Mbp	546 kbp	Draft	32%
Aqua	Vernal Pool	60 cm	839 kbp	77 kbp	Draft	34%
Black	Vernal Pool	40 cm	902 kbp	902 kbp	Complete	32%
Blue	Vernal Pool	5 cm	806 kbp	188 kbp	Draft	34%
Rose	Vernal Pool	1 cm	619 kbp	454 kbp	Draft	33%
Apricot	Vernal Pool	1 cm	811 kbp	134 kbp	Draft	34%
Olive	Vernal Pool	1 cm	974 kbp	188 kbp	Draft	34%
White	Riverbed	50 cm	157 kbp	23 kbp	Partial	33%
Red	Corona Mine	Surface water	223 kbp	27 kbp	Partial	32%

Figure 1: Borgs share overall genomic features. (A) Genome replichores (arrows) and coding strands (black bars) for aligned pairs of the four complete (Black, Purple, Sky and Lilac) and one near-complete (Orange) Borg. Blocks of sequence with identifiable similarity are shown in between each pair (colored graphs linked by lines, y-axes show similarity). (B) Genome overviews showing the distribution of three or more perfect tandem direct repeats (gold rods) along the complete genomes. Insets provide examples of local elevated GC content associated with certain gene clusters and within gene and intergenic tandem direct repeats (gold arrows).

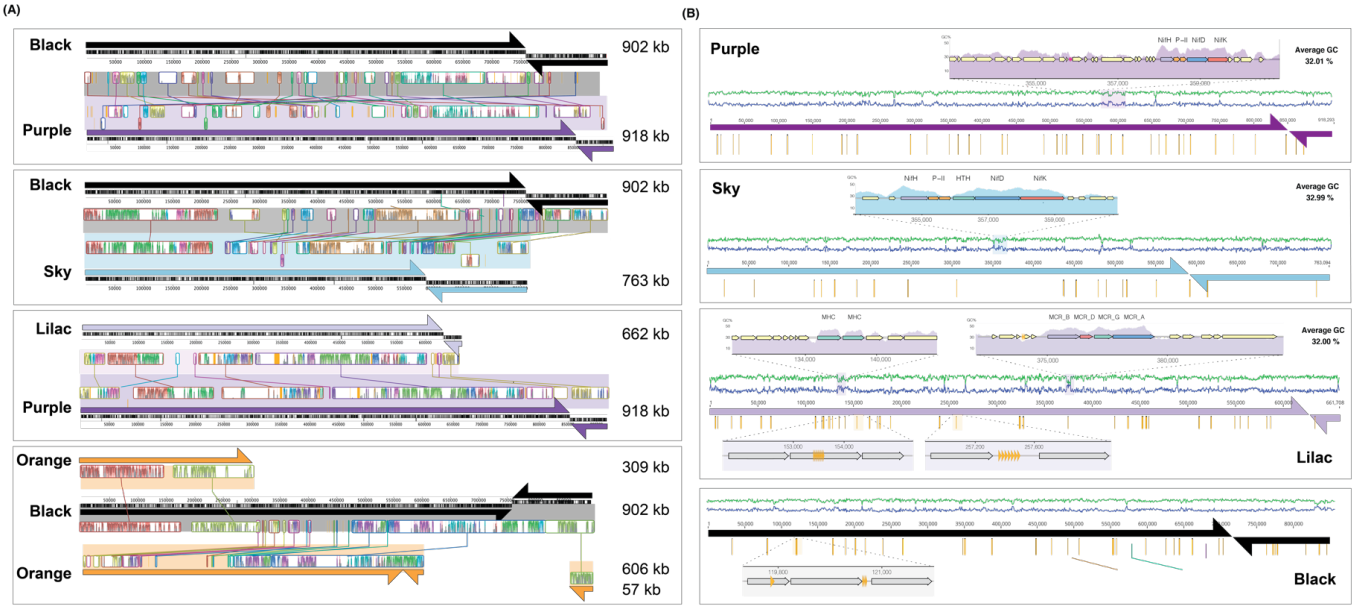


Figure 2: **Borg and *Methanoperedens spp.* genomic features and abundance patterns**. **A.** The average genome GC contents of Borgs and *Methanoperedens spp.* are distinct. **B.** Groups of related *Methanoperedens spp.* (rows) correlate with groups of Borgs (columns) across a set of 50 samples. Asterisks indicate Pearson correlations above 0.92 with FDR-corrected p-values below 2.0E-20 that suggest that Brown, Green, Orange, Beige and Ochre Borgs associate with one group of *Methanoperedens* spp., Olive, Cyan, Gold, Apricot and Rose with a second group, and Black with a third group. **C.** Frequency of genes in different functional groups in the four complete Borg genomes. **D.** Comparison of the protein family composition of Borgs and *Methanoperedens spp.* Clustering based on shared protein family content highlights groups of Borg-specific protein families (blue shading) and protein families shared with their hosts (orange shading). The full clustering, including diverse archaeal mobile elements, is shown in **Figure S6.**

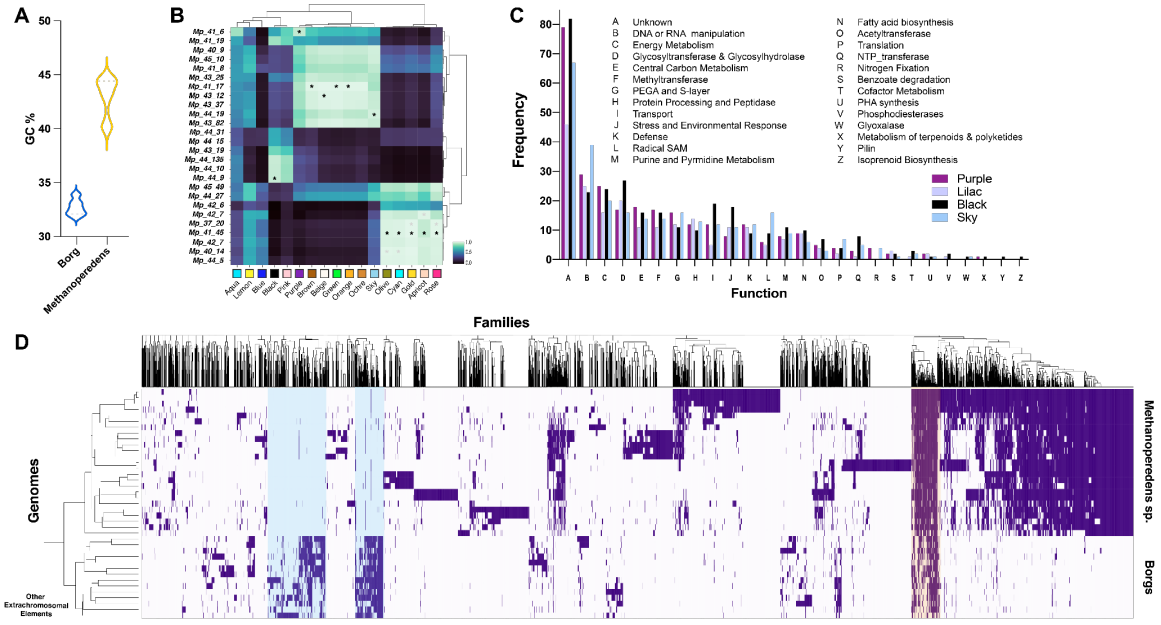
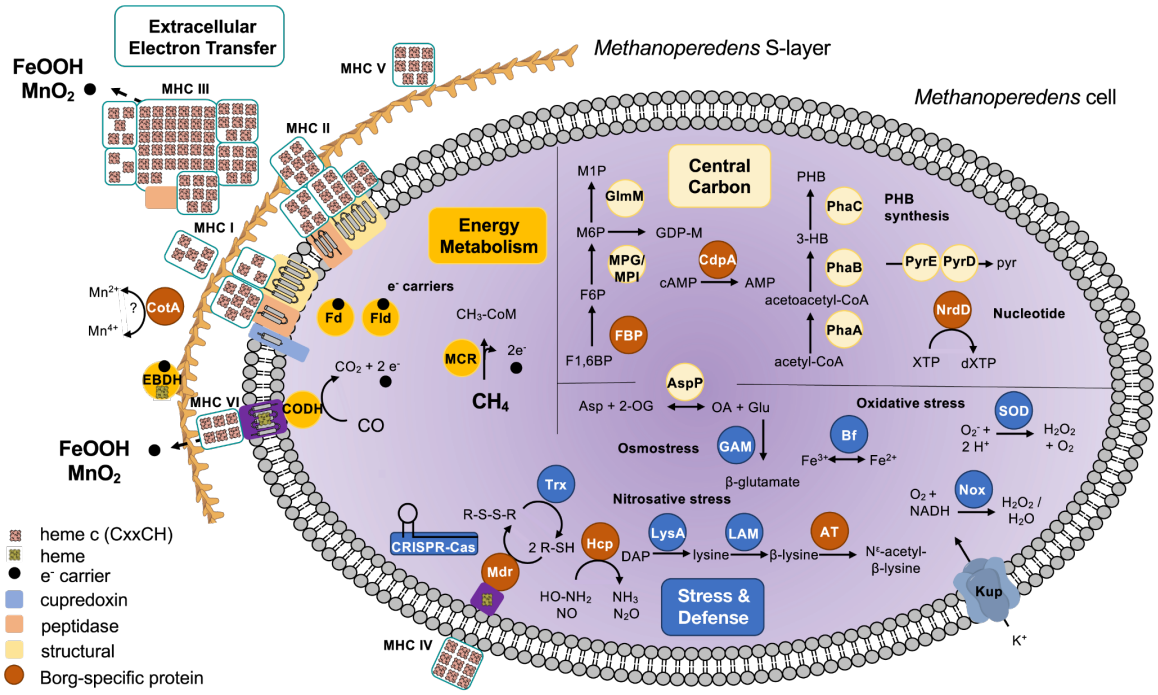


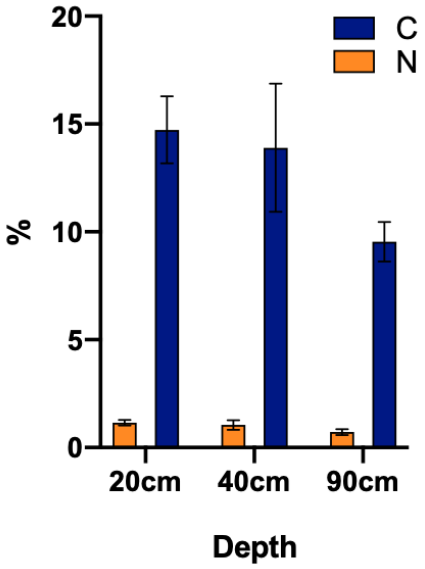
Figure 3: Cell cartoon illustrating capacities inferred to be provided to *Methanoperedens spp.* by the coexisting Lilac Borg. Like all Borgs, this Borg lacks the capacity for independent existence, and we infer that it replicates within host *Methanoperedens spp.* cells. Borg-specific proteins (red circles) are those that were not identified in the genome of coexisting *Methanoperedens spp.* Borg-encoded capacities are grouped into the major categories of energy metabolism (including the MCR complex involved in methane oxidation), extracellular electron transfer (including multiheme cytochromes, MHC) involved in electron transport to external electron acceptors), central carbon metabolism (including genes that enable production of polyhydroxybutyrate, PHB), and stress response/defense (including production of compatible solutes). Locus codes are listed in **Table S7**.



EXTENDED DATA

Figure S1. Geochemical profiles of the permanently moist and organic-rich wetland soils. (A) The concentrations of total carbon, nitrogen as well as (B) iron and manganese in wetland soils. Deeper soils, where these extrachromosomal elements are most abundant, are somewhat depleted in carbon, iron and manganese compared to shallow soils. Error bars denote SD.

(A)



(B)

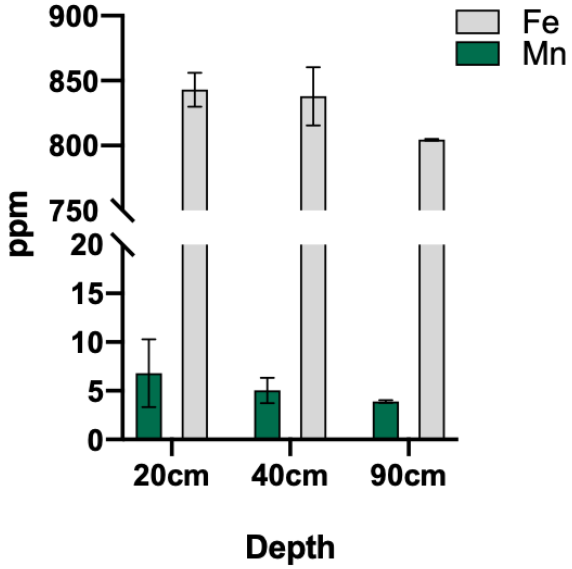


Figure S2. Sets of three or more perfect tandem direct repeats (TDR) are a characteristic feature of the Borg genomes. Up to 54 instances occur in the four complete Borg genomes, with, on average, one repeat every 12 (Lilac) - 31 (Sky) kbp. These repeat regions fragment assemblies and cause local assembly errors, which we resolved by manual curation (Methods). Within the TDR regions of the four curated, complete genomes, the unit repeats occur up to 20 times and unit repeats are up to 54 bps in length **(Table S2)**. Between54 and 64% of these perfect TDRs are encoded in intergenic regions, although part or all of the first repeat may occur within the C- terminus of a protein-coding gene. When the TDRs occur within proteins, the unit lengths are almost always divisible by 3, so they introduce perfect amino acid repeats. TDR sequences within a single Borg genome are almost always unique. Repeat sequence comparison from the four complete curated Borgs highlights the novelty of almost all TDR sequences (both within and across genomes).

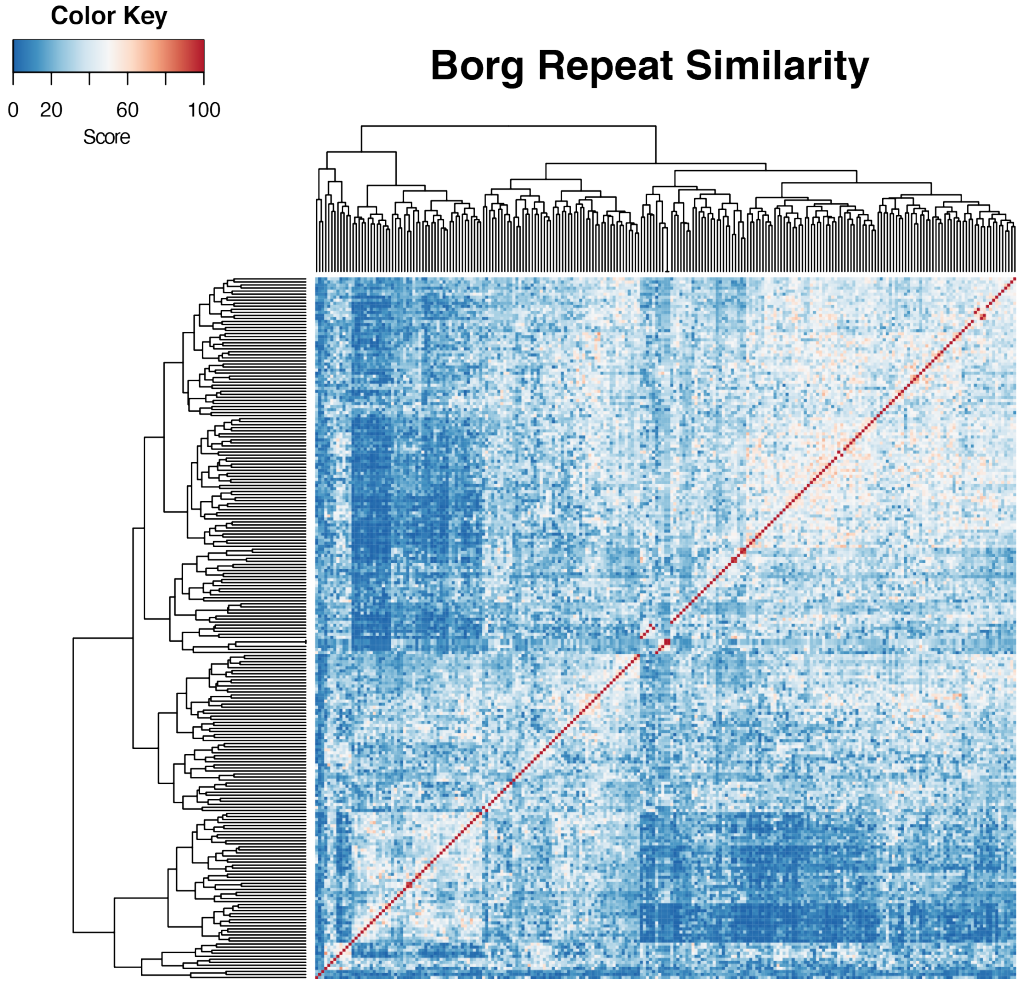


Figure S3. All genomes have two replichores of unequal lengths. GC skew (grey plots) and cumulative GC skew (green lines) across the four complete Borg genomes, all of which end in long inverted terminal repeats (1.4 - 2.7 kbp in length). The cumulative GC skew plots indicate replication is initiated in these terminal repeats (red lines). Blue lines mark the predicted replication termini. The red and blue lines define two replichores of unequal length that correspond almost completely to distinct coding strands (almost all genes on the +ve strand of the large replichore and on the -ve strand of the small replichore).

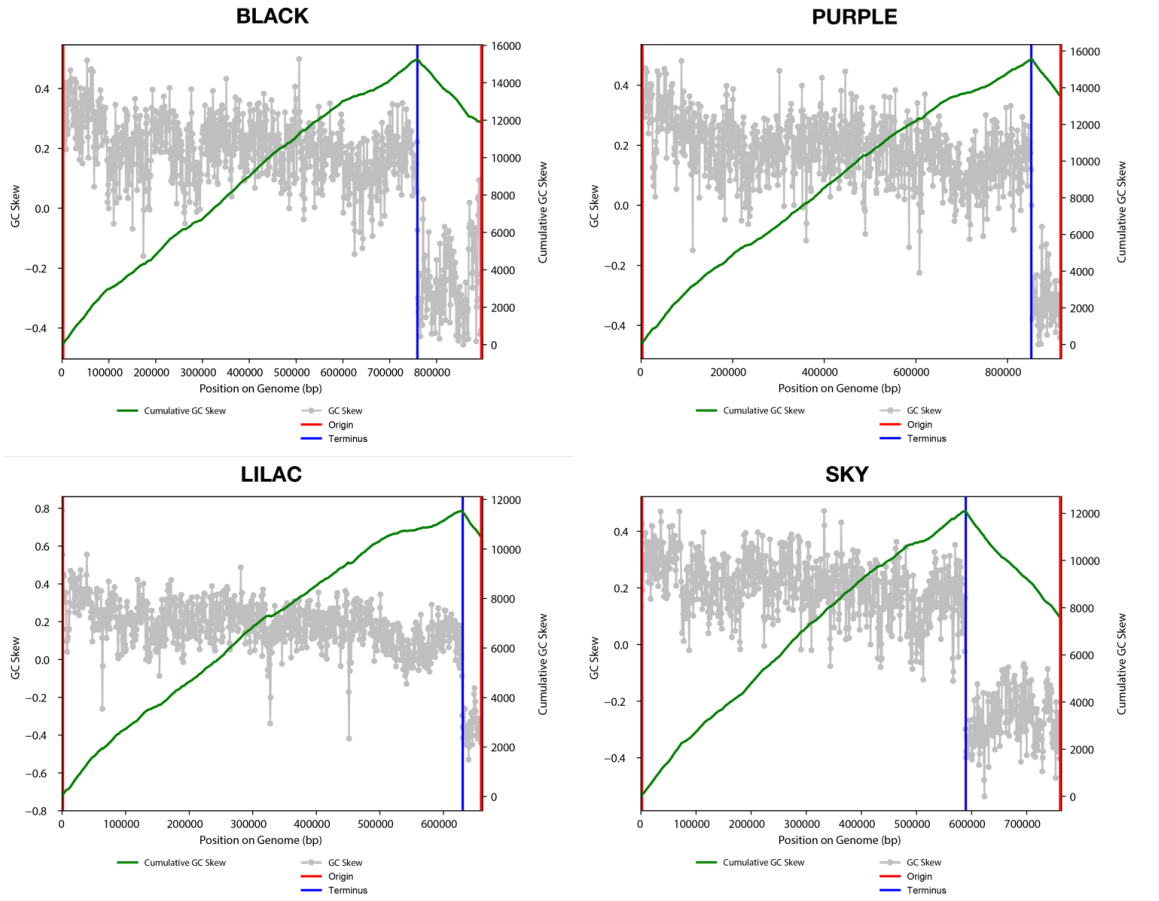


Figure S4. Taxonomic profiles of the four complete Borg genomes. **A.** In all cases, the majority of proteins have no similarity to proteins in the reference database ("Unknown"; e-value of >0.0001). For the cases where a protein has an identifiable hit (blue and red bars in A), the plots in **B.** show the taxonomy of the organisms in which those hits were identified. Only cases where the same organism accounted for hits for >0.5% of genes are shown. The results clearly indicate that the vast majority of cases where proteins have identifiable matches involve matches to proteins of *Methanoperedens* (gold bars)

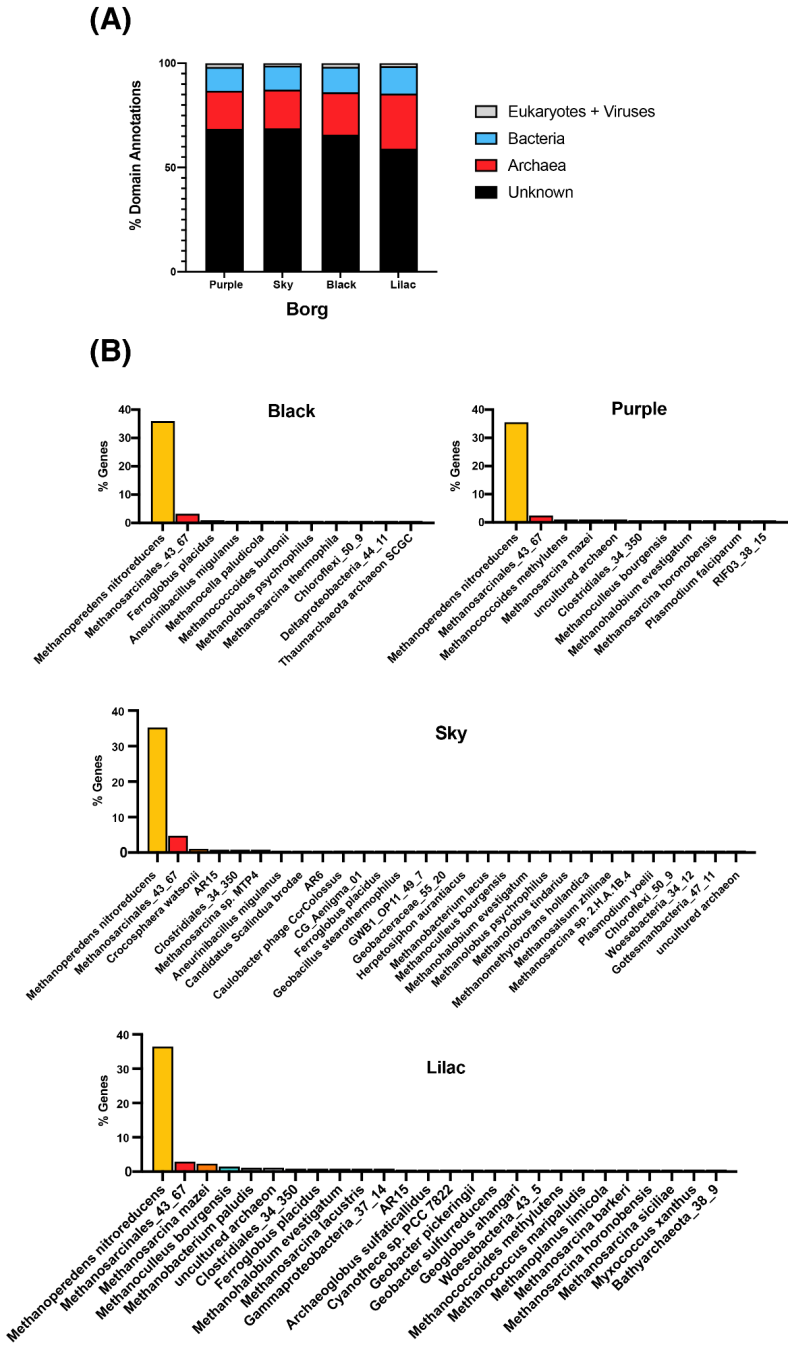


Figure S5: The clustering based on protein family content demonstrates that the *Methanoperedens*, Borgs, archaeal viruses and plasmids/minichromosomes are distinct from each other. (A) Colored blocks indicate presence of each protein family in the corresponding genome. The blue highlight at the top indicates the *Methanoperedens* (top) and Borg (bottom) protein family profiles. For details see **Fig. 2D**. We note that archaeal plasmids are highly undersampled. If Borgs are ultimately classified as plasmids, they dramatically expand the known characteristics (e.g., size, linear genomes) and diversity of archaeal plasmids. **(B)** Borg protein inventories (purple highlight) compared to giant linear bacterial plasmids. **(C)** Protein families occurring in more than 5 genomes of Borgs and giant linear bacterial plasmids. Few protein families are shared between Borgs and linear plasmids in bacteria beyond methyltransferases, histidine kinases, and other enzymes unrelated to replication. **(D)** Average Nucleotide Identity of different *Methanoperedens* species that coexist with Borgs (red) and previously reported genomes (gray) and the 95% species threshold shown with a dashed line.


A
Genomes
Protein Families
Archaeal Viruses
Archaeal Extra-chromosomal Elements
B
Protein Families
Linear Bacterial Plasmids
Borgs
C
Protein Families in >5 genomes
Borgs
D
Comp method: ANImf Clust method: average Min cov: 0.5
Candidatus_Methanoperedens_spp_UBA453
Candidatus_Methanoperedens_spp_ANME-2d
Candidatus_Methanoperedens_spp_Minv1
Candidatus_Methanoperedens_spp_DP16D
Candidatus_Methanoperedens_spp_MnB21
Candidatus_Methanoperedens_spp_SB2
Candidatus_Methanoperedens_spp_Ru_MN
Candidatus_Methanoperedens_spp_BLZ2
Candidatus_Methanoperedens_spp_BLZ1
Candidatus_Methanoperedens_spp_MnB20
Candidatus_Methanoperedens_spp_Rif_45_12
Candidatus_Methanoperedens_spp_Rif_45_8
Candidatus_Methanoperedens_spp_Rif_44_10_1
Candidatus_Methanoperedens_spp_Rif_44_10_2
Candidatus_Methanoperedens_spp_Rif_44_10_3
Candidatus_Methanoperedens_spp_Rif_45_7
Candidatus_Methanoperedens_spp_VP_41_45
Candidatus_Methanoperedens_spp_contig_VP_37_20
Candidatus_Methanoperedens_spp_VP_43_19
Candidatus_Methanoperedens_spp_VP_44_135
Candidatus_Methanoperedens_spp_VP_44_10
Candidatus_Methanoperedens_spp_VP_44_9
Candidatus_Methanoperedens_spp_VP_41_17
Candidatus_Methanoperedens_spp_VP_44_31
Candidatus_Methanoperedens_spp_VP_44_15
Candidatus_Methanoperedens_spp_VP_41_6
Candidatus_Methanoperedens_spp_VP_44_19
Candidatus_Methanoperedens_spp_contig_VP_40_9
Candidatus_Methanoperedens_spp_VP_43_82
Candidatus_Methanoperedens_spp_VP_43_12
Candidatus_Methanoperedens_spp_VP_43_37
Candidatus_Methanoperedens_spp_VP_45_49
Candidatus_Methanoperedens_spp_VP_44_27
100 90 80 70 60 50
Average Nucleotide Identity (ANI)

Figure S6: (A) The array of single-copy archaeal ribosomal genes (columns) vs. Borg (blue) and *Methanoperedens* (gold) genomes illustrating that although Borgs often have rpL11 and occasionally, other ribosomal proteins, they do not have the gene inventory needed to construct ribosomes. **(B) Left;** Dendrogram of hierarchical clustering of all-vs-all pearson correlation values between all Borgs and Methanoperedens from the wetland**. Right;** Maximum Likelihood Phylogeny of concatenated ribosomal proteins from *Methanoperedens* species that do and do not coexist with Borgs and previously reported genomes. We found no data indicating the presence of Borgs in samples containing previously reported *Methanoperedens* genomes. We searched for Borgs in the samples highlighted in blue using the same methods used to detect Borgs in this study and concluded that they do not contain Borgs. A subset of the Borg-free samples contain *Methanoperedens* at very high abundance levels.

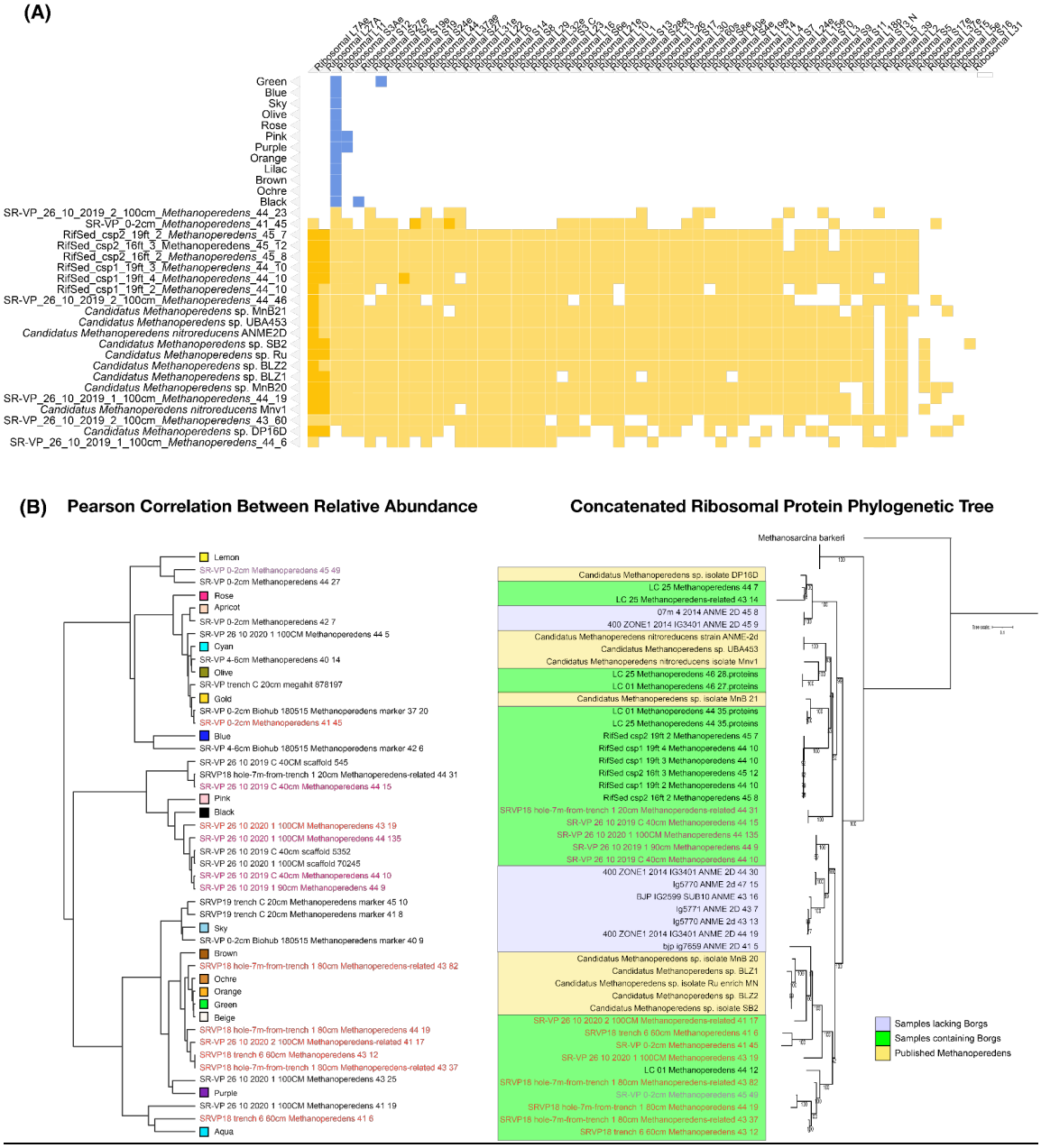

(A)
Green
Blue
Sky
Olive
Rose
Pink
Purple
Orange
Lilac
Brown
Ochre
Black
SR-VP_26_10_2019_2_100cm_Methanoperedens_44_23
SR-VP_0-2cm_Methanoperedens_41_45
RifSed_csp2_19ft_2_Methanoperedens_45_7
RifSed_csp2_16ft_3_Methanoperedens_45_12
RifSed_csp2_16ft_2_Methanoperedens_45_8
RifSed_csp1_19ft_3_Methanoperedens_44_10
RifSed_csp1_19ft_4_Methanoperedens_44_10
RifSed_csp1_19ft_2_Methanoperedens_44_10
SR-VP_26_10_2019_2_100cm_Methanoperedens_44_46
Candidatus Methanoperedens sp. MnB21
Candidatus Methanoperedens sp. UBA453
Candidatus Methanoperedens nitroreducens ANME2D
Candidatus Methanoperedens sp. SB2
Candidatus Methanoperedens sp. Ru
Candidatus Methanoperedens sp. BLZ2
Candidatus Methanoperedens sp. BLZ1
Candidatus Methanoperedens sp. MnB20
SR-VP_26_10_2019_1_100cm_Methanoperedens_44_19
Candidatus Methanoperedens nitroreducens Mnv1
SR-VP_26_10_2019_2_100cm_Methanoperedens_43_60
Candidatus Methanoperedens sp. DP16D
SR-VP_26_10_2019_1_100cm_Methanoperedens_44_6
(B) Pearson Correlation Between Relative Abundance
Lemon
SR-VP 0-2cm Methanoperedens 45 49
SR-VP 0-2cm Methanoperedens 44 27
Rose
Apricot
SR-VP 0-2cm Methanoperedens 42 7
SR-VP 26 10 2020 1 100CM Methanoperedens 44 5
Cyan
SR-VP 4-6cm Methanoperedens 40 14
Olive
SR-VP trench C 20cm megahit 878197
Gold
SR-VP 0-2cm Biohub 180515 Methanoperedens marker 37 20
SR-VP 0-2cm Methanoperedens 41 45
Blue
SR-VP 4-6cm Biohub 180515 Methanoperedens marker 42 6
SR-VP 26 10 2019 C 40CM scaffold 545
SRVP18 hole-7m-from-trench 1 20cm Methanoperedens-related 44 31
SR-VP 26 10 2019 C 40cm Methanoperedens 44 15
Pink
Black
SR-VP 26 10 2020 1 100CM Methanoperedens 43 19
SR-VP 26 10 2020 1 100CM Methanoperedens 44 135
SR-VP 26 10 2019 C 40cm scaffold 5352
SR-VP 26 10 2020 1 100CM scaffold 70245
SR-VP 26 10 2019 C 40cm Methanoperedens 44 10
SR-VP 26 10 2019 1 90cm Methanoperedens 44 9
SRVP19 trench C 20cm Methanoperedens marker 45 10
SRVP19 trench C 20cm Methanoperedens marker 41 8
Sky
SR-VP 0-2cm Biohub 180515 Methanoperedens marker 40 9
Brown
SRVP18 hole-7m-from-trench 1 80cm Methanoperedens-related 43 82
Ochre
Orange
Green
Beige
SRVP18 hole-7m-from-trench 1 80cm Methanoperedens 44 19
SR-VP 26 10 2020 2 100CM Methanoperedens-related 41 17
SRVP18 trench 6 60cm Methanoperedens 43 12
SRVP18 hole-7m-from-trench 1 80cm Methanoperedens-related 43 37
SR-VP 26 10 2020 1 100CM Methanoperedens 43 25
Purple
SR-VP 26 10 2020 1 100CM Methanoperedens 41 19
SRVP18 trench 6 60cm Methanoperedens 41 6
Aqua
Concatenated Ribosomal Protein Phylogenetic Tree
Methanosarcina barkeri
Candidatus Methanoperedens sp. isolate DP16D
LC 25 Methanoperedens 44 7
LC 25 Methanoperedens-related 43 14
07m 4 2014 ANME 2D 45 8
400 ZONE1 2014 IG3401 ANME 2D 45 9
Candidatus Methanoperedens nitroreducens strain ANME-2d
Candidatus Methanoperedens sp. UBA453
Candidatus Methanoperedens nitroreducens isolate Mnv1
LC 25 Methanoperedens 46 28.proteins
LC 01 Methanoperedens 46 27.proteins
Candidatus Methanoperedens sp. isolate MnB 21
LC 01 Methanoperedens 44 35.proteins
LC 25 Methanoperedens 44 35.proteins
RifSed csp2 19ft 2 Methanoperedens 45 7
RifSed csp1 19ft 4 Methanoperedens 44 10
RifSed csp1 19ft 3 Methanoperedens 44 10
RifSed csp2 16ft 3 Methanoperedens 45 12
RifSed csp1 19ft 2 Methanoperedens 44 10
RifSed csp2 16ft 2 Methanoperedens 45 8
SRVP18 hole-7m-from-trench 1 20cm Methanoperedens-related 44 31
SR-VP 26 10 2019 C 40cm Methanoperedens 44 15
SR-VP 26 10 2020 1 100CM Methanoperedens 44 135
SR-VP 26 10 2019 1 90cm Methanoperedens 44 9
SR-VP 26 10 2019 C 40cm Methanoperedens 44 10
400 ZONE1 2014 IG3401 ANME 2D 44 30
Ig5770 ANME 2d 47 15
BJP IG2599 SUB10 ANME 43 16
Ig5771 ANME 2D 43 7
Ig5770 ANME 2d 43 13
400 ZONE1 2014 IG3401 ANME 2D 44 19
bjp ig7659 ANME 2D 41 5
Candidatus Methanoperedens sp. isolate MnB 20
Candidatus Methanoperedens sp. BLZ1
Candidatus Methanoperedens sp. isolate Ru enrich MN
Candidatus Methanoperedens sp. BLZ2
Candidatus Methanoperedens sp. isolate SB2
SR-VP 26 10 2020 2 100CM Methanoperedens-related 41 17
SRVP18 trench 6 60cm Methanoperedens 41 6
SR-VP 0-2cm Methanoperedens 41 45
SR-VP 26 10 2020 1 100CM Methanoperedens 43 19
LC 01 Methanoperedens 44 12
SRVP18 hole-7m-from-trench 1 80cm Methanoperedens-related 43 82
SR-VP 0-2cm Methanoperedens 45 49
SRVP18 hole-7m-from-trench 1 80cm Methanoperedens 44 19
SRVP18 hole-7m-from-trench 1 80cm Methanoperedens-related 43 37
SRVP18 trench 6 60cm Methanoperedens 43 12
Samples lacking Borgs
Samples containing Borgs
Published Methanoperedens

Figure S7: Abundance and distribution of Borgs and *Methanoperedens spp* in the wetland soil and Rifle aquifer. A. Relative abundances of *Methanoperedens spp.* and Borgs in samples collected over time and arrayed by sample collection depth from the wetland soils, sediments and groundwater. The absolute abundances of Borgs are far greater in the deeper compared to shallower soils B. Although some Borgs can substantially exceed all the combined abundance of *Methanoperedens spp*, no Borgs were detected in some *Methanoperedens*-bearing samples. "W" indicates that the sample was pumped groundwater.

