Cryptic inoviruses revealed as pervasive in bacteria and archaea across Earth’s biomes

Simon Roux1,*, Mart Krupovic2, Rebecca A. Daly3, Adair L. Borges4, Stephen Nayfach1, Frederik Schulz1, Allison Sharrar5, Paula B. Matheus Carnevali6, Jan-Fang Cheng1, Natalia N. Ivanova1, Joseph Bondy-Denomy4,6, Kelly C. Wrighton3, Tanja Woyke1, Allison Sharrar5, Paula B. Matheus Carnevali2, Rebecca A. Daly3, Adair L. Borges4, Stephen Nayfach1, Nikos C. Kyripides1 and Emiley A. Eloe-Fadrosh3,1*

Bacteriophages from the Inoviridae family (inoviruses) are characterized by their unique morphology, genome content and infection cycle. One of the most striking features of inoviruses is their ability to establish a chronic infection whereby the viral genome resides within the cell in either an exclusively episomal state or integrated into the host chromosome and virions are continuously released without killing the host. To date, a relatively small number of inovirus isolates have been extensively studied, either for biotechnological applications, such as phage display, or because of their effect on the toxicity of known bacterial pathogens including Vibrio cholerae and Neisseria meningitidis. Here, we show that the current 56 members of the Inoviridae family represent a minute fraction of a highly diverse group of inoviruses. Using a machine learning approach leveraging a combination of marker gene and genome features, we identified 10,295 inovirus-like sequences from microbial genomes and metagenomes. Collectively, our results call for reclassification of the current Inoviridae family into a viral order including six distinct proposed families associated with nearly all bacterial phyla across virtually every ecosystem. Putative inoviruses were also detected in several archaeal genomes, suggesting that, collectively, members of this supergroup infect hosts across the domains Bacteria and Archaea. Finally, we identified an expansive diversity of inovirus-encoded toxin–antitoxin and gene expression modulation systems, alongside evidence of both synergistic (CRISPR evasion) and antagonistic (superinfection exclusion) interactions with co-infecting viruses, which we experimentally validated in a Pseudomonas model. Capturing this previously obscured component of the global virosphere may spark new avenues for microbial manipulation approaches and innovative biotechnological applications.

Inoviruses, bacteriophages from the Inoviridae family, exhibit unique morphological and genetic features. While the vast majority of known bacteriophages carry double-stranded DNA (dsDNA) genomes encapsidated into icosahedral capsids, inoviruses are instead characterized by rod-shaped or filamentous virions, circular single-stranded DNA genomes of ~5–15 kb and a chronic infection cycle1–3 (Fig. 1a). Owing to their unique morphology and simple genome amenable to genetic engineering, several inoviruses are widely used for biotechnological applications, including phage display or as drug delivery nanocarriers4–7. Ecologically, cultivated inoviruses are known to infect hosts from only 5 bacterial phyla and 10 genera but can have significant effect on the growth and pathogenicity of their host8–10. For instance, an inovirus prophage, CTXϕ, encodes and expresses the major virulence factor, CTXϕ, that indirectly influence pathogenicity by altering biofilm formation and host colonization abilities11–16.

Despite these remarkable properties, their elusive life cycle and peculiar genomic and morphological properties have hampered systematic discovery of additional inoviruses: to date, only 56 inovirus genomes have been described17. Most inoviruses do not elicit negative effects on the growth of their hosts when cultivated in the laboratory and can thus easily evade detection. Furthermore, established computational approaches for the detection of virus sequences in whole-genome shotgun sequencing data are not efficient for inoviruses because of their unique and diverse gene content18–20 (Fig. 1b). Finally, inoviruses are probably undersampled in viral metagenomes due to their long, flexible virions with low buoyant density11,21.

Here, we unveil a substantial diversity of 10,295 inovirus sequences derived from a broad range of bacterial and archaeal hosts, and identified through an exhaustive search of 56,868 microbial genomes and 6,412 shotgun metagenomes using a custom computational approach to identify putative inovirus genomes. These sequences reveal that inoviruses are far more widespread, diverse and ecologically pervasive than previously appreciated, and provide a robust foundation to further characterize their biology across multiple hosts and environments.