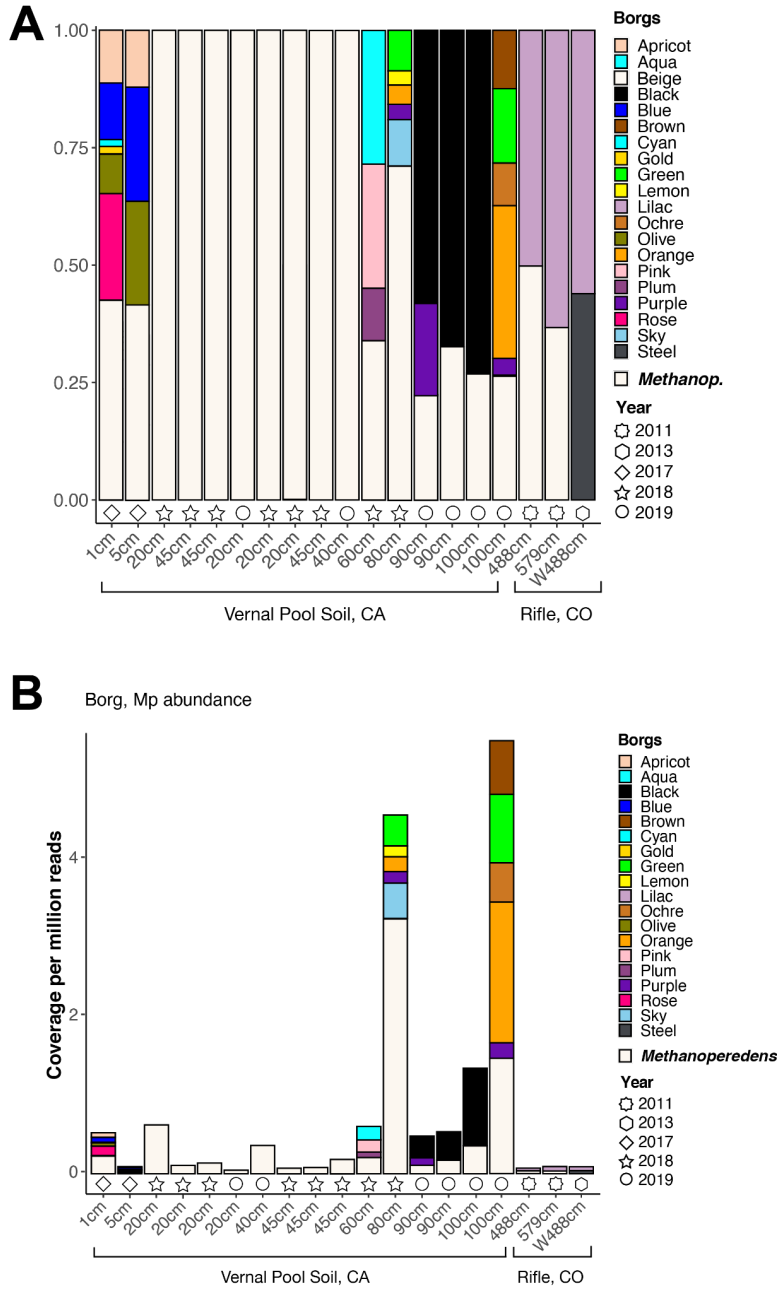


Figure S8: (A) Genome-to-genome comparisons provide evidence for recombination between two of the mostly closely related Borgs, Sky and Rose. These Borgs share only moderate overall genomic nucleic acid identity although, as is the case for other Borgs (**Figure 1A**), have blocks of partially alignable sequence throughout their genomes. Notable, and indicating recent homologous recombination, are 100% identical regions of up to ~11 kbp in length (B). Although not fully manually curated to completion, the relevant Rose Borg genome regions were carefully checked by inspection of the mapped reads to rule out chimeric assembly that could otherwise explain perfect identity with the Sky Borg sequence (Sky is one of the four curated complete genomes). (C) Read coverages over the Rose and Sky genomes are consistent throughout, with the regions in B noted with green boxes (D) Diagram illustrating the organization of the Type III-A CRISPR-Cas system variant (lacking acquisition machinery and Csm6) in the Orange Borg. One spacer from the CRISPR array targets a small protein with a ribbon-helix-helix motif, a common transcriptional regulator in archaeal mobile elements, in a mobile region of a *Methanoperedens* genome bin from the same wetland site.

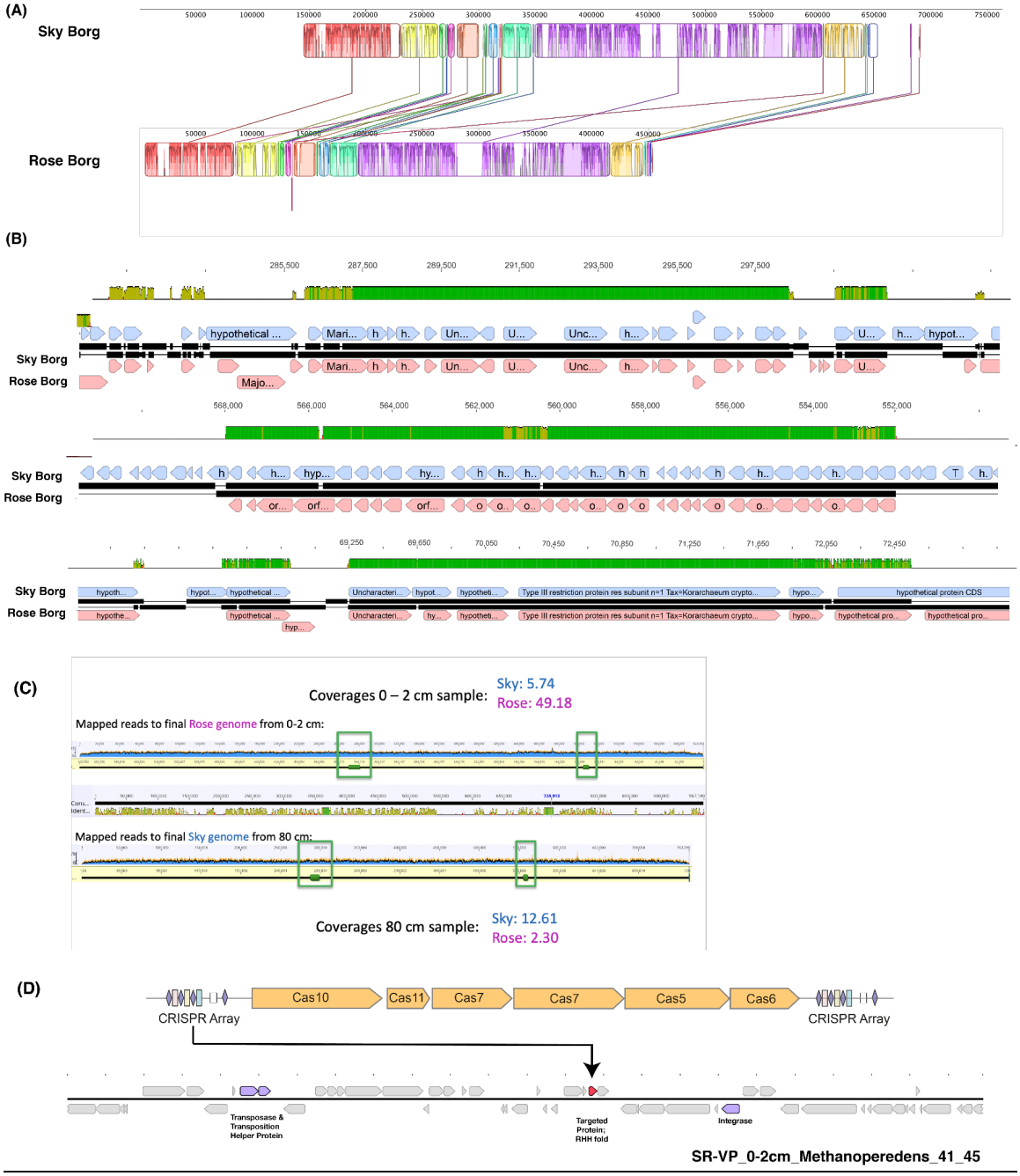

(A)
Sky Borg
Rose Borg
(B)
Sky Borg
Rose Borg
(C)
Coverages 0 – 2 cm sample:
Sky: 5.74
Rose: 49.18
Mapped reads to final Rose genome from 0-2 cm:
Mapped reads to final Sky genome from 80 cm:
Coverages 80 cm sample:
Sky: 12.61
Rose: 2.30
(D)
Cas10
Cas11
Cas7
Cas7
Cas5
Cas6
CRISPR Array
CRISPR Array
Transposase & Transposition Helper Protein
Targeted Protein; RHH fold
Integrase
SR-VP_0-2cm_Methanoperedens_41_45

Figure S9: The Borg ribosomal sequences form monophyletic groups that cluster adjacent to those from *Methanoperedens*. Phylogenetic tree constructed using the protein sequences for (A) ribosomal protein L11 (rpL11), (B) Ribosomal protein S2 (C) Ribosomal protein 3ae.

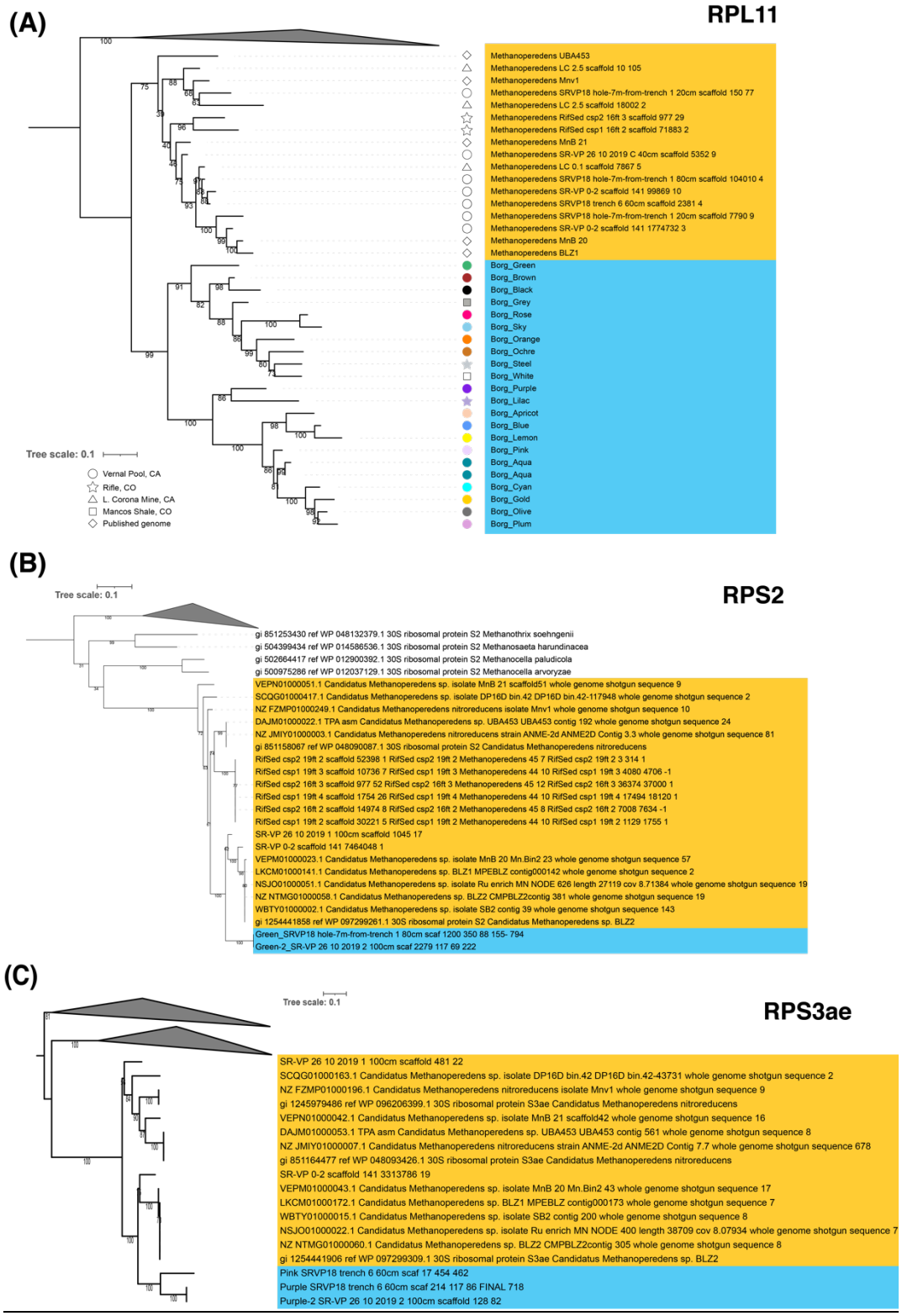
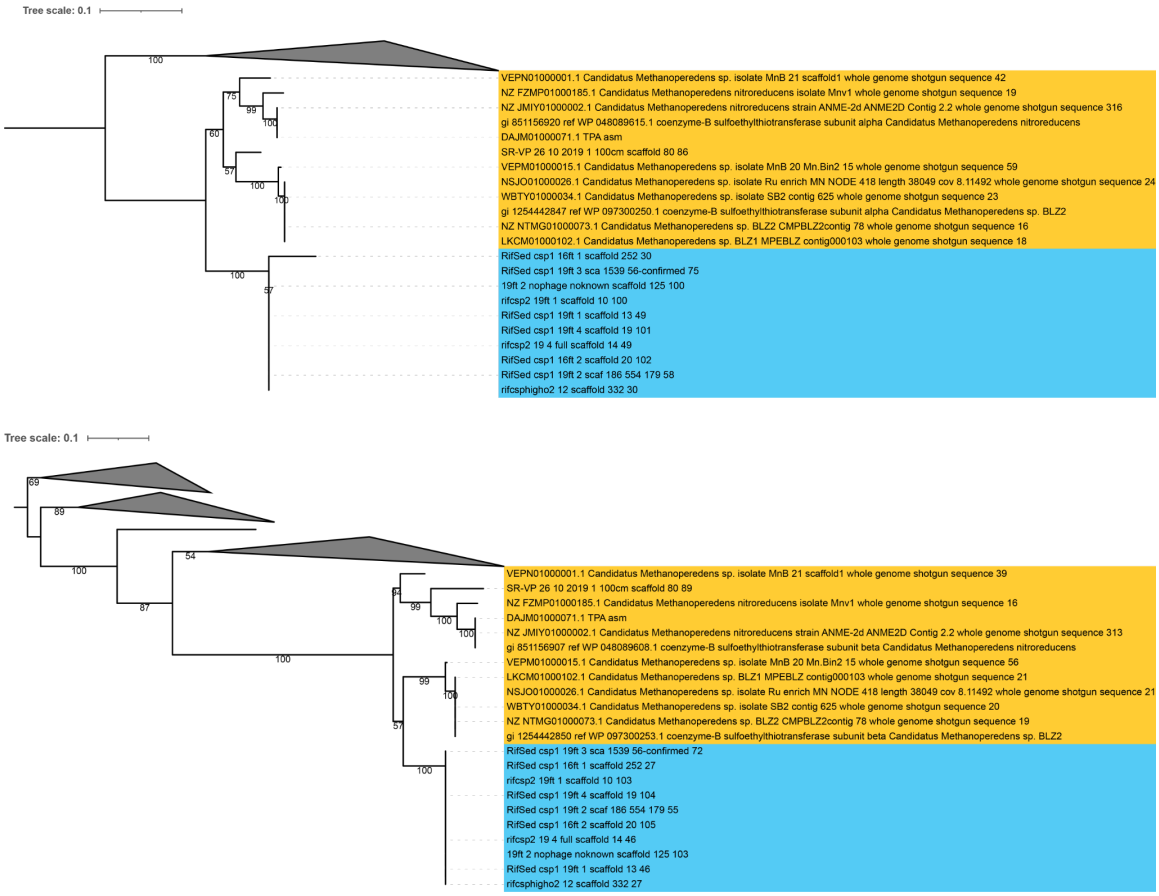
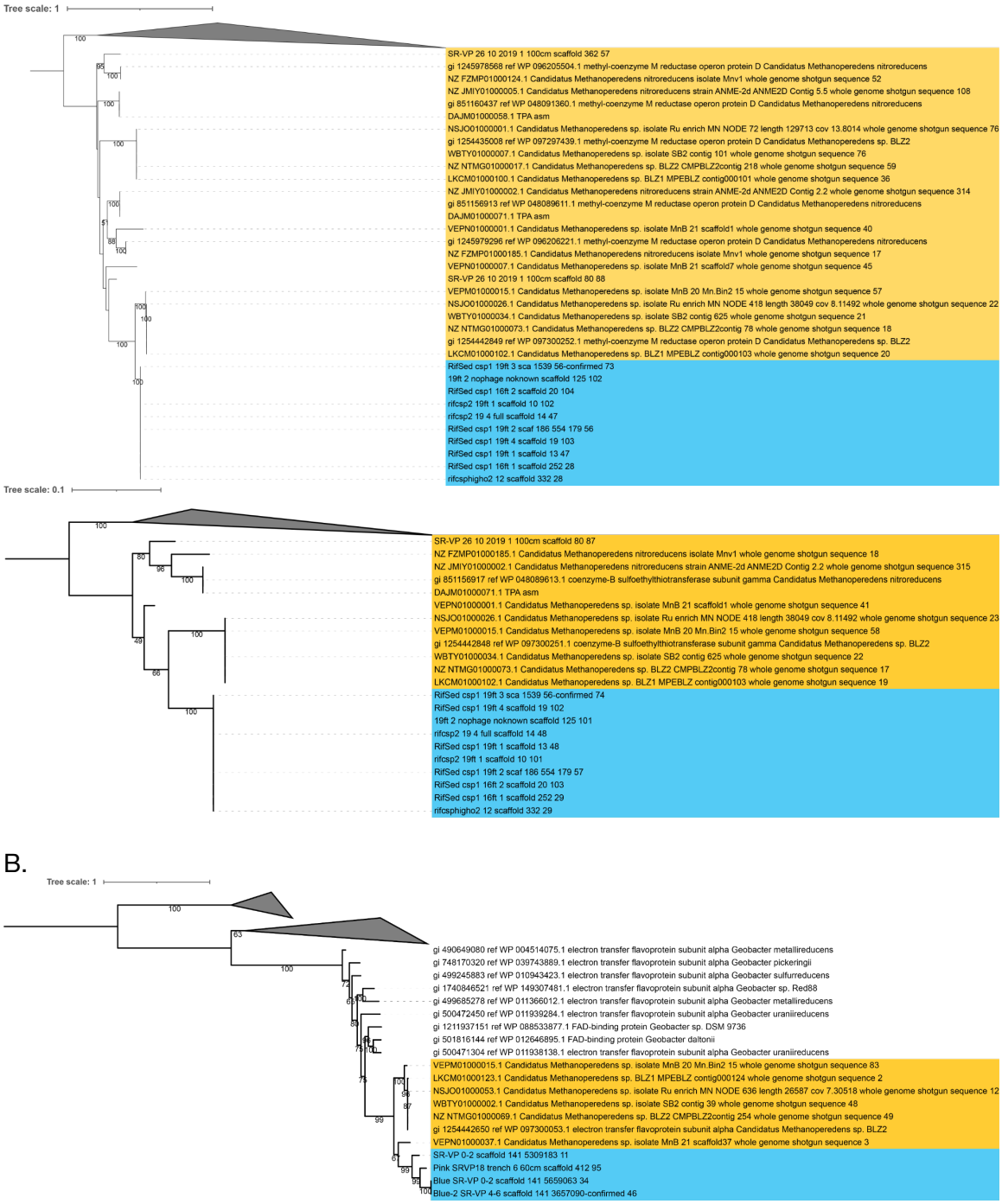


Figure S10: Phylogenetic trees for key Borg genes with functional predictions showing

that the protein sequences (blue) cluster sister to or within those from Methanoperedens (gold), with GC contents that approach those of Methanoperedens. **(**A) MCR_A, _B, _D, _G. (B) ETF Alpha, Beta. (C) Aconitase. (D) NifH and NifDK trees with alignments showing conservation of cysteine motifs that are involved in the attachment of the P-clusters.

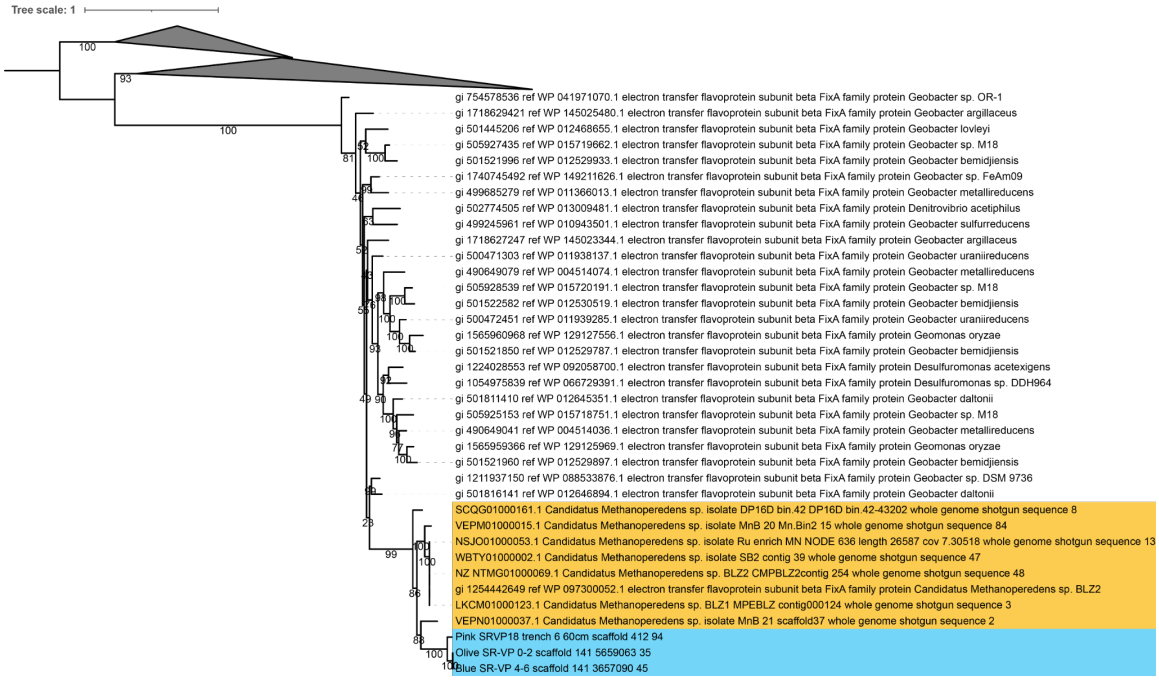
(A)



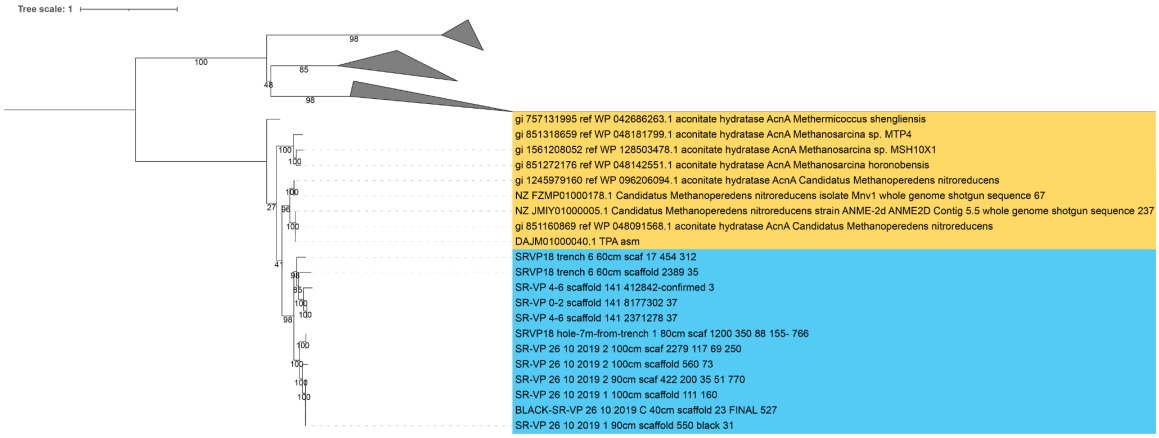

Tree scale: 1
SR-VP 26 10 2019 1 100cm scaffold 362 57
gi 1245978568 ref WP 096205504.1 methyl-coenzyme M reductase operon protein D Candidatus Methanoperedens nitroreducens
NZ FZMP01000124.1 Candidatus Methanoperedens nitroreducens isolate Mnv1 whole genome shotgun sequence 52
NZ JMIY01000005.1 Candidatus Methanoperedens nitroreducens strain ANME-2d ANME2D Contig 5.5 whole genome shotgun sequence 108
gi 851160437 ref WP 048091360.1 methyl-coenzyme M reductase operon protein D Candidatus Methanoperedens nitroreducens
DAJM01000058.1 TPA asm
NSJO01000001.1 Candidatus Methanoperedens sp. isolate Ru enrich MN NODE 72 length 129713 cov 13.8014 whole genome shotgun sequence 76
gi 1254435008 ref WP 097297439.1 methyl-coenzyme M reductase operon protein D Candidatus Methanoperedens sp. BLZ2
WBTY01000007.1 Candidatus Methanoperedens sp. isolate SB2 contig 101 whole genome shotgun sequence 76
NZ NTMG01000017.1 Candidatus Methanoperedens sp. BLZ2 CMPBLZ2contig 218 whole genome shotgun sequence 59
LKCM01000100.1 Candidatus Methanoperedens sp. BLZ1 MPEBLZ contig000101 whole genome shotgun sequence 36
NZ JMIY01000002.1 Candidatus Methanoperedens nitroreducens strain ANME-2d ANME2D Contig 2.2 whole genome shotgun sequence 314
gi 851156913 ref WP 048089611.1 methyl-coenzyme M reductase operon protein D Candidatus Methanoperedens nitroreducens
DAJM01000071.1 TPA asm
VEPN01000001.1 Candidatus Methanoperedens sp. isolate MnB 21 scaffold1 whole genome shotgun sequence 40
gi 1245979296 ref WP 096206221.1 methyl-coenzyme M reductase operon protein D Candidatus Methanoperedens nitroreducens
NZ FZMP01000185.1 Candidatus Methanoperedens nitroreducens isolate Mnv1 whole genome shotgun sequence 17
VEPN01000007.1 Candidatus Methanoperedens sp. isolate MnB 21 scaffold7 whole genome shotgun sequence 45
SR-VP 26 10 2019 1 100cm scaffold 80 88
VEPM01000015.1 Candidatus Methanoperedens sp. isolate MnB 20 Mn.Bin2 15 whole genome shotgun sequence 57
NSJO01000026.1 Candidatus Methanoperedens sp. isolate Ru enrich MN NODE 418 length 38049 cov 8.11492 whole genome shotgun sequence 22
WBTY01000034.1 Candidatus Methanoperedens sp. isolate SB2 contig 625 whole genome shotgun sequence 21
NZ NTMG01000073.1 Candidatus Methanoperedens sp. BLZ2 CMPBLZ2contig 78 whole genome shotgun sequence 18
gi 1254442849 ref WP 097300252.1 methyl-coenzyme M reductase operon protein D Candidatus Methanoperedens sp. BLZ2
LKCM01000102.1 Candidatus Methanoperedens sp. BLZ1 MPEBLZ contig000103 whole genome shotgun sequence 20
RifSed csp1 19ft 3 sca 1539 56-confirmed 73
19ft 2 nophage noknown scaffold 125 102
RifSed csp1 16ft 2 scaffold 20 104
rifcsp2 19ft 1 scaffold 10 102
rifcsp2 19 4 full scaffold 14 47
RifSed csp1 19ft 2 scaf 186 554 179 56
RifSed csp1 19ft 4 scaffold 19 103
RifSed csp1 19ft 1 scaffold 13 47
RifSed csp1 16ft 1 scaffold 252 28
rifcsphigho2 12 scaffold 332 28
Tree scale: 0.1
SR-VP 26 10 2019 1 100cm scaffold 80 87
NZ FZMP01000185.1 Candidatus Methanoperedens nitroreducens isolate Mnv1 whole genome shotgun sequence 18
NZ JMIY01000002.1 Candidatus Methanoperedens nitroreducens strain ANME-2d ANME2D Contig 2.2 whole genome shotgun sequence 315
gi 851156917 ref WP 048089613.1 coenzyme-B sulfoethylthiotransferase subunit gamma Candidatus Methanoperedens nitroreducens
DAJM01000071.1 TPA asm
VEPN01000001.1 Candidatus Methanoperedens sp. isolate MnB 21 scaffold1 whole genome shotgun sequence 41
NSJO01000026.1 Candidatus Methanoperedens sp. isolate Ru enrich MN NODE 418 length 38049 cov 8.11492 whole genome shotgun sequence 23
VEPM01000015.1 Candidatus Methanoperedens sp. isolate MnB 20 Mn.Bin2 15 whole genome shotgun sequence 58
gi 1254442848 ref WP 097300251.1 coenzyme-B sulfoethylthiotransferase subunit gamma Candidatus Methanoperedens sp. BLZ2
WBTY01000034.1 Candidatus Methanoperedens sp. isolate SB2 contig 625 whole genome shotgun sequence 22
NZ NTMG01000073.1 Candidatus Methanoperedens sp. BLZ2 CMPBLZ2contig 78 whole genome shotgun sequence 17
LKCM01000102.1 Candidatus Methanoperedens sp. BLZ1 MPEBLZ contig000103 whole genome shotgun sequence 19
RifSed csp1 19ft 3 sca 1539 56-confirmed 74
RifSed csp1 19ft 4 scaffold 19 102
19ft 2 nophage noknown scaffold 125 101
rifcsp2 19 4 full scaffold 14 48
RifSed csp1 19ft 1 scaffold 13 48
rifcsp2 19ft 1 scaffold 10 101
RifSed csp1 19ft 2 scaf 186 554 179 57
RifSed csp1 16ft 2 scaffold 20 103
RifSed csp1 16ft 1 scaffold 252 29
rifcsphigho2 12 scaffold 332 29

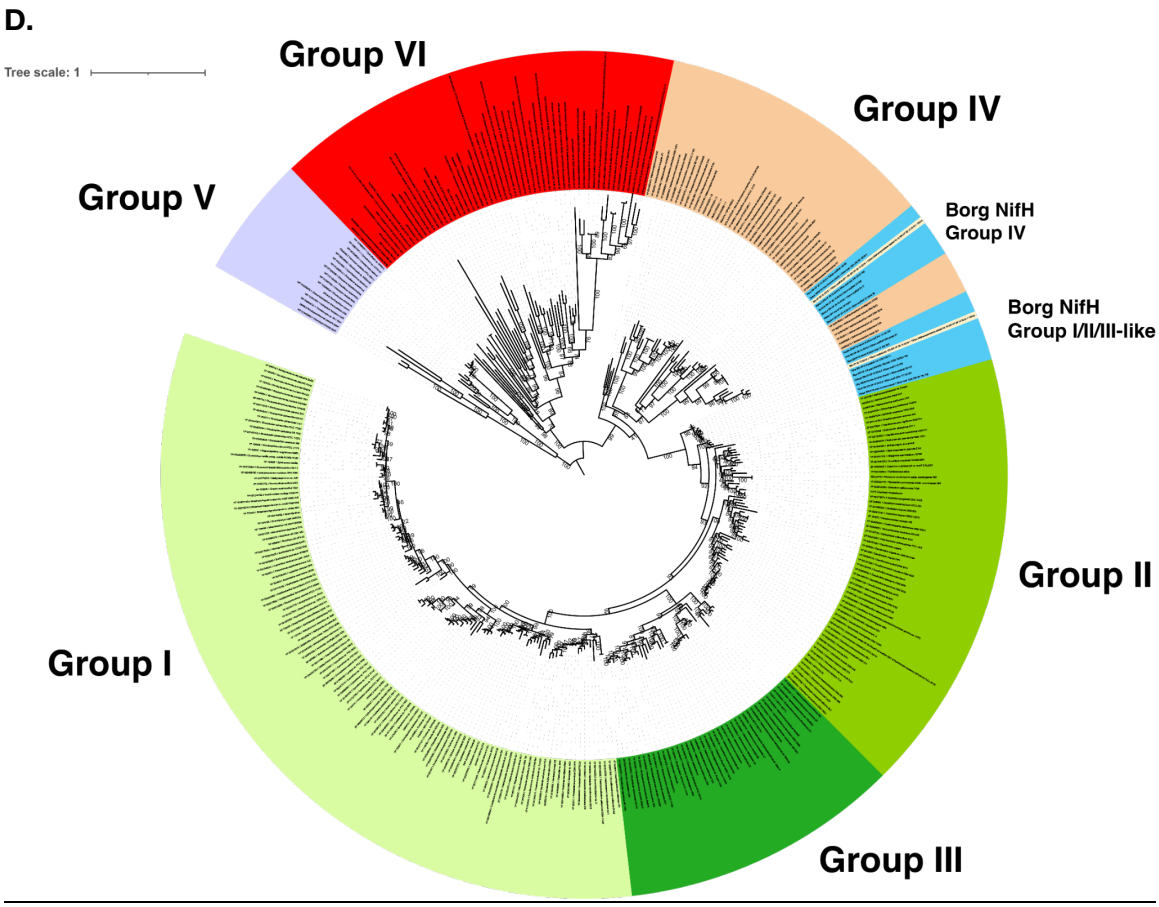
B.

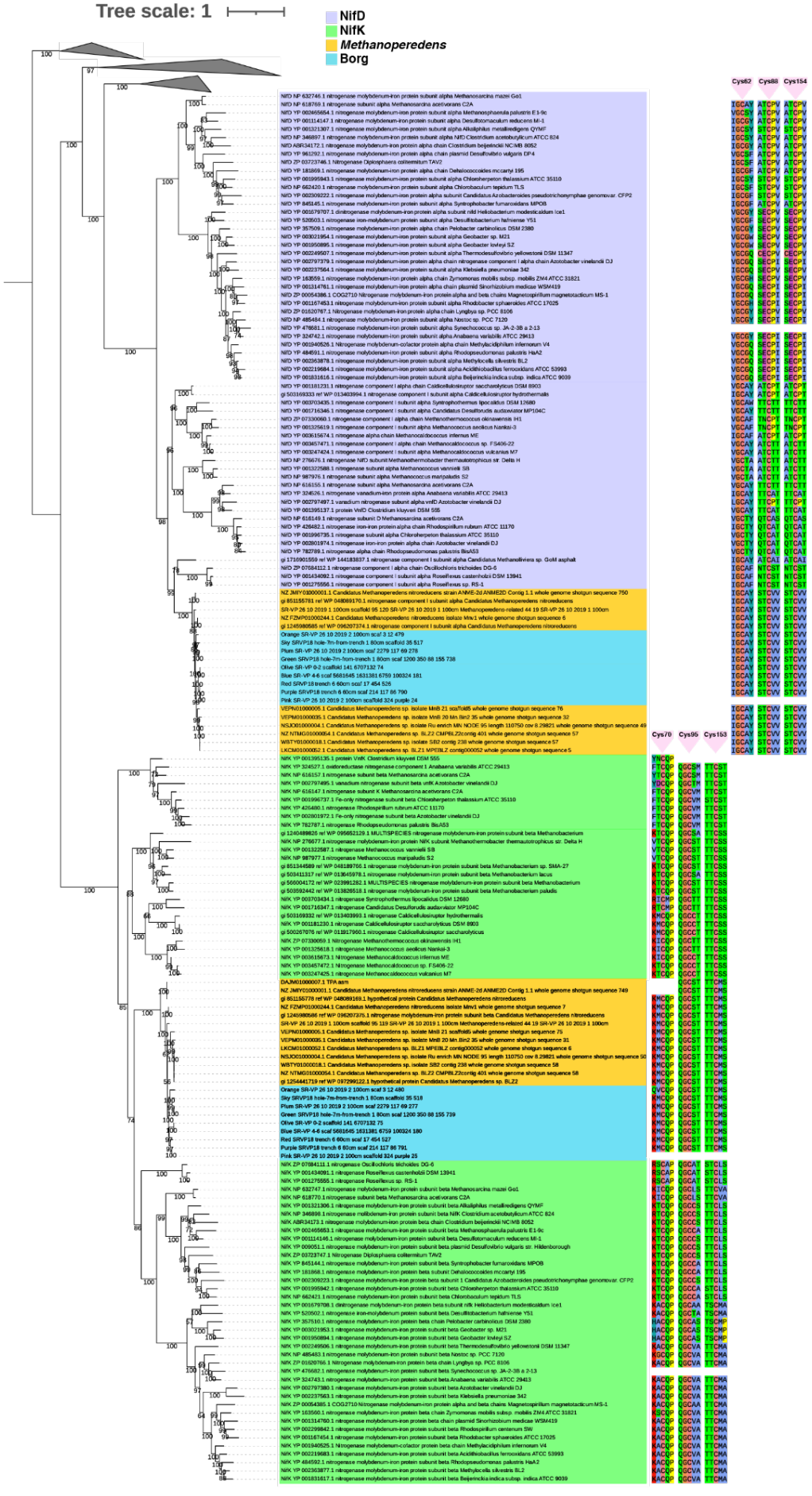
Tree scale: 1
gi 490649080 ref WP 004514075.1 electron transfer flavoprotein subunit alpha Geobacter metallireducens
gi 748170320 ref WP 039743889.1 electron transfer flavoprotein subunit alpha Geobacter pickeringii
gi 499245883 ref WP 010943423.1 electron transfer flavoprotein subunit alpha Geobacter sulfurreducens
gi 1740846521 ref WP 149307481.1 electron transfer flavoprotein subunit alpha Geobacter sp. Red88
gi 499685278 ref WP 011366012.1 electron transfer flavoprotein subunit alpha Geobacter metallireducens
gi 500472450 ref WP 011939284.1 electron transfer flavoprotein subunit alpha Geobacter uraniireducens
gi 1211937151 ref WP 088533877.1 FAD-binding protein Geobacter sp. DSM 9736
gi 501816144 ref WP 012646895.1 FAD-binding protein Geobacter daltonii
gi 500471304 ref WP 011938138.1 electron transfer flavoprotein subunit alpha Geobacter uraniireducens
VEPM01000015.1 Candidatus Methanoperedens sp. isolate MnB 20 Mn.Bin2 15 whole genome shotgun sequence 83
LKCM01000123.1 Candidatus Methanoperedens sp. BLZ1 MPEBLZ contig000124 whole genome shotgun sequence 2
NSJO01000053.1 Candidatus Methanoperedens sp. isolate Ru enrich MN NODE 636 length 26587 cov 7.30518 whole genome shotgun sequence 12
WBTY01000002.1 Candidatus Methanoperedens sp. isolate SB2 contig 39 whole genome shotgun sequence 48
NZ NTMG01000069.1 Candidatus Methanoperedens sp. BLZ2 CMPBLZ2contig 254 whole genome shotgun sequence 49
gi 1254442650 ref WP 097300053.1 electron transfer flavoprotein subunit alpha Candidatus Methanoperedens sp. BLZ2
VEPN01000037.1 Candidatus Methanoperedens sp. isolate MnB 21 scaffold37 whole genome shotgun sequence 3
SR-VP 0-2 scaffold 141 5309183 11
Pink SRVP18 trench 6 60cm scaffold 412 95
Blue SR-VP 0-2 scaffold 141 5659063 34
Blue-2 SR-VP 4-6 scaffold 141 3657090-confirmed 46


Tree scale: 1
gi 754578536 ref WP 041971070.1 electron transfer flavoprotein subunit beta FixA family protein Geobacter sp. OR-1
gi 1718629421 ref WP 145025480.1 electron transfer flavoprotein subunit beta FixA family protein Geobacter argillaceus
gi 501445206 ref WP 012468655.1 electron transfer flavoprotein subunit beta FixA family protein Geobacter lovleyi
gi 505927435 ref WP 015719662.1 electron transfer flavoprotein subunit beta FixA family protein Geobacter sp. M18
gi 501521996 ref WP 012529933.1 electron transfer flavoprotein subunit beta FixA family protein Geobacter bemidjiensis
gi 1740745492 ref WP 149211626.1 electron transfer flavoprotein subunit beta FixA family protein Geobacter sp. FeAm09
gi 499685279 ref WP 011366013.1 electron transfer flavoprotein subunit beta FixA family protein Geobacter metallireducens
gi 502774505 ref WP 013009481.1 electron transfer flavoprotein subunit beta FixA family protein Denitrovibrio acetiphilus
gi 499245961 ref WP 010943501.1 electron transfer flavoprotein subunit beta FixA family protein Geobacter sulfurreducens
gi 1718627247 ref WP 145023344.1 electron transfer flavoprotein subunit beta FixA family protein Geobacter argillaceus
gi 500471303 ref WP 011938137.1 electron transfer flavoprotein subunit beta FixA family protein Geobacter uraniireducens
gi 490649079 ref WP 004514074.1 electron transfer flavoprotein subunit beta FixA family protein Geobacter metallireducens
gi 505928539 ref WP 015720191.1 electron transfer flavoprotein subunit beta FixA family protein Geobacter sp. M18
gi 501522582 ref WP 012530519.1 electron transfer flavoprotein subunit beta FixA family protein Geobacter bemidjiensis
gi 500472451 ref WP 011939285.1 electron transfer flavoprotein subunit beta FixA family protein Geobacter uraniireducens
gi 1565960968 ref WP 129127556.1 electron transfer flavoprotein subunit beta FixA family protein Geomonas oryzae
gi 501521850 ref WP 012529787.1 electron transfer flavoprotein subunit beta FixA family protein Geobacter bemidjiensis
gi 1224028553 ref WP 092058700.1 electron transfer flavoprotein subunit beta FixA family protein Desulfuromonas acetexigens
gi 1054975839 ref WP 066729391.1 electron transfer flavoprotein subunit beta FixA family protein Desulfuromonas sp. DDH964
gi 501811410 ref WP 012645351.1 electron transfer flavoprotein subunit beta FixA family protein Geobacter daltonii
gi 505925153 ref WP 015718751.1 electron transfer flavoprotein subunit beta FixA family protein Geobacter sp. M18
gi 490649041 ref WP 004514036.1 electron transfer flavoprotein subunit beta FixA family protein Geobacter metallireducens
gi 1565959366 ref WP 129125969.1 electron transfer flavoprotein subunit beta FixA family protein Geomonas oryzae
gi 501521960 ref WP 012529897.1 electron transfer flavoprotein subunit beta FixA family protein Geobacter bemidjiensis
gi 1211937150 ref WP 088533876.1 electron transfer flavoprotein subunit beta FixA family protein Geobacter sp. DSM 9736
gi 501816141 ref WP 012646894.1 electron transfer flavoprotein subunit beta FixA family protein Geobacter daltonii
SCQG01000161.1 Candidatus Methanoperedens sp. isolate DP16D bin.42 DP16D bin.42-43202 whole genome shotgun sequence 8
VEPM01000015.1 Candidatus Methanoperedens sp. isolate MnB 20 Mn.Bin2 15 whole genome shotgun sequence 84
NSJO01000053.1 Candidatus Methanoperedens sp. isolate Ru enrich MN NODE 636 length 26587 cov 7.30518 whole genome shotgun sequence 13
WBTY01000002.1 Candidatus Methanoperedens sp. isolate SB2 contig 39 whole genome shotgun sequence 47
NZ NTMG01000069.1 Candidatus Methanoperedens sp. BLZ2 CMPBLZ2contig 254 whole genome shotgun sequence 48
gi 1254442649 ref WP 097300052.1 electron transfer flavoprotein subunit beta FixA family protein Candidatus Methanoperedens sp. BLZ2
LKCM01000123.1 Candidatus Methanoperedens sp. BLZ1 MPEBLZ contig000124 whole genome shotgun sequence 3
VEPN01000037.1 Candidatus Methanoperedens sp. isolate MnB 21 scaffold37 whole genome shotgun sequence 2
Pink SRVP18 trench 6 60cm scaffold 412 94
Olive SR-VP 0-2 scaffold 141 5659063 35
Blue SR-VP 4-6 scaffold 141 3657090 45

C.


Tree scale: 1
gi 757131995 ref WP 042686263.1 aconitate hydratase AcnA Methermicoccus shengliensis
gi 851318659 ref WP 048181799.1 aconitate hydratase AcnA Methanosarcina sp. MTP4
gi 1561208052 ref WP 128503478.1 aconitate hydratase AcnA Methanosarcina sp. MSH10X1
gi 851272176 ref WP 048142551.1 aconitate hydratase AcnA Methanosarcina horonobensis
gi 1245979160 ref WP 096206094.1 aconitate hydratase AcnA Candidatus Methanoperedens nitroreducens
NZ FZMP01000178.1 Candidatus Methanoperedens nitroreducens isolate Mnv1 whole genome shotgun sequence 67
NZ JMIY01000005.1 Candidatus Methanoperedens nitroreducens strain ANME-2d ANME2D Contig 5.5 whole genome shotgun sequence 237
gi 851160869 ref WP 048091568.1 aconitate hydratase AcnA Candidatus Methanoperedens nitroreducens
DAJM01000040.1 TPA asm
SRVP18 trench 6 60cm scaf 17 454 312
SRVP18 trench 6 60cm scaffold 2389 35
SR-VP 4-6 scaffold 141 412842-confirmed 3
SR-VP 0-2 scaffold 141 8177302 37
SR-VP 4-6 scaffold 141 2371278 37
SRVP18 hole-7m-from-trench 1 80cm scaf 1200 350 88 155- 766
SR-VP 26 10 2019 2 100cm scaf 2279 117 69 250
SR-VP 26 10 2019 2 100cm scaffold 560 73
SR-VP 26 10 2019 2 90cm scaf 422 200 35 51 770
SR-VP 26 10 2019 1 100cm scaffold 111 160
BLACK-SR-VP 26 10 2019 C 40cm scaffold 23 FINAL 527
SR-VP 26 10 2019 1 90cm scaffold 550 black 31


D.
Tree scale: 1
Group VI
Group IV
Group V
Borg NifH
Group IV
Borg NifH
Group I/II/III-like
Group II
Group I
Group III


Tree scale: 1
NifD
NifK
Methanoperedens
Borg
Cys62
Cys88
Cys154
Cys70
Cys95
Cys153

Methods

Sampling and creation of metagenomic datasets

We analyzed sequences from sediments of an aquifer in Rifle, Colorado that were retrieved from cores from depths of 5 and 6 m below the surface(Hug et al., 2015) in July 2011, and cell concentrates from pumped groundwater from the same aquifer collected at a time of elevated O_2 concentration in May 2013. Discharge from the Corona Mine, Napa County, California was sampled in December, 2019. Shallow pore water was collected from the riverbed at the East River, Crested Butte, Colorado sampled in August 2016. Soil was sampled from depth intervals between 1 cm to 1 m from a permanently moist wetland located in Lake County, California. Wetland soils were sampled in late October and early November of 2017, 2018 and 2019. DNA was extracted from each sample (DNeasy PowerSoil Pro) and submitted for Illumina sequencing (150 bp or 250 bp reads) at the QB3 facility, University of California, Berkeley. Reads were adapter and quality trimmed using BBduk(Bushnell, 2014a) and sickle(Joshi and Sickle, 2011). Filtered reads were assembled using IDBA-UD(Peng et al., 2012) and MEGAHIT, gene predictions were established using Prodigal(Hyatt et al., 2010) and USEARCH(Edgar, 2010) was used for initial annotations(Edgar, 2010; Joshi and Sickle, 2011; Li et al., 2015; Peng et al., 2012). Functional predictions and predictions of tRNAs followed previously reported methods(Al-Shayeb et al., 2020).

Genome identification, binning, and curation

Hundreds of kbp *de novo* assembled sequences were identified to be of interest as potential novel extrachromosomal elements first based on their taxonomic profile. The taxonomic profiles were determined through a voting scheme in which the taxonomy is assigned at the species to domain level (Bacteria, Archaea, Eukaryotes, no Domain) by comparison with a sequence database (protein annotations in the UniProt and ggKbase: https://ggkbase.berkeley.edu/) when the same taxonomic assignment received >50% votes. Assembled sequences selected for further analysis had no taxonomic profile, even at the Domain level. The majority of contigs of interest had more genes with similarity to those of archaea of the genus *Methanoperedens spp.* than to any other genus (see **Fig. S4**). The second feature of interest was dominance by hypothetical proteins yet absence of genes that would indicate identification as phage or viruses or plasmids.

These initially identified large fragments were manually curated to remove scaffolding gaps and local assembly errors, to extend and join contigs with the same profile, GC, and coverage, and then to extend the near-complete sequences fully into their long terminal repeats. The last step required reassignment of reads mapped at one end and at double depth to both ends. The fully extended sequences had no unplaced reads extending outwards, despite genome-wide deep coverage. Given this, and the absence of any fragments that could potentially be part of a larger genome, it was concluded that sequences represented linear genomes.

In more detail, our curation method involved mapping of reads to the *de novo* fragments and extension within gaps and at termini using previously unplaced reads that we added based on overlap or by the relocation of misplaced reads (these could often be identified based on improper paired reads distances and/or wrong orientation). Local assembly errors were sought by visualization of the reads mapped throughout the assembly and identified based on imperfect read support, or where a subset of reads was partly discrepant and discrepancies involved sequences that were shared by tandem direct repeats of the same region (i.e., the tandem direct repeat regions were collapsed during assembly). *De novo* assembled sequences often ended in tandem direct repeat regions because repeats fragment assemblies. To resolve local assembly errors, gaps were inserted and reads relocated to generate the sequence required to fill the gaps. This ensured comprehensive essentially perfect agreement between reads and the final consensus sequence. In some cases, the tandem direct repeat regions had greater than the expected depth of mapped reads and no reads spanned the flanking unique sequences. In these cases, the repeat number was approximated to achieve the expected read depth, but some arrays may be larger than shown. GC skew and cumulative GC skew were calculated using iRep(Brown et al., 2016) for the fully manually curated complete genomes and the patterns were used to identify the origins and terminus of replication. The pattern of use of coding strands for genes (predicted in Bacterial Code 11) was compared to these origin and terminus predictions to resolve genome organization. The curated sequences were searched for perfect repeats of lengths ≥ 50 nucleotides using Repeat Finder in Geneious. When repeat sequences overlapped, the unit of direct repeat was identified and the length of that repeat, number of repeats, location (within gene vs. intergenic), and genome position were tabulated. Once the features characteristic of the extrachromosomal elements of interest had been determined, we sought related elements. Sequences of interest were identified based on (1) credible partial alignment with the complete sequences, (2) no Domain level profile, (3) GC content 30 - 35%, (4) regions with three or more direct tandem repeats scattered throughout the genome fragment and (5) more best hits to *Methanoperedens spp.* proteins than to proteins from any other organisms. If scaffolds met criterion (1) they were immediately classified as targets. If they met most or all of the other criteria and had similar coverage values, they were binned together with other scaffolds from the same sample with these features. Often, ends of some of the contigs in the same bin overlapped perfectly and could be joined, increasing confidence in the bin quality. Genome sequences were aligned to each other using Mauve(Darling et al., 2010). Where anomalously high (perfect) sequence identity suggestive of recent recombination was detected between Borgs, reads mapped to the region were visualized to verify that the assembly was correct (i.e., not chimeric; also see information in the extended data).
Genome fragments were phylogenetically profiled to establish relatedness to sequences in public databases. Sequences were classified as having no detectable hit if the protein had no similar database sequence with an e-value of <0.0001.

Correlation Analyses
Reads from each sample were aligned to each Methanoperedens and Borg genome. Alignments were performed using bbmap.sh(Bushnell, 2014b) using the following parameters: editfilter=5, minid=0.96, idfilter=0.97, ambiguous=random. The number of reads aligning to each genome was then parsed into a matrix and the correlation

between abundance patterns for *Methanoperedens* and Borg genomes was then calculated using Pearson correlation metric as implemented in scipy(Virtanen et al., 2020). Correlation between a *Methanoperedens* genome and Borg genome was deemed significant if the Pearson correlation between the two genomes was higher than 0.92

CRISPR-Cas analysis

Borg and *Methanoperedens*-encoded CRISPR repeats and spacers were identified using CRISPRDetect(Biswas et al., 2016). The coding sequences from this study were searched against Cas gene sequences reported from previous studies(Makarova et al., 2019) using hmmsearch with $E < 1 \times 10{-}5$ to identify the full locus. Matches were checked using a combination of hmmscan and BLAST searches against the NCBI nr database and manually verified by identifying colocated CRISPR arrays and Cas genes. Spacers extracted from between repeats of the CRISPR locus were compared to sequence assemblies from the sites where Borgs were identified using BLASTN-short(Altschul et al., 1990). Matches with alignment length >24 bp and ≤1 mismatch were retained and targets were classified as bacteria, phage or other. CRISPR arrays that had ≤1 mismatch, were further searched for more spacer matches in the target sequence by finding more hits with ≤3 mismatches.

Protein and gene content analysis

After the identification and curation of Borg genomes and accumulation of usearch annotations for coding sequences, functional annotations were further assigned by searching against PFAM r32, KEGG, pVOG. Transmembrane regions in proteins were predicted with TMHMM. All *Methanoperedens* genomes and genome assemblies as well as 1153 archaeal viruses and extrachromosomal elements were downloaded from the NCBI RefSeq database. Open reading frames were predicted using Prodigal, and all proteins from Borg genomes and the reconstructed ECE database were clustered into protein families and compared across genomes as previously described(Al-Shayeb et al., 2020). Briefly, the coding sequences were clustered into families using a two-step procedure; first an all-versus-all sequence search was performed using an E-value cut-off of $1 \times 10{-}3$, sensitivity of 7.5 and coverage of 0.5, and a sequence similarity network was built on the basis of the pairwise similarities and the greedy set cover algorithm to define protein subclusters. The resulting subclusters were grouped into protein families using a comparison of hidden Markov models (HMMs). For subfamilies with probability scores of at least 95% and coverage at least 0.50, a similarity score (probability × coverage) was used as weight of the input network in the final clustering using the Markov clustering algorithm, with 2.0 as the inflation parameter. These clusters were defined as the protein families.

Functional annotation

Genes of interest were further verified and compared using NCBI's conserved domain search and InterproScan(McWilliam et al., 2013) to identify conserved motifs within the amino acid sequence. Multiheme cytochromes were identified based on ≥3 CxxCH motifs within one gene. The cellular localization of proteins was predicted with Psort (v3.0.3) using archaea as organism type. Proteins were compared using blastp and aligned using MAFFT v.7.407(Katoh and Standley, 2013) to visualize homologous

regions and check conserved amino acid residues that constitute the active site or are required for cofactor/ligand binding.

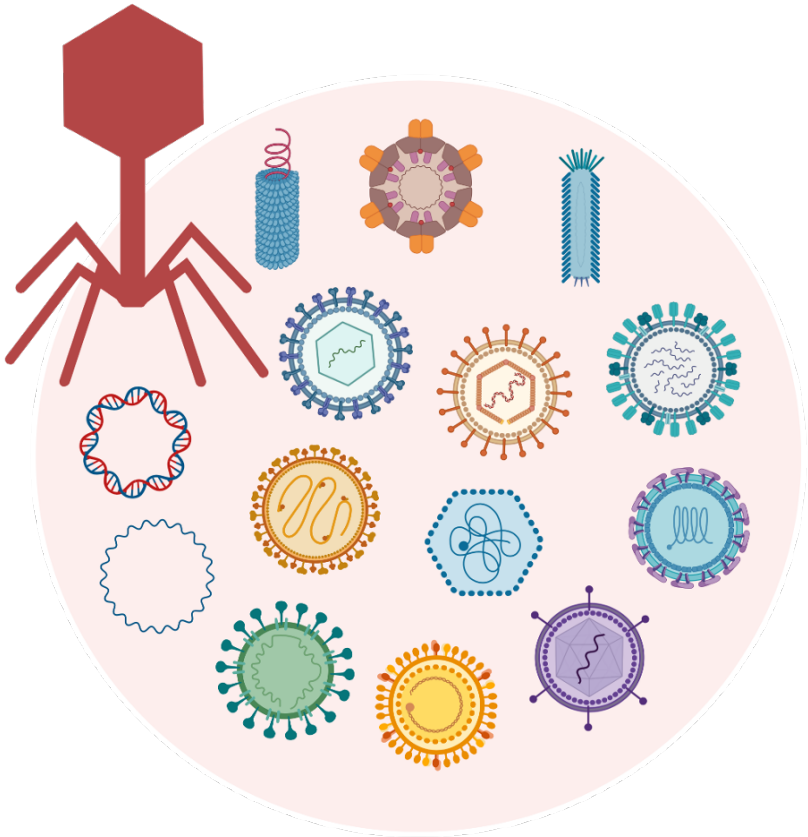
Phylogenetic trees

For each gene, references were compiled by BLASTing the corresponding gene against the NCBI nr database, and their top 50 hits clustered by CD-HIT using a 90% similarity threshold(Huang et al., 2010). The final set of genes was aligned using MAFFT v.7.407 and a phylogenetic tree was inferred using IQTREE v.1.6.6 using automatic model selection(Nguyen et al., 2015) and visualized using iTOL(Letunic and Bork, 2007). Synteny plots were generated using Mauve(Darling et al., 2004), and gene clusters through Adobe Illustrator and ggenes.

4 Chapter 4: Petabase-scale sequence alignment catalyses viral discovery

Robert C. Edgar #, Jeff Taylor #, Victor Lin #, Tomer Altman #, Pierre Barbera #, Dmitry Meleshko #, Dan Lohr #, Gherman Novakovsky #, Benjamin Buchfink #, Basem Al-Shayeb #, Jillian F. Banfield #, Marcos de la Peña #, Anton Korobeynikov #, Rayan Chikhi # & Artem Babaian #

Published in Nature, 2022.



4.1 **Abstract**

Public databases contain a planetary collection of nucleic acid sequences, but their systematic exploration has been inhibited by a lack of efficient methods for searching this corpus, now exceeding multiple petabases and growing exponentially (Leinonen et al., 2011). We developed a cloud computing infrastructure, Serratus, to enable ultra-high throughput sequence alignment at the petabase scale. We searched 5.7 million biologically diverse samples (10.2 petabases) for the hallmark gene RNA dependent RNA polymerase, identifying well over 10^5 novel RNA viruses and thereby expanding the number of known species by roughly an order of magnitude. We characterised novel viruses related to coronaviruses, hepatitis δ virus, and huge phages respectively and explored their environmental reservoirs. To catalyse the ongoing revolution of viral discovery, we established a free and comprehensive database of these data and tools. Expanding the known sequence diversity of viruses can reveal the evolutionary origins of emerging pathogens and improve pathogen surveillance for the anticipation and mitigation of future pandemics.

N.B. All main figures for this manuscript can be found below in their dedicated section. All supplementary files (including figures and tables) can be found online with the published manuscript.

4.2 Introduction

Viral zoonotic disease has had a major impact on human health over the past century, notably the 1918 Spanish influenza, AIDS, SARS, Ebola, and COVID-19. There are an estimated 3×10^5 mammalian virus species from which infectious diseases in humans may arise (Anthony et al., 2013), of which only a fraction is currently known. Global surveillance of virus diversity is required for improved prediction and prevention of future epidemics and is the focus of international consortia and hundreds of research laboratories (Carroll et al., 2018; Johnson et al., 2020).

Pioneering works expanding Earth's virome have each uncovered thousands of novel viruses, with the rate of virus discovery increasing exponentially and driven largely by the increased availability of high-throughput sequencing (Camarillo-Guerrero et al., 2021; Chen et al., 2021; Mitchell et al., 2020; Nayfach et al., 2021a; Shi et al., 2018; Wahba et al., 2020; Wolf et al., 2020). Sequence analysis remains computationally expensive, in particular the assembly of short reads into contigs, limiting the breadth of samples analysed. Here, we propose an alternative alignment-based strategy which is significantly cheaper than assembly and enables processing of massive datasets.

Petabases (1×10^{15} bases) of sequencing data are freely available in public databases such as the Sequence Read Archive (SRA) (Leinonen et al., 2011) where viral nucleic acids are often captured incidental to the goals of the original studies (Moore et al., 2011). To catalyse global virus discovery, we developed the Serratus cloud computing infrastructure for ultra-high throughput sequence alignment, screening 5.7 million ecologically diverse sequencing libraries or 10.2 petabases of data.

Identification of Earth's virome is a fundamental step in preparing for the next pandemic. We lay the foundations for years of future research by enabling direct access to 883,502 RNA dependent RNA polymerase (RdRP) containing sequences, including the RdRP from 132,260 novel RNA viruses (sequences with >10% divergence from a known RdRP), including nine novel coronaviruses. Altogether this captures the collective efforts of over a decade of sequencing studies in a free repository, available at https://serratus.io.

4.3 Results and Discussion

Accessing the planetary virome

Serratus is a free, open-source cloud-computing infrastructure optimised for petabase-scale sequence alignment against a set of query sequences. Using Serratus, we aligned in excess of one million short-read sequencing datasets per day for under 1 US cent per dataset (Extended Figure 1). We used a widely available commercial computing service to deploy up to 22,250 virtual CPUs simultaneously (see Methods), leveraging SRA data mirrored onto cloud platforms as part of the NIH STRIDES initiative.

Our search space spans data deposited over 13 years from every continent and ocean, and all kingdoms of life (Figure 1). We applied Serratus in two of many possible configurations. First, to identify libraries containing known or closely related viruses we

searched 3,837,755 (ca. May 2020) public RNA-seq, meta-genome, meta-transcriptome, and meta-virome datasets (termed sequencing runs) against a nucleotide pangenome of all coronavirus sequences and reference vertebrate viruses. We then aligned 5,686,715 runs (ca. January 2021) against all known viral RdRP amino acid sequences using a specially-optimised version of DIAMOND v2 (Buchfink et al., 2021)[, Methods]), completing this search within 11 days, for a cost of 23,980 USD (Figure 1a and Methods).

Previous approaches for identifying sequences across the entire SRA rely on pre-computed indexes (Karasikov et al., 2020; Katz et al., 2021) requiring exact substring or hash-based matches which limits sensitivity to diverged sequences (Extended Figure 1f). Pre-assembled reads (e.g. NCBI Transcriptome Shotgun Assembly database) enable efficient alignment-based searches (Shi et al., 2018), but are currently available only for a small fraction of the SRA. Serratus aligns a query of up to hundreds of Mb against unassembled libraries, achieving greater sensitivity to diverged viruses compared to substring (k-mer) indexes while using far less computational resources than de novo assembly (Figure 1g and Methods).

A sketch of RNA dependent RNA polymerase

Viral RdRP is a hallmark gene of RNA viruses which lack a DNA stage of replication (Koonin and Dolja, 2014). We identified RdRP by a well-conserved amino acid sub-sequence we call the “palmprint”. Palmprints are delineated by three essential motifs which together form the catalytic core in the RdRP structure (Figure 2 and (Babaian and Edgar, 2021)). We constructed species-like operational taxonomic units (sOTUs) by clustering palmprints at a threshold of 90% amino-acid identity, chosen to approximate taxonomic species (Babaian and Edgar, 2021).

3,376,880 (59.38%) sequencing runs contained ≥ 1 reads mapping to the RdRP query (E-value ≤1e-4). We assembled RdRP aligned reads from each library (and their mate-pairs when available), yielding 4,261,616 “microassembly” contigs. 881,167 (20.7%) contained a high-confidence palmprint identified by Palmscan (false discovery rate = 0.001, (Babaian and Edgar, 2021)), representing 260,808 unique palmprints. Applying Palmscan to reference databases, (Wolf et al., 2018, 2020) we obtained 45,824 unique palmprints, which clustered into 15,016 known sOTUs. If a newly acquired palmprint aligned to a known palmprint at ≥90% identity, it was assigned membership to that reference sOTU, otherwise it was designated novel. We clustered novel palmprints at 90% identity, obtaining 131,957 novel sOTUs, representing an increase of known RNA viruses by a factor ~9.8. Clustering novel palmprints at genus-like 75% and family-like 40% thresholds yielded 78,485 and 3,599 novel OTUs, representing increases of 8.0x and 1.9x, respectively (Figure 2b).

We extracted host, geospatial, and temporal metadata for each biological sample when available (Figure 1c), noting that the majority (88%) of novel RdRP sOTUs were observed from metagenomic or environmental runs, where accurate host inference is challenging. Mapping observations of virus marker genes across time and space suggests ecological niches for these viruses, while improved characterisation of sequence diversity can improve PCR primer design for *in situ* virus identification.

We estimate that ~1% of sOTUs are endogenous virus elements (EVEs), i.e. viral RdRPs which have serendipitously reverse-transcribed into a host germline. We did not attempt to systematically distinguish EVEs from virus RdRPs, noting that EVEs with intact catalytic motifs are likely to be recent insertions which can serve as a representative sequence for related exogenous viruses. Most (60.5%) recovered palmprints were found in exactly one run (singletons), and are observed within the expected frequency range predicted by extrapolating from more abundant sequences (Figure 2b).

The abundance distribution of distinct palmprints is consistent with log-log-linear for each year from 2015 to 2020 (Extended Figure 2e), and over time, singletons are confirmed by subsequent runs at an approximately constant rate (Extended Figure 2g). The majority of novel viruses will be singletons until the diversity represented by the search query and the fraction of the planetary virome sampled in the SRA both approach saturation. Extrapolating one year forward, when the SRA is expected to double in size, we project 430,000 (95% CI [330K, 561K]) additional unique palmprints will be identified by running Serratus with its current query (Figure 2b).

RNA viruses have highly divergent sequences, even within the conserved RdRP (Koonin and Dolja, 2014). Amino acid sequence alignment can recover the majority of RdRP short reads above 60% identity, but sensitivity falls as sequences diverge further (Extended Figure 2f). Subsequent microassembly fragmentation can in part account for the decreased abundance of novel sOTU below 60% identity (Figure 2b), thus the sensitivity to highly diverged (<50% identity) RdRP sequences is limited in the present study. Saturation of virus discovery within the SRA is far from complete, even if data-growth rates are ignored. Intensive search for so called highly diverged or ”dark” viruses (Obbard et al., 2020), in combination with iterative re-analysis (conceptually similar to PSI-BLAST (Altschul et al., 1997)) are likely to yield further expansion of the known virome.

The total number of virus species is estimated to be 10^8 to 10^{12} (Koonin et al., 2020), thus our data captured at most 0.1% of the global virome. However, if exponential data growth combined with increased search sensitivity continues, we are at the cusp of identifying a significant fraction of Earth’s total genetic diversity with tools such as Serratus.

Expanding the scope of *Coronaviridae*

The SARS-CoV-2 pandemic has significantly impacted human society. We further exemplify the potential of Serratus for virus discovery with *Coronaviridae* (CoV), including a recently proposed sub-family (Bukhari et al., 2018) which contains a CoV-like virus, Microhyla alphaletovirus 1 (MLeV), in the frog *Microhyla fissipes*, and Pacific salmon nidovirus (PsNV) described in the endangered *Oncorhynchus tshawytscha* (Mordecai et al., 2019).

First, we identified 52,772 runs containing ≥10 CoV-aligned reads or ≥2 CoV k-mers (32-mer, (Katz et al., 2021)). These runs were *de novo* assembled with a new version of synteny-informed SPAdes called coronaSPAdes (discussed in a companion manuscript

(Meleshko et al., 2021)). This yielded 11,120 identifiable CoV contigs which we annotated for a comprehensive assemblage of *Coronaviridae* in the SRA (see Methods for discussion). With this training data we defined a scoring function to predict subsequent success of assembly (Extended Figure 3b).

CoV and neighbouring palmprints comprise 70 sOTUs, 44 of which are described in public databases. 17 CoV sOTUs contained partial RdRP (inclusive of full palmprint) from an amplicon-based virus discovery study not yet publicly deposited (Tao et al., 2020). The remaining 9 sOTUs are novel viruses, with protein domains consistent with a CoV or CoV-like genome organisation (Extended Figure 4).

We operationally designate MLeV, PsNV and the nine novel viruses broadly as group E, noting that all were found in samples from non-mammalian aquatic vertebrates (Figure 3). Notably, *Ambystoma mexicanum* (axolotl) nidovirus (AmexNV) was assembled in 18 runs, 11 of which yielded common ~19 kb contigs. Easing the criteria of requiring an RdRP match in a contig, 28/44 (63.6%) of the runs from the associated studies were AmexNV positive (Tsai et al., 2020). Consistent assembly breakpoints in AmexNV, PsNV and similar viruses suggests that the viral genomes of this clade of CoV-like viruses are organised in at least two segments, one containing ORF1ab with RdRP, and a shorter segment containing a lamin-associated domain protein, spike and N' accessory genes (Figure 3). An assembly gap with common breakpoints is present in the published PsNV genome (Mordecai et al., 2019). Together these seven monophyletic species possibly represent a distinct clade of segmented CoV-like nidoviruses, although molecular validation of this hypothesis is required.

While our manuscript was under review, public transcriptome screening by Miller *et al.* (Miller et al., 2021), identified three group E CoV sequences not included in our sOTU analysis. One CoV+ library had failed at the alignment step, and microassembly from two others yielded incomplete palmprint sub-sequences, thus lacking the required specificity for the systematic palmprint classification. A high sensitivity re-analysis of microassemblies for any group E RdRP sequence fragment captured the two missing Miller *et al.* CoV, and found another approximately 25 putative-novel CoV species from 53 fragmented contigs (Supplementary Table 1e).

In addition to identifying genetic diversity within CoV, we cross-referenced CoV+ library meta-data to identify possible zoonoses and vectors of transmission. Discordant libraries, one in which a CoV is identified and the viral expected host (Mukherjee et al., 2021) does not match the sequencing library source taxa, were rare, accounting for only 0.92% of cases (Supplementary Table 1f).

An important limitation for these analyses is that the nucleic acid reads do not prove viral infection has occurred in the nominal host species. For example, we identified five libraries in which a porcine, avian or bat coronavirus were found in plant samples. The parsimonious explanation is that CoV was present in faeces/fertiliser originating from a mammalian or avian host applied to these plants. However, this exemplifies a merit of exhaustive search in identifying transmission vectors and for monitoring the geo-temporal distribution of viruses.

Rapid expansion into the viral unknowns

The global mortality from viral hepatitis exceeds that of HIV/AIDS, tuberculosis or malaria (Stanaway et al., 2016). Hepatitis δ virus (HDV) has a small circular RNA genome (~1.7 knt) which folds into a rod-like shape and encodes three genes: a delta antigen protein, and two self-cleaving ribozymes (drbz) (Taylor, 2020).

Prior to 2018, HDV was the sole known member of its genus; 13 drbz-containing members have since been characterised (Bergner et al., 2021; Chang et al., 2019; Iwamoto et al., 2021; Paraskevopoulou et al., 2020; Szirovicza et al., 2020; Wille et al., 2018), and recently a second class of ribozyme (known as hammerhead or hhrbz) characteristic of plant viroids was identified in delta-like viruses we refer to as epsilon viruses (de la Peña et al., 2021). By sequence search for the delta antigen protein and ribozymes, we identified 14 delta viruses, 39 epsilon viruses and 311 enigmatic sequences with deltavirus-like synteny we term zeta viruses (Figure 4, Extended Figure 5). The evolutionary histories of these mammalian delta viruses are explored further in a companion paper (Bergner et al., 2021).

The zeta virus circular genomes are highly compressed, ranging from 324-789 nt and predicted to fold into rod-like structures. They contain a hhrbz in each orientation and encode two ORFs, one sense and one anti-sense. Both ORFs generally lack stop codons and encompass the entire genome, potentially producing an endless tandem-repeat of antigen. The atypical coiled-coil domain of the HDV antigen (Zuccola et al., 1998) is conserved in the antigens of new delta and epsilon viruses, whereas epsilon and zeta genomes show analogous hhrbzs (Extended Figure 6), supporting that these sequences may share common ancestry. These abundant elements may help to solve a long-standing question about the origins of circular RNA subviral agents in higher eukaryotes (Extended Figure 6), historically regarded as molecular fossils of a prebiotic RNA world (Flores et al., 2014).

To evaluate the feasibility of applying Serratus in the context of microbiome research, we sought to locate bacteriophages related to recently reported huge phages (Al-Shayeb et al., 2020), searching for terminase amino acid sequences. Targeted assembly of 287 high-scoring runs returned 252 terminase-containing contigs ≥140 kbp. Phylogenetics of these sequences resolved new groups of phages with large genomes (Figure 4e). While most phages were from a single animal genus, we identified closely related phages crossing animal orders, including related phages in a human from Bangladesh (ERR866585) and groups of cats (PRJEB9357) and dogs (PRJEB34360) from England, sampled 5 years apart. Similarly, we recovered two ~554 kbp Lak megaphage genomes (among the largest animal microbiome phages reported to date) that are extremely closely related to sequences previously reported from pigs, baboons and humans (Extended Figure 7) (Devoto et al., 2019). These two genomes were circularised and manually curated to completion. The large carrying capacity of such phages and broad distribution underlines their potential for extensive lateral gene transfer amongst animal microbiomes and modification of host bacterial function. The newly-recovered sequences substantially expand and augment the inventory of phages with genomes whose length range overlaps with those of bacteria.

4.4 Conclusions

Since the completion of the human genome, growth of DNA sequencing databases has outpaced Moore's Law. Serratus provides rapid and focused access to genomic sequences captured over more than a decade by the global research community which would otherwise be inaccessible in practice. This work and further extensions of petabase scale genomics (Bradley et al.; Karasikov et al., 2020; Katz et al., 2021) are shaping a new era in computational biology, enabling expansive gene discovery, pathogen surveillance, and pangenomic evolutionary analyses.

Optimal translation of such massive datasets into meaningful biomedical advances requires free and open collaboration amongst scientists (Baker et al., 2020). The current pandemic underscores the need for prompt, unrestricted and transparent data sharing. With these goals in mind, we deposited 7.3 terabytes of virus alignments and assemblies into an open-access database which can be explored via a graphical web interface at https://serratus.io or programmatically through the Tantalus R package and its PostgreSQL interface.

The "metagenomics revolution" of virus discovery is accelerating (Nayfach et al., 2021a; Wolf et al., 2020). Innovative fields such as high-throughput viromics (Letko et al., 2020a) can leverage vast collections of virus sequences to inform policies that predict and mitigate emerging pandemics (Letko et al., 2020b). Combining ecoinformatics with virus, host, and geotemporal metadata offers a proof of concept for a global pathogen surveillance network, arising as a byproduct of centralised and open data sharing.

Human population growth and encroachment on animal habitats is bringing more species into proximity, leading to increased rate of zoonosis (Anthony et al., 2013) and accelerating the Anthropocene mass extinction (Chase et al., 2020). While Serratus enhances our capability to chronicle the full genetic diversity of our planet, the genetic diversity of the biosphere is diminishing. Thus investment in collection and curation of biologically diverse samples, with emphasis on geographically under-represented regions, has never been more pressing. If not for the conservation of endangered species, then to better conserve our own.

4.5 Figures

Figure 1: Searching the planetary virome

a Total bases searched from the 5,686,715 SRA sequencing runs analysed in the viral RdRP search grouped by sample taxonomy, where available (see Extended Figures 1 and 3, and Supplementary Table 1). 8,871/15,016 (59%) of known RdRP species-like operational taxonomic units (sOTUs) were observed in the SRA, and 131,957 unique and novel RdRP sOTUs were identified (see Extended Figure 2). sOTUs identified in multiple taxonomic groups are counted in each group separately, numbers shown indicate the number of novel sOTUs in each group. b Release dates of the runs included in the analysis reflecting the growth rate of available data. c Sample locations for 635,656 RdRP-containing contigs (27.8% of samples lacked geographic metadata). The high density of RdRP seen in North America, Western Europe and Eastern Asia reflects the substantial acquisition bias for samples originating from these regions. Interactive RdRP map is available at https://serratus.io/geo.

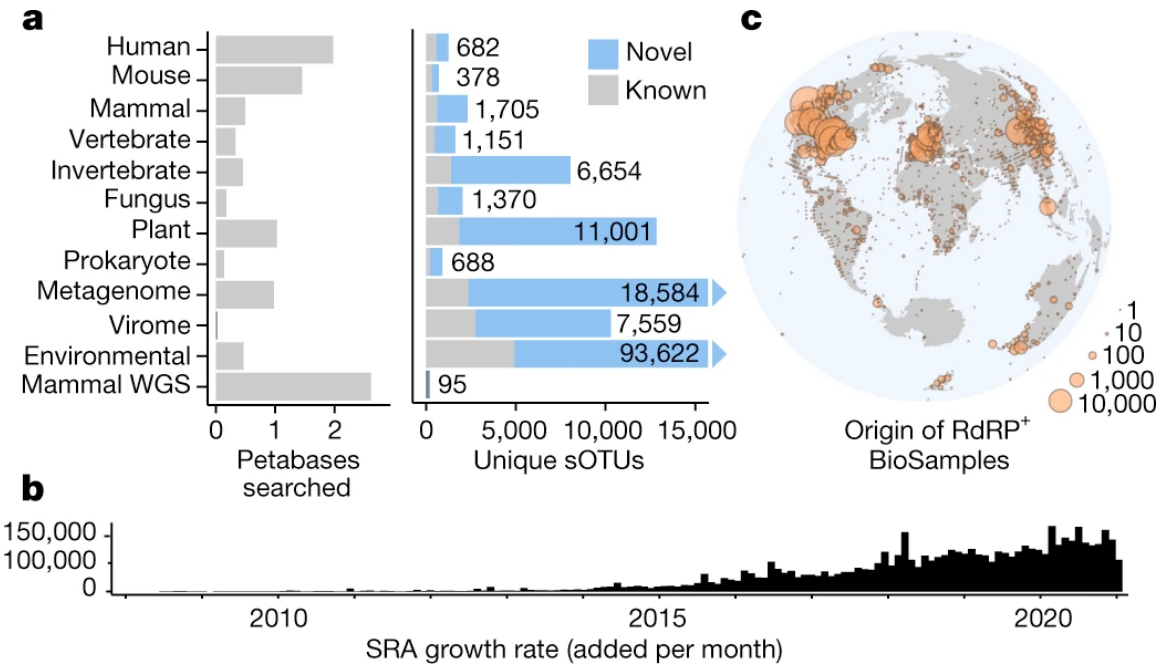


Figure 2: RNA dependent RNA Polymerase in the Sequence Read Archive

a The RdRP palmprint is the protein sequence spanning three well-conserved sequence motifs (A, B, and C), including intervening variable regions, exemplified within full-length poliovirus RdRP structure with essential aspartic acid residues(*) (pdb: 1RA6 (Thompson and Peersen, 2004)). Conservation was calculated from RdRP alignment in(Wolf et al., 2018), trimmed to the poliovirus sequence; motif sequence logos are shown below. b Per-phylum histogram of amino acid identity of novel species-like operational taxonomic units (sOTUs) aligned to the NCBI non-redundant protein database and Extended Figure 3c shows per-order distribution. inlay Preston plot and linear regression of palmprint abundances indicates that singleton palmprints (i.e., observed in exactly one run) occur within 95% confidence intervals of the value predicted by extrapolation from high-abundance palmprints (linear regression applied to log-transformed data), and this distribution is consistent through time (Extended Figure 2).