Results
Inoviruses are highly diverse and globally prevalent. To evaluate the global diversity of inoviruses, an analysis of all publicly available inovirus genomes was first conducted to identify characteristic traits that would enable automatic discovery of divergent inovirus sequences (Supplementary Table 1). Across the 56 known Inoviridae genomes, the gene encoding the morphogenesis (pI) protein, an ATPase of the FtsK–HerA superfamily, represented the only

1DOE Joint Genome Institute, Walnut Creek, CA, USA. 2Department of Microbiology, Institut Pasteur, Paris, France. 3Department of Soil and Crop Sciences, Colorado State University, Fort Collins, CO, USA. 4Department of Microbiology and Immunology, University of California, San Francisco, San Francisco, CA, USA. 5Department of Earth & Planetary Sciences, University of California, Berkeley, Berkeley, CA, USA. 6Quantitative Biosciences Institute, University of California, San Francisco, San Francisco, CA, USA. *e-mail: sroux@lbl.gov; eaeloefadrosh@lbl.gov
Inoviruses infect a broad diversity of bacterial hosts. To examine the host range of these inoviruses, we focused on the 2,284 inovirus species directly associated with a host, that is, proviruses identified using genome-wide average nucleotide identity (ANI), and only 38 of these included isolate inovirus genomes. About one-third of these species (30%) encoded an ‘atypical’ morphogenesis gene, with an amino-terminal instead of carboxy-terminal TMD (Supplementary Fig. 3). Although this atypical domain organization has been observed in four isolate species currently classified as inoviruses, some of these inovirus-like sequences might eventually be considered as entirely separate groups of viruses. Sequence accumulation curves did not reach saturation, highlighting the large diversity of inoviruses yet to be sampled (Supplementary Fig. 4).

Inovirus sequences were identified in 6% of bacterial and archaeal genomes (3,609 of 56,868) and 35% of metagenomes (2,249 of 6,412). More than half of the species (n = 3,675) were exclusively composed of sequences assembled from metagenomes. These revealed that inoviruses are found in every major microbial habitat whether aquatic, soil or human associated, and throughout the entire globe (Fig. 2 and Supplementary Notes). Hence, inoviruses are much more diverse than previously estimated and globally distributed.

Conserved marker gene (Fig. 1a,b and Supplementary Fig. 1). However, three additional features specific of inovirus genomes could be defined: (1) short structural proteins (30–90 amino acids) with a single predicted transmembrane domain (TMD; Supplementary Table 1), (2) genes either functionally uncharacterized or similar to other inoviruses, and (3) shorter genes than those in typical bacterial or archaeal genomes (Supplementary Fig. 2A). These features were used to automatically detect inovirus sequences through a two-step process (Fig. 1b). First, pI-like proteins are detected through a standard hidden Markov model (HMM)-based similarity search. Then, a random forest classifier trained on genomes of isolate inoviruses and manually curated prophages used these genome features to identify inovirus sequences from the background host genome. This approach yielded 92.5% recall and 99.8% precision on our manually curated reference set (Fig. 1c, Supplementary Fig. 2 and Supplementary Notes).

This detection approach was applied to 56,868 bacterial and archaeal genomes and 6,412 metagenomes publicly available from the Integrated Microbial Genomes (IMG) database (Supplementary Table 2). After manual curation of edge cases and removal of detections not based on a clear inovirus-like ATPase, a total of 10,295 sequences were recovered (Fig. 1d, Supplementary Fig. 3 and Supplementary Notes). From these, 5,964 distinct species were identified using genome-wide average nucleotide identity (ANI), and only 38 of these included isolate inovirus genomes. About one-third of these species (30%) encoded an ‘atypical’ morphogenesis gene, with an amino-terminal instead of carboxy-terminal TMD (Supplementary Fig. 3). Although this atypical domain organization has been observed in four isolate species currently classified as inoviruses, some of these inovirus-like sequences might eventually be considered as entirely separate groups of viruses. Sequence accumulation curves did not reach saturation, highlighting the large diversity of inoviruses yet to be sampled (Supplementary Fig. 4).
derived from a microbial genome (Fig. 3). The majority (90%) of these species were associated with Gammaproteobacteria and Betaproteobacteria, from which most known inoviruses were previously isolated (Supplementary Table 1). However, the range of host genera within these groups was vastly expanded, including clinically and ecologically relevant microorganisms such as Azotobacter, Haemophilus, Kingella or Nitrosomonas (Supplementary Table 3). The remaining 412 species strikingly increased the potential host range of inoviruses to 22 additional phyla, including the Candidate Phyla Radiation (Fig. 3). For three of these (Acidobacteria, Chlamydiae and Spirochaetes), only short contigs were detected, lacking host flanking regions, which would provide confident host linkages. Hence, these contigs could potentially derive from sample contamination (for example, from reagents), and inovirus presence within these phyla remains uncertain (Supplementary Table 4). The notable host expansion is consistent with reported experimental observations of filamentous virus particles induced from a broad range of bacteria, for example, Mesorhizobium, Clostridium, Flavobacterium, Bacillus and Arthrobacter\(^{2,23}\) (Fig. 3).

This large-scale detection of inovirus sequences in microbial genomes also enabled a comprehensive assessment of co-infection, both between different inoviruses and with other types of viruses. In the majority of cases, a single inovirus sequence was detected per genome, with multiple detections mostly found within Gammaproteobacteria, Betaproteobacteria and Spiroplasma genomes (Supplementary Fig. 5). Conversely, inovirus prophages were frequently detected along and sometimes colocalized with Caudovirales prophages, suggesting that these two types of phages frequently co-infect the same host cell (Supplementary Fig. 5 and Supplementary Notes). Overall, the broad range of bacteria and archaea infected by inoviruses combined with their propensity to co-infect a microbial cell with other viruses and their global distribution indicate that inoviruses probably play an important ecological role in all types of microbial ecosystems.

Inoviruses sporadically transferred from bacterial to archaeal hosts. Although no archaea-infecting inoviruses have been reported so far\(^{26-28}\), some inovirus sequences were associated with members of two archaeal phyla (Euryarchaeota and Aenigmarchaeota), which suggests that inoviruses infect hosts across the entire prokaryotic diversity (Fig. 3). These putative archaeal proviruses encoded the full complement of genes expected in an active inovirus (Fig. 4a and Supplementary Notes). Using PCR, we further confirmed the presence of a circular, excised form of the complete inovirus genome for the provirus identified in the Methanobacterium profundi genome (Fig. 4b, Supplementary Fig. 6 and Supplementary Notes). This indicates that our predictions in archaeal genomes are probably genuine inoviruses.

Few groups of viruses include both bacteriophages and archaeoviruses. Such evolutionary relationships between viruses infecting hosts from different domains of life might signify either descent from an ancestral virus that infected the common ancestor of bacteria and archaea, or horizontal virus transfer from one host domain to the other\(^{26-28}\). Here, the four archaea-associated inoviruses were clearly distinct from most other inoviruses and clustered only with metagenomic sequences in 16S phylogeny (Fig. 4c). In addition, they were classified into two different proposed families (see below) corresponding to the two host groups, reflecting clear differences in their gene content (Fig. 4a,c and Supplementary Notes). The high genetic diversity of these archaea-associated inoviruses, combined with the lack of similarity to bacteria-infesting species, suggest that they are not derived from a recent host switch event.

A possible scenario would involve an ancestral group of inoviruses infecting the common ancestor of archaea, as postulated for the double-jelly-roll virus lineage\(^{26}\). However, to be confirmed, this hypothesis would require the detection of additional inoviruses in other archaeal clades or an explanation as to why inoviruses were retained only in a handful of archaeal hosts. Instead, on the basis of the current data, a more likely scenario involves ancient and rare events of interdomain inovirus transfer from bacteria to archaea.
including possibly to a *Methanosarcina* host for which substantive horizontal transfers of bacterial genes have already been reported\(^{39}\).

**Gene content classification reveals six distinct inovirus families.** The vast increase of inovirus sequences provided a great opportunity for re-evaluation of the inovirus classification and the development of an expanded taxonomic framework for the large number of inovirus species identified. Similar to other bacterial viruses, especially temperate phages\(^{30}\), inovirus genomes display modular organization and are prone to recombination and horizontal gene transfers\(^{31}\) (Supplementary Fig. 7). Hence, we opted to apply a bipartite network approach, in which genomes are connected to gene families, enabling a representation and clustering of the diversity based on shared gene content. A similar approach has been previously employed for the analysis of DNA and RNA viruses, and was shown to be efficient in cases in which the genomes to be clustered share only a handful of genes\(^{26,32–34}\). Here, this approach yielded 6 distinct groups of genomes divided into 212 subgroups (Fig. 5a and Supplementary Table 3).