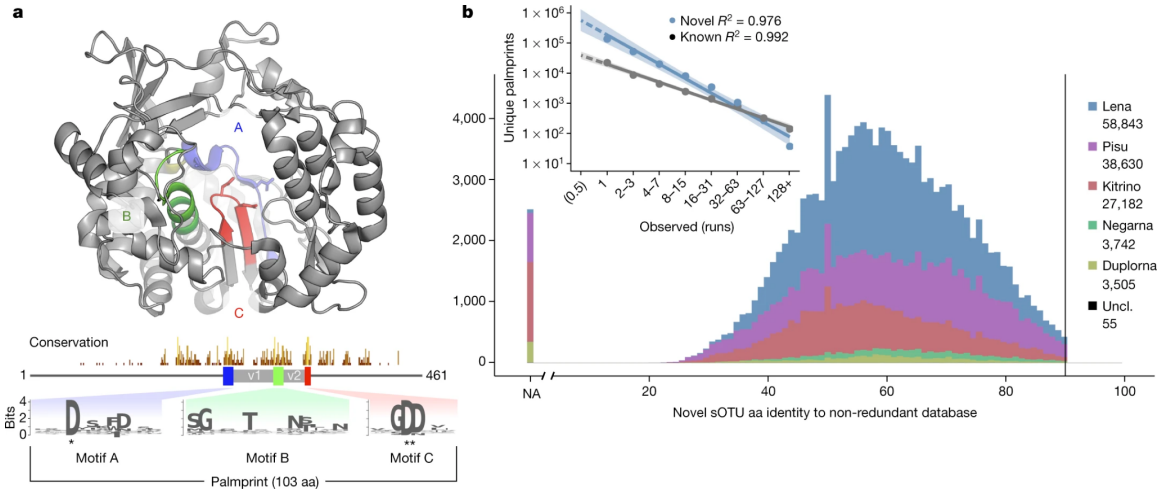


Figure 3: Expanding Coronaviridae

a Phylogram for group E sequences. Six viruses were similar to PsNV in Ambystoma mexicanum (axolotl; AmexNV), Puntigrus tetrazona (tiger barb; PtetNV), Hippocampus kuda (seahorse; HkudNV), Syngnathus typhle (broad-nosed pipefish; StypNV), Takifugu pardalis (fugu fish; TparNV), and the Acanthemblemaria sp. (blenny; AcaNV). More distant members identified were in Hypomesus transpacificus (the endangered delta smelt; HtraNV), Silurus sp. (catfish) SilNV, and Monopterus albus (asian swamp eel) MalbNV. b Unrooted phylogram for Coronaviridae annotated with genera (Greek letters) and group E CoV-like nidoviruses (see also Extended Figure 4). Maximum likelihood tree generated by clustering the RdRP amino acid sequences at 97% identity to show sub-species variability. c Genome structure of AmexNV and the contigs recovered from group E CoV-like viruses annotated with hidden-Markov model matches. AmexNV contigs contain an identical 129 nt trailing sequence (Tr). All the putatively segmented CoV-like are monophyletic with PsNV. A gap in the PsNV reference sequence (Mordecai et al., 2019) is shown with circles, overlapping the common contig ends seen in these viruses.

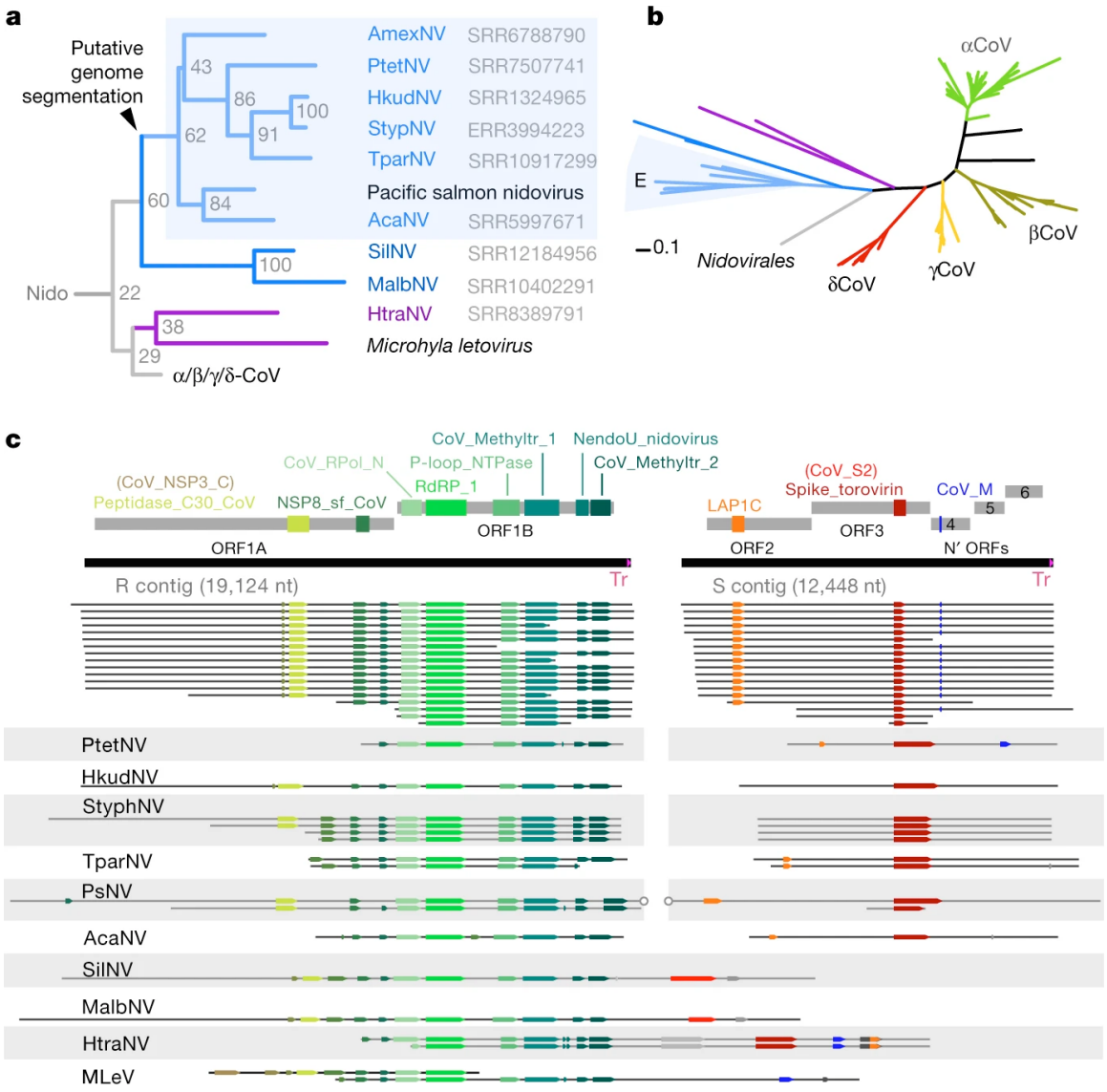
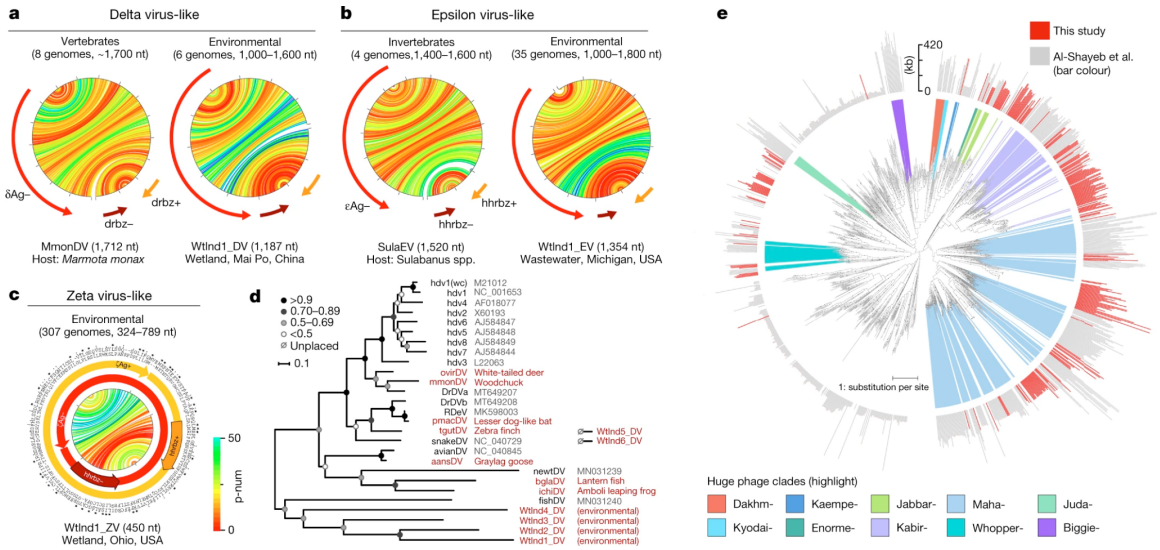

a
Putative genome segmentation
43
86
100
91
62
60
84
100
Nido
22
38
29
α/β/γ/δ-CoV
AmexNV SRR6788790
PtetNV SRR7507741
HkudNV SRR1324965
StypNV ERR3994223
TparNV SRR10917299
Pacific salmon nidovirus
AcaNV SRR5997671
SilNV SRR12184956
MalbNV SRR10402291
HtraNV SRR8389791
Microhyla letovirus
b
αCoV
E
0.1
Nidovirales
δCoV
γCoV
βCoV
c
CoV_Methyltr_1
NendoU_nidovirus
CoV_RPol_N
P-loop_NTPase
CoV_Methyltr_2
(CoV_NSP3_C)
Peptidase_C30_CoV
NSP8_sf_CoV
RdRP_1
ORF1B
ORF1A
R contig (19,124 nt)
Tr
(CoV_S2)
Spike_torovirin
CoV_M
6
5
LAP1C
ORF3
4
ORF2
N′ ORFs
S contig (12,448 nt)
Tr
PtetNV
HkudNV
StyphNV
TparNV
PsNV
AcaNV
SilNV
MalbNV
HtraNV
MLeV

Figure 4: Expanding deltaviruses and huge phages

a Genome structure for the Marmota monax Delta virus (MmonDV) and a DV-like genome detected in an environmental dataset each containing a negative-sense delta-antigen (δAg) ORF; two delta ribozymes (dvrbz); and characteristic rod-like folding, where each line shows the predicted base-pairing within the RNA genome, coloured by base-pairing confidence score (p-num) (Zuker, 2003). b Similar genome structure for the Sulabanus spp. Epsilon virus-like (SulaEV) and an EV-like genome from an environmental dataset each containing a negative-sense epsilon-antigen (δAg) ORF; two hammerhead ribozymes (hhrbz); and rod-like folding. c Example of the compact genome structure of a Zeta virus-like from an environmental dataset containing two predicted zeta-antigen (ζAg+/-, protein alignment is shown in the outer circles) ORFs without stop codons; two hhrbz overlapping with the ORFs; and rod-like folding. Further novel genomes are shown in Extended Figures 5 and 6. d Maximum-likelihood phylogenetic tree of DVs derived from a delta-antigen protein alignment with bootstrap values. Two divergent environmental DV could not yet be placed. e Tree showing huge phage clade expansion. Black dots indicate branches with bootstrap values >90. Outer ring indicates genome or genome fragment length: gray are sequences from (Al-Shayeb et al., 2020) and reference sequences, shadings indicate previously defined clades of phages with very large genomes (200-735 kbp). The Kabirphages (light purple) are shown in expanded view in Extended Figure 7.



4.6 **Methods**

1.1 Serratus alignment architecture

Serratus (v0.3.0) (https://github.com/ababaian/serratus) is an open-source cloud-infrastructure designed for ultra-high throughput sequence alignment against a query sequence or pangenome (Extended Figure 1). Serratus compute costs are dependent on search parameters (expanded discussion available: https://github.com/ababaian/serratus/wiki/pangenome_design). The nucleotide vertebrate viral pangenome search (bowtie2, database size: 79.8 Mb) reached processing rates of 1.29 million SRA runs in 24-hours at a cost of $0.0062 US dollars per dataset (Extended Figure 1). The translated-nucleotide RdRP search (DIAMOND (Buchfink et al., 2021), database size: 7.1 Mb) reached processing rates exceeding 0.5 million SRA runs in 12-hours at a cost of $0.0042 per dataset. All 5,686,715 runs analysed in the RdRP search were completed within 11 days for a total cost of $23,980 or ~$2,350 per petabase. For a detailed breakdown of Serratus project costs and recommendations for managing cloud-computing costs, see Serratus wiki: https://github.com/ababaian/serratus/wiki/budget. Tutorials on how to find particular novel viruses using Serratus data is available at https://github.com/ababaian/serratus /wiki/Find_novel_viruses.

1.1.1 Computing cluster architecture

The processing of each sequencing library is split into three modules *dl* (download), *align*, and *merge*. The *dl* module acquires compressed data (.sra format) via prefetch (v2.10.4), from the Amazon Web Services (AWS) Simple Storage Service (S3) mirror of the Sequence Read Archive (SRA), decompresses to FASTQ with fastq-dump (v2.10.4), and splits the data into chunks of 1 million reads or read-pairs (*fq-blocks*) into a temporary S3 cache bucket. To mitigate excessive disk usage caused by a few large datasets, a total limit of 100 million reads per dataset was imposed. The *align* module reads individual fq-blocks and aligns to an indexed database of user-provided query sequences using either bowtie2 (v2.4.1*,*--very-sensitive-local) (Langmead and Salzberg, 2012) for nucleotide search, or DIAMOND (v2.0.6 development version, --mmap-target-index --target-indexed --masking 0 --mid-sensitive -s 1 -c1 -p1 -k1 -b 0.75) (Buchfink et al., 2021) for translated-protein search. Finally, the *merge* module concatenates the aligned blocks into a single output file (.bam for nucleotide, or .pro for protein) and generates alignment statistics with a Python script (see Summarizer below).

1.1.2 Computing resource allocation

Each component is launched from a separate AWS autoscaling group with its own launch template, allowing the user to tailor instance requirements per task. This enabled us to minimise the use of costly block storage during compute-bound tasks such as alignment. We used the following Spot instance types; *dl*: 250GB SSD block storage, 8vCPUs, 32GB RAM (*r5.xlarge*) ~1300 instances; align: 10GB SSD block storage, 8vCPUs, 8GB RAM (*c5.xlarge*) ~4,300 instances; *merge*: 150GB SSD block storage, 4vCPUs, 4GB RAM (*c5.large*) ~60 instances. Users should note that it may be necessary to submit a service ticket to access more than the default EC2 instance limit.

AWS Elastic Compute Cloud (EC2) instances have higher network bandwidth (up to 1.25 GB/s) than block storage bandwidth (250 MB/s). To exploit this, we used S3 buckets as a data buffering and streaming system to transfer data between instances following methods developed in a previous cloud architecture (https://github.com/FredHutch/sra-pipeline). This, combined with splitting of FASTQ files into individual blocks, effectively eliminated file input/output (i/o) as a bottleneck, since the available i/o is multiplied per running instance (conceptually analogous to a RAID0 configuration or a Hadoop distributed filesystem (Schatz, 2009)).

Using S3 as a buffer also allowed us to decouple the input and output of each module. S3 storage is cheap enough that in the event of unexpected issues (e.g., exceeding EC2 quotas) we could resolve system problems in realtime and resume data processing. For example, shutting down the align modules to hotfix a genome indexing problem without having to re-run the *dl* modules, or if an alignment instance is killed by a Spot termination, only that block needs to be reprocessed instead of the entire sequencing run.

1.1.3 Work queue and scheduling

The Serratus scheduler node controls the number of desired instances to be created for each component of the workflow, based on the available work queue. We implemented a pull-based work queue. Upon boot-up each instance launches a number of worker threads equal to the number of CPU available. Each *worker* independently manages itself via a boot script, and queries the scheduler for available tasks. Upon completion of the task, the worker updates the scheduler of the result: success, or fail, and queries for a new task. Under ideal conditions, this allows for a worst-case response rate in the hundreds of milliseconds, keeping cluster throughput high. Each task typically lasts several minutes depending on the pangenome.

The scheduler itself was implemented using Postgres (for persistence and concurrency) and Flask (to pool connections and translate REST queries into SQL). The Flask layer allowed us to scale the cluster past the number of simultaneous sessions manageable by a single Postgres instance. The work queue can also be managed manually by the user, to perform operations such as re-attempt downloading of an SRA accession upon a failure or to pause an operation while debugging. Up to 300,000 SRA jobs can be processed in the work queue per batch process.

The system is designed to be fully self-scaling. An “autoscaling controller” was implemented which scales-in or scales-out the desired number of instances per task every five minutes based on the work queue. As a backstop, when all workers on an instance fail to receive work instructions from the scheduler, the instance self shuts-down. Finally a “job cleaner” component checks the active jobs against currently running instances. If an instance has disappeared due to SPOT termination or manual shutdown, it resets the job allowing it to be processed up by the next available instance.

To monitor cluster performance in real-time, we used Prometheus (v2.5.0) and node exporter to retrieve CPU, disk, memory, and networking statistics from each instance, to expose performance information about the work queue, and Python exporter to export information from the Flask server. This allowed us to identify and diagnose performance problems within minutes to avoid costly overruns.

1.1.4 Generating viral summary reports

We define a viral pangenome as the entire collection of reference sequences belonging to a taxonomic viral family, which may contain both full-length genomes and sequence fragments such as those obtained by RdRP amplicon sequencing.

We developed a Summarizer module written in Python to provide a compact, human- and machine-readable synopsis of the alignments generated for each SRA dataset. The method was implemented in Serratus_summarizer.py for nucleotide alignment and Serratus_psummarizer.py for amino acid alignments. Reports generated by the Summarizer are text files with three sections described in detail online (https://github.com/ababaian/serratus/wiki/.summary-Reports). In brief, each contains a header section with alignment meta-data and one-line summaries for each virus family pangenome, reference sequence and gene respectively, with gene summaries provided for protein alignments only.

For each summary line we include descriptive statistics gathered from the alignment data such as the number of aligned reads, estimated read depth, mean alignment identity, and coverage, i.e. the distribution of reads across each reference sequence or pangenome. Coverage is measured by dividing a reference sequence into 25 equal bins and depicted as an ASCII text string of 25 symbols, one per bin; for example oaooomoUU:oWWUUWOWamWAAUW. Each symbol represents $log_2(n + 1)$ where n is the number of reads aligned to a bin in this order _.:uwaomUWAOM^. Thus, '_' indicates no reads, '.' exactly one read, ':' two reads, 'u' 3-4 reads, 'w' 5-7 reads and so on; '^' represents $>2^{13} = 8{,}192$ reads in the bin. For a pangenome, alignments to its reference sequences are projected onto a corresponding set of 25 bins. For a complete genome, the projected pangenome bin number 1,2,...,25 is the same as the reference sequence bin number. For a fragment, a bin is projected onto the pangenome bin implied by the alignment of the fragment to a complete genome. For example, if the start of a fragment aligns half way into a complete genome, bin 1 of the fragment is projected to bin floor(25/2) = 12 of the pangenome. The introduction of pangenome bins was motivated by the observation that bowtie2 selects an alignment at random when there are two or more top-scoring alignments, which tends to distribute coverage over several reference sequences when a single viral genome is present in the reads. Coverage of a single reference genome may therefore be fragmented, and binning to a pangenome better assesses coverage over a putative viral genome in the reads while retaining pangenome sequence diversity for detection.

1.1.5 Identification of viral families within a sequencing dataset

The Summarizer implements a binary classifier predicting the presence or absence of each virus family in the query based on pangenome-aligned short reads. For a given family *F*, the classifier reports a score in the range [0,100] with the goal of assigning a high score to a dataset if it contains *F* and a low score if it does not. Setting a threshold on the score divides datasets into disjoint subsets representing predicted positive and negative detections of family *F*. The choice of threshold implies a trade-off between false positives and false negatives. Sorting by decreasing score ranks datasets in decreasing order of confidence that *F* is present in the reads.

Naively, a natural measure of the presence of a virus family is the number of alignments to its reference sequences. However, alignments may be induced by non-homologous sequence similarity, for example low-complexity sequence.

The score for a family was therefore designed to reflect the overall coverage of a pangenome because coverage across all or most of a pangenome is more likely to reflect true homology, i.e. the presence of a related virus. Ideally, coverage would be measured individually for each base in the reference sequence, but this could add undesirable overhead in compute time and memory for a process which is executed in the Linux alignment pipe (FASTQ decompression → aligner → Summarizer → alignment file compression). Coverage was therefore measured by binning as described above, which can be implemented with minimal overhead.

A virus that is present in the reads with coverage too low to enable an assembly may have less practical value than an assembled genome. Also, genomes with lower identity to previously known sequences will tend to contain more novel biological information than genomes with high identity and will tend to have fewer alignments highly diverged segments. With these considerations in mind, the classifier was designed to give higher scores when coverage is high, read depth is high, and/or identity is low. This was accomplished as follows. Let H be the number of bins with at least 8 alignments to F, and L be the number of bins with from 1 to 7 alignments. Let S be the mean alignment percentage identity, and define the identity weight $w = (S/100)^{-3}$, which is designed to give higher weight to lower identities, noting that w is close to one when identity is close to 100% and increases rapidly at lower identities. The classification score for family F is calculated as $Z_F = max(w(4H + L)),100)$. By construction, Z_F has a maximum of 100 when coverage is consistently high across a pangenome, and is also high when identity is low and coverage is moderate, which may reflect high read depth but many false negative alignments due to low identity. Thus, Z_F is greater than zero when there is at least one alignment to F and assigns higher scores to SRA datasets which are more likely to support successful assembly of a virus belonging to F.

1.1.6 Sensitivity to novel viruses as a function of identity

We aimed to assess the sensitivity of our pipeline as a function of sequence identity by asking what fraction of novel viruses is detected at increasingly low identities compared to the reference sequences used for the search. Several variables other than identity affect sensitivity, including read length, whether reads are mate-paired, sequencing error rate, coverage bias, and presence of other similar viruses which may cause some variants to be unreported in the contigs. Coverage bias can render a virus with high average read depth undetectable, in particular if the query is RdRP-only and the RdRP gene has low coverage or is absent from the reads. Successful detection might be defined in different ways, depending on the goals of the search; e.g. a single local alignment of a reference to a read (maximising sensitivity, but not always useful in practice); a micro-assembled palmprint; a full assembly contig that contains a complete palmprint or otherwise classifiable fragment of a marker gene; or an assembly of a complete genome. We assessed alignment sensitivity of bowtie2 --very-sensitive-local and Serratus-optimised DIAMOND (Buchfink et al., 2015) as a function of identity by simulating typical examples in representative scenario: unpaired reads of length 100 with a base call error rate of 1%. We manually selected test-reference pairs of RefSeq complete *Ribovirus* genomes at RdRP aa identities 100%, 95% ... 20%, generating

simulated length-100 reads at uniformly-distributed random locations in the test genome with a mean coverage of 1000x. For bowtie2, the complete reference genome was used as a reference; for *DIAMOND* the reference was the translated amino acid sequence of the RdRP gene (400aa), which was identified by aligning to the *Wolf18* dataset. These choices model the coronavirus pan-genome used as a bowtie2 query and the *rdrp1* protein reference used as a DIAMOND query, respectively. Sensitivity was assessed as the fraction of reads aligned to the reference. With bowtie2, the number of unmapped reads reflects a combination of lack of alignment sensitivity and divergence in gene content as some regions of the genome may lack homology to the reference. With DIAMOND, the number of unmapped reads reflects a combination of lack of alignment sensitivity and the fraction of the genome which is not RdRP, which varies by genome length 1g. They show that the fraction of aligned reads by bowtie2 drops to around 2% to 4% at 90% RdRP aa identity, and maps no reads for most of the lower identity test-reference pairs. DIAMOND maps around 5% to 10% of reads down to 50% RdRP aa identity, then less than 1% at lower identities; around 30% to 35% is the lower limit of practical detection.

1.2 Defining viral pangenomes and the SRA search space

1.2.1 Nucleotide search pangenomes

To create a collection of viral pangenomes, a comprehensive set of complete and partial genomes representing the genetic diversity of each viral family, we used two approaches.

For *Coronaviridae*, we combined all RefSeq (n = 64) and GenBank (n = 37,451) records matching the NCBI Nucleotide (Coordinators and NCBI Resource Coordinators, 2012) server query "txid11118[Organism:exp]" (date accessed: June 1st 2020). Sequences <200 nt were excluded as well as sequences identified to contain non-CoV contaminants during preliminary testing (such as plasmid DNA or ribosomal RNA fragments). Remaining sequences were clustered at 99% identity with UCLUST (USEARCH:v11.0.667) (Edgar, 2010) and masked by Dustmasker (ncbi-blast:2.10.0) (--window 30 and --window 64) (Morgulis et al., 2006). The final query contained 10,101 CoV sequences (accessions in Supplementary Table 1a, masked coordinates in Supplementary Table 1b). SeqKit (v.0.15) was used for working with fasta files (Shen et al., 2016).

For all other vertebrate viral family pangenomes, RefSeq sequences (n = 2,849) were downloaded from the NCBI Nucleotide server with the query "Viruses[Organism] AND srcdb refseq[PROP] NOT wgs[PROP] NOT cellular organisms[ORGN] NOT AC 000001:AC 999999[PACC] AND ("vhost human"[Filter] AND "vhost vertebrates"[Filter])" (date accessed: May 17th 2020). Retroviruses (n = 80) were excluded as preliminary testing yielded excessive numbers of alignments to transcribed endogenous retroviruses. Each sequence was annotated with its taxonomic family according to its RefSeq record; those for which no family was assigned by RefSeq (n = 81) were designated as “unknown”.

The collection of these pangenomes was termed *cov3m*, and was the nucleotide sequence reference used for this study.

1.2.2 Amino acid viral RNA-dependent RNA polymerase search panproteome

For the translated-nucleotide search of viral RNA-dependent RNA polymerase (RdRP; hereinafter viral RdRP is implied) we combined sequences from several sources. 1) The '*wolf18*' collection is a curated snapshot (ca. 2018) of RdRP from GenBank ((Wolf et al., 2018) accessed: ftp://ftp.ncbi.nlm.nih.gov/pub/wolf/_suppl/rnavir18/RNAvirome.S2.afa) 2) The '*wolf20*' collection is RdRPs from assembled from marine metagenomes (accessed: ftp://ftp.ncbi.nlm.nih.gov/pub/wolf/_suppl/yangshan/gb_rdrp.afa) 3) All viral GenBank protein sequences were aligned with DIAMOND --ultra-sensitive against the combined '*wolf18*' and '*wolf20*' sequences (E-value <1e-6). These produced local alignments which contained truncated RdRP, so each RdRP-containing GenBank sequence was then re-aligned to the '*wolf18*' and *'wolf20'* collection to "trim" them to '*wolf*' RdRP boundaries. 4) The above algorithm was also applied to all viral GenBank nucleotide records to capture additional RdRP not annotated as such by GenBank . A region of HCV capsid protein shares similarity to HCV RdRP, sequences annotated as HCV-capsid were therefore removed. Eight novel coronavirus RdRP sequences identified in a pilot experiment were added manually. The combined RdRP sequences from the above collections were clustered (UCLUST) at 90% amino acid identity and the resulting representative sequences (centroids, N = 14 653) used as the *rdrp1* search query.

In addition, we added Deltavirus antigen proteins from NC 001653, M21012, X60193, L22063, AF018077, AJ584848, AJ584847, AJ584844, AJ584849, MT649207, MT649208, MT649206, NC 040845, NC 040729, MN031240, MN031239, MK962760, MK962759, and eight additional homologs we identified in a pilot experiment.

1.2.3 SRA search space and queries

To run Serratus, a target list of SRA run accessions is required. We defined eleven (not-mutually exclusive) queries as our search space which were named human, mouse, mammal, vertebrate, invertebrate, eukaryotes, prokaryotes/others, bat (including genomic sequences), virome, metagenome and mammalian genome (Supplementary Table 1c). Our search was restricted to Illumina sequencing technologies and to RNA-seq, meta-genomic, and meta-transcriptome library types for these organisms (except for mammalian genome query which was genome or exome). Prior to each Serratus deployment, target lists were depleted of accessions already analysed. Reprocessing of a failed accession was attempted at least twice. In total, we aligned 3,837,755/4,059,695 (94.5%) of the runs in our nucleotide-pangenome search (ca. May 2020) and 5,686,715/5,780,800 (98.37%) of the runs in our translated-nucleotide RdRP search (ca. January 2021).

1.3 User interfaces for the Serratus databases

We implemented an on-going, multi-tiered release policy for code and data generated by this study, as follows. All code, electronic notebooks and raw data is immediately available at https://github.com/ababaian/serratus and on the s3://serratus -public/ bucket, respectively. Upon completion of a project milestone, a structured data release is issued containing raw data into our viral data warehouse s3://lovelywater/. For example, the .bam nucleotide alignment files from 3.84 million SRA runs are stored in s3://lovelywater/bam/X.bam; the protein .summary files are in

s3://lovelywater/psummary/X.psummary, where X is a SRA run accession. These FAIR and structured releases enable downstream and third-party programmatic access to the data.

Summary files for every searched SRA dataset are parsed into a publicly accessible AWS Relational Database (RDS) instance which can be queried remotely via any PostgreSQL client. This enables users and programs to perform complex operations such as retrieving summaries and meta-data for all SRA runs matching a given reference sequence with above a given classifier score threshold. For example, one can query for all records containing at least 20 aligned reads to Hepatitis Delta Virus (NC 001653.2) and the associated host taxonomy for the corresponding SRA datasets:

```
SELECT sequence_accession, run_id, tax_id, n_reads
FROM nsequence
JOIN srarun ON (nsequence.run_id = srarun.run) WHERE n_reads >= 20
```

For users unfamiliar with SQL, we developed Tantalus (https://github.com/serratus-bio/tantalus, an R programming-language package which directly interfaces the Serratus PostgreSQL database to retrieve summary information as data-frames. Tantalus also offers functions to explore and visualise the data.

Finally, the Serratus data can be explored via a graphical web interface by accession, virus, or viral family at https://serratus.io/explorer. Under the hood, we developed a REST API to query the database from the website. The website uses React+D3.js to serve graphical reports with an overview of viral families found in each SRA accession matching a user query.

All four data access interfaces are under ongoing development, receiving community feedback via their respective GitHub issue trackers to facilitate the translation of this data collection into an effective viral discovery resource. Documentation for data access methods is available at https://serratus.io/access.

1.3.1 Geocoding BioSamples

To generate the map in Figure 1c, we parsed and extracted geographic information from all 16 million BioSample XML submissions. Geographic information is either in the form of coordinates (latitude/longitude) or freeform text (e.g. “France”, “Great Lakes”). For each BioSample, coordinate extraction was attempted using regular expressions. If that failed, text extraction was attempted using a manually curated list of keywords that capture BioSample attribute names likely to contain geographic information. If that failed, then we were unable to extract geographic information for that BioSample. Geocoding the text to coordinates was done using Amazon Location Service on a reduced set of distinct filtered text values (52,028 distinct values from 2,760,241 BioSamples with potential geographic text). BioSamples with geocoded coordinates were combined with BioSamples with submitted coordinate information to form a set of 5,325,523 geospatial BioSamples. This is then cross-referenced with our subset of SRA accessions with an RdRP match to generate the figure.

All intermediate and resulting data from this step is stored on the SQL database described in 1.3. Development work is public at https://github.com/serratus-bio/biosample-sql.

1.4 Viral alignment, assembly and annotation

Upon identification of CoV reads in a run from alignment, we assembled 52,772 runs containing at ≥ 10 reads which aligned to our CoV pan-genome or ≥ 2 reads with CoV-positive *k*-mers. (Katz et al., 2021). 11,120 of the resulting assemblies contained identifiable CoV contigs, of which only 4,179 (37.58%) contained full-length CoV RdRP (Supplementary Table 1d). The discrepancy between alignment-positive, assembly-positive and RdRP-positive libraries arises due random sampling of viral reads and assembly fragmentation. In this respect, alignment or k-mer based methods are more sensitive than assembly in detecting for the presence of low-abundance viruses (genome coverage <1) with high identity to a reference sequence. Scoring libraries for genome-coverage and depth is a good predictor of ultimate assembly success (Extended Figure 3) thus, it can be used to efficiently prioritise computationally expensive assembly in the future, as has been previously demonstrated for large-scale SRA alignment-analyses (Levi et al., 2018).

1.4.1 DIAMOND optimisation and output

To optimise DIAMOND (Buchfink et al., 2021) for small (< 10 Mb) databases such as the RdRP search database, we built a probabilistic hash set which stores 8-bit hash values for the database seeds, using SIMD instructions for fast probing. This index is loaded as a memory mapped file to be shared among processes and allows us to filter the query reads for seeds contained in the database, thus omitting the full construction of the query seed table. We also eliminated the overhead of building seed distribution histograms that is normally required to allocate memory and construct the query table in a single pass over the data using a deque-like data structure. In addition, query reads were not masked for simple repeats, as the search database is already masked. These features are available starting from DIAMOND v2.0.8 with the command line flags --target-indexed --masking 0. In a benchmark of 4 sets of 1 million reads from a bat metagenome (ERR2756788), the implemented optimisation produced a speed-up of x1.47 and reduced memory use by 64%, compared to the public unmodified DIAMOND v2.0.6, using our optimised set of parameters in both cases (see 1.1.1). Together, the optimised parameters and implementation reduced DIAMOND runtime against RdRP-search from 197.96s (s.d = 0.18s), to 21.29s (s.d=0.23s) per million reads, a speed-up of a factor of 9.3. This effectively reduced the computational cost of translated-nucleotide search for Serratus from $0.03, to $0.0042 per library.

DIAMOND output files (we label .pro) were specified with the command -f 6 qseqid qstart qend qlen qstrand sseqid sstart send slen pident evalue cigar qseq_translated full_qseq full_qseq_mate.

1.4.2 coronaSPAdes

RNA viral genome assembly faces several distinct challenges stemming from technical and biological bias in sequencing data. During library preparation, reverse transcription introduces 50 end coverage bias, and GC-content skew and secondary structures lead to unequal PCR amplification (Hunt et al., 2015). Technical bias is confounded by biological complexity such as intra-sample sequence variation due to transcript isoforms and/or to presence of multiple strains.

To address the assembly challenges specific to RNA viruses, we developed coronaSPAdes (v3.15.3), described in detail in a companion manuscript (Meleshko et al., 2021). In brief, rnaviralSPAdes and the more specialized variant, coronaSPAdes, combines algorithms and methods from several previous approaches based on metaSPAdes(Nurk et al., 2017), rnaSPAdes (Bushmanova et al., 2019) and metaviralSPAdes (Antipov et al., 2020) with a HMMPathExtension step. coronaSPAdes constructs an assembly graph from a RNA-sequencing dataset (transcriptome, meta-transcriptome, and meta-virome are supported), removing expected sequencing artifacts such as low-complexity (poly-A / poly-T) tips, edges, single-strand chimeric loops or double-strand hairpins (Bushmanova et al., 2019) and subspecies-bases variation (Antipov et al., 2020).

To deal with possible misassemblies and high-covered sequencing artefacts, a secondary HMMPathExtension step is performed to leverage orthogonal information about the expected viral genome. Protein domains are identified on all assembly graphs using a set of viral hidden Markov models (HMMs), and similar to biosyntheticSPAdes (Meleshko et al., 2019), HMMPathExtension attempts to find paths on the assembly graph which pass through significant HMM matches in order.

coronaSPAdes is bundled with the Pfam SARS-CoV-2 set of HMMs, although these may be substituted by the user. This latter feature of coronaSPAdes was utilized for HDV assembly, where the HMM model of HDAg, the Hepatitis Delta Antigen, was used instead of Pfam SARS-CoV-2 set. Note that despite the name, the HMMs from this set are quite general, modeling domains found in all coronavirus genera in addition to RdRP, which is found in many RNA virus families. Hits from these HMMs cover most bases in most known coronavirus genomes, enabling the recovery of strain mixtures and splice variants.

1.4.3 Micro-assembly of RdRP-aligned reads

Reads aligned by DIAMOND in the translated-nucleotide RdRP search are stored in the .pro alignment file. All sets of mapped reads (3,379,127 runs) were extracted, and each non-empty set was assembled with rnaviralSPAdes (v3.15.3) using default parameters. This process is referred to as "*micro-assembly"* since a collection of DIAMOND hits is orders of magnitude smaller than the original SRA accession (40±534 KB compressed size, ranging from a single read up to 53 MB). Then bowtie2 (Langmead and Salzberg, 2012) (default parameters) was used to align the DIAMOND read hits of an accession back to the micro-assembled contigs of that accession. Palmscan (v1.0.0 -rdrp -hicon) (Babaian and Edgar, 2021) was run on microassembled contigs, resulting in high-confidence palmprints for 337,344 contigs. Finally mosdepth (v0.3.1) (Pedersen and Quinlan, 2018) was used to calculate a coverage pileup for each palmprint hit region within micro-assembled contigs.

1.4.4 Classification of assembled RdRP sequences

Our methods for RdRP classification are described and validated in a companion paper [18]. Briefly, we defined a barcode sequence, the polymerase palmprint (PP), as a ~100 aa segment of the RdRP palm sub-domain delineated by well-conserved catalytic motifs. We implemented an algorithm, palmscan, to identify palmprint sequences and discriminate RdRPs from reverse transcriptases. The combined set of RdRP palmprints

from public databases and our assemblies were classified by clustering into operational taxonomic units (OTUs) at 90%, 75% and 40% identity, giving species-like, genus-like and family-like clusters (sOTUs, gOTUs and fOTUs), respectively. Tentative taxonomy of novel OTUs was assigned by aligning to palmprints of named viruses and taking a consensus of the top hits above the identity threshold for each rank.

1.4.5 Quality control of assembled RdRP sequences

Our goal was to identity novel viral RdRP sequences and novel sOTUs in SRA libraries. From this perspective, we considered the following to be erroneous to varying degrees: sequences which are (a) not polymerases, (b) not viral, (c) with differences due to experimental artefacts, or (d) with sufficient differences to cause a spurious inference of a novel sOTU. We categorised potential sources of such errors and implemented quality control procedures to identify and mitigate them, as follows.

Point errors are single-letter substitution and indel errors which may be caused by PCR or sequencing *per se*. Random point errors are not reproduced in multiple non-PCR duplicate reads and are unlikely to assemble because such errors almost always induce identifiable structures in the assembly graph (tips and bubbles) which are pruned during graph simplification. In rare cases, a contig may contain a read with random point errors. Such contigs will have low coverage ~1, and we therefore recorded coverage as a QC metric and assessed whether low-coverage assemblies were anomalous compared to high-coverage assemblies by measures such as the frequencies with which they are reproduced in multiple libraries compared to exactly one library, finding no noticeable difference when coverage is low.