A comparison of marker gene conservation between these groups and established viral taxa suggested that the former *Inoviridae* family should be reclassified as an order, provisionally divided into 6 candidate families and 212 candidate subfamilies, with few shared genes across candidate families (Fig. 5a, Supplementary Fig. 7 and Supplementary Notes). Beyond gene content, these proposed families also displayed clearly distinct host ranges as well as specific genome features, particularly in terms of genome size and coding density (Supplementary Fig. 7). Thus, we propose to establish these as candidate families named ‘Protoinoviridae,’ ‘Vespertilinoviridae,’ ‘Amplinoviridae,’ ‘Paulinoviridae,’ ‘Densinoviridae’ and ‘Photinoviridae,’ on the basis of their isolate members and characteristics (see Supplementary Notes). If confirmed, and compared with currently recognized inoviruses, the genomes reported here would increase diversity by 3 families and 198 subfamilies.

The host envelope organization seems to play an important role in the evolution of inoviruses, which is reflected in their classification: members of the ‘Protoinoviridae’ and ‘Amplinoviridae’ are associated with diderm hosts—that is, Gram-negative bacteria with an outer membrane—whereas the other candidate families are associated with monoderm hosts or hosts without a cell wall (Supplementary Fig. 7). Conversely, no structuring by biome was observed and all proposed families were broadly detected across multiple types of ecosystems. Hence, we propose here a classification of inovirus diversity into six families based on gene content with coherent host ranges and...
specific genomic features, which strongly suggests that they represent ecologically and evolutionarily meaningful units.

**Inovirus genomes encode an extensive functional repertoire.** The extended catalogue of inovirus genomes offers an unprecedented window into the diversity of their genes and predicted functions. Overall, 68,912 proteins were predicted and clustered into 3,439 protein families and 13,714 singletons. This is on par with the functional diversity observed in known Caudovirales protein families (iPFs). Most of these (~95%) could be linked to virion structure, virion extrusion, DNA replication and integration, stress response and transcription regulation (Supplementary Table 5). A total of 51 and 47 distinct iPFs could be annotated as major and minor coat proteins, respectively, with an additional 934 iPFs identified as potentially structural based on their size and presence of a TMD (see Methods). Notably, each candidate inovirus family seemed to be associated with a specific set of structural proteins, including distinct major coat iPFs (Supplementary Fig. 8). Conversely, genome replication and integration-associated iPFs were broadly shared across candidate families (Fig. 5b). This confirms that replication-associated and integration-associated genes are among the most frequently exchanged among viral genomes and with other mobile genetic elements, especially in small single-stranded DNA viruses.

In addition, 15 distinct sets of iPFs representing potential toxin–antitoxin pairs were identified across 181 inovirus genomes, including 10 unaffiliated iPFs that were predicted as putative antitoxins through co-occurrence with a toxin iPF (Fig. 5b and Supplementary Table 5; see Methods). These genes typically stabilize plasmids or prophages in host cell populations, although alternative roles in stress response and transcription regulation...
have been reported. In addition, toxin–antitoxin systems often affect host cell phenotypes, such as motility or biofilm formation. Here, similar toxin proteins could be associated with distinct and seemingly unrelated antitoxins and vice versa, suggesting that gene shuffling and lateral transfer occur even within these tightly linked gene pairs (Supplementary Fig. 9). All but one toxin–antitoxin pairs were detected in proteobacteria-associated inoviruses, most likely because of a database bias. Thus, numerous uncharacterized iPFs across other candidate families of inoviruses may also encode previously undescribed toxin–antitoxin systems and, more generally, host manipulation mechanisms.

Inoviruses can both leverage and restrict co-infecting viruses. Finally, we investigated potential interactions between persistently infecting inoviruses, other co-infecting viruses, and the host clustered regularly interspaced short palindromic repeats (CRISPR)–CRISPR-associated (Cas) immunity systems. CRISPR–Cas systems typically target bacteriophages, plasmids and other mobile genetic elements. We detected 1,150 inovirus-matching CRISPR spacers across 42 bacterial and 1 archaeal families. These spacers were associated with three types and eight subtypes of CRISPR–Cas systems, indicating that inoviruses are broadly targeted by antiviral defences (Fig. 6a, Supplementary Table 6 and Supplementary Notes).
host groups, most notably *Neisseria meningitidis*, were clear outliers, that is, they displayed a particularly high ratio of inovirus-derived spacers suggesting a uniquely high level of spacer acquisition and inovirus infection (Fig. 6a). This is particularly notable because inoviruses were recently suggested to increase *N. meningitidis* pathogenicity\(^1\) and hints at conflicting host–inovirus interactions in this specific group.

Next, we examined instances of ‘self-targeting’, that is, CRISPR spacers matching an inovirus integrated in the same host genome. Among the 1,429 genomes that included both a CRISPR–Cas system and an inovirus prophage, only 45 displayed a spacer match(es) to a resident prophage (Supplementary Table 6), suggesting that self-targeting of these integrated elements is lethal and strongly counter-selected\(^1\). This was confirmed experimentally using the *Pseudomonas aeruginosa* strain PA14 harbouring an integrated inovirus prophage (Pfi), for which the introduction of a plasmid carrying Pfi-targeting CRISPR spacers was lethal (Supplementary Fig. 10a). In the 45 cases of observed self-targeting, the corresponding CRISPR–Cas system is thus probably non-functional or inhibited via an anti-CRISPR (acr) locus, as recently described in dsDNA phages\(^3\). We first evaluated ten hypothetical proteins, and hence candidate Acr proteins, from self-targeted inoviruses infecting *P. aeruginosa*; however, none showed Acr activity (Supplementary Notes and Supplementary Fig. 10b). Alternatively, inoviruses could leverage the Acr activity of a co-integrated virus. This hypothesis was further reinforced by the fact that 43 of the 45 self-targeted inoviruses were detected alongside co-infecting dsDNA phages, with 5 of these encoding known acr genes (Supplementary Table 6).

We confirmed experimentally cross-protection by trans-acting Acr in the *P. aeruginosa* PA14 model, and observed that co-infection with an acr-encoding dsDNA bacteriophage rescued the lethality caused by self-targeted inoviruses (Supplementary Notes and Supplementary Fig. 10a).

While this represents an instance of beneficial co-infection for inoviruses, we also uncovered evidence of antagonistic interactions between inoviruses and dsDNA bacteriophages. Specifically, 2 of the 10 inovirus-encoded hypothetical proteins tested strongly limited infection of *Pseudomonas* cells by different bacteriophages (Fig. 6b, Supplementary Figs. 10c and 12 and Supplementary Notes). This superinfection exclusion effect was found to be host and virus strain dependent, which could drive intricate tripartite coevolutionary dynamics. Thus, these preliminary observations indicate that inoviruses may not only evade CRISPR–Cas immunity by leveraging the Acr activity of co-integrated phages, but also significantly influence the infection dynamics of unrelated co-infecting viruses through superinfection exclusion (Fig. 6c). Multiple effects of virus–virus interactions on host ecology and evolution have been recently highlighted or proposed, and are the main focus...
of a nascent ‘sociovirology’ field\(^{\text{a}}\). Given their broad host range (Fig. 3), frequent detection alongside non-inovirus prophages (Supplementary Fig. 5), extended host cell residence time and the experimental results presented here, inoviruses could be driving many of these interactions and are undeniably important to consider in this framework.

**Discussion**

Taken together, the results presented here call for a complete re-evaluation of the diversity and role of inoviruses in nature. Collectively, inoviruses are distributed across all biomes and display an extremely broad host range spanning both prokaryotic domains of life. Comparative genomics revealed evidence of long-standing virus–host codiversification, leading to strong partitioning of inovirus diversity by host taxonomy, high inovirus prevalence in several microbial groups, including major pathogens, and potential interdomain transfer. Even though small (5–20 kb), their genomes encode a large functional diversity shaped by frequent gene exchange with unrelated groups of viruses, plasmids and transposable elements. Some of the many uncharacterized inovirus genes probably encode molecular mechanisms at the interface of virus–host and virus–virus interactions, such as modulators of the CRISPR–Cas systems, superinfection exclusion genes or toxin–antitoxin modules. This expanded and restructured catalogue of 5,964 distinct inovirus genomes thus provides a renewed framework for further investigation of the different effects that inoviruses have on microbial ecosystems, and exploration of their unique potential for biotechnological applications and manipulation of microorganisms.