Chimeras of polymerases from different species could arise from PCR amplification or assembly. We used the UCHIME2 (usearch v8.0.1623) algorithm (Edgar) to screen assembled palmprint sequences, finding no high-scoring putative chimeras. Mosaic sequences formed by joining a polymerase to unrelated sequence would either have an intact palmprint, in which case the mosaic would be irrelevant to our analysis, or would be rejected by Palmscan due to the lack of delimiting motifs.

Reverse transcriptases (RTs) are homologous to RdRP. Retroviral insertions into host genomes induce ubiquitous sequence similarity between host genomes and viral RdRP. Palmscan was designed to discriminate RdRP from sequences of RT origin. Testing on a large decoy set of non-RdRP sequences with recognisable sequence similarity showed that the Palmscan false discovery rate for RdRP identification is 0.001. We estimated the probability of false positives matches in unrelated sequence by generating sufficient random nucleotide and amino acid sequences to show that the expected number of false positive palmprint identifications is zero in a dataset of comparable size to our assemblies. We also regard the low observed frequency of palmprints in DNA WGS data (in 2.6 Pbp or 25.8% of reads, accounted for 100 known palmprints and 95 novel palmprints or 0.13% of the total identified) as a *de facto* confirmation of the low probability false positives in unrelated sequence.

Endogenous viral elements (EVEs, i.e. insertions of viral sequence into host genomes which are potentially degraded and non-functional) cannot be distinguished from viral genomes on the basis of the palmprint sequence alone. To assess the frequency of EVEs in our data, we re-assembled 890 randomly-chosen libraries yielding one or more palmprints using all reads, extracted the 23 530 resulting contigs with a positive

palmprint hit by palmscan, and classified them using Virsorter2 (v2.1) (Guo et al., 2021). Of these contigs, 11,914 were classified as viral, confirming the palmscan identification; 49 as *Viridiplantae* (green plants); 46 as *Metazoa*; 25 as Fungi and the remainder were unclassified. Thus, 120/12034 = 1% of the classified contigs were predicted as non-viral, suggesting that the frequency of EVEs in the reported palmprints is ~1%.

1.4.6 Annotation of CoV assemblies

Accurate annotation of CoV genomes is challenging due to ribosomal frameshifts and polyproteins which are cleaved into maturation proteins (Thiel et al., 2003), and thus previously-annotated viral genomes offer a guide to accurate gene-calls and protein functional predictions. However, while many of the viral genomes we were likely to recover would be similar to previously-annotated genomes in Refseq or GenBank, we anticipated that many of the genomes would be taxonomically distant from any available reference. To address these constraints, we developed an annotation pipeline called DARTH (version maul) which leverages both reference-based and *ab initio* annotation approaches.

In brief, DARTH consists of the following phases: standardise the ordering and orientation of assembly contigs using conserved domain alignments, perform reference-based annotation of the contigs, annotate RNA secondary structure, *ab initio* gene-calling, generate files for aiding assembly and annotation diagnostics, and generate a master annotation file. It is important to put the contigs in the “expected” orientation and ordering to facilitate comparative analysis of synteny and as a requirement for genome deposition. To perform this standardisation, DARTH generates the six-frame translation of the contigs using the transeq (v EMBOSS:6.6.0.0) (Rice et al., 2000) and uses HMMER*3* (v3.3.2) (Eddy, 2011) to search the translations for Pfam domain models specific to CoV [64]. DARTH compares the Pfam accessions from the HMMER alignment to the NCBI SARS-CoV-2 reference genome (NCBI Nucleotide accession NC_045512.2) to determine the correct ordering and orientation, and produces an updated assembly FASTA file. DARTH performs reference-based annotation using VADR (v1.1) (Schäffer et al., 2020), which provides a set of genome models for all CoV RefSeq genomes (Nawrocki). VADR provides annotations of gene coordinates, polyprotein cleavage sites, and functional annotation of all proteins. DARTH supplements the VADR annotation by using *Infernal* (Nawrocki and Eddy, 2013) to scan the contigs against the SARS-CoV-2 Rfam release which provides updated models of CoV 50 and 30 untranslated regions (UTRs) along with stem-loop structures associated with programmed ribosomal frame-shifts. While VADR provides reference-based gene-calling, DARTH also provides *ab initio* gene-calling by using FragGeneScan (v1.31) (Rho et al., 2010), a frameshift-aware gene caller. DARTH also generates auxiliary files which are useful for assembly quality and annotation diagnostics, such as indexed BAM files created with SAMtools (v1.7) (Li et al., 2009) representing self-alignment of the trimmed reads to the canonicalized assembly using bowtie2, and variant-calls using bcftools from SAMtools. DARTH generates these files so that the can be easily loaded into a genome browser such as JBrowse (Buels et al., 2016) or IGV (Robinson et al., 2017). As the final step DARTH generates a single Generic Feature Format (GFF) 3.0 file (Reeves et al., 2008) containing combined set of annotation information described above, ready for use in a genome browser, or for submitting the annotation and sequence to a genome repository.

1.4.7 Phage assembly

Each metagenomic dataset was individually *de novo*-assembled using MEGAHIT (v 1.2.9) (Li et al., 2016), and filtered to remove contigs smaller than 1 kbp in size. ORFs were then predicted on all contigs using Prodigal (v2.6.3) (Hyatt et al., 2012) with the following parameters: -m -p meta. Predicted ORFs were initially annotated using USEARCH to search all predicted ORFs against UniProt (The UniProt Consortium and The UniProt Consortium, 2017), UniRef90 and KEGG (Altman et al., 2013). Sequencing coverage of each contig was calculated by mapping raw reads back to assemblies using bowtie2 (Langmead and Salzberg, 2012). Terminase sequences from Al-Shayeb *et al. (Al-Shayeb et al., 2020)* were clustered at 90% amino acid identity to reduce redundancy using CD-HIT (v4.8.1) (Li et al., 2012), and HMM models were built with hmmbuild (from the HMMER3 suite (Eddy, 2011)) from the resulting set. Terminases in the assemblies from Serratus were identified using hmmsearch, retaining representatives from contigs greater than 140 kbp in size. Some examples of prophage and large phages that did not co-cluster with the sequences from Al-Shayeb *et al.*, were also recovered because they were also present in a sample that contained the expected large phages. The terminases were aligned using MAFFT (v.7.407) and filtered by TrimAL (v1.14) to remove columns comprised of more than 50% gaps, or 90% gaps, or using the automatic gappyout setting to retain the most conserved residues. Maximum likelihood trees were built from the resulting alignments using IQTREE (v.1.6.6) (Nguyen et al., 2015).

1.4.8 Deploying the assembly and annotation workflow

The Serratus search for known or closely related viruses identified 37,131 libraries (14,304 by nucleotide and 23,898 by amino acid) as potentially positive for CoV (score ≥20 and ≥10 reads). To supplement this search we also employed a recently developed index of the SRA called STAT with which identified an additional 18,584 SRA datasets not in the defined SRA search space. The STAT BigQuery (accessed June 24th 2020) was: WHERE tax id=11118 AND total count >1.

We used AWS Batch to launch thousands of assemblies of NCBI accessions simultaneously. The workflow consists of four standard parts: a job queue, a job definition, a compute environment, and finally, the jobs themselves. A CloudFormation template (https://gitlab.pasteur.fr/rchikhi_pasteur/serratus-batch-assembly/-/blob/10934001/template/template.yaml) was created for building all parts of the cloud infrastructure from the command line. The job definition specifies a Docker image, and asks for 8 virtual CPUs (vCPUs, corresponding to threads) and 60 GB of memory per job, corresponding to a reasonable allocation for coronaSPAdes. The compute environment is the most involved component. We set it to run jobs on cost-effective Spot instances (optimal setting) with an additional cost-optimization strategy (SPOT_CAPACITY_OPTIMIZED setting), and allowing up to 40,000 vCPUs total. In addition, the compute environment specifies a launch template which, on each instance, i) automatically mounts an exclusive 1 TB EBS volume, allowing sufficient disk space for several concurrent assemblies, and ii) downloads the 5.4 GB CheckV (v0.6.0) (Nayfach et al., 2021b) database, to avoid bloating the Docker image.

The peak AWS usage of our Batch infrastructure was ~28,000 vCPUs, performing ~3,500 assemblies simultaneously. A total of 46,861 accessions out of 55,715 were

assembled in a single day. They were then analysed by two methods to detect putative CoV contigs. The first method is CheckV, followed selecting contigs associated to known CoV genomes. The second method is a custom script (https://gitlab.pasteur.fr/rchikhi_pasteur/serratus-batch-assembly/-/blob/10934001/stats/bgc_parse_and_extract.py) that parses coronaSPAdes BGC candidates and keeps contigs containing CoV domain(s). For each accession, we kept the set of contigs obtained by the first method (CheckV) if it is non-empty, and otherwise we kept the set of contigs from the second method (BGC).

A majority (76%) of the assemblies were discarded for one of the following reasons: i) no CoV contigs were found by either filtering method, ii) reads were too short to be assembled, iii) Batch job or SRA download failed, or iv) coronaSPAdes ran out of memory. A total of 11,120 assemblies were considered for further analysis.

The average cost of assembly was between $0.30-$0.40 per library, varying depending on library-type (RNA-seq versus metagenomic). This places an estimate of 46-95 fold higher cost for assembly alone compared to a cost of $0.0042 or $0.0065 for an alignment based search.

1.5 Taxonomic and phylogenetic analyses

1.5.1 Taxonomy prediction for coronavirus genomes

We developed a module, SerraTax, to predict taxonomy for CoV genomes and assemblies (https://github.com/ababaian/serratus/tree1f92d7e4/containers/serratax). SerraTax was designed with the following requirements in mind: provide taxonomy predictions for fragmented and partial assemblies in addition to complete genomes; report best-estimate predictions balancing over-classification and under-classification (too many and too few ranks, respectively); and assign an NCBI Taxonomy Database (Schoch et al., 2020) identifier (TaxID).

Assigning a best-fit TaxID was not supported by any previously published taxonomy prediction software to the best of our knowledge; this requires assignment to intermediate ranks such as sub-genus and ranks below species (commonly called strains, but these ranks are not named in the Taxonomy database), and to unclassified taxa, e.g. TaxID 2724161, unclassified Buldecovirus, in cases where the genome is predicted to fall inside a named clade but outside all named taxa within that clade.

SerraTax uses a reference database containing domain sequences with TaxIDs. This database was constructed as follows. Records annotated as CoV were downloaded from UniProt (The UniProt Consortium and The UniProt Consortium, 2017), and chain sequences were extracted. Each chain name, e.g. Helicase, was considered to be a separate domain. Chains were aligned to all complete coronavirus genomes in GenBank using UBLAST (usearch: v11.0.667) to expand the repertoire of domain sequences. The reference sequences were clustered using UCLUST (Edgar, 2010) at 97% sequence identity to reduce redundancy.

For a given query genome, open reading frames (ORFs) are extracted using the getorf (EMBOSS:6.6.0) software (Rice et al., 2000). ORFs are aligned to the domain references and the top 16 reference sequences for each domain are combined with the best-matching query ORF. For each domain, a multiple alignment of the top 16 matches plus

query ORF is constructed on the fly by MUSCLE (v3.8.31(Edgar, 2004)) and a neighbour-joining tree is inferred from the alignment, also using MUSCLE. Finally, a consensus prediction is derived from the placement of the ORF in the domain trees. Thus, the presence of a single domain in the assembly suffices to enable a prediction; if more domains are present they are combined into a consensus.

1.5.2 Taxonomic assignment by phylogenetic placement

To generate an alternate taxonomic annotation of an assembled genome, we created a pipeline based on phylogenetic placement, SerraPlace.

To perform phylogenetic placement, a reference phylogenetic tree is required. To this end, we collected 823 reference amino acid RdRP sequences, spanning all *Coronaviridae*. To this set we added an outgroup RdRP sequence from the Torovirus family (NC 007447). We clustered the sequences to 99% identity using USEARCH ((Edgar, 2010) UCLUST algorithm, v11.0.667), resulting in 546 centroid sequences. Subsequently we performed multiple sequence alignment on the clustered sequences using MUSCLE. We then performed maximum likelihood tree inference using RAxML-NG ((Kozlov et al., 2019) '*PROTGTR+FO+G4*', v0.9.0), resulting in our reference tree.

To apply SerraPlace to a given genome, we first use HMMER ((Eddy, 2011), v3.3) to generate a reference HMM, based on the reference alignment. We then split each contig into ORFs using esl-translate, and use hmmsearch (p-value cutoff 0.01) and seqtk (commit 7c04ce7) to identify those query ORFs that align with sufficient quality to the previously generated reference HMM. All ORFs that pass this test are considered valid input sequences for phylogenetic placement. This produces a set of likely placement locations on the tree, with an associated likelihood weight. We then use Gappa (v0.6.1,(Czech et al., 2020)) to assign taxonomic information to each query, using the taxonomic information for the reference sequences. Gappa assigns taxonomy by first labeling the interior nodes of the reference tree by a consensus of the taxonomic labels of all descendant leaves of that node. If 66% of leaves share the same taxonomic label up to some level, then the internal node is assigned that label. Then, the likelihood weight associated with each sequence is assigned to the labels of internal nodes of the reference tree, according to where the query was placed.

From this result, we select that taxonomic label that accumulated the highest total likelihood weight as the taxonomic label of a sequence. Note that multiple ORFs of the same genome may result in a taxonomic label, in which case, we select the longest sequence as the source of the taxonomic assignment of the genome.

1.5.3 Phylogenetic inference

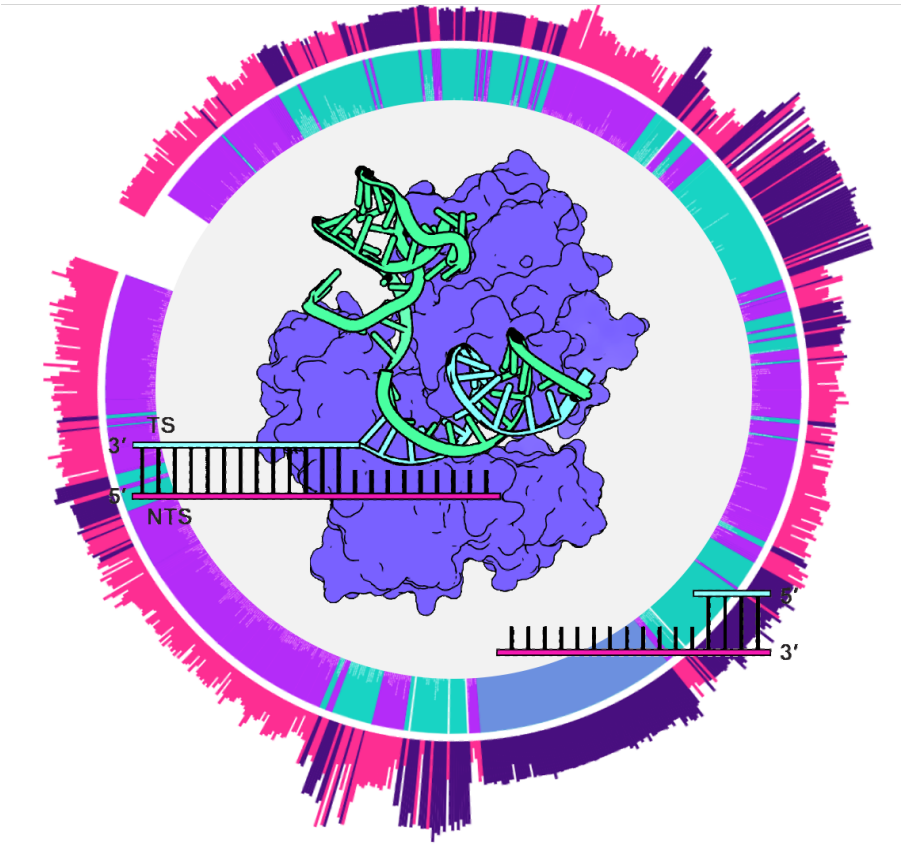
We performed phylogenetic inferences using a custom snakemake (v6.6.0) pipeline (available at https://github.com/lczech/nidhoggr), using ParGenes (v1.1.2)(Morel et al., 2019). ParGenes is a tree search orchestrator, combining ModelTestNG (v0.1.3) (Darriba et al., 2020)and RAxML-NG, and enabling higher levels of parallelisation for a given tree search.

To infer the maximum likelihood phylogenetic trees, we performed a tree search comprising 100 distinct starting trees (50 random, 50 parsimony), as well as 1000 bootstrap searches. We used ModelTest-NG to automatically select the best

evolutionary model for the given data. The pipeline also automatically produces versions of the best maximum likelihood tree annotated with Felsenstein's Bootstrap (Felsenstein, 1985) support values, and Transfer Bootstrap Expectation values (Lemoine et al., 2018).

5 Chapter 5: Diverse virus-encoded CRISPR-Cas systems include streamlined genome editors

Basem Al-Shayeb, Petr Skopintsev, Katarzyna Soczek, Elizabeth Stahl, Zheng Li, Evan Groover, Dylan Smock, Amy R Eggers, Patrick Pausch, Brady F Cress, Carolyn J Huang, Brian Staskawicz, David F Savage, Steven E Jacobsen, Jillian F Banfield, Jennifer A Doudna



5.1 **Abstract**

CRISPR-Cas systems protect microbes from viral infection using an adaptive RNA-guided mechanism that recognizes and cuts foreign genetic material. Using genome-resolved metagenomics, we find that these systems are also encoded in diverse classes of bacteriophages, where they occur as divergent and hypercompact anti-viral systems. More than 6000 bacteriophages, <1% of all phages examined, encode CRISPR systems spanning both Class 1 and 2, including all six of the known CRISPR-Cas types. Many of these systems target competing mobile elements predicted to infect the same bacterial hosts, and RNA-targeting systems often lack crucial components that would, in their bacterial counterparts, result in acute abortive infection, suggesting alternate targeting outcomes or complementation by host factors. We describe multiple new Cas9-like protein families and 44 families related to type V CRISPR-Cas systems that occur on phage genomes and provide the first biochemical and structural insights into the Casλ family. Casλ recognizes double-stranded DNA using a unique structured crRNA reminiscent of engineered sgRNA and generated by 3' end cleavage, unlike any previously described single-RNA CRISPR-Cas system. The Casλ-RNA-DNA structure determined by cryo-electron microscopy reveals a compact architecture capable of robust RNA-guided DNA cutting. Despite its significant divergence in sequence, domain organization and crRNA production, Casλ possesses a bi-lobed structure reminiscent of Cas9 and Cas12, exemplifying the convergent evolution of RNA-guided enzymes. Remarkably, Casλ induces efficient genome editing of endogenous genes in mammalian, *Arabidopsis*, and hexaploid wheat cells on par with, or in some cases, exceeding Cas12a-mediated genome editing.

N.B. All main figures for this manuscript can be found below in their dedicated section. All supplementary files (including figures and tables) can be found online with the published manuscript.

5.2 **Introduction**

CRISPR-Cas systems confer resistance in prokaryotes against invading extra-chromosomal elements including viruses and plasmids (Barrangou et al., 2007). To generate immunological memory, microbes capture fragments of foreign genetic elements and incorporate them into their genomic CRISPR array using the Cas1-Cas2 integrase. Subsequent transcription of the array creates CRISPR RNAs (crRNAs) that bind to and direct CRISPR-associated (Cas) nucleases to target complementary nucleic acids. These systems comprise two classes, each with three different types, defined by the architectures of their nuclease effector modules involved in crRNA processing and DNA or RNA interference.

Reports of CRISPR-Cas loci encoded in bacteriophages (phage) that infect *Vibrio cholera (O'Hara et al., 2017; Seed et al., 2013)* or in huge phage genomes reconstructed from microbial community DNA sequences (Al-Shayeb et al., 2020) hinted at a wider distribution of phage-encoded CRISPR systems that might play as-yet-unknown roles in prokaryotic biology. These observations motivated us to perform a comprehensive study of the abundance, distribution, and diversity of CRISPR-Cas systems encoded throughout the virosphere and to begin to explore the biochemical activity of novel systems.

Here we report the widespread occurrence of diverse as well as compact CRISPR-Cas systems encoded in phage genomes identified by metagenomic analysis of microbial samples isolated from soil, aquatic, human and animal microbiomes, demonstrating an unexpected biological reservoir of anti-viral machinery within infectious agents. Phage-encoded CRISPR-Cas systems include members of all six CRISPR types (types I-VI) as defined by bacterially-encoded examples. We found evidence for new or alternative modes of nucleic acid interference involving phage-encoded type I, III, IV, and VI systems. In addition, the phage and phage-like sequences result in a severalfold expansion of CRISPR-Cas9 and -Cas12 enzymes belonging to the type II and type V families that are widely deployed for genome editing applications. Casλ, the most divergent of the phage-encoded type V enzymes identified in this study, was found to have robust biochemical activity as an RNA-guided double-stranded DNA cutter. Its cryo-EM-determined molecular structure explains its use of a natural single-guide RNA for DNA binding, and cell-based experiments demonstrated robust endogenous genome editing activity in plant and mammalian cells. The compact architecture of Casλ and other phage-encoded CRISPR-Cas proteins holds significant promise for vector-based and direct delivery into cells for wide-ranging biotechnological applications.

5.3 **Results and Discussion**

A wide diversity of phages across many bacterial phyla encode divergent CRISPR-Cas systems

Using genome-resolved metagenomics, we analyzed over 660 Gigabasepairs of assembled genomic DNA from both environmental and animal-associated microbiomes to reveal a surprising diversity of over 6000 CRISPR-encoding phages (Fig. 1A). Our analysis of publicly available phage genomes revealed that CRISPR-Cas systems occur in only 0.3% of Genbank-recorded phages, 0.8% of complete RefSeq-recorded phages, and 0.4% of IMG-VR-recorded phages, making them exceptionally rare compared to their abundance in prokaryotic genomes where they occur in 40% of bacteria and 85% of archaea. The majority of CRISPR-containing phages formed distinct clusters relative to reference genomes based on their protein repertoire. At least two phages harboring CRISPR arrays were alternatively coded such that the TAG stop codon was recoded to glutamine. Although circularized CRISPR-encoding phages included huge phages such as a >620 kbp megaphage (Fig. 1B), most had a genome size close to the average of 52 kbp. Rather than being constrained to a specific bacterial phylum, CRISPR-encoding phages are predicted to predate most major bacterial phyla including Firmicutes, Proteobacteria, Bacteroidetes, and Actinobacteria (Fig. 1C). Notably, however, relatively few phages encode complete CRISPR-Cas systems. Fewer than 10% of CRISPR-encoding phages were found to contain machinery for the acquisition of new spacer sequences into their CRISPR arrays, consistent with observations in huge phages (Al-Shayeb et al., 2020). Many phages encode CRISPR arrays, but few include Cas effectors encoded nearby (Fig. 1C). In such situations, phages may produce their own guide RNAs but hijack the Cas effectors provided by their hosts. Consistent with this possibility, >50 phages encode only the Cas1-Cas2 integrase used for the acquisition of new spacers, but no other Cas enzymes. In some cases, phage-encoded Cas1 contained a fusion to another protein such as reverse transcriptase, suggesting the possibility of the acquisition of RNA protospacers into the phage array.

Phage-encoded RNA-targeting CRISPR-Cas systems are rare

Out of the thousands of phage-encoded CRISPR-Cas loci identified in this study, only 27 represent known RNA-targeting systems. Some of these are type III systems associated with CRISPR arrays targeting vital or highly abundant RNA transcripts of other mobile elements (Fig. 2). In typical Type III systems, the Cas10 protein converts ATP into a cyclic oligoadenylate (cOA) product, which allosterically activates an auxiliary Csm6 ribonuclease (Niewoehner et al. 2017). The activated Csm6 amplifies the immune response by degrading RNA transcripts indiscriminately, thereby destroying the invasive transcriptome or inducing host cell dormancy or death, aborting the phage infectious cycle (Jiang et al., 2016; Kazlauskiene et al., 2017; Koonin and Zhang, 2017; Niewoehner et al., 2017). Interestingly, in huge phage-encoded type III systems, the Cas10 subunit contains multiple mutations consistent with an inability to produce cOA (Fig. S1) and Csm6 or a related CARF-domain ribonuclease is absent. Notably, the key residues for DNA cleavage in the Cas10 HD domain, and for RNA cleavage in Cas7 remain intact (Fig S1, S2). Unless the cOA production and Csm6 RNase functionalities are complemented by orthogonal type III systems from the host genome, this suggests that the type III phage systems may be capable of targeting and cis-cleavage of key RNA transcripts and genomic DNA of competing mobile elements to interfere with their infectious cycle, but circumventing the abortive infection mechanism activated by cOA signaling and subsequent trans-cleavage of transcripts in

the host cell, which may deplete the CRISPR phage population itself and inadvertently provide herd immunity to uninfected bacteria.

In addition to type III systems, we found the first examples of phage-encoded type VI (Cas13) ribonucleases, most of which belong to the Cas13b and the relatively small Cas13d superfamilies. Analogously to the findings above with type III systems in abortive infection, the lack of signature csx27 and csx28 proteins, which are transmembrane factors that enhance abortive infection mechanisms (VanderWal et al., 2021), may indicate the absence of an abortive infection pathway unless supplemented by the host.

Miniature single-effector CRISPR-Cas systems are enriched in phage genomes

Class 2 CRISPR-Cas systems, including types II, V, and VI, generally employ single subunit RNA-guided, nucleic acid-targeting interference enzymes. In addition to new Cas9 (a, b, c) and Cas12 (a, b, c, f, i) enzyme variants, we identified miniature CRISPR-associated nucleases in phages harboring both HNH and RuvC catalytic domains characteristic of Cas9 and Cas12. These miniature nucleases constitute phylogenetically distinct clades denoted as types II-X, -Y, and -Z (Fig. 1C). These systems lack the Cas1, Cas2, or Csn2 sequence acquisition machinery (Fig. 1C) and have distinct domain organizations compared to previously studied Cas9 orthologs with significant deletions across the proteins in comparison.

Furthermore, we observed that bacteriophage genomes harbor an unusual enrichment of hypercompact type V effectors (Fig. 1B, D), including hundreds of variants comprising 44 protein families that are evolutionarily distant from previously reported and experimentally validated miniature type V CRISPR-Cas nucleases including Cas12f and CasΦ (Fig. 1E). Evolutionary analysis suggests that distinct type V nuclease subtypes may have evolved multiple times from separate transposon-encoded TnpB families, which have recently been shown to be RNA-guided nucleases themselves (Karvelis et al., 2021), and we observe that TnpB is also widely encoded on phages.

CRISPR arrays associated with the Type V families contained spacer sequences targeting competing double-stranded DNA (dsDNA)-based extrachromosomal elements that are predicted to infect the same host (Fig. 2). We found that in multiple related Biggiephages, miniature type V families including Casμ and CasΦ co-occurred with a Type I system that we term Type I-X, bearing similarities to Type I-C CRISPR systems but featuring a distinct helicase in place of the processive nuclease Cas3. Biggiephage genomes were recovered over a four-year timespan, and remained identical save for their CRISPR arrays, which were nevertheless remarkably similar over time (Fig. S3). While we were unable to validate DNA cleavage by this system, it is possible that double-stranded DNA binding silences the expression of target genes (Fig. S4). In some cases, the arrays of the type I-X system target the same circular extrachromosomal element, albeit with distinct spacers, as the array associated with co-occurring type V systems. One such cryptic element harbored restriction enzymes and retron-based anti-phage defense systems that could limit Biggiephage infectivity, underscoring the dynamic nature of the evolutionary arms race between mobile elements in competition for host resources.

We also found the first type IV systems encoded in lytic phage genomes. Type IV systems are predominantly found on plasmids, where their mechanisms of action are poorly understood and they sometimes lack a CRISPR array (Pinilla-Redondo et al.,

2020). We report a Type IV subtype that lacks the DinG hallmark gene and encodes in its place a CysH-like protein bearing limited similarity to non-CRISPR associated CysH phosphoadenosine 5'-phosphosulfate reductases. Remarkably, the CRISPR array associated with this type IV-F system and a neighboring type V targets the type V Cas gene encoded in a competing cyanophage (Fig. 2).

Casλ is a divergent phage-specific CRISPR-Cas enzyme with a unique guide RNA

A distinctive phage-encoded enzyme family, Casλ, exists within huge bacteriophages that are evolutionarily linked to the recently reported Mahaphage clade (Al-Shayeb et al., 2020). This family of 33 compact homologs exhibited such sequence divergence that it had negligible sequence identity (<5%) to, and clustered separately from, type V and type II enzymes (Fig. S5). The protein is not encoded along with any other Cas proteins, and the RuvC nuclease was not immediately identifiable from the sequence. Difficulty in aligning this system to reported enzymes via remote homology (Fig S5) further suggested that a direct evolutionary relationship with known Cas superfamilies was questionable. CRISPR arrays associated with Casλ contain spacer sequences complementary to double-stranded DNA (dsDNA)-based extrachromosomal elements predicted to infect the same Bacteroidetes host (Fig. 2). These observations implied that Casλ may be targeting dsDNA in native contexts of the host similarly to Cas9 or Cas12 systems.

In any CRISPR-Cas system, processing of CRISPR array transcripts, consisting of repeats and spacer sequences acquired from previously encountered mobile genetic elements (MGEs) (McGinn and Marraffini, 2019), is essential to generating mature crRNAs that guide Cas proteins (Hille et al., 2018) to destroy foreign viruses. Similarly to the distinct nature of the protein, the Casλ crRNA is predicted to form an elongated hairpin secondary structure not previously observed in guide RNAs associated with Cas12 (Fig. 3A). Despite their divergent nucleotide sequences, crRNAs retain a similar predicted hairpin structure across the protein family (Fig. S5B). Furthermore, Casλ crRNAs contain conserved sequences at their 5’ and 3’ ends and in the center of the RNA (Fig. 3B). The overall sequence divergence of the protein, its putative RuvC domain, and the encoded crRNA prompted us to further analyze this protein family.

RuvC-mediated crRNA processing in the spacer region by Casλ

The lack of a detectable tracrRNA encoded within the genomic locus begged the question of how this aberrant RNA, akin to a naturally occurring crRNA-tracrRNA hybrid, may be processed by the CRISPR-Cas system or host factors to produce mature crRNA. Using radiolabeled precursor crRNAs as substrates, we first tested whether purified Casλ protein catalyzes RNA cleavage. Surprisingly, analytical denaturing gel electrophoresis showed that pre-crRNAs are cut by Casλ in the spacer region as opposed to the 5’ end of the RNA, where cutting has been observed in all self-processing single-effector systems analyzed previously (Fig. 3C, D, S6). The Casλ-induced pre-crRNA processing yields a crRNA spacer sequence that is complementary to DNA target sites 14-17 nucleotides (nt) in length.

The fact that Casλ can process its own pre-crRNA obviates the need for Ribonuclease III or other host factors required for the function of most known Cas9 and Cas12 family members. While some CRISPR-Cas proteins process pre-crRNAs using an internal active site distinct from the RuvC domain (Fonfara et al., 2016) or by recruiting Ribonuclease III to cleave a pre-crRNA:tracrRNA duplex (Deltcheva et al.,

2011), phage-encoded Casλ, like phage-encoded CasΦ, processes pre-crRNA using its RuvC active site. We thus tested Mg2+ dependence and showed that Casλ is indeed reliant on the presence of Mg2+ and thus, by extension, the RuvC active site for crRNA maturation (Fig. 3D).

CRISPR–Cas systems target DNA sequences following or preceding a 2–5 base pair (bp) Protospacer Adjacent Motif (PAM) for self-versus-non-self discrimination (Westra et al., 2013). We determined the sequence requirements for DNA targeting by Casλ by depleting plasmids harboring functional PAMs. This revealed the crRNA-guided double-strand DNA (dsDNA) targeting capability of Casλ and the lack of requirement for additional RNA components (Fig. S7). Casλ with GFP-targeting guides showed a reduction in colony-forming units (as a proxy for cell viability) of multiple orders of magnitude, in comparison to negative control of Casλ with a non-targeting guide (Fig. 3E).

Incubation of purified Casλ with crRNAs along with linear dsDNA target generated cleavage products with surprisingly pronounced staggered 5′-overhangs of 11–16 nt (Fig. 3F, 3G). Type V CRISPR-Cas enzymes such as Cas12a have also been observed to generate staggered overhangs, albeit smaller. Furthermore, the non-target strand (NTS) was cleaved faster than the target strand (TS) within the RuvC active site over a 2-hour time period (Fig. 3H).

Casλ induces genome editing in endogenous genes in human and plant cells

The development of single-effector CRISPR-Cas systems for editing eukaryotic cells has revolutionized genome engineering (Jinek et al., 2012). However, the large sizes of Cas9 and Cas12a enzymes can inhibit delivery into many cell types for which hypercompact genome editors with favorable kinetics imply great promise as an alternative. We conducted a head-to-head comparison of insertion and deletion efficiencies using Casλ and Cas12a ribonucleoproteins (RNPs) with identical guide RNA spacers targeting sequences recognizing VEGF and EMX1 genes in HEK293T cells. Despite their miniature size, Casλ RNPs generated promising genome-editing outcomes compared to Cas12a, and in at least one case, exceeded Cas12a indel percentages (Fig. 4A). Extending these experiments to *Arabidopsis thaliana*, we confirmed that Casλ exhibited editing efficiencies of up to 18% at the endogenous *PDS3* gene (Fig. 4B), notably higher than observed previously using CasΦ (Pausch & Al-Shayeb et al., 2020). Furthermore, we were able to achieve editing in the endogenous disease resistance gene *Snn5* in hexaploid wheat protoplasts, where six concurrent edits were required for successful editing (Fig. 4C). Next-generation sequencing revealed indel profiles in both mammalian and plant cells exhibited large deletions (Fig. 4D, Fig. S8), consistent with the staggered cuts observed *in vitro* at the PAM distal region.

Casλ protein structure explains recognition and interference mechanisms

CryoEM maps of a Casλ-crRNA-dsDNA ternary complex revealed a bi-lobed architecture analogous to Cas9 and Cas12 enzymes (Fig. 5A, B, S9, S10A). The 3 Å resolution structure revealed the shape and domain organization of the protein and the structure of the guide RNA (Fig. 5A-D, S9, Table S1). Notably, the RuvC domain, which in Casλ spans half the protein, is split into four parts within the sequence, likely hindering reliable alignment and clustering with reported Cas12 systems (Fig. 5D). The REC I and REC II domains are also segmented in the protein sequence, with the PAM-

interacting domain wedged within REC I as opposed to the N terminus of the protein as seen in CasΦ and other type V systems. Unlike CasΦ, Casλ contains a Target Strand Loading (TSL) domain that likely functions to bind the Target Strand, in a position analogous to the "Nuc" domain that was incorrectly hypothesized in other type V CRISPR–Cas enzymes to be a second nuclease domain responsible for DNA cleavage (Liu et al., 2019a). Casλ also exhibits a distinct structure in the REC I domain compared to CasΦ (Fig. S10B).

The crRNA assumes an unexpected shape that blankets the protein, with a divergent recognition lobe in Casλ that binds to distinct sequences and structural features of the guide RNA (Fig. 5C). Specifically, we observed possible interactions between primarily polar or charged residues within the REC II domain in Casλ with the conserved motifs of the crRNA hairpin (Fig. 3A, 3B, 5C). These residues are conserved across the protein family and likely interact either directly with the RNA nucleobases (Q452, N510), or with the RNA phosphate backbone to stabilize the guide (S451, K596, E444, N445, K503, Y619) (Fig. 5C).

CRISPR–Cas proteins initiate the unwinding of target double-stranded DNA through PAM recognition. In Casλ, this is achieved via interactions with the OBD, REC I, and a five α-helical bundle referred to as the PAM-interacting domain; PID. Residues within the three domains interact with the sugar-phosphate backbone of the target DNA (Fig. 5B) and, in some cases such as residue N102, interact directly with the nucleobases. The interaction between N102 and nucleobase G(-1) may explain the preference for purines in this position as opposed to pyrimidines since a pyrimidine substitution would result in a base that is too distant from the interacting asparagine. In examining the aftermath of cis-cleavage of DNA, we found that Casλ had a very low level of ssDNA or ssRNA cleavage in trans upon DNA recognition in cis (Fig. S7). Incubation of the Casλ protein with non-cognate guides from other orthologs within the protein family replicated the ssDNA trans cleavage effect despite differences in their sequence (Fig 5E, S7), confirming that guides within the Casλ family may be interchangeable, unlike Cas9. Single mismatches across the ssDNA target revealed that the seed region of the target DNA (1-5) and the region extending from bases 7-13 are required to match the spacer sequence of the guide RNA for efficient cleavage (Fig. 5F). Investigation of positions that possibly interact with the DNA in these regions (Fig. 5G) or the corresponding RNA revealed conserved residues in REC, OBD, PID, and RuvC domains that may account for the complex's intolerance to target mismatches, and, therefore, the possibility of relatively high fidelity in the context of gene editing. Overall, the domains within Casλ exhibit unexpected segmentation and rearrangement compared to known type V systems, and our structural understanding of Casλ provides a starting point for the future design of variants with expanded target space and improved specificity.

5.4 **Conclusions**

CRISPR-Cas systems are rarely found on viral genomes. Nevertheless, analysis of CRISPR-encoding phages demonstrated unexpected sequence and biochemical diversity across CRISPR-Cas types in both classes. Spacer sequences associated with CRISPR arrays in phages show significant complementarity to double-stranded DNA (dsDNA)-based viruses and extrachromosomal elements predicted to infect the same hosts, further elucidating the broad strategy of phages to protect their hosts against superinfection by competing elements. We also find that phage genomes may harbor CRISPR-Cas systems of all six known types. However, the absence of components related to abortive infection in RNA-targeting systems, the absence of a processive nuclease in some type I systems, and the presence of CysH in type IV systems, suggest some alternate outcomes to nucleic acid-targeting within a host cell compared to well-studied mechanisms. The analysis of known systems that are well-studied in a microbial cell context, within the context of phage predation, allows us to explore different possibilities and outcomes of phage variants. The targeting of abundant or essential transcripts of competing phages, such as phage tail proteins or transposases, by RNA-targeting CRISPR-Cas systems without inducing abortive infection, suggests a strategy to avoid self-destruction of transcripts of the CRISPR-encoding phage or induction of a dormant state in the host that may be disadvantageous to the phage lifecycle. In this biological context, phages may be using the RNA-targeting system in a way that best suit their needs. The lack of a Cas3 nuclease in the type I system targeting plasmid-like elements suggests a gene silencing mechanism that precludes DNA cutting to augment the activity of the co-occurring type Casμ system in the same genome. The observation that the targeted plasmid-like elements harbor restriction enzymes and retron-based anti-phage defense systems that could limit the infectivity of the CRISPR-encoding phage, combined with the targeting of a CRISPR-encoding phage by another phage encoding type IV and V CRISPR-Cas systems, underscores the dynamic nature of the evolutionary arms race between mobile elements in competition for host resources.