**Methods**

**Construction of an Inoviridae genome reference set.** Genome sequences affiliated to *Inoviridae* and ≥2.5 kb were downloaded from NCBI GenBank and RefSeq on 14 July 2017 (refs. \(^{\text{a,40}}\)). These were clustered at 98% ANI to remove duplicates and screened for cloning vectors and partial genomes (Supplementary Table 1). Two of these genomes (Stenotrophomonas phage phiSM9, NC_007189, and *Rutilonia* phage RSS30, NC_021862) presented an unusually long section (≥1 kb) without any predicted gene, associated with a lack of short genes that are typical of *Inoviridae*. For these, genes were predicted de novo using Glimmer\(^{\text{a,42}}\) trained on their host genomes (NC_019043 for phiSM9 and NC_003295 for RSS30) with standard genetic code. Similarly, genes for *Acholeplasma* phage MV-L1 (NC_001341) were predicted de novo using Glimmer with genetic code 4 (Mycoplasma/Spiroplasma) and trained on the host genome (NC_010163), followed by a manual curation step to integrate both RefSeq-annotated genes and these newly predicted CDS.

Protein clusters (PCs) were computed from these genomes from an all-versus-all blastp of predicted CDS (thresholds: e ≤ 0.001, bit score ≥ 30) and clustered with InfernalMap\(^{\text{b}}\). Sequences from these PCs were then aligned with MUSCLE\(^{\text{a,43}}\), transformed into an HMM profile and compared with each other using HHSearch\(^{\text{a,44}}\) (cut-offs: probability ≥ 90% and coverage ≥ 50%, or probability ≥ 99%, coverage ≥ 20% and hit length ≥ 100). The larger clusters generated through this second step are designated here as iPfs. Only ten PCs were clustered into larger iPfs, but these were consistent with the functional annotation of these proteins. For instance, one iPf combined two PCs both composed of replication initiation proteins. Marker genes were identified from a bipartite network linking *Inoviridae* genomes to iPfs (Supplementary Fig. 1). Only the genes encoding the morphogenesis (pI) protein represented good candidates for a universally conserved gene across all members of the *Inoviridae*, and HMM profiles were built for the three pI iPfs. To optimize these profiles, sequences were first clustered at 90% amino acid identity with cd-hit\(^{\text{a,45}}\), then aligned with MUSCLE\(^{\text{a,43}}\) and the profile generated with hmmbuild\(^{\text{a}}\).

These reference genomes were also used to evaluate the detection of the *Inoviridae* structural proteins based on protein features beyond sequence similarity (see Supplementary Notes). Here, signal peptides were predicted using SignalP\(^{\text{a,46}}\) for the three pI iPfs. To optimize these profiles, sequences were first clustered at 90% amino acid identity with cd-hit\(^{\text{a,45}}\), then aligned with MUSCLE\(^{\text{a,43}}\) and the profile generated with hmmbuild\(^{\text{a}}\).

**Search for inovirus in microbial genomes and metagenomes.** Proteins predicted from 56,868 microbial genomes publicly available in the IMG as of October 2017 (Supplementary Table 2) were compared with the reference morphogenesis (pI) proteins with hmmscan\(^{\text{a}}\) (hmmer.org, score ≥ 30 and e ≤ 0.001) for the pI-like iPfs and blastp\(^{\text{a}}\) (bit score ≥ 50) for the singleton pI protein (*Acholeplasma* phage MV-L1). These included 54,405 bacterial genomes, 1,304 archaeal genomes and 1,149 plasmid sequences. A total of 6,819 hits were detected, from which 795 corresponded to complete inovirus genomes. These included 213 circular contigs, that is, likely complete genomes, and 582 integrated prophages with canonical attachment (att) sites, that is, direct repeats of ≥10 bp in a tRNA or outside of an integrated gene. All sequences were manually inspected, and 561 were plausible inovirus genomes (see Supplementary Notes). The predicted pl proteins from the curated genomes were then added to the references to generate new improved HMM models. Using these improved models, an additional set of 639 putative pl proteins was identified. New models were built from these proteins and used in a third round of searches, which did not yield any additional genuine inovirus sequence after manual inspection.