We further establish that phage genomes are a natural reservoir of novel miniature single-effector CRISPR-Cas systems such as DNA targeting type II and type V systems, and refer to these systems using greek nomenclature, such as Casμ, CasΩ, and Casλ, extending the tradition established by phage-encoded CasΦ. In contrast to the multi-subunit class 1 systems (namely type I and type III) being the most prevalent CRISPR-Cas systems in prokaryotic genomes (Makarova et al., 2019), we find a remarkable abundance of miniature class 2 systems on phages. As phages are known to be relatively fast-evolving with relatively significant genome size limitations compared to prokaryotes, they may be a potential incubator in which divergent or hypercompact systems emerge. Some systems such as Casλ bear such sequence-level divergence that they cluster separately from Cas12 and Cas9 systems using our methods, indicating that a direct evolutionary relationship with known Cas superfamilies is unclear. This suggests that distinct type V nuclease subtypes such as Casλ may have evolved multiple times, perhaps within phage genomes, from distinct transposon-encoded TnpB families, which have recently been shown to be RNA-guided nucleases themselves (Karvelis et al., 2021). We observed very minimal collateral cleavage of ssDNA and RNA supplied in trans, which was exhibited only for some targets, but not endogenous ones, suggesting that Casλ may have minimal ability to target single-stranded MGE intermediates. Despite being from very different clades of phages, with

CasΦ originally found in Biggiephages and Casλ in Mahaphages, likely originating from different ancestral protein families, and having divergent sequence and domain organizations, we observe a convergent evolution of Cas12-like structure, possibly on the phage genomes. Interestingly, both phage-encoded Casλ and CasΦ process their own pre-crRNA, particularly relying on the same RuvC active site used for DNA cleavage for crRNA maturation which has not been observed in other prokaryotic-borne CRISPR-Cas proteins, albeit with distinct protein sequences and structures. In both cases, if the phages encode their own Cas variants without reliance on host factors, this may provide some benefit by eliminating the possibility that the ongoing evolution of an essential host protein will render it incompatible with phage-encoded antiviral systems.

The bi-lobed CryoEM structure of the Casλ-crRNA-dsDNA complex analogous to Cas9 and Cas12 enzymes exhibits an interesting case of convergent evolution of RNA-guided effectors despite extreme sequence divergence and distinct ancestral protein families. The domain architecture exhibits unexpected segmentation and rearrangement compared to known type V systems, with the RECI, RECII, OBD, RuvC, and TSL domains are all is split at the sequence level into multiple parts, and the unique organization of the PAM interacting domain and REC I explain the difficulty in accurately aligning Casλ to reported enzymes, despite overall structural similarity.

In addition to their streamlined nature that is advantageous for cellular delivery (Fig. S12), hypercompact phage systems such as Casλ can induce remarkably efficient genome editing of endogenous genes in mammalian, *Arabidopsis*, and wheat cells on par with, or in some cases, exceeding Cas12a-mediated genome editing, showing that there isn't necessarily a tradeoff between Cas effector size and molecular function or utility in mammalian cells

5.5 **Figures**

Figure 1. **Diversity of Phages, Hosts, and Class 2 CRISPR-Cas systems across viruses.** (A) Protein-clustering network analysis based on the number of shared protein clusters between the CRISPR-encoding phages in this study and RefSeq phages. Each node represents a genome and each edge is the hypergeometric similarity between genomes based on shared protein clusters. (B) Genome size distribution of circularized CRISPR phages from this study (n=152). (C) A heat map showing the number of CRISPR phage genomes containing each CRISPR type with respect to major bacterial phyla. 'Unknown' indicates CRISPR phages that could not be assigned to any of the known types. (D) Maximum likelihood phylogenetic tree of phage and bacterial encoded type II nucleases and respective predicted ancestral IscB nucleases. Bootstrap and approximate likelihood-ratio test values ≥ 50 are denoted on the branches. Bottom illustration of genomic CRISPR-Cas loci of type II and representative type V systems previously employed in genome editing applications. (E) Maximum likelihood phylogenetic tree of phage and previously reported bacterially-encoded type V nucleases and respective predicted ancestral TnpB nucleases. (F) Maximum likelihood phylogenetic tree of phage and previously reported bacterially- encoded type VI nucleases.

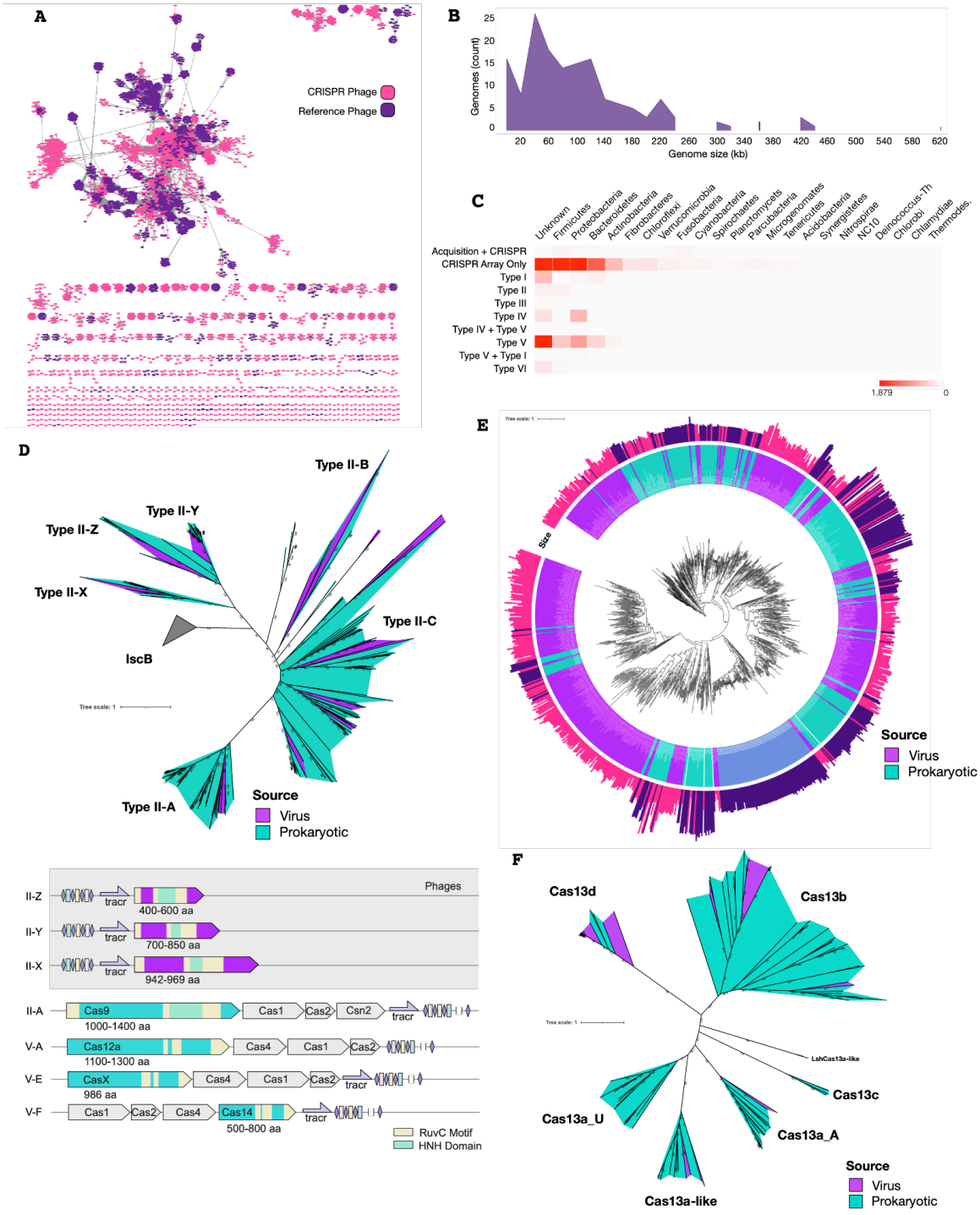

A
CRISPR Phage
Reference Phage
B
Genomes (count)
Genome size (kb)
C
Unknown
Firmicutes
Proteobacteria
Bacteroidetes
Actinobacteria
Fibrobacteres
Chloroflexi
Verrucomicrobia
Fusobacteria
Cyanobacteria
Spirochaetes
Planctomycets
Parcubacteria
Microgenomates
Tenericutes
Acidobacteria
Synergistetes
Nitrospirae
NC10
Deinococcus-Th
Chlorobi
Chlamydiae
Thermodes.
Acquisition + CRISPR
CRISPR Array Only
Type I
Type II
Type III
Type IV
Type IV + Type V
Type V
Type V + Type I
Type VI
1,879
0
D
Type II-B
Type II-Y
Type II-Z
Type II-X
Type II-C
IscB
Type II-A
Tree scale: 1
Source
Virus
Prokaryotic
E
Tree scale: 1
Size
Source
Virus
Prokaryotic
Phages
II-Z
tracr
400-600 aa
II-Y
tracr
700-850 aa
II-X
tracr
942-969 aa
II-A
Cas9
Cas1
Cas2
Csn2
tracr
1000-1400 aa
V-A
Cas12a
Cas4
Cas1
Cas2
1100-1300 aa
V-E
CasX
Cas4
Cas1
Cas2
tracr
986 aa
V-F
Cas1
Cas2
Cas4
Cas14
tracr
500-800 aa
RuvC Motif
HNH Domain
F
Cas13d
Cas13b
Tree scale: 1
LshCas13a-like
Cas13c
Cas13a_U
Cas13a_A
Cas13a-like
Source
Virus
Prokaryotic

Figure 2. **Mobile genetic element targeting by diverse mechanisms in various phyla to abrogate superinfection.** Graphical illustrations of representative phage CRISPR loci harboring novel subtypes and their proposed mechanisms and functions as determined via spacer targeting and protein sequence analysis. Special consideration is given to phages carrying multiple loci.

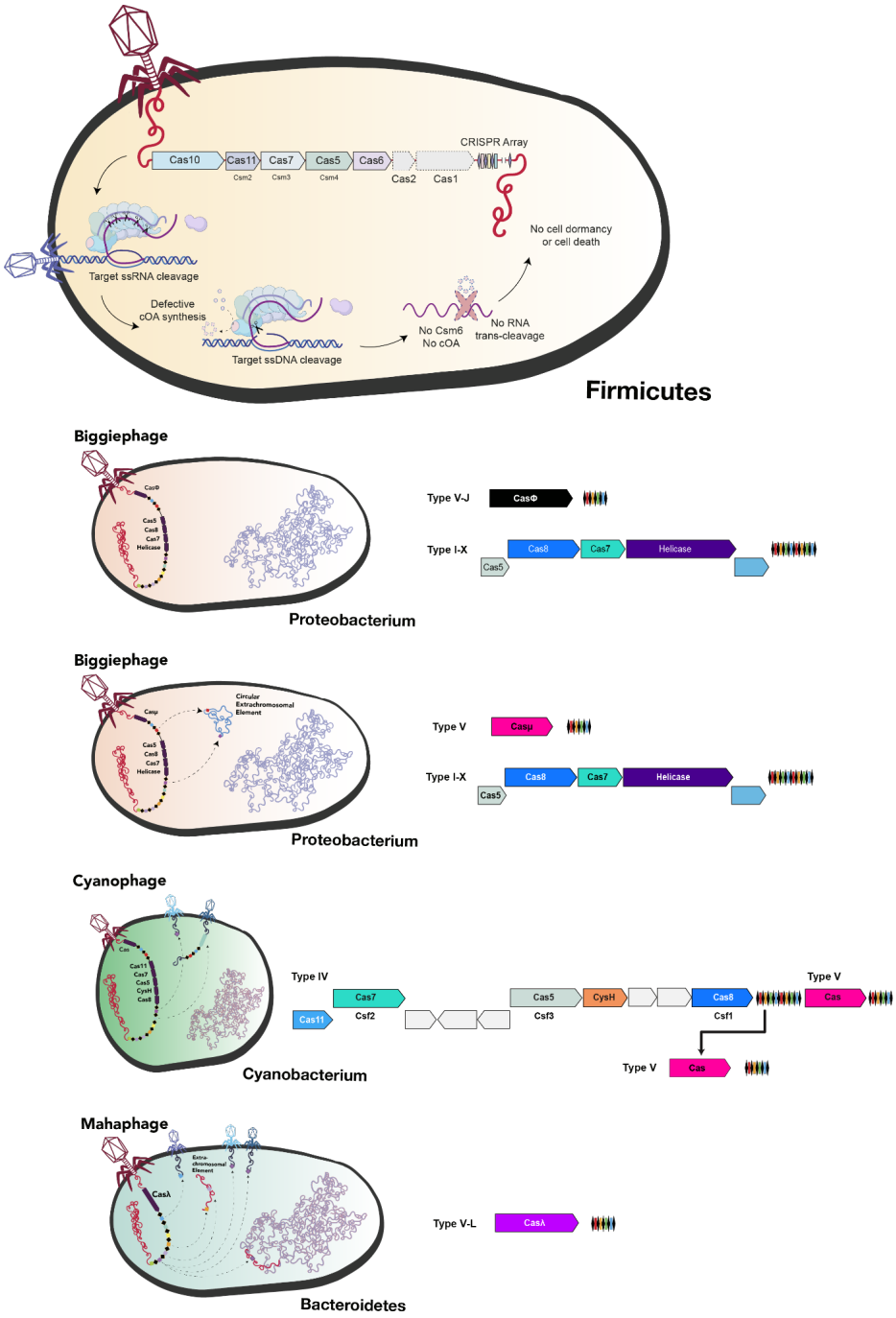


Figure 3. Casλ processes its own crRNA and cleaves dsDNA. (A) Casλ repeats uniquely display highly conserved nucleotide sequences at the 5', 3', and center of the RNA (B) Casλ1 from Huge Mahaphages displays a unique crRNA hairpin compared to known Cas12 enyzmes, and is reminiscent of stem loop 1 of the engineered SpyCas9 single gRNA (sgRNA). (C) 5' radiolabelling of crRNAs indicate that Casλ1 uniquely processes its own crRNA in the spacer region (or 5' end). OH-ladder enables the pre-crRNA processing sites (red triangles) to be derived. (D) Processing of the Repeat-Spacer-Repeat pre-crRNA substrate occurs similarly to (C) in the spacer region, and does not occur in the absence of Mg2+, indicating a role for the RuvC in the processing mechanism. (E) Casλ with targeting or non-targeting guides validate its capacity to cleave DNA flanking experimentally determined PAMs in E. coli. (F) Cleavage assay targeting dsDNA for mapping of the cleavage structure. (G) Efficiency and kinetics of DNA cleavage of NTS and TS (n = 3 each, mean ± s.d.) (H) Scheme illustrating the DNA cleavage pattern

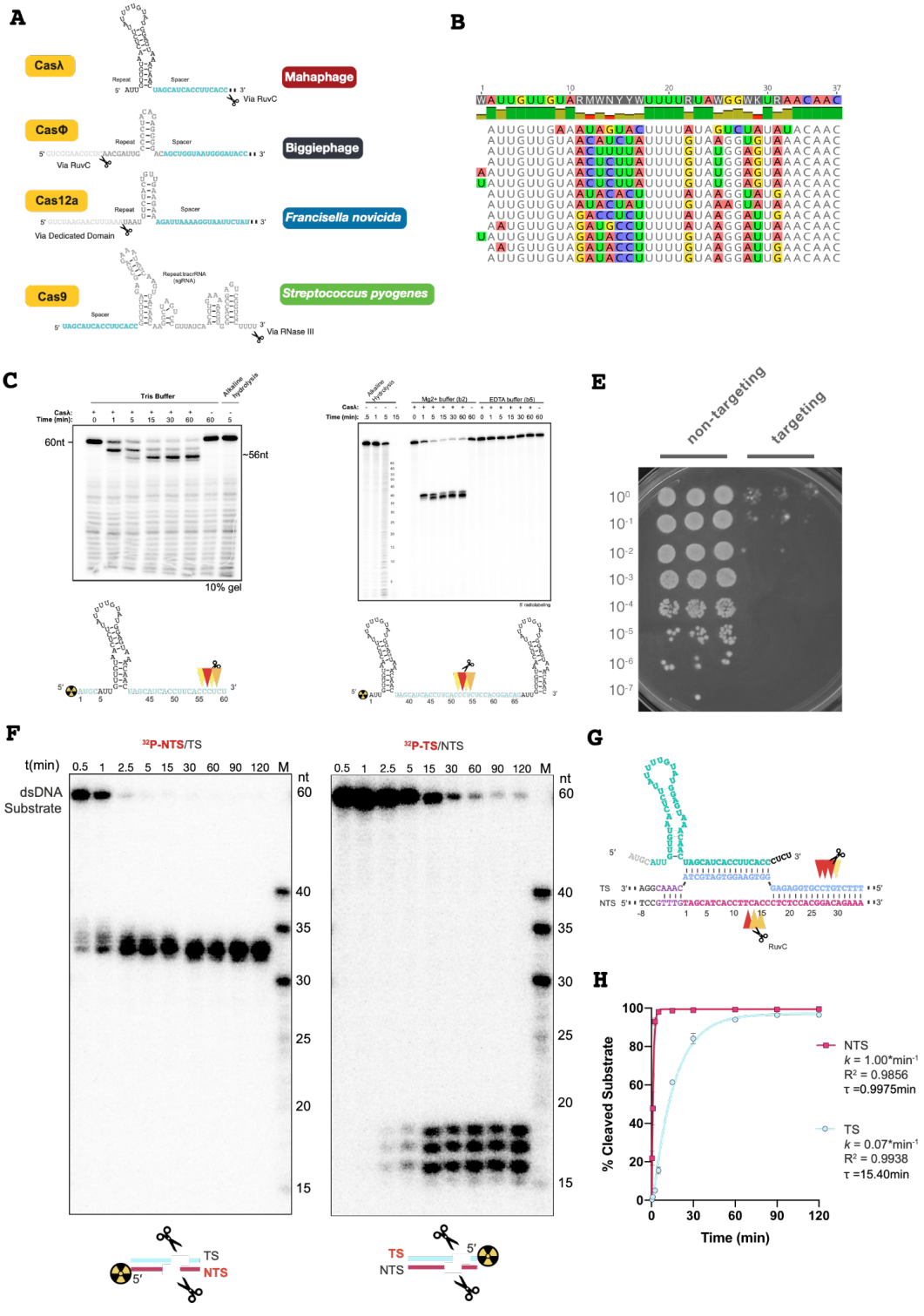

A
CasΛ
Mahaphage
Via RuvC
CasΦ
Biggiephage
Via RuvC
Cas12a
Francisella novicida
Via Dedicated Domain
Cas9
Streptococcus pyogenes
Via RNase III
B
C
Tris Buffer
Alkaline hydrolysis
60nt
~56nt
10% gel
Mg2+ buffer (b2)
EDTA buffer (b5)
5' radiolabeling
E
non-targeting
targeting
F
32P-NTS/TS
32P-TS/NTS
t(min)
dsDNA Substrate
nt
TS
NTS
G
RuvC
H
% Cleaved Substrate
Time (min)
NTS
k = 1.00*min-1
R² = 0.9856
τ =0.9975min
TS
k = 0.07*min-1
R² = 0.9938
τ =15.40min

Figure 4. Casλ is functional for editing of endogenous genes in human, *Arabidopsis*, and wheat cells with large deletion profiles. (A) Indel efficiency using Casλ and Cas12a RNPs with identical spacers targeting VEGF and EMX1 genes in HEK293T cells, and a schematic of the model of DNA cleavage outcomes following DNA cleavage by Casλ. (B) Indel efficiencies in Arabidopsis thaliana protoplasts showing significantly higher levels of editing than previously achieved by CasΦ for the same *PDS3* gene, and (C) in wheat protoplasts targeting the disease resistance gene *Snn5*. (D) Indel profiles generated by Casλ RNP administration show primarily large deletions, and little change without Casλ.

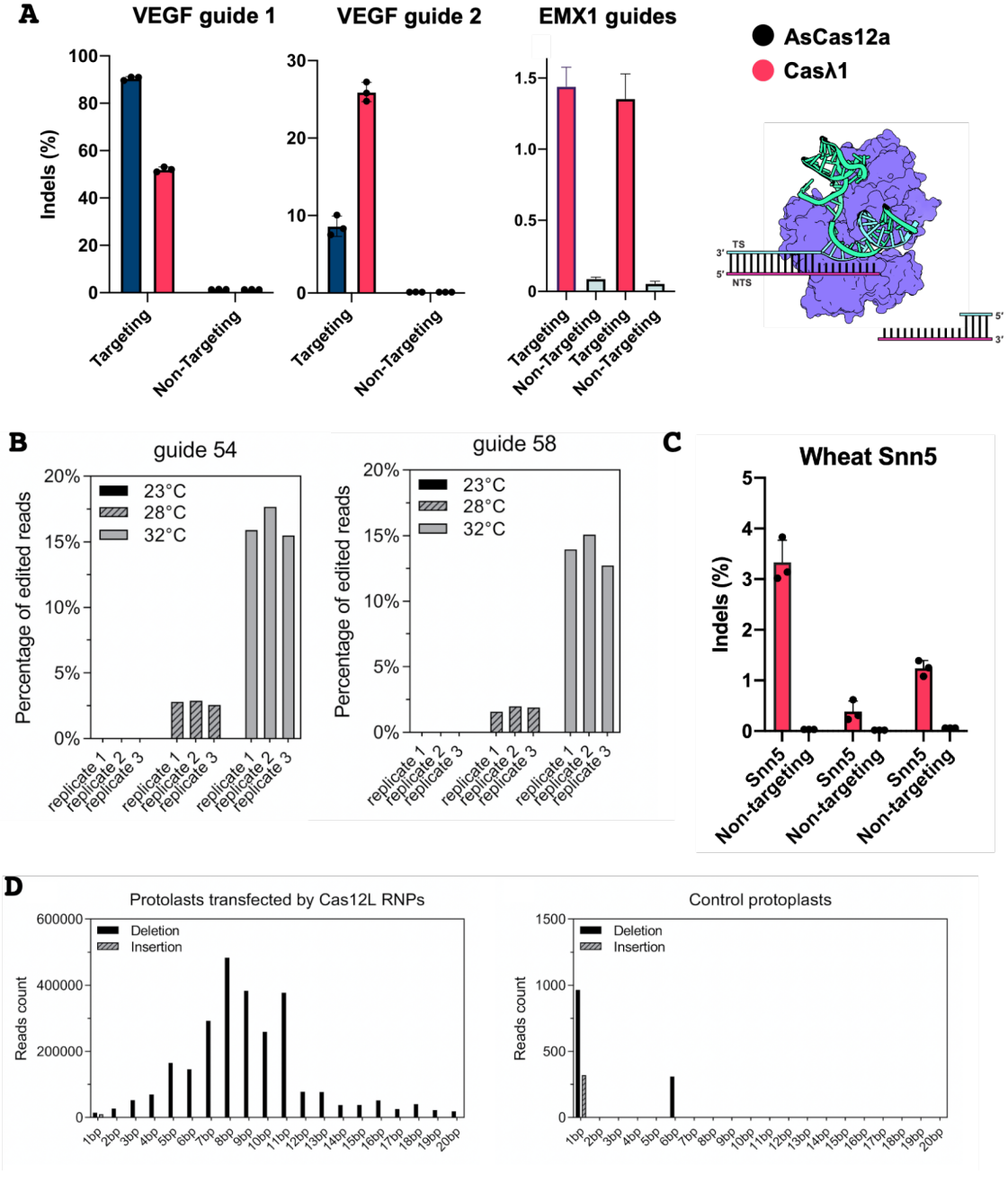
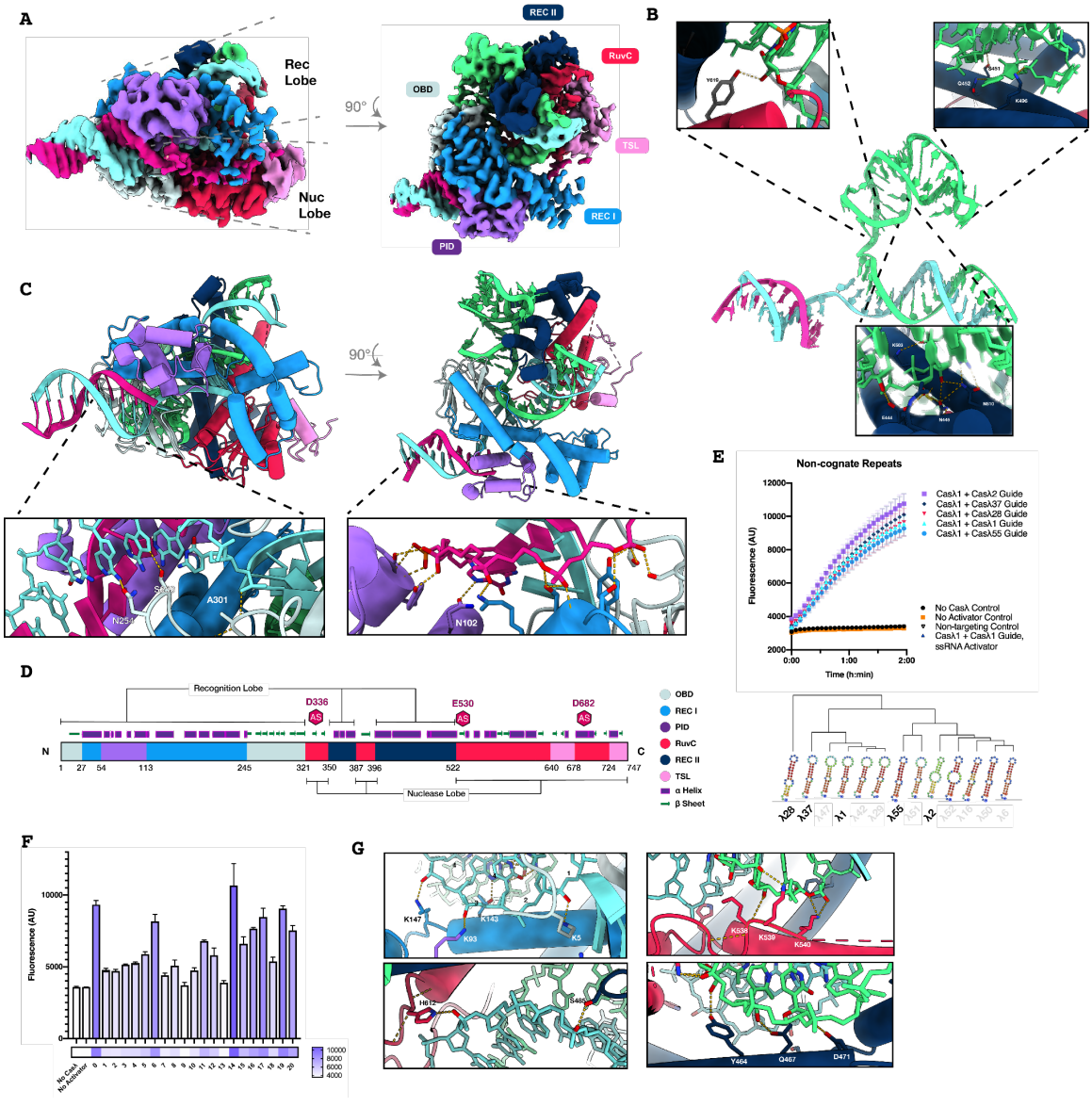


Figure 5. (A) Cryo-EM maps of the Casλ- guide RNA- DNA complex. The target strand is shown in cyan and the non-target strand is shown in magenta. (B) Cylinder representation of the Casλ–gRNA–DNA complex in two 90°-rotated orientations. Disordered linkers are shown as dotted lines. Insets highlight residues responsible for PAM recognition. Hydrogen bonds are shown as dashed lines. (C) Model of guide RNA–target DNA complex, with insets highlighting residues interacting with the RNA. (D) Schematic of the domain organization of Casλ. (E) Casλ can still cleave ssDNA in trans with guide RNAs consisting of non-cognate repeats that are divergent at the sequence level. Hierarchical clustering dendrogram of different repeats with their predicted secondary structures. (F) Fluorescence output using oligonucleotide activators with mismatches at each respective position along the target DNA. (G) Close-up views of the residues responsible for recognition of the seed and low mismatch tolerance regions



5.6 Methods

Phylogenetic analysis

Cas protein sequences and representatives from the TnpB superfamily were collected and the resulting set was clustered at 90% amino acid identity to reduce redundancy. A new alignment of Casλ with the resulting sequence set was generated using MAFFT with 1000 iterations and filtered to remove columns composed of gaps in 95% of sequences. The phylogenetic tree was inferred using IQTREE v1.6.6 using automatic model selection and 1000 bootstraps.

crRNA sequence analysis

CRISPR-RNA (crRNA) repeats from Phage-encoded CRISPR loci were identified using MinCED (github.com/ctSkennerton/minced). The repeats were compared by generating pairwise similarity scores using the Needleman-Wunsch algorithm. A heatmap was built using the similarity score matrix and hierarchical clustering produced dendrograms that were overlaid onto the heatmap to delineate different clusters of repeats.

PAM depletion analysis

PAM depletion assays were performed with both, Casλ plasmids that either carried the whole Casλ locus as derived from metagenomics (pPP049, pPP056 and pPP062), or with plasmids that contained only the casL gene and a mini CRISPR (pPP097, pPP102 and pPP107). Assays were performed as three individual biological replicates. Plasmids containing casL and mini CRISPRs were transformed into E. coli BL21(DE3) (NEB) and constructs containing Casλ genomic loci were transformed into E. coli DH5α (QB3-Macrolab, UC Berkeley). Subsequently, electrocompetent cells were prepared by ice cold H20 and 10 % glycerol washing. A plasmid library was constructed with 8 randomized nucleotides upstream (5') end of the target sequence. Competent cells were transformed in triplicate by electroporation with 200 ng library plasmids (0.1 mm electroporation cuvettes (Bio-Rad) on a Micropulser electroporator (Bio-Rad)). After a two-hour recovery period, cells were plated on selective media and colony forming units were determined to ensure appropriate coverage of all possible combinations of the randomized 5' PAM region. Strains were grown at 25 °C for 48 hours on media containing appropriate antibiotics (either 100 µg/mL carbenicillin and 34 µg/mL chloramphenicol, or 100 µg/mL carbenicillin and 50 µg/mL kanamycin) and 0.05 mM isopropyl-β-D-thiogalactopyranoside (IPTG), or 200 nM anhydrotetracycline (aTc), depending on the vector to ensure propagation of plasmids and Casλ effector production. Subsequently, propagated plasmids were isolated using a QIAprep Spin Miniprep Kit (Qiagen).

PAM depletion sequencing analysis

Amplicon sequencing of the targeted plasmid was used to identify PAM motifs that are preferentially depleted. Sequencing reads were mapped to the respective plasmids and PAM randomized regions were extracted. The abundance of each possible 8 nucleotide combination was counted from the aligned reads and normalized to the total reads for each sample. Enriched PAMs were computed by calculating the log ratio compared to the abundance in the control plasmids, and were used to produce sequence logos.

Programmable DNA targeting

A flp recombination assay was performed in E. coli to eliminate the Kanamycin resistance cassette from *E. coli* strains that contain GFP and RFP expression cassettes integrated into the genome. Individual colonies of the E. coliΔKan were picked to inoculate three 5 mL (LB) starter cultures to prepare electrocompetent cells the following day. 100 mL (LB) main cultures were inoculated from the starter cultures and grown vigorously shaking at 37 °C to an OD600 of 0.6-0.7 before preparation of electrocompetent cells by repeated ice-cold H20 and 10% glycerol washes. Cells were resuspended in 10% glycerol and 50µL aliquots were flash frozen in liquid nitrogen and stored at -80 °C. Casλ vectors were generated containing codon optimized Casλ1 gene and a guide comprised of its cognate repeat element and selections of spacers targeting the GFP DNA within the resulting E. coliΔKan strain (pBAS41, pBAS42, pBAS43, pBAS44) were subcloned from pBAS12. Casλ vectors containing Casλ1 and a guide composed of a non-cognate repeat unit from Casλ2 and a GFP-targeting spacer (TAGCATCACCTTCACCCTCTCCACGGACAG) guide were also subcloned to form pBAS40. The Casλ vectors and Casλ vectors with a non-targeting guide control plasmid were transformed into 25 µL of electrocompetent cells with 100 ng of plasmid via electroporation in 0.1 mm electroporation cuvettes (Bio-Rad) on a Micropulser electroporator (Bio-Rad), cells were recovered in 1 mL recovery medium (Lucigen) shaking at 37 °C for one hour. 10-fold dilution series were then prepared and 3.5 µL of the respective dilutions were spot-plated on LB-Agar containing the appropriate antibiotics and IPTG inducer. Plates were incubated overnight at 37 °C and colonies were counted the following day to determine the transformation efficiency. To assess the transformation efficiency, the mean and standard deviations were calculated from the cell forming units per ng transformed plasmids for the electroporation triplicates. The experiment showed marked reduction of GFP E. coli using Casλ vectors with their cognate guides (pBAS44) in comparison to the non-targeting control, indicating a double-stranded DNA break at the target region. The growth of primarily RFP-positive/ GFP-negative colonies under blue light further supports the ability to confer targeted programmable genome editing to result in strains lacking GFP production. Growth inhibition using Casλ vectors with guides from a separate Casλ ortholog (pBAS40), with colonies observed expressing primarily RFP and no GFP, also indicate that Casλ orthologs may function using guides from related CRISPR-Cas systems to confer editing in cells, with a precise ablation of GFP production. This can be further expanded to HEK293T mammalian cells with integrated GFP, which indicate activity in mammalian cells. The sickly phenotype of E. coli colonies that have grown in both cases even in undiluted samples is also indicative of possible trans-cleavage of nucleic acids (RNA or DNA), which can be used for diagnostic purposes by providing a sample containing the target nucleic acid with the Casλ RNP and a single-stranded DNA fluorophore-quencher (ssDNA-FQ) reporter or RNA fluorophore-quencher (ssRNA-FQ) reporter molecule, generating a strong fluorescence signal in the presence of the target nucleic acid compared to a markedly lower fluorescence signal in its absence.

Protein Purification

Casλ overexpression vectors containing a His-Tag were transformed into chemically competent E. coli BL21(DE3)-Star (QB3-Macrolab, UC Berkeley) and incubated overnight at 37°C on LB-Kan agar plates (50 µg/mL Kanamycin). Single colonies were picked to inoculate 50 mL (LB, Kanamycin 50 µg/mL) starter cultures which were

incubated at 37 °C shaking vigorously overnight. The following day, 2 750 mL TB-Kan media (50 µg/mL Kanamycin) were inoculated with 40 mL starter culture and grown at 37 °C to an OD600 of 0.6, cooled down on ice, and gene expression was subsequently induced with 0.5 mM IPTG followed by incubation overnight at 16 °C.
The cells were harvested by centrifugation and resuspended in low salt buffer, and then subsequently lysed by sonication. The soluble fraction was loaded on a 5 mL Ni-NTA Superflow Cartridge (Qiagen) pre-equilibrated in wash buffer. Bound proteins were washed with 20 column volumes (CV) wash buffer and subsequently eluted in 5 CV elution buffer (50 mM HEPES-Na pH 7.5 RT, 500 mM NaCl, 500 mM imidazole, 5 % glycerol, and 0.5 mM TCEP). The eluted proteins were concentrated to 1 mL before injection into a HiLoad 16/600 Superdex 200pg column (GE Healthcare) pre-equilibrated in size-exclusion chromatography buffer (20 mM HEPES-Na pH 7.5 RT, 500 mM NaCl, 5 % glycerol, and 0.5 mM TCEP). Peak fractions were concentrated to 1 mL and concentrations were determined using a NanoDrop 8000 Spectrophotometer (Thermo Scientific). Proteins were purified at a constant temperature of 4 °C and concentrated proteins were kept on ice to prevent aggregation, snap-frozen in liquid nitrogen, and stored at -80 °C. **SDS-PAGE gel electrophoresis of Casλ at varying stages of protein purification showed a protein size in line with computationally predicted values of ~70-85 kDa.**

Pre-crRNA processing assays

The reactions were carried out in RNA cleavage buffer containing 20 mM Tris-Cl (pH 7.5 at 37°C), 150 mM KCl, 5 mM $MgCl_2$, 1 mM TCEP, and 5% (v/v) glycerol. Pre-crRNA substrates were 5′-radiolabeled with T4 PNK (NEB) in the presence of gamma ^{32}P-ATP. In a typical pre-crRNA processing reaction, the concentrations of Casλ and ^{32}P-labeled pre-crRNA substrates were 100 nM and 3 nM, respectively. Reactions were incubated at 37°C, and an aliquot of each reaction was quenched with 2x Quench Buffer (94% (v/v) formamide, 30 mM EDTA, 400 µg/mL heparin, 0.2% SDS, and 0.025% (w/v) bromophenol blue) at 0, 1, 5, 15, 30, and 60 min. RNA hydrolysis ladders were prepared by incubating RNA probes in 1X RNA Alkaline Hydrolysis Buffer (Invitrogen) at 95°C before the addition of 2x Quench Buffer. Quenched reactions were incubated at 95°C for 3 min, and products were then resolved by denaturing PAGE (10% or 20% acrylamide:bis-acrylamide 19:1, 7 M urea, 1X TBE). Gels were dried (3 hr, 80°C) on a Model 583 Gel Dryer (Bio-Rad) and exposed to a phosphor screen. Phosphor screens were imaged on an Amersham Typhoon phosphorimager (GE Healthcare). For assays in an EDTA-containing buffer, 25 mM EDTA was substituted for 5 mM $MgCl_2$.

***In vitro* cleavage assays - radiolabeled nucleic acids**

crRNA oligonucleotides were manufactured synthetically and dissolved in DEPC-treated ddH20 to a concentration of 0.5 mM. Subsequently, the crRNA was heated to 65 °C for 3 min and allowed to cool down to room temperature. Casλ RNP complexes were reconstituted at a concentration of 10 µM by incubation of 10 µM Casλ and 12 µM crRNA for 10 min at RT in 2x cleavage buffer (20 mM Hepes-Na pH 7.5, 300 mM KCl, 10 mM MgCl2, 20 % glycerol, 1 mM TCEP). RNPs were aliquoted to a volume of 10 µL, flash-frozen in liquid nitrogen, and stored at −80 °C. RNP aliquots were thawed on ice before experimental use. Substrates were 5′-end-labelled using T4-PNK (NEB) in the presence of 32P-γ-ATP. Oligonucleotide-duplex targets were generated by combining

32P-labelled and unlabelled complementary oligonucleotides in a 1:1.5 molar ratio. Oligos were hybridized to a DNA-duplex concentration of 50 nM in hybridization buffer (10 mM Hepes-Na pH 7.5 RT, 150 mM NaCl), by heating for 5 min to 95 °C and a slow cool down to RT in a heating block. Cleavage reactions were initiated by combining 200 nM RNP with 2 nM substrate in CB buffer and subsequently incubated at 37 °C. Reactions were stopped by the addition of two volumes of formamide loading buffer (96 % formamide, 100 µg/mL bromophenol blue, 50 µg/mL xylene cyanol, 10 mM EDTA, 50 µg/mL heparin), heated to 95 °C for 5 min, and cooled down on ice before separation on a 12.5 % denaturing urea-PAGE. Gels were dried for 4 h at 80°C before phosphor-imaging visualization using an Amersham Typhoon scanner, v2.0.0.6 firmware version 208 (GE Healthcare). Bands were quantified using ImageQuant TL 8.1 (Cytivia) and the cleaved fraction was calculated as the product intensity sum divided by the combined substrate and product intensity sum. Curves were fitted to a One-Phase-Decay model to derive the rate of cleavage.

Fluorophore quencher and DNA mismatch tolerance assay

DNA oligo activators were ordered from IDT to contain mismatches at each respective position, (A->C, T->G, C->A, G->T). Casλ RNPs were prepared as described above. Reactions were started by combining 100 nM RNP (100 nM Casλ, 120 nM crRNA), 100 nM DNase Alert (IDT) FQ probe, with and without activator ssDNA and with the addition of a non-targeting guide or activator control in cleavage buffer in a 384 well flat bottom black polystyrene assay plate (#3820, Corning). Three replicates for each reaction were monitored (λex: 530 nm; λex: 590 nm) in a Cytation 5 plate reader (BioTek, software Gen v3.04) at 37 °C every 1.5 min for the activator titration experiment. For the FQ-mismatch-assay, 2 nM activator oligonucleotides were used in singlicates. The data were background-subtracted using the mean values of the measurements taken for three no-activator controls at the respective time point.

Mammalian Genome Editing

RNPs were formed in the SF nucleofection buffer with 100pmol protein & 120pmol crRNA in up to 10uL (10uM concentration) for 10’ at RT. 78 pmol (1uL) of IDT Cas12a electroporation enhancer was then added. HEK293T cells were added in a 10uL SF nucleofection buffer at 200,000 cells per nucleofection. 21uL reactions were loaded into cuvettes. Pulse code used was DS-150. Cells were grown in duplicate from each nucleofection in 24-well plates. gDNA was collected after 72 hours in Quick Extract. PCR1 was performed followed by bead clean-up to remove primers and submitted for PCR2, bead clean up, and iSeq.

Plant Genome Editing

Guides were designed to target the PDS3 gene in plant protoplasts, incubated with protein as described for in vitro assays, and 10uM of RNP was transfected onto Arabidopsis protoplasts as previously described.

Ternary complex reconstitution for cryo-EM.

Casλ was produced as described above. crRNA (rBAS80) was ordered as a synthetic RNA oligonucleotide from IDT and dissolved in DEPC-treated ddH2O to a concentration of 0.5mM. Subsequently, the crRNA was heated to 65 °C for 3 min and cooled down to RT to allow for hairpin formation. DNA oligonucleotides (dBAS608, dBAS609) were

designed to contain a non complementary protospacer segment to produce 'bubbled' substrates and facilitate rapid R-loop formation during ternary complex reconstitution. Oligonucleotides were ordered from and synthesized by IDT. DNA oligonucleotides were combined in a 1:1.2 molar ratio (target strand:nontarget strand) and annealed to form a DNA duplex in hybridization buffer (10mM Hepes-Na pH 7.5 RT, 150mM NaCl) by heating for 5min at 95 °C and a subsequent slow cool down in a thermocycler.

Prior to reconstitution, thawed Casλ protein was incubated with crRNA in 1:1.1 ratio for 10 min at room temperature, and the DNA duplex was added. The ternary complex was reconstituted with the final Casλ : crRNA : TS : NTS strands stoichiometry of 1 : 1.1 : 1.2 : 1.44, for another 10 min at RT, and further injected into a Superdex 200 prep grade 10/300 column (GE Healthcare) pre-equilibrated in low salt buffer (10mM Hepes-Na pH 7.5, 150mM NaCl) at 4 °C to separate complexes from excess nucleic acids. Peak fractions were pooled and concentrated down to ~20 uM with a centrifugal filter device (Millipore 10 kDa Mw cutoff), as measured by absorbance at 260 nm with NanoDrop 8000 Spectrophotometer (Thermo Scientific), and kept on ice before plunge-freezing.

Electron microscopy grid preparation and data collection.

The resulting sample was frozen using FEI Vitrobot Mark IV, cooled to 8 °C at 100% humidity. 1.2/1.3 300 mesh UltrAuFoil gold grids (Electron Microscopy Sciences #Q350AR13A), were glow discharged at 15 mA for 25 s using PELCO easyGLOW. Total volume of 4 uL sample was applied to the grid and immediately blotted for 5 s with a blot force of 8 units. Micrographs were collected on a Talos Arctica operated at 200 kV and x36,000 magnification (1.115 A pixel size), in the super-resolution setting of K3 Direct Electron Detector. Cryo-EM data was collected using SerialEM v.3.8.7 software. Images were obtained in a series of exposures generated by the microscope stage and beam shifts.

Single-particle cryo-EM data processing and 3D volume reconstruction.

In total, 2795 movies were collected with a defocus range of -0.8 to -2.2 um. Data processing was further performed in cryoSPARC v3.2.0. Movies were corrected for beam-induced motion using patch motion, and CTF parameters were calculated using patch CTF. Two rounds of Topaz training were applied to the data to enrich the amounts of Casλ ternary complex particles picked as follows. In the first round, as a result of initial curation, a subset of 562 micrographs with seemingly best ice quality and CTF fit were selected. Further, 3931 particles were manually picked and submitted to Topaz particle training. The resulting Topaz model was used to pick particles from the micrographs, and a total of 153,537 particles were extracted with bin factor 2**,** and applied to 2D classification. Following the selection of the best classes, 113,638 particles were used for *ab initio* reconstruction with three classes. The 55,587 particles constituting the best class in terms of resolution and resemblance to an RNP were subject to non-uniform map refinement, and an initial complex map was obtained. In the second round, the latter particles were used to train a new Topaz model. Following the second round of curation, a total of 1931 micrographs were selected, and the new Topaz model was applied to pick and extract the particles. In total, 884,595 particles were subject to a round of 2D classification. After excluding a minor subset of classes, a total of 874,119 particles were selected and submitted to *ab initio* reconstruction with three classes. Three resulting maps and all particles were applied to a round of heterogeneous refinement. Particles

constituting the best class in terms of resolution were subject to the remove duplicates procedure, and further to non-uniform map refinement. As a result, a 2.99 A map reconstructed from a total of 369,389 particles was obtained. This map was further used for model building.

Model building and refinement.

The initial model of the Casλ protein was obtained with the AlphaFold program. The predicted model was split into two parts (eventually constituting REC and Nuc lobes), and each was docked independently into the map with the fitmap tool in ChimeraX. The dsDNA and crRNA models were built de novo. The combined ternary complex model was refined using the real-space refinement and rigid body fit tools in Coot v0.9.4.1. Finally, the model was subject to a round of real_space_refine tool in Phenix v 1.19.2-4158-000, using secondary structure, Ramachandran, and rotamer restraints.

Data deposition and figure preparation.

Cryo-EM maps and model coordinates were deposited to the EMDB (code EMD-NNNNN) and PDB (code NNNN). The structure figures were generated in UCSF ChimeraX v1.2.5. Cryo-EM map σ levels were calculated as: map level/root mean square deviation from zero. The orientation distribution plots were either obtained from CryoSPARC or generated using pyem csparc2star.py and star2bild.py programs. Map versus model Fourier shell correlation (FSC) graphs were calculated in Mtriage, as implemented in Phenix. Gold standard FSC plot was generated in cryoSPARC.

6 Contributed work

6.1 Summary

Throughout my time at Berkeley, I have had the pleasure of collaborating with colleagues and providing insights or preliminary data for their projects, or helped generate interesting ideas, hypotheses, and valuable computational and intellectual contributions for fellow lab members that have instigated fulfilling collaborations. These projects revolved around the biology of CRISPR-Cas and antiCRISPR systems, host-phage interactions, structural biology of phage proteins, or regional wastewater surveillance and analysis of SARS-CoV-2 viral strains. The latter included the collaborative development and testing of economical methods for coronavirus surveillance that can be developed in-house. As a result, I am a co-author on the following manuscripts at the conclusion of this PhD.

6.2 List of Publications

21. Joy Wang, Petr Skopintsev, **Basem Al-Shayeb**, Jennifer A. Doudna. “Molecular mechanisms of prespacer processing during generation of CRISPR immunological memory”. *In prep*.

20. **Basem Al-Shayeb,** Petr Skopintsev, Kasia Soczek, Elizabeth Stahl, Zheng Li, Evan Groover, Dylan Smock, Amy Eggers, Patrick Pausch, Brady Cress, Carolyn Huang, Brian Staskawicz, David Savage, Steven Jacobsen, Jillian F. Banfield, Jennifer A. Doudna. “Diverse virus-encoded CRISPR-Cas systems include streamlined genome editors”. *In prep*.

19. Robert C. Edgar*, Jeff Taylor*, Tomer Altman*, Pierre Barbera*, Dmitry Meleshko*, Victor Lin*, Dan Lohr*, Gherman Novakovsky*, **Basem Al-Shayeb***, Jillian F. Banfield*, Anton Korobeynikov*, Rayan Chikhi*, and Artem Babaian*. 2020. "Petabase-scale sequence alignment catalyses viral discovery". *Nature.*

18. Adair L Borges, Yue Clare Lou, Rohan Sachdeva, **Basem Al-Shayeb**, Alexander L. Jaffe, Shufei Lei, Joanne M. Santini, Jillian F. Banfield. "Stop codon recoding is widespread in diverse phage lineages and has the potential to regulate translation of late stage and lytic genes". *BioRxiv*.

17. Jacob A. West-Roberts, Paula B. Matheus-Carnevali, **Basem Al-Shayeb**, Marie Charlotte Schoelmerich, Alex D. Thomas, Allison Sharrar, Christine He, Lin-Xing Chen, Adi Lavy, Ray Keren, Yuki Amano, Jillian F Banfield. “The Chloroflexi supergroup is metabolically diverse and representatives have novel genes for non-photosynthesis based CO2 fixation”. *BioRxiv*.

16. **Basem Al-Shayeb**, Marie C Schoelmerich, Jacob West-Roberts, Luis E. Valentin-Alvarado, Rohan Sachdeva, Susan Mullen, Alexander Crits-Christoph, Michael J. Wilkins, Kenneth H. Williams, Jennifer Doudna, Jillian F. Banfield. 2021. "Borgs are giant extrachromosomal elements with the potential to augment methane oxidation". *BioRxiv.*

15. Alexander Crits-Christoph, Spencer Diamond, **Basem Al-Shayeb**, Luis Valentin-Alvarado, and Jillian F. Banfield. 2021. "A widely distributed genus of soil Acidobacteria genomically enriched in biosynthetic gene clusters". *BioRxiv*.

14. Hannah D. Greenwald, Lauren L. Kennedy, Adrian Hinkle, Oscar Whitney, Vinson Fan, Alexander Crits-Christoph, Sasha Harris-Lovett, Avi Flamholz, **Basem Al-Shayeb**, L. Liao, M. Beyers, D. Brown, A. Chakrabarti, J. Dow, D. Frost, M. Koekemer, C. Lynch, P. Sarkar, E. White, R. Kantor, Kara L. Nelson. 2021. "Interpretation of spatial and temporal trends of SARS-CoV-2 RNA in San Francisco Bay Area wastewater". *Water Research X.*

13. Patrick Pausch*, Kasia M. Soczek*, D. A. Herbst, **Basem Al-Shayeb**, Jillian F. Banfield, Eva Nogales, Jennifer A. Doudna. "DNA interference states of the hypercompact CRISPR-CasΦ effector". 2021. *Nature Structural & Molecular Biology.*

12. Joy Y. Wang, Christopher M. Hoel, **Basem Al-Shayeb**, Jillian F. Banfield, Stephen G. Brohawn, Jennifer A. Doudna. "Structural coordination between active sites of a Cas6-reverse transcriptase-Cas1—Cas2 CRISPR integrase complex". 2021. *Nature Communications*

11. Kris Saha, Erik J. Sontheimer, P. J. Brooks, ... , & **The SCGE Consortium**. 2021. "The NIH Somatic Cell Genome Editing program". *Nature*.

10. Oscar N. Whitney, Lauren Kennedy, Vinson Fan, Adrian Hinkle, Rose Kantor, Hannah Greenwald, Alexander Crits-Christoph, **Basem Al-Shayeb**, Mira Chaplin, Anna Maurer, Robert Tjian, Kara L. Nelson. 2021. "Sewage, Salt, Silica and SARS-CoV-2 (4S): An economical kit-free method for direct capture of SARS-CoV-2 RNA from wastewater". *Environmental Science & Technology.*

9. Alexander Crits-Christoph, Rose S. Kantor, Matthew R. Olm, Oscar N. Whitney, **Basem Al-Shayeb**, Yue C. Lou, Avi Flamholz, Lauren C. Kennedy, Hannah Greenwald, Adrian Hinkle, Jonathan Hetzel, Sara Spitzer, Jeffery Koble, Asako Tan, Fred Hyde, Gary Schroth, Scott Kuersten, Jillian F. Banfield, and Kara L. Nelson. 2021. "Genome sequencing of sewage detects regionally prevalent SARS-CoV-2 variants". *mBio*.

8. Patrick Pausch*, **Basem Al-Shayeb*,** Ezra Bisom-Rapp, Connor A. Tsuchida, Zheng Li, Brady F. Cress, Gavin J. Knott, Steven E. Jacobsen, Jillian F. Banfield, Jennifer. A. Doudna. 2020. "CRISPR-CasΦ from huge phages is a hypercompact genome editor". *Science*. https://doi.org/10.1126/science.abb1400

7. Lucas B. Harrington*, Enbo Ma*, Janice S. Chen, Isaac P. Witte, Dov Gertz, David Paez-Espino, **Basem Al-Shayeb**, Nikos C. Krypides, David Burstein, Jillian F.

Banfield, Jennifer A. Doudna. 2020. "A scoutRNA Is Required for Some Type V CRISPR-Cas Systems." *Molecular Cell*. https://doi.org/10.1016/j.molcel.2020.06.022

6. **Basem Al-Shayeb***, Rohan Sachdeva*, Lin-Xing Chen, Fred Ward, Patrick Munk, Audra Devoto, Cindy J. Castelle, Matthew R. Olm, Keith Bouma-Gregson, Yuki Amano, Christine He, Raphaël Méheust, Brandon Brooks, Alex Thomas, Adi Lavy, Paula Matheus-Carnevali, Christine Sun, Daniela S. A. Goltsman, Mikayla A. Borton, Allison Sharrar, Alexander L. Jaffe, Tara C. Nelson, Rose Kantor, Ray Keren, Katherine R. Lane, Ibrahim F. Farag, Shufei Lei, Kari Finstad, Ronald Amundson, Karthik Anantharaman, Jinglie Zhou, Alexander J. Probst, Mary E. Power, Susannah G. Tringe, Wen-Jun Li, Kelly Wrighton, Sue Harrison, Michael Morowitz, David A. Relman, Jennifer A. Doudna, Anne-Catherine Lehours, Lesley Warren, Jamie H. D. Cate, Joanne M. Santini, & Jillian F. Banfield. 2020. “Clades of Huge Phages from across Earth’s Ecosystems.” *Nature*, February. https://doi.org/10.1038/s41586-020-2007-4.

5. Gavin J. Knott, Brady F. Cress, Jun-Jie Liu, Brittney W. Thornton, Rachel J. Lew, **Basem Al-Shayeb**, Daniel J. Rosenberg, Michal Hammel, Benjamin A Adler, Marco J Lobba, Michael Xu, Adam P Arkin, Christof Fellmann, Jennifer A Doudna. 2019. “Structural Basis for AcrVA4 Inhibition of Specific CRISPR-Cas12a.” *eLife* 8 (August): e49110. http://doi.org/10.7554/eLife.49110

4. Lin-Xing Chen, **Basem Al-Shayeb**, Raphaël Méheust, Wen-Jun Li, Jennifer A. Doudna, & Jillian F. Banfield. 2019. “Candidate Phyla Radiation Roizmanbacteria From Hot Springs Have Novel and Unexpectedly Abundant CRISPR-Cas Systems.” *Frontiers in Microbiology* 10: 928. https://doi.org/10.3389/fmicb.2019.00928

3. Gavin J. Knott, Brittney W. Thornton, Marco J. Lobba, Jun-Jie Liu, **Basem Al-Shayeb**, Kyle E. Watters, and Jennifer A. Doudna. 2019. “Broad-Spectrum Enzymatic Inhibition of CRISPR-Cas12a.” *Nature Structural & Molecular Biology*, April, 1. https://doi.org/10.1038/s41594-019-0208-z

2. Jun-Jie Liu*, Natalia Orlova*, Benjamin L. Oakes*, Enbo Ma, Hannah B. Spinner, Katherine L. M. Baney, Jonathan Chuck, Dan Tan, Gavin J. Knott, Lucas B. Harrington, **Basem Al-Shayeb**, Alexander Wagner, Julian Brötzmann, Brett T. Staahl, Kian L. Taylor, John Desmarais, Eva Nogales & Jennifer A. Doudna. (2019). “CasX Enzymes Comprise a Distinct Family of RNA-Guided Genome Editors.” *Nature*, February, 1. https://doi.org/10.1038/s41586-019-0908-x

1. Aunica L Kane, **Basem Al-Shayeb**, Patrick V Holec, Srijay Rajan, Nicholas E Le Mieux, Stephen C Heinsch, Sona Psarska, Kelly G Aukema, Casim A Sarkar, Edward A Nater, Jeffrey A Gralnick (2016). "Toward Bioremediation of Methylmercury Using Silica Encapsulated Escherichia coli Harboring the mer Operon." *Plos One* 11,1. https://doi.org/10.1371/journal.pone.0147036

Protocols

2. Oscar N Whitney, **Basem Al-Shayeb**, Alex Crits-Cristoph, Mira Chaplin, Vinson Fan, Hannah Greenwald, Adrian Hinkle, Rose Kantor, Lauren Kennedy, Anna Maurer,

Robert Tjian, Kara L. Nelson, UC Berkeley Wastewater-based epidemiology consortium 2020. Direct wastewater RNA extraction via the "Milk of Silica (MoS)" method - A companion method to "Sewage, Salt, Silica and SARS-CoV-2 (4S)". protocols.io. https://dx.doi.org/10.17504/protocols.io.biwfkfbn

1. Oscar N Whitney, **Basem Al-Shayeb**, Alex Crits-Cristoph, Mira Chaplin, Vinson Fan, Hannah Greenwald, Adrian Hinkle, Rose Kantor, Lauren Kennedy, Anna Maurer, Robert Tjian, Kara L. Nelson, UC Berkeley Wastewater-based epidemiology consortium 2020. V.2 - Direct wastewater RNA capture and purification via the "Sewage, Salt, Silica and SARS-CoV-2 (4S)" method. protocols.io. https://dx.doi.org/10.17504/protocols.io.bjr9km96

7 Concluding Remarks

We show that phages with huge genomes are widespread across Earth's ecosystems. We manually completed 35 genomes, distinguishing them from prophage, providing accurate genome lengths and complete inventories of genes, including those encoded in complex repeat regions that break automated assemblies. Even closely related phages have diversified across habitats. Host and phage migration could transfer genes relevant in medicine and agriculture (e.g., pathogenicity factors and antibiotic resistance, SI). Additional medical significance could involve direct or indirect activation of immune responses. For example, some phages directly stimulate IFN-g via a TLR9-dependent pathway and exacerbate colitis(Gogokhia et al., 2019). Huge phage may represent a reservoir of novel nucleic acid manipulation tools with applications in genome editing and might be harnessed to improve human and animal health. For instance, huge phages equipped with CRISPR-Cas systems might be tamed and used to modulate bacterial microbiome function or eliminate unwanted bacteria.

The huge phages define massive clades, suggesting that a gene inventory comparable in size to those of many symbiotic bacteria is a conserved strategy for phage survival. Overall, their genes appear to redirect the host's protein production capacity to favor phage genes by first intercepting the earliest steps of translation and then ensuring efficient protein production thereafter. These inferences are aligned with findings for some eukaryotic viruses, which control every phase of protein synthesis(Jaafar and Kieft, 2019). Some acquired CRISPR-Cas systems with unusual compositions that may function to control host genes and eliminate competing phages.

More broadly, huge phages represent little-known biology, the platforms for which are distinct from those of small phages and partially analogous to those of symbiotic bacteria, somewhat blurring the distinctions between life and non-life. Given phylogenetic evidence for large radiations of huge phages, we wonder if they are ancient and arose simultaneously with free-living cells, their symbionts, and other phages from a pre-life (protogenote) state (Woese, 1998) rather than appearing more recently via episodes of genome expansion.

Three well-characterized Cas enzymes Cas9, Cas12a, and CasX, use one (Cas12a and CasX) or two active sites (Cas9) for DNA cutting and rely on a separate active site (Cas12a) or additional factors (CasX and Cas9) for crRNA processing (Fig. 4C). The finding that a single RuvC active site in Chapter 2's CasΦ from huge phages is capable of crRNA processing and DNA cutting suggests that size limitations of phage genomes, possibly in combination with large population sizes and higher mutation rates in phages compared to prokaryotes (Duffy et al., 2008; Lee and Marx, 2012; Lynch, 2006), led to a consolidation of chemistries within one catalytic center. Further work in Chapter 5 shows that such compact proteins may be particularly amenable to engineering and laboratory evolution to create new functionalities for genome manipulation, and highlight phages as an exciting forefront for discovery and biotechnological applications for human health.

Chapter 3's Borgs are enigmatic extrachromosomal elements that can approach (and likely exceed) 1 Mbp in length. We can neither prove that they are archaeal viruses or plasmids or mini-chromosomes, nor can we prove that they are not. Although they may ultimately be classified as megaplasmids, they are clearly different from anything that has been reported previously. It is fascinating to ponder their possible evolutionary origins. Borg homologous recombination may indicate movement among hosts, thus their possible roles as gene transfer agents. It has been noted that *Methanoperedens spp.* have been particularly open to gene acquisition from diverse bacteria and archaea(Leu et al., 2020), and Borgs may have contributed to this. The existence of Borgs encoding MCR demonstrates for the first time that MCR and MCR-like proteins for metabolism of methane and short-chain hydrocarbons can exist on extrachromosomal elements and thus could potentially be dispersed across lineages, as is inferred to have occurred several times over the course of archaeal evolution(Boyd et al., 2019; Hua et al., 2019). Borgs carry numerous metabolic genes, some of which produce variants of *Methanoperedens spp.* proteins that could have distinct biophysical and biochemical properties. Assuming that these genes either augment *Methanoperedens spp.* energy metabolism or extend the conditions under which they can function, Borgs may have far-reaching biogeochemical consequences, with important and unanticipated climate implications.

Since the completion of the human genome, growth of DNA sequencing databases has outpaced Moore's Law. Chapter 4's Serratus provides rapid and focused access to genomic sequences captured over more than a decade by the global research community which would otherwise be inaccessible in practice. This work and further extensions of petabase scale genomics [15, 16, 44] are shaping a new era in computational biology, enabling expansive gene discovery, pathogen surveillance, and pangenomic evolutionary analyses.

References

Al-Shayeb, B., Sachdeva, R., Chen, L.-X., Ward, F., Munk, P., Devoto, A., Castelle, C.J., Olm, M.R., Bouma-Gregson, K., Amano, Y., et al. (2020). Clades of huge phages from across Earth's ecosystems. Nature.

Altman, T., Travers, M., Kothari, A., Caspi, R., and Karp, P.D. (2013). A systematic comparison of the MetaCyc and KEGG pathway databases. BMC Bioinformatics *14*, 112.

Altschul, S.F., Gish, W., Miller, W., Myers, E.W., and Lipman, D.J. (1990). Basic local alignment search tool. J. Mol. Biol. *215*, 403–410.

Altschul, S.F., Madden, T.L., Schäffer, A.A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D.J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs Nucleic Acids Res 25: 3389--3402.

Anantharaman, K., Duhaime, M.B., Breier, J.A., Wendt, K.A., Toner, B.M., and Dick, G.J. (2014). Sulfur oxidation genes in diverse deep-sea viruses. Science *344*, 757–760.

Anthony, S.J., Epstein, J.H., Murray, K.A., Navarrete-Macias, I., Zambrana-Torrelio, C.M., Solovyov, A., Ojeda-Flores, R., Arrigo, N.C., Islam, A., Ali Khan, S., et al. (2013). A strategy to estimate unknown viral diversity in mammals. MBio *4*, e00598–13.

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

Ausiannikava, D., Mitchell, L., Marriott, H., Smith, V., Hawkins, M., Makarova, K.S., Koonin, E.V., Nieduszynski, C.A., and Allers, T. (2018). Evolution of Genome Architecture in Archaea: Spontaneous Generation of a New Chromosome in Haloferax volcanii. Mol. Biol. Evol. *35*, 1855–1868.

Babaian, A., and Edgar, R.C. (2021). Ribovirus classification by a polymerase barcode sequence.

Baker, D., van den Beek, M., Blankenberg, D., Bouvier, D., Chilton, J., Coraor, N., Coppens, F., Eguinoa, I., Gladman, S., Grüning, B., et al. (2020). No more business as usual: Agile and effective responses to emerging pathogen threats require open data and open analytics. PLoS Pathog. *16*, e1008643.

Balcazar, J.L. (2014). Bacteriophages as vehicles for antibiotic resistance genes in the environment. PLoS Pathog. *10*, e1004219.

Barrangou, R., Fremaux, C., Deveau, H., Richards, M., Boyaval, P., Moineau, S., Romero, D.A., and Horvath, P. (2007). CRISPR provides acquired resistance against viruses in prokaryotes. Science *315*, 1709–1712.

Berger, W., Steiner, E., Grusch, M., Elbling, L., and Micksche, M. (2009). Vaults and the major vault protein: novel roles in signal pathway regulation and immunity. Cell. Mol. Life Sci. *66*, 43–61.

Bergner, L.M., Orton, R.J., Broos, A., Tello, C., Becker, D.J., Carrera, J.E., Patel, A.H., Biek, R., and Streicker, D.G. (2021). Diversification of mammalian deltaviruses by host shifting. Proc. Natl. Acad. Sci. U. S. A. *118*.

Bernardes, J.S., Vieira, F.R.J., Zaverucha, G., and Carbone, A. (2016). A multi-objective optimization approach accurately resolves protein domain architectures. Bioinformatics *32*, 345–353.

Biswas, A., Staals, R.H.J., Morales, S.E., Fineran, P.C., and Brown, C.M. (2016). CRISPRDetect: A flexible algorithm to define CRISPR arrays. BMC Genomics *17*, 356.

Boetius, A., Ravenschlag, K., Schubert, C.J., Rickert, D., Widdel, F., Gieseke, A., Amann, R., Jørgensen, B.B., Witte, U., and Pfannkuche, O. (2000). A marine microbial consortium apparently mediating anaerobic oxidation of methane. Nature *407*, 623–626.

Bolduc, B., Jang, H.B., Doulcier, G., You, Z.-Q., Roux, S., and Sullivan, M.B. (2017). vConTACT: an iVirus tool to classify double-stranded DNA viruses that infect Archaea and Bacteria. PeerJ *5*, e3243.

Bondy-Denomy, J., Garcia, B., Strum, S., Du, M., Rollins, M.F., Hidalgo-Reyes, Y., Wiedenheft, B., Maxwell, K.L., and Davidson, A.R. (2015). Multiple mechanisms for CRISPR-Cas inhibition by anti-CRISPR proteins. Nature *526*, 136–139.

Boyd, J.A., Jungbluth, S.P., Leu, A.O., Evans, P.N., Woodcroft, B.J., Chadwick, G.L., Orphan, V.J., Amend, J.P., Rappé, M.S., and Tyson, G.W. (2019). Divergent methyl-coenzyme M reductase genes in a deep-subseafloor Archaeoglobi. ISME J. *13*, 1269–1279.

Bradley, P., Den Bakker, H.C., Rocha, E.P.C., McVean, G., and Iqbal, Z. Real-time search of all bacterial and viral genomic data.

Breitbart, M., Bonnain, C., Malki, K., and Sawaya, N.A. (2018). Phage puppet masters of the marine microbial realm. Nat Microbiol *3*, 754–766.

Brown, K.L., and Hughes, K.T. (1995). The role of anti-sigma factors in gene

regulation. Mol. Microbiol. *16*, 397–404.

Brown, C.T., Olm, M.R., Thomas, B.C., and Banfield, J.F. (2016). Measurement of bacterial replication rates in microbial communities. Nat. Biotechnol. *34*, 1256–1263.

Brown-Jaque, M., Rodriguez Oyarzun, L., Cornejo-Sánchez, T., Martín-Gómez, M.T., Gartner, S., de Gracia, J., Rovira, S., Alvarez, A., Jofre, J., González-López, J.J., et al. (2018). Detection of Bacteriophage Particles Containing Antibiotic Resistance Genes in the Sputum of Cystic Fibrosis Patients. Front. Microbiol. *9*, 856.

Buchfink, B., Xie, C., and Huson, D.H. (2015). Fast and sensitive protein alignment using DIAMOND. Nat. Methods *12*, 59–60.

Buchfink, B., Reuter, K., and Drost, H.-G. (2021). Sensitive protein alignments at tree-of-life scale using DIAMOND. Nat. Methods *18*, 366–368.

Buels, R., Yao, E., Diesh, C.M., Hayes, R.D., Munoz-Torres, M., Helt, G., Goodstein, D.M., Elsik, C.G., Lewis, S.E., Stein, L., et al. (2016). JBrowse: a dynamic web platform for genome visualization and analysis. Genome Biol. *17*, 66.

Bukhari, K., Mulley, G., Gulyaeva, A.A., Zhao, L., Shu, G., Jiang, J., and Neuman, B.W. (2018). Description and initial characterization of metatranscriptomic nidovirus-like genomes from the proposed new family Abyssoviridae, and from a sister group to the Coronavirinae, the proposed genus Alphaletovirus. Virology *524*, 160–171.

Burstein, D., Harrington, L.B., Strutt, S.C., Probst, A.J., Anantharaman, K., Thomas, B.C., Doudna, J.A., and Banfield, J.F. (2017). New CRISPR-Cas systems from uncultivated microbes. Nature *542*, 237–241.

Bushmanova, E., Antipov, D., Lapidus, A., and Prjibelski, A.D. (2019). rnaSPAdes: a de novo transcriptome assembler and its application to RNA-Seq data. GigaScience *8*.

Bushnell, B. (2014a). BBTools software package. URL Http://sourceforge. Net/projects/bbmap *578*, 579.

Bushnell, B. (2014b). BBMap: A fast, accurate, splice-aware aligner (Lawrence Berkeley National Lab. (LBNL), Berkeley, CA (United States)).

Bushnell, B. (2016). BBMap short read aligner.

Cai, C., Leu, A.O., Xie, G.-J., Guo, J., Feng, Y., Zhao, J.-X., Tyson, G.W., Yuan,

Z., and Hu, S. (2018). A methanotrophic archaeon couples anaerobic oxidation of methane to Fe(III) reduction. ISME J. *12*, 1929–1939.

Camarillo-Guerrero, L.F., Almeida, A., Rangel-Pineros, G., Finn, R.D., and Lawley, T.D. (2021). Massive expansion of human gut bacteriophage diversity. Cell *184*, 1098–1109.e9.

Carroll, D., Daszak, P., Wolfe, N.D., Gao, G.F., Morel, C.M., Morzaria, S., Pablos-Méndez, A., Tomori, O., and Mazet, J.A.K. (2018). The Global Virome Project. Science *359*, 872–874.

Castelle, C.J., Brown, C.T., Anantharaman, K., Probst, A.J., Huang, R.H., and Banfield, J.F. (2018). Biosynthetic capacity, metabolic variety and unusual biology in the CPR and DPANN radiations. Nat. Rev. Microbiol. *16*, 629–645.

Chaikeeratisak, V., Nguyen, K., Khanna, K., Brilot, A.F., Erb, M.L., Coker, J.K.C., Vavilina, A., Newton, G.L., Buschauer, R., Pogliano, K., et al. (2017a). Assembly of a nucleus-like structure during viral replication in bacteria. Science *355*, 194–197.

Chaikeeratisak, V., Nguyen, K., Egan, M.E., Erb, M.L., Vavilina, A., and Pogliano, J. (2017b). The Phage Nucleus and Tubulin Spindle Are Conserved among Large Pseudomonas Phages. Cell Rep. *20*, 1563–1571.

Chang, W.-S., Pettersson, J.H.-O., Le Lay, C., Shi, M., Lo, N., Wille, M., Eden, J.-S., and Holmes, E.C. (2019). Novel hepatitis D-like agents in vertebrates and invertebrates. Virus Evol *5*, vez021.

Chase, J.M., Blowes, S.A., Knight, T.M., Gerstner, K., and May, F. (2020). Ecosystem decay exacerbates biodiversity loss with habitat loss. Nature *584*, 238–243.

Chen, I.-M.A., Chu, K., Palaniappan, K., Ratner, A., Huang, J., Huntemann, M., Hajek, P., Ritter, S., Varghese, N., Seshadri, R., et al. (2021). The IMG/M data management and analysis system v.6.0: new tools and advanced capabilities. Nucleic Acids Res. *49*, D751–D763.

Chen, J.S., Ma, E., Harrington, L.B., Da Costa, M., Tian, X., Palefsky, J.M., and Doudna, J.A. (2018). CRISPR-Cas12a target binding unleashes indiscriminate single-stranded DNase activity. Science *360*, 436–439.

Chen, L.-X., Méheust, R., Crits-Christoph, A., McMahon, K.D., Nelson, T.C., Slater, G.F., Warren, L.A., and Banfield, J.F. (2020). Large freshwater phages with the potential to augment aerobic methane oxidation. Nat Microbiol *5*, 1504–1515.

Cofsky, J.C., Karandur, D., Huang, C.J., Witte, I.P., Kuriyan, J., and Doudna, J.A. (2020). CRISPR-Cas12a exploits R-loop asymmetry to form double-strand breaks.

Cole, C., Barber, J.D., and Barton, G.J. (2008). The Jpred 3 secondary structure prediction server. Nucleic Acids Res. *36*, W197–W201.

Coordinators, N.R., and NCBI Resource Coordinators (2012). Database resources of the National Center for Biotechnology Information. Nucleic Acids Research *41*, D8–D20.

Czech, L., Barbera, P., and Stamatakis, A. (2020). Genesis and Gappa: processing, analyzing and visualizing phylogenetic (placement) data. Bioinformatics *36*, 3263–3265.

Darling, A.C.E., Mau, B., Blattner, F.R., and Perna, N.T. (2004). Mauve: multiple alignment of conserved genomic sequence with rearrangements. Genome Res. *14*, 1394–1403.

Darling, A.E., Mau, B., and Perna, N.T. (2010). progressiveMauve: multiple genome alignment with gene gain, loss and rearrangement. PLoS One *5*, e11147.

Darriba, D., Posada, D., Kozlov, A.M., Stamatakis, A., Morel, B., and Flouri, T. (2020). ModelTest-NG: A New and Scalable Tool for the Selection of DNA and Protein Evolutionary Models. Molecular Biology and Evolution *37*, 291–294.

DasSarma, S., Capes, M., and DasSarma, P. (2009). Haloarchaeal Megaplasmids. In Microbial Megaplasmids, E. Schwartz, ed. (Berlin, Heidelberg: Springer Berlin Heidelberg), pp. 3–30.

Delcher, A.L., Harmon, D., Kasif, S., White, O., and Salzberg, S.L. (1999). Improved microbial gene identification with GLIMMER. Nucleic Acids Res. *27*, 4636–4641.

Deltcheva, E., Chylinski, K., Sharma, C.M., Gonzales, K., Chao, Y., Pirzada, Z.A., Eckert, M.R., Vogel, J., and Charpentier, E. (2011). CRISPR RNA maturation by trans-encoded small RNA and host factor RNase III. Nature *471*, 602–607.

Devoto, A.E., Santini, J.M., Olm, M.R., Anantharaman, K., Munk, P., Tung, J., Archie, E.A., Turnbaugh, P.J., Seed, K.D., Blekhman, R., et al. (2019). Megaphages infect Prevotella and variants are widespread in gut microbiomes. Nat Microbiol.

Duffy, S., Shackelton, L.A., and Holmes, E.C. (2008). Rates of evolutionary change in viruses: patterns and determinants. Nat. Rev. Genet. *9*, 267–276.

Dutilh, B.E., Jurgelenaite, R., Szklarczyk, R., van Hijum, S.A.F.T., Harhangi, H.R., Schmid, M., de Wild, B., Françoijs, K.-J., Stunnenberg, H.G., Strous, M., et al. (2011). FACIL: Fast and Accurate Genetic Code Inference and Logo. Bioinformatics *27*, 1929–1933.

East-Seletsky, A., O'Connell, M.R., Knight, S.C., Burstein, D., Cate, J.H.D., Tjian, R., and Doudna, J.A. (2016). Two Distinct RNase Activities of CRISPR-C2c2 Enable Guide RNA Processing and RNA Detection. Nature *538*, 270–273.

Eddy, S.R. (2011). Accelerated Profile HMM Searches. PLoS Comput. Biol. *7*, e1002195.

Edgar, R. (2015). USEARCH: ultra-fast sequence analysis.

Edgar, R.C. (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Research *32*, 1792–1797.

Edgar, R.C. (2010). Search and clustering orders of magnitude faster than BLAST. Bioinformatics *26*, 2460–2461.

Edgar, R.C. UCHIME2: improved chimera prediction for amplicon sequencing.

Emerson, J.B., Roux, S., Brum, J.R., Bolduc, B., Woodcroft, B.J., Jang, H.B., Singleton, C.M., Solden, L.M., Naas, A.E., Boyd, J.A., et al. (2018). Host-linked soil viral ecology along a permafrost thaw gradient. Nat Microbiol *3*, 870–880.

Enright, A.J., Van Dongen, S., and Ouzounis, C.A. (2002). An efficient algorithm for large-scale detection of protein families. Nucleic Acids Res. *30*, 1575–1584.

Ettwig, K.F., Zhu, B., Speth, D., Keltjens, J.T., Jetten, M.S.M., and Kartal, B. (2016). Archaea catalyze iron-dependent anaerobic oxidation of methane. Proc. Natl. Acad. Sci. U. S. A. *113*, 12792–12796.

Farwell, M.A., Roberts, M.W., and Rabinowitz, J.C. (1992). The effect of ribosomal protein S1 from Escherichia coli and Micrococcus luteus on protein synthesis in vitro by E. coli and Bacillus subtilis. Mol. Microbiol. *6*, 3375–3383.

Felsenstein, J. (1985). CONFIDENCE LIMITS ON PHYLOGENIES: AN APPROACH USING THE BOOTSTRAP. Evolution *39*, 783–791.

Finn, R.D., Bateman, A., Clements, J., Coggill, P., Eberhardt, R.Y., Eddy, S.R., Heger, A., Hetherington, K., Holm, L., Mistry, J., et al. (2014). Pfam: the protein families database. Nucleic Acids Res. *42*, D222–D230.

Flores, R., Gago-Zachert, S., Serra, P., Sanjuán, R., and Elena, S.F. (2014). Viroids: Survivors from the RNA World? Annual Review of Microbiology *68*, 395–

414.

Fonfara, I., Richter, H., Bratovič, M., Le Rhun, A., and Charpentier, E. (2016). The CRISPR-associated DNA-cleaving enzyme Cpf1 also processes precursor CRISPR RNA. Nature *532*, 517–521.

Frank, J.A., Lorimer, D., Youle, M., Witte, P., Craig, T., Abendroth, J., Rohwer, F., Edwards, R.A., Segall, A.M., and Burgin, A.B., Jr (2013). Structure and function of a cyanophage-encoded peptide deformylase. ISME J. *7*, 1150–1160.

Gleditzsch, D., Pausch, P., Müller-Esparza, H., Özcan, A., Guo, X., Bange, G., and Randau, L. (2019). PAM identification by CRISPR-Cas effector complexes: diversified mechanisms and structures. RNA Biology *16*, 504–517.

Gogokhia, L., Buhrke, K., Bell, R., Hoffman, B., Brown, D.G., Hanke-Gogokhia, C., Ajami, N.J., Wong, M.C., Ghazaryan, A., Valentine, J.F., et al. (2019). Expansion of Bacteriophages Is Linked to Aggravated Intestinal Inflammation and Colitis. Cell Host Microbe *25*, 285–299.e8.

Gootenberg, J.S., Abudayyeh, O.O., Lee, J.W., Essletzbichler, P., Dy, A.J., Joung, J., Verdine, V., Donghia, N., Daringer, N.M., Freije, C.A., et al. (2017). Nucleic acid detection with CRISPR-Cas13a/C2c2. Science *356*, 438–442.

Grazziotin, A.L., Koonin, E.V., and Kristensen, D.M. (2017). Prokaryotic Virus Orthologous Groups (pVOGs): a resource for comparative genomics and protein family annotation. Nucleic Acids Res. *45*, D491–D498.

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

Haft, D.H., Selengut, J.D., Richter, R.A., Harkins, D., Basu, M.K., and Beck, E. (2013). TIGRFAMs and Genome Properties in 2013. Nucleic Acids Res. *41*, D387–D395.

Hall, J.P.J., Botelho, J., Cazares, A., and Baltrus, D.A. (2022). What makes a megaplasmid? Philos. Trans. R. Soc. Lond. B Biol. Sci. *377*, 20200472.

Hallam, S.J., Girguis, P.R., Preston, C.M., Richardson, P.M., and DeLong, E.F. (2003). Identification of methyl coenzyme M reductase A (mcrA) genes associated with methane-oxidizing archaea. Appl. Environ. Microbiol. *69*, 5483–5491.

Hanson, R.S., and Hanson, T.E. (1996). Methanotrophic bacteria. Microbiol. Rev. *60*, 439–471.

Harrington, L.B., Burstein, D., Chen, J.S., Paez-Espino, D., Ma, E., Witte, I.P., Cofsky, J.C., Kyrpides, N.C., Banfield, J.F., and Doudna, J.A. (2018). Programmed DNA destruction by miniature CRISPR-Cas14 enzymes. Science *362*, 839–842.

Hauser, M., Steinegger, M., and Söding, J. (2016). MMseqs software suite for fast and deep clustering and searching of large protein sequence sets. Bioinformatics *32*, 1323–1330.

Heider, J., Szaleniec, M., Sünwoldt, K., and Boll, M. (2016). Ethylbenzene Dehydrogenase and Related Molybdenum Enzymes Involved in Oxygen-Independent Alkyl Chain Hydroxylation. J. Mol. Microbiol. Biotechnol. *26*, 45–62.

Hille, F., Richter, H., Wong, S.P., Bratovič, M., Ressel, S., and Charpentier, E. (2018). The Biology of CRISPR-Cas: Backward and Forward. Cell *172*, 1239–1259.

Hua, Z.-S., Wang, Y.-L., Evans, P.N., Qu, Y.-N., Goh, K.M., Rao, Y.-Z., Qi, Y.-L., Li, Y.-X., Huang, M.-J., Jiao, J.-Y., et al. (2019). Insights into the ecological roles and evolution of methyl-coenzyme M reductase-containing hot spring Archaea. Nat. Commun. *10*, 4574.

Huang, Y., Niu, B., Gao, Y., Fu, L., and Li, W. (2010). CD-HIT Suite: a web server for clustering and comparing biological sequences. Bioinformatics *26*, 680–682.

Hug, L.A., Thomas, B.C., Brown, C.T., Frischkorn, K.R., Williams, K.H., Tringe, S.G., and Banfield, J.F. (2015). Aquifer environment selects for microbial species cohorts in sediment and groundwater. ISME J. *9*, 1846–1856.

Hunt, M., Gall, A., Ong, S.H., Brener, J., Ferns, B., Goulder, P., Nastouli, E., Keane, J.A., Kellam, P., and Otto, T.D. (2015). IVA: accurate*de novo*assembly of RNA virus genomes. Bioinformatics *31*, 2374–2376.

Hyatt, D., Chen, G.-L., Locascio, P.F., Land, M.L., Larimer, F.W., and Hauser, L.J. (2010). Prodigal: prokaryotic gene recognition and translation initiation site identification. BMC Bioinformatics *11*, 119.

Hyatt, D., LoCascio, P.F., Hauser, L.J., and Uberbacher, E.C. (2012). Gene and translation initiation site prediction in metagenomic sequences. Bioinformatics *28*, 2223–2230.

Ivanova, N.N., Schwientek, P., Tripp, H.J., Rinke, C., Pati, A., Huntemann, M., Visel, A., Woyke, T., Kyrpides, N.C., and Rubin, E.M. (2014). Stop codon reassignments in the wild. Science *344*, 909–913.

Iwamoto, M., Shibata, Y., Kawasaki, J., Kojima, S., Li, Y.-T., Iwami, S.,

Muramatsu, M., Wu, H.-L., Wada, K., Tomonaga, K., et al. (2021). Identification of novel avian and mammalian deltaviruses provides new insights into deltavirus evolution. Virus Evol *7*, veab003.

Jaafar, Z.A., and Kieft, J.S. (2019). Viral RNA structure-based strategies to manipulate translation. Nat. Rev. Microbiol. *17*, 110–123.

Janssen, B.D., and Hayes, C.S. (2012). The tmRNA ribosome-rescue system. Adv. Protein Chem. Struct. Biol. *86*, 151–191.

Jinek, M., Chylinski, K., Fonfara, I., Hauer, M., Doudna, J.A., and Charpentier, E. (2012). A Programmable Dual-RNA–Guided DNA Endonuclease in Adaptive Bacterial Immunity. Science *337*, 816–821.

Johnson, C.K., Hitchens, P.L., Pandit, P.S., Rushmore, J., Evans, T.S., Young, C.C.W., and Doyle, M.M. (2020). Global shifts in mammalian population trends reveal key predictors of virus spillover risk. Proc. Biol. Sci. *287*, 20192736.

Joshi, N.A., and Fass, J.N. (2011). Sickle: A sliding-window, adaptive, quality-based trimming tool for FastQ files (Version 1.33)[Software].

Joshi, N., and Sickle, F.J. (2011). A sliding-window, adaptive, quality-based trimming tool for FastQ files.

Kanehisa, M., Sato, Y., Kawashima, M., Furumichi, M., and Tanabe, M. (2016). KEGG as a reference resource for gene and protein annotation. Nucleic Acids Res. *44*, D457–D462.

Karasikov, M., Mustafa, H., Danciu, D., Zimmermann, M., Barber, C., Rätsch, G., and Kahles, A. (2020). MetaGraph: Indexing and Analysing Nucleotide Archives at Petabase-scale (bioRxiv).

Karsenti, E., Acinas, S.G., Bork, P., Bowler, C., De Vargas, C., Raes, J., Sullivan, M., Arendt, D., Benzoni, F., Claverie, J.-M., et al. (2011). A holistic approach to marine eco-systems biology. PLoS Biol. *9*, e1001177.

Karvelis, T., Druteika, G., Bigelyte, G., Budre, K., Zedaveinyte, R., Silanskas, A., Kazlauskas, D., Venclovas, Č., and Siksnys, V. (2021). Transposon-associated TnpB is a programmable RNA-guided DNA endonuclease. Nature *599*, 692–696.

Katoh, K., and Standley, D.M. (2013). MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol. Biol. Evol. *30*, 772–780.

Katz, K.S., Shutov, O., Lapoint, R., Kimelman, M., Brister, J.R., and O'Sullivan, C. (2021). STAT: a fast, scalable, MinHash-based k-mer tool to assess

Sequence Read Archive next-generation sequence submissions. Genome Biol. *22*, 270.

Knott, G.J., and Doudna, J.A. (2018). CRISPR-Cas guides the future of genetic engineering. Science *361*, 866–869.

Knott, G.J., Thornton, B.W., Lobba, M.J., Liu, J.-J., Al-Shayeb, B., Watters, K.E., and Doudna, J.A. (2019). Broad-spectrum enzymatic inhibition of CRISPR-Cas12a. Nat. Struct. Mol. Biol. 1.

Koonin, E.V., and Dolja, V.V. (2014). Virus world as an evolutionary network of viruses and capsidless selfish elements. Microbiol. Mol. Biol. Rev. *78*, 278–303.

Koonin, E.V., Dolja, V.V., Krupovic, M., Varsani, A., Wolf, Y.I., Yutin, N., Zerbini, F.M., and Kuhn, J.H. (2020). Global Organization and Proposed Megataxonomy of the Virus World. Microbiol. Mol. Biol. Rev. *84*.

Kozlov, A.M., Darriba, D., Flouri, T., Morel, B., and Stamatakis, A. (2019). RAxML-NG: a fast, scalable and user-friendly tool for maximum likelihood phylogenetic inference. Bioinformatics *35*, 4453–4455.

Krogh, A., Larsson, B., von Heijne, G., and Sonnhammer, E.L. (2001). Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. J. Mol. Biol. *305*, 567–580.

Kurtz, S. (2003). The Vmatch large scale sequence analysis software. Ref Type: Computer Program *412*, 297.

Langmead, B., and Salzberg, S.L. (2012). Fast gapped-read alignment with Bowtie 2. Nat. Methods *9*, 357–359.

Laslett, D., and Canback, B. (2004). ARAGORN, a program to detect tRNA genes and tmRNA genes in nucleotide sequences. Nucleic Acids Res. *32*, 11–16.

Lawrence, J.G., and Ochman, H. (1997). Amelioration of bacterial genomes: rates of change and exchange. J. Mol. Evol. *44*, 383–397.

Lee, M.-C., and Marx, C.J. (2012). Repeated, selection-driven genome reduction of accessory genes in experimental populations. PLoS Genet. *8*, e1002651.

Lee, S., Sieradzki, E.T., Nicolas, A.M., Walker, R.L., Firestone, M.K., Hazard, C., and Nicol, G.W. (2021). Methane-derived carbon flow through host-virus trophic networks in soil.

Leinonen, R., Sugawara, H., Shumway, M., and on behalf of the International

Nucleotide Sequence Database Collaboration (2011). The Sequence Read Archive. Nucleic Acids Research *39*, D19–D21.

Lemoine, F., Domelevo Entfellner, J.-B., Wilkinson, E., Correia, D., Dávila Felipe, M., De Oliveira, T., and Gascuel, O. (2018). Renewing Felsenstein's phylogenetic bootstrap in the era of big data. Nature *556*, 452–456.

Letko, M., Seifert, S.N., Olival, K.J., Plowright, R.K., and Munster, V.J. (2020a). Bat-borne virus diversity, spillover and emergence. Nat. Rev. Microbiol. *18*, 461–471.

Letko, M., Marzi, A., and Munster, V. (2020b). Functional assessment of cell entry and receptor usage for SARS-CoV-2 and other lineage B betacoronaviruses. Nature Microbiology *5*, 562–569.

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

Leu, A.O., McIlroy, S.J., Ye, J., Parks, D.H., Orphan, V.J., and Tyson, G.W. (2020). Lateral Gene Transfer Drives Metabolic Flexibility in the Anaerobic Methane-Oxidizing Archaeal Family Methanoperedenaceae. MBio *11*.

Levi, K., Rynge, M., Abeysinghe, E., and Edwards, R.A. (2018). Searching the Sequence Read Archive using Jetstream and Wrangler. In Proceedings of the Practice and Experience on Advanced Research Computing, (New York, NY, USA: Association for Computing Machinery), pp. 1–7.

Li, D., Liu, C.-M., Luo, R., Sadakane, K., and Lam, T.-W. (2015). MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. Bioinformatics *31*, 1674–1676.

Li, D., Luo, R., Liu, C.-M., Leung, C.-M., Ting, H.-F., Sadakane, K., Yamashita, H., and Lam, T.-W. (2016). MEGAHIT v1.0: A fast and scalable metagenome assembler driven by advanced methodologies and community practices. Methods *102*, 3–11.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., and 1000 Genome Project Data Processing Subgroup (2009). The Sequence Alignment/Map format and SAMtools. Bioinformatics *25*, 2078–2079.

Li, W., Fu, L., Niu, B., Wu, S., and Wooley, J. (2012). Ultrafast clustering algorithms for metagenomic sequence analysis. Briefings in Bioinformatics *13*, 656–668.

Lindell, D., Sullivan, M.B., Johnson, Z.I., Tolonen, A.C., Rohwer, F., and

Chisholm, S.W. (2004). Transfer of photosynthesis genes to and from Prochlorococcus viruses. Proc. Natl. Acad. Sci. U. S. A. *101*, 11013–11018.

Lindell, D., Jaffe, J.D., Johnson, Z.I., Church, G.M., and Chisholm, S.W. (2005). Photosynthesis genes in marine viruses yield proteins during host infection. Nature *438*, 86–89.

Liu, J.-J., Orlova, N., Oakes, B.L., Ma, E., Spinner, H.B., Baney, K.L.M., Chuck, J., Tan, D., Knott, G.J., Harrington, L.B., et al. (2019a). CasX enzymes comprise a distinct family of RNA-guided genome editors. Nature *566*, 218–223.

Liu, Z., Zhu, Z., Yang, J., Wu, S., Liu, Q., Wang, M., Cheng, H., Yan, J., and Wang, L. (2019b). Domain-centric dissection and classification of prokaryotic poly(3-hydroxyalkanoate) synthases.

Lobry, J.R. (1996). Asymmetric substitution patterns in the two DNA strands of bacteria. Mol. Biol. Evol. *13*, 660–665.

Loveland, A.B., and Korostelev, A.A. (2018). Structural dynamics of protein S1 on the 70S ribosome visualized by ensemble cryo-EM. Methods *137*, 55–66.

Lowe, T.M., and Eddy, S.R. (1997). tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. Nucleic Acids Res. *25*, 955–964.

Luo, M.L., Mullis, A.S., Leenay, R.T., and Beisel, C.L. (2015). Repurposing endogenous type I CRISPR-Cas systems for programmable gene repression. Nucleic Acids Res. *43*, 674–681.

Lynch, M. (2006). Streamlining and simplification of microbial genome architecture. Annu. Rev. Microbiol. *60*, 327–349.

Makarova, K.S., Wolf, Y.I., Alkhnbashi, O.S., Costa, F., Shah, S.A., Saunders, S.J., Barrangou, R., Brouns, S.J.J., Charpentier, E., Haft, D.H., et al. (2015). An updated evolutionary classification of CRISPR-Cas systems. Nat. Rev. Microbiol. *13*, 722–736.

Makarova, K.S., Wolf, Y.I., Iranzo, J., Shmakov, S.A., Alkhnbashi, O.S., Brouns, S.J.J., Charpentier, E., Cheng, D., Haft, D.H., Horvath, P., et al. (2019). Evolutionary classification of CRISPR–Cas systems: a burst of class 2 and derived variants. Nat. Rev. Microbiol. 1–17.

Mali, P., Yang, L., Esvelt, K.M., Aach, J., Guell, M., DiCarlo, J.E., Norville, J.E., and Church, G.M. (2013). RNA-guided human genome engineering via Cas9. Science *339*, 823–826.

McGinn, J., and Marraffini, L.A. (2019). Molecular mechanisms of CRISPR–Cas spacer acquisition. Nat. Rev. Microbiol. *17*, 7–12.

McGlynn, S.E., Chadwick, G.L., Kempes, C.P., and Orphan, V.J. (2015). Single cell activity reveals direct electron transfer in methanotrophic consortia. Nature *526*, 531–535.

McWilliam, H., Li, W., Uludag, M., Squizzato, S., Park, Y.M., Buso, N., Cowley, A.P., and Lopez, R. (2013). Analysis Tool Web Services from the EMBL-EBI. Nucleic Acids Res. *41*, W597–W600.

Medema, M.H., Trefzer, A., Kovalchuk, A., van den Berg, M., Müller, U., Heijne, W., Wu, L., Alam, M.T., Ronning, C.M., Nierman, W.C., et al. (2010). The sequence of a 1.8-mb bacterial linear plasmid reveals a rich evolutionary reservoir of secondary metabolic pathways. Genome Biol. Evol. *2*, 212–224.

Meleshko, D., Mohimani, H., Tracanna, V., Hajirasouliha, I., Medema, M.H., Korobeynikov, A., and Pevzner, P.A. (2019). BiosyntheticSPAdes: reconstructing biosynthetic gene clusters from assembly graphs. Genome Research *29*, 1352–1362.

Meleshko, D., Hajirasouliha, I., and Korobeynikov, A. (2021). coronaSPAdes: from biosynthetic gene clusters to RNA viral assemblies. Bioinformatics.

Mendoza, S.D., Berry, J.D., Nieweglowska, E.S., Leon, L.M., Agard, D., and Bondy-Denomy, J. (2018). A nucleus-like compartment shields bacteriophage DNA from CRISPR-Cas and restriction nucleases.

Miller, A.K., Mifsud, J.C.O., Costa, V.A., Grimwood, R.M., Kitson, J., Baker, C., Brosnahan, C.L., Pande, A., Holmes, E.C., Gemmell, N.J., et al. (2021). Slippery when wet: cross-species transmission of divergent coronaviruses in bony and jawless fish and the evolutionary history of the Coronaviridae. Virus Evol *7*, veab050.

Mitchell, A.L., Almeida, A., Beracochea, M., Boland, M., Burgin, J., Cochrane, G., Crusoe, M.R., Kale, V., Potter, S.C., Richardson, L.J., et al. (2020). MGnify: the microbiome analysis resource in 2020. Nucleic Acids Res. *48*, D570–D578.

Mizuno, C.M., Guyomar, C., Roux, S., Lavigne, R., Rodriguez-Valera, F., Sullivan, M.B., Gillet, R., Forterre, P., and Krupovic, M. (2019). Numerous cultivated and uncultivated viruses encode ribosomal proteins. Nat. Commun. *10*, 752.

Moore, R.A., Warren, R.L., Freeman, J.D., Gustavsen, J.A., Chénard, C., Friedman, J.M., Suttle, C.A., Zhao, Y., and Holt, R.A. (2011). The sensitivity of massively parallel sequencing for detecting candidate infectious agents

associated with human tissue. PLoS One *6*, e19838.

Mordecai, G.J., Miller, K.M., Di Cicco, E., Schulze, A.D., Kaukinen, K.H., Ming, T.J., Li, S., Tabata, A., Teffer, A., Patterson, D.A., et al. (2019). Endangered wild salmon infected by newly discovered viruses. Elife *8*.

Morel, B., Kozlov, A.M., and Stamatakis, A. (2019). ParGenes: a tool for massively parallel model selection and phylogenetic tree inference on thousands of genes. Bioinformatics *35*, 1771–1773.

Morgulis, A., Michael Gertz, E., Schäffer, A.A., and Agarwala, R. (2006). A Fast and Symmetric DUST Implementation to Mask Low-Complexity DNA Sequences. Journal of Computational Biology *13*, 1028–1040.

Mukherjee, S., Stamatis, D., Bertsch, J., Ovchinnikova, G., Sundaramurthi, J.C., Lee, J., Kandimalla, M., Chen, I.-M.A., Kyrpides, N.C., and Reddy, T.B.K. (2021). Genomes OnLine Database (GOLD) v.8: overview and updates. Nucleic Acids Research *49*, D723–D733.

Nakabachi, A., Yamashita, A., Toh, H., Ishikawa, H., Dunbar, H.E., Moran, N.A., and Hattori, M. (2006). The 160-kilobase genome of the bacterial endosymbiont Carsonella. Science *314*, 267.

Nawrocki, E.P. Faster SARS-CoV-2 sequence validation and annotation for GenBank using VADR.

Nawrocki, E.P., and Eddy, S.R. (2013). Infernal 1.1: 100-fold faster RNA homology searches. Bioinformatics *29*, 2933–2935.

Nayfach, S., Roux, S., Seshadri, R., Udwary, D., Varghese, N., Schulz, F., Wu, D., Paez-Espino, D., Chen, I.-M., Huntemann, M., et al. (2021a). A genomic catalog of Earth's microbiomes. Nat. Biotechnol. *39*, 499–509.

Nayfach, S., Camargo, A.P., Schulz, F., Eloe-Fadrosh, E., Roux, S., and Kyrpides, N.C. (2021b). CheckV assesses the quality and completeness of metagenome-assembled viral genomes. Nature Biotechnology *39*, 578–585.

Ng, W.V., Ciufo, S.A., Smith, T.M., Bumgarner, R.E., Baskin, D., Faust, J., Hall, B., Loretz, C., Seto, J., Slagel, J., et al. (1998). Snapshot of a large dynamic replicon in a halophilic archaeon: megaplasmid or minichromosome? Genome Res. *8*, 1131–1141.

Nguyen, L.-T., Schmidt, H.A., von Haeseler, A., and Minh, B.Q. (2015). IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. Mol. Biol. Evol. *32*, 268–274.

Nowotny, M. (2009). Retroviral integrase superfamily: the structural perspective. EMBO Rep. *10*, 144–151.

Nurk, S., Meleshko, D., Korobeynikov, A., and Pevzner, P.A. (2017). metaSPAdes: a new versatile metagenomic assembler. Genome Res. *27*, 824–834.

Obbard, D.J., Shi, M., Roberts, K.E., Longdon, B., and Dennis, A.B. (2020). A new lineage of segmented RNA viruses infecting animals. Virus Evol *6*, vez061.

O'Hara, B.J., Barth, Z.K., McKitterick, A.C., and Seed, K.D. (2017). A highly specific phage defense system is a conserved feature of the Vibrio cholerae mobilome. PLoS Genet. *13*, e1006838.

Olm, M.R., Brown, C.T., Brooks, B., and Banfield, J.F. (2017). dRep: a tool for fast and accurate genomic comparisons that enables improved genome recovery from metagenomes through de-replication. ISME J. *11*, 2864–2868.

Paez-Espino, D., Eloe-Fadrosh, E.A., Pavlopoulos, G.A., Thomas, A.D., Huntemann, M., Mikhailova, N., Rubin, E., Ivanova, N.N., and Kyrpides, N.C. (2016). Uncovering Earth's virome. Nature *536*, 425–430.

Paez-Espino, D., Chen, I.-M.A., Palaniappan, K., Ratner, A., Chu, K., Szeto, E., Pillay, M., Huang, J., Markowitz, V.M., Nielsen, T., et al. (2017). IMG/VR: a database of cultured and uncultured DNA Viruses and retroviruses. Nucleic Acids Res. *45*, D457–D465.

Paraskevopoulou, S., Pirzer, F., Goldmann, N., Schmid, J., Corman, V.M., Gottula, L.T., Schroeder, S., Rasche, A., Muth, D., Drexler, J.F., et al. (2020). Mammalian deltavirus without hepadnavirus coinfection in the neotropical rodent Proechimys semispinosus. Proc. Natl. Acad. Sci. U. S. A. *117*, 17977–17983.

Pausch, P., Al-Shayeb, B., Bisom-Rapp, E., Tsuchida, C.A., Li, Z., Cress, B.F., Knott, G.J., Jacobsen, S.E., Banfield, J.F., and Doudna, J.A. (2020). CRISPR-CasΦ from huge phages is a hypercompact genome editor. Science *369*, 333–337.

Pawluk, A., Staals, R.H.J., Taylor, C., Watson, B.N.J., Saha, S., Fineran, P.C., Maxwell, K.L., and Davidson, A.R. (2016). Inactivation of CRISPR-Cas systems by anti-CRISPR proteins in diverse bacterial species. Nat Microbiol *1*, 16085.

Peabody, M.A., Laird, M.R., Vlasschaert, C., Lo, R., and Brinkman, F.S.L. (2016). PSORTdb: expanding the bacteria and archaea protein subcellular localization database to better reflect diversity in cell envelope structures. Nucleic Acids Res. *44*, D663–D668.

Pedersen, B.S., and Quinlan, A.R. (2018). Mosdepth: quick coverage calculation for genomes and exomes. Bioinformatics *34*, 867–868.

de la Peña, M., Ceprián, R., Casey, J.L., and Cervera, A. (2021). Hepatitis delta virus-like circular RNAs from diverse metazoans encode conserved hammerhead ribozymes. Virus Evol *7*, veab016.

Penadés, J.R., Chen, J., Quiles-Puchalt, N., Carpena, N., and Novick, R.P. (2015). Bacteriophage-mediated spread of bacterial virulence genes. Curr. Opin. Microbiol. *23*, 171–178.

Peng, Y., Leung, H.C.M., Yiu, S.M., and Chin, F.Y.L. (2012). IDBA-UD: a de novo assembler for single-cell and metagenomic sequencing data with highly uneven depth. Bioinformatics *28*, 1420–1428.

Pérez-Brocal, V., Gil, R., Ramos, S., Lamelas, A., Postigo, M., Michelena, J.M., Silva, F.J., Moya, A., and Latorre, A. (2006). A small microbial genome: the end of a long symbiotic relationship? Science *314*, 312–313.

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

Pettersen, E.F., Goddard, T.D., Huang, C.C., Couch, G.S., Greenblatt, D.M., Meng, E.C., and Ferrin, T.E. (2004). UCSF Chimera--a visualization system for exploratory research and analysis. J. Comput. Chem. *25*, 1605–1612.

Pinilla-Redondo, R., Mayo-Muñoz, D., Russel, J., Garrett, R.A., Randau, L., Sørensen, S.J., and Shah, S.A. (2020). Type IV CRISPR-Cas systems are highly diverse and involved in competition between plasmids. Nucleic Acids Res. *48*, 2000–2012.

Rascovan, N., Duraisamy, R., and Desnues, C. (2016). Metagenomics and the Human Virome in Asymptomatic Individuals. Annu. Rev. Microbiol. *70*, 125–141.

Raveh-Sadka, T., Thomas, B.C., Singh, A., Firek, B., Brooks, B., Castelle, C.J., Sharon, I., Baker, R., Good, M., Morowitz, M.J., et al. (2015). Gut bacteria are rarely shared by co-hospitalized premature infants, regardless of necrotizing enterocolitis development. Elife *4*.

Reeves, G.A., Eilbeck, K., Magrane, M., O'Donovan, C., Montecchi-Palazzi, L., Harris, M.A., Orchard, S., Jimenez, R.C., Prlic, A., Hubbard, T.J.P., et al. (2008). The Protein Feature Ontology: a tool for the unification of protein feature annotations. Bioinformatics *24*, 2767–2772.

Remmert, M., Biegert, A., Hauser, A., and Söding, J. (2011). HHblits: lightning-

fast iterative protein sequence searching by HMM-HMM alignment. Nat. Methods *9*, 173–175.

Rho, M., Tang, H., and Ye, Y. (2010). FragGeneScan: predicting genes in short and error-prone reads. Nucleic Acids Research *38*, e191–e191.

Rice, P., Longden, I., and Bleasby, A. (2000). EMBOSS: The European Molecular Biology Open Software Suite. Trends in Genetics *16*, 276–277.

Richardson, C.D., Ray, G.J., DeWitt, M.A., Curie, G.L., and Corn, J.E. (2016). Enhancing homology-directed genome editing by catalytically active and inactive CRISPR-Cas9 using asymmetric donor DNA. Nat. Biotechnol. *34*, 339–344.

Robinson, J.T., Thorvaldsdóttir, H., Wenger, A.M., Zehir, A., and Mesirov, J.P. (2017). Variant Review with the Integrative Genomics Viewer. Cancer Research *77*, e31–e34.

Roux, S., Brum, J.R., Dutilh, B.E., Sunagawa, S., Duhaime, M.B., Loy, A., Poulos, B.T., Solonenko, N., Lara, E., Poulain, J., et al. (2016). Ecogenomics and potential biogeochemical impacts of globally abundant ocean viruses. Nature *537*, 689–693.

Schäffer, A.A., Hatcher, E.L., Yankie, L., Shonkwiler, L., Brister, J.R., Karsch-Mizrachi, I., and Nawrocki, E.P. (2020). VADR: validation and annotation of virus sequence submissions to GenBank. BMC Bioinformatics *21*, 211.

Schatz, M.C. (2009). CloudBurst: highly sensitive read mapping with MapReduce. Bioinformatics *25*, 1363–1369.

Scheller, S., Yu, H., Chadwick, G.L., McGlynn, S.E., and Orphan, V.J. (2016). Artificial electron acceptors decouple archaeal methane oxidation from sulfate reduction. Science *351*, 703–707.

Schoch, C.L., Ciufo, S., Domrachev, M., Hotton, C.L., Kannan, S., Khovanskaya, R., Leipe, D., Mcveigh, R., O'Neill, K., Robbertse, B., et al. (2020). NCBI Taxonomy: a comprehensive update on curation, resources and tools. Database *2020*.

Schoelmerich, M.C., Oubouter, H.T., Sachdeva, R., Penev, P., Amano, Y., West-Roberts, J., Welte, C.U., and Banfield, J.F. (2022). A widespread group of large plasmids in methanotrophic Methanoperedens archaea.

Seed, K.D., Lazinski, D.W., Calderwood, S.B., and Camilli, A. (2013). A bacteriophage encodes its own CRISPR/Cas adaptive response to evade host innate immunity. Nature *494*, 489–491.

Shen, W., Le, S., Li, Y., and Hu, F. (2016). SeqKit: A Cross-Platform and Ultrafast Toolkit for FASTA/Q File Manipulation. PLoS One *11*, e0163962.

Shi, M., Lin, X.-D., Chen, X., Tian, J.-H., Chen, L.-J., Li, K., Wang, W., Eden, J.-S., Shen, J.-J., Liu, L., et al. (2018). The evolutionary history of vertebrate RNA viruses. Nature *556*, 197–202.

Shkoporov, A.N., and Hill, C. (2019). Bacteriophages of the Human Gut: The "Known Unknown" of the Microbiome. Cell Host Microbe *25*, 195–209.

Shmakov, S., Abudayyeh, O.O., Makarova, K.S., Wolf, Y.I., Gootenberg, J.S., Semenova, E., Minakhin, L., Joung, J., Konermann, S., Severinov, K., et al. (2015). Discovery and Functional Characterization of Diverse Class 2 CRISPR-Cas Systems. Mol. Cell *60*, 385–397.

Shmakov, S., Smargon, A., Scott, D., Cox, D., Pyzocha, N., Yan, W., Abudayyeh, O.O., Gootenberg, J.S., Makarova, K.S., Wolf, Y.I., et al. (2017). Diversity and evolution of class 2 CRISPR–Cas systems. Nat. Rev. Microbiol. *15*, 169–182.

Smargon, A.A., Cox, D.B.T., Pyzocha, N.K., Zheng, K., Slaymaker, I.M., Gootenberg, J.S., Abudayyeh, O.A., Essletzbichler, P., Shmakov, S., Makarova, K.S., et al. (2017). Cas13b is a Type VI-B CRISPR-associated RNA-Guided RNAse differentially regulated by accessory proteins Csx27 and Csx28. Mol. Cell *65*, 618–630.e7.

Smoot, M.E., Ono, K., Ruscheinski, J., Wang, P.-L., and Ideker, T. (2011). Cytoscape 2.8: new features for data integration and network visualization. Bioinformatics *27*, 431–432.

Sørensen, M.A., Fricke, J., and Pedersen, S. (1998). Ribosomal protein S1 is required for translation of most, if not all, natural mRNAs in Escherichia coli in vivo. J. Mol. Biol. *280*, 561–569.

Stachler, A.-E., and Marchfelder, A. (2016). Gene Repression in Haloarchaea Using the CRISPR (Clustered Regularly Interspaced Short Palindromic Repeats)-Cas I-B System. J. Biol. Chem. *291*, 15226–15242.

Stamatakis, A. (2014). RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics *30*, 1312–1313.

Stanaway, J.D., Flaxman, A.D., Naghavi, M., Fitzmaurice, C., Vos, T., Abubakar, I., Abu-Raddad, L.J., Assadi, R., Bhala, N., Cowie, B., et al. (2016). The global burden of viral hepatitis from 1990 to 2013: findings from the Global Burden of Disease Study 2013. Lancet *388*, 1081–1088.

Steinegger, M., Meier, A., and Biegert, A. HH-suite for sensitive protein sequence searching based on HMM-HMM alignment. Bioinformatics *21*, 951–960.

Subramanian, A.R. (1983). Structure and functions of ribosomal protein S1. Prog. Nucleic Acid Res. Mol. Biol. *28*, 101–142.

Swarts, D.C., and Jinek, M. (2019). Mechanistic Insights into the cis- and trans-Acting DNase Activities of Cas12a. Mol. Cell *73*, 589–600.e4.

Swarts, D.C., van der Oost, J., and Jinek, M. (2017). Structural Basis for Guide RNA Processing and Seed-Dependent DNA Targeting by CRISPR-Cas12a. Mol. Cell *66*, 221–233.e4.

Szirovicza, L., Hetzel, U., Kipar, A., Martinez-Sobrido, L., Vapalahti, O., and Hepojoki, J. (2020). Snake Deltavirus Utilizes Envelope Proteins of Different Viruses To Generate Infectious Particles. MBio *11*.

Tao, Y., Paden, C.R., Queen, K., Zhang, J., Tyagi, E., and Tong, S. (2020). Broad-Range Virus Detection and Discovery Using Microfluidic PCR Coupled with High-throughput Sequencing.

Taylor, J.M. (2020). Infection by Hepatitis Delta Virus. Viruses *12*.

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

The UniProt Consortium, and The UniProt Consortium (2017). UniProt: the universal protein knowledgebase. Nucleic Acids Research *45*, D158–D169.

Thiel, V., Ivanov, K.A., Putics, Á., Hertzig, T., Schelle, B., Bayer, S., Weißbrich, B., Snijder, E.J., Rabenau, H., Doerr, H.W., et al. (2003). Mechanisms and enzymes involved in SARS coronavirus genome expression. Journal of General Virology *84*, 2305–2315.

Thompson, A.A., and Peersen, O.B. (2004). Structural basis for proteolysis-dependent activation of the poliovirus RNA-dependent RNA polymerase. EMBO J. *23*, 3462–3471.

Toms, A., and Barrangou, R. (2017). On the global CRISPR array behavior in class I systems. Biol. Direct *12*, 20.

Tsai, S.L., Baselga-Garriga, C., and Melton, D.A. (2020). Midkine is a dual regulator of wound epidermis development and inflammation during the initiation of limb regeneration. Elife *9*.

VanderWal, A.R., Park, J.-U., Polevoda, B., Kellogg, E.H., and O'Connell, M.R. (2021). CRISPR-Csx28 forms a Cas13b-activated membrane pore required for robust CRISPR-Cas adaptive immunity.

Van Duin, J., and Wijnands, R. (1981). The function of ribosomal protein S21 in protein synthesis. Eur. J. Biochem. *118*, 615–619.

Virtanen, P., Gommers, R., Oliphant, T.E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., et al. (2020). SciPy 1.0: fundamental algorithms for scientific computing in Python. Nat. Methods *17*, 261–272.

Wagenknecht, M., Dib, J.R., Thürmer, A., Daniel, R., Farías, M.E., and Meinhardt, F. (2010). Structural peculiarities of linear megaplasmid, pLMA1, from Micrococcus luteus interfere with pyrosequencing reads assembly. Biotechnol. Lett. *32*, 1853–1862.

Wahba, L., Jain, N., Fire, A.Z., Shoura, M.J., Artiles, K.L., McCoy, M.J., and Jeong, D.-E. (2020). An Extensive Meta-Metagenomic Search Identifies SARS-CoV-2-Homologous Sequences in Pangolin Lung Viromes. mSphere *5*.

Wallenius, A.J., Dalcin Martins, P., Slomp, C.P., and Jetten, M.S.M. (2021). Anthropogenic and Environmental Constraints on the Microbial Methane Cycle in Coastal Sediments. Front. Microbiol. *12*, 631621.

Wang, H., Peng, N., Shah, S.A., Huang, L., and She, Q. (2015). Archaeal extrachromosomal genetic elements. Microbiol. Mol. Biol. Rev. *79*, 117–152.

Wang, Q., Alowaifeer, A., Kerner, P., Balasubramanian, N., Patterson, A., Christian, W., Tarver, A., Dore, J.E., Hatzenpichler, R., Bothner, B., et al. (2021). Aerobic bacterial methane synthesis. Proceedings of the National Academy of Sciences *118*, e2019229118.

Westra, E.R., Semenova, E., Datsenko, K.A., Jackson, R.N., Wiedenheft, B., Severinov, K., and Brouns, S.J.J. (2013). Type I-E CRISPR-cas systems discriminate target from non-target DNA through base pairing-independent PAM recognition. PLoS Genet. *9*, e1003742.

Wille, M., Netter, H., Littlejohn, M., Yuen, L., Shi, M., Eden, J.-S., Klaassen, M., Holmes, E., and Hurt, A. (2018). A Divergent Hepatitis D-Like Agent in Birds. Viruses *10*, 720.

Woese, C. (1998). The universal ancestor. Proc. Natl. Acad. Sci. U. S. A. *95*, 6854–6859.

Wolf, Y.I., Kazlauskas, D., Iranzo, J., Lucía-Sanz, A., Kuhn, J.H., Krupovic, M.,

Dolja, V.V., and Koonin, E.V. (2018). Origins and Evolution of the Global RNA Virome. MBio *9*.

Wolf, Y.I., Silas, S., Wang, Y., Wu, S., Bocek, M., Kazlauskas, D., Krupovic, M., Fire, A., Dolja, V.V., and Koonin, E.V. (2020). Doubling of the known set of RNA viruses by metagenomic analysis of an aquatic virome. Nat Microbiol *5*, 1262–1270.

Wrighton, K.C., Castelle, C.J., Wilkins, M.J., Hug, L.A., Sharon, I., Thomas, B.C., Handley, K.M., Mullin, S.W., Nicora, C.D., Singh, A., et al. (2014). Metabolic interdependencies between phylogenetically novel fermenters and respiratory organisms in an unconfined aquifer. ISME J. *8*, 1452–1463.

Yan, W.X., Chong, S., Zhang, H., Makarova, K.S., Koonin, E.V., Cheng, D.R., and Scott, D.A. (2018). Cas13d Is a Compact RNA-Targeting Type VI CRISPR Effector Positively Modulated by a WYL-Domain-Containing Accessory Protein. Mol. Cell *70*, 327–339.e5.

Yan, W.X., Hunnewell, P., Alfonse, L.E., Carte, J.M., Keston-Smith, E., Sothiselvam, S., Garrity, A.J., Chong, S., Makarova, K.S., Koonin, E.V., et al. (2019). Functionally diverse type V CRISPR-Cas systems. Science *363*, 88–91.

Yoo, S.-D., Cho, Y.-H., and Sheen, J. (2007). Arabidopsis mesophyll protoplasts: a versatile cell system for transient gene expression analysis. Nat. Protoc. *2*, 1565–1572.

Yuan, Y., and Gao, M. (2017). Jumbo Bacteriophages: An Overview. Front. Microbiol. *8*.

Zetsche, B., Gootenberg, J.S., Abudayyeh, O.O., Slaymaker, I.M., Makarova, K.S., Essletzbichler, P., Volz, S.E., Joung, J., van der Oost, J., Regev, A., et al. (2015). Cpf1 is a single RNA-guided endonuclease of a class 2 CRISPR-Cas system. Cell *163*, 759–771.

Zuccola, H.J., Rozzelle, J.E., Lemon, S.M., Erickson, B.W., and Hogle, J.M. (1998). Structural basis of the oligomerization of hepatitis delta antigen. Structure *6*, 821–830.

Zuker, M. (2003). Mfold web server for nucleic acid folding and hybridization prediction. Nucleic Acids Res. *31*, 3406–3415.