An automatic classifier was trained on this extended inovirus genome catalogue, that is, the reference genomes and the 795 manually curated genomes, to detect putative inovirus fragments around pl-like genes, based on 10 distinctive features of inovirus genomes (Supplementary Fig. 2 and Supplementary Notes). These 795 manually curated genomes were identified from 17 host phyla (or class for Proteobacteria) and were later classified into 5 proposed families and 245 proposed subfamilies (see below ‘Gene-content-based clustering of inovirus genomes’). Three types of classifiers were tested: random forest (function *randomForest* from R package *randomForest*\(^{\text{a}}\) using 2,000 trees, other parameters left as default), random forest with conditional inference (function *forest* from R package *randomForest*) using 2,000 trees, other parameters left as default)) and a generalized linear model with lasso regularization (function *glmnet* from R package *glmnet*\(^{\text{a}}\)). The efficiency of classifiers was evaluated using a tenfold cross-validation in which the input data set was partitioned into ten equal-sized subsets, with one subset reserved for validation and the other nine used for training through the ten possible permutations. Results were visualized as a ROC curve generated with ggplot2 (refs. \(^{\text{48,49}}\)). The importance of features in the random forest classifier was evaluated using the function ‘importance’, from the package ‘randomForest’.

On the basis of the inflection point observed on the ROC curves, the random forest classifier was selected as the optimal method as it provided the highest true-positive rate (>92%) for false-positive rates of <1%. (Supplementary Fig. 2). This model was then used to classify all putative inovirus fragments that had not been identified as complete genomes previously, using a sliding window approach (up to 30 genes around the putative pl protein), and looking for the fragment with the maximal score in the random forest model (if ≥0.99). For the predicted integrated prophages, putative non-canonical att sites were next searched as direct repeats (10 bp or longer) around the fragment. Overall, 3,908 additional putative inovirus sequences were detected, including 738 prophages flanked by direct repeats. A similar approach was used to search for inovirus sequences in 6,412 metagenome assemblies (Supplementary Table 2). Predicted proteins were compared with the 4 HMM profiles as well as to the *Acholeplasma* phage MV-L1 singleton sequence, which led to 27,037 putative pl proteins using the same thresholds as for isolate genomes. The final data set of inovirus sequences predicted from these metagenome assemblies consisted of 6,094 sequences, including 922 circular contigs, 44 prophages with canonical att sites (direct repeats of 10 bp or longer in a tRNA or next to an integrase) and 994 prophages with non-canonical att sites (direct repeats of 10 bp or longer).

**Clustering of inovirus genomes in putative species.** Next, we sought to cluster these putative inovirus genomes along with the previously collected reference genomes to remove duplicated sequences and to select only one representative per species. This clustering was conducted according to the latest guidelines submitted to the International Committee on Taxonomy of Viruses (ICTV) for *Inoviridae*, that is, “95% DNA sequence identity as the criterion for demarcation of species”\(^{\text{50}}\) (https://talk.ictvonline.org/files/ictv_official_taxonomy_updates_since_the_8th_report/m/prokaryote-official/6774/download), and included our 10,295 sequences. Notably, however, predictions spanning multiple tandemly integrated inovirus prophages had to be processed separately, otherwise they could lead to clusters gathering multiple species. To detect these cases of tandem insertions, we searched for and clustered separately all predictions with multiple pl proteins, as this gene is expected to be present in single copy in inoviruses (=800 sequences).

These 795 manually curated proteins were first clustered incrementally with priority given to complete genomes over partial genomes as well as fragments identified in microbial genomes over fragments from metagenomes. First, circular contigs and prophages with canonical att sites identified in a microbial genome were clustered, and all other fragments were affiliated to these seed sequences. Next, unaffiliated fragments detected in microbial genomes and with non-canonical att sites (that is, similar proteins with a repeat) were clustered together, and other fragments were affiliated to this second set of seed sequences. Finally, the remaining unaffiliated sequences detected in microbial genomes were clustered together. This allowed us to use the more ‘certain’ predictions (that is, circular sequences and prophages with identified att sites) preferentially as seeds of putative species.

Then, putative inovirus genomes were clustered with a maximum of 100% identity to cluster genomes identified from metagenomes, as well as to separately cluster putative tandem fragments, that is, those including multiple pl proteins. All the clustering and affiliation was done with a threshold of 95% ANI on 100% of alignment fraction (according to the ICTV guidelines), with sequence similarity computed using mummer\(^{\text{a}}\).
Clustering of predicted proteins from non-redundant inovirus sequences. Predicted proteins from the representative genome of each putative species were next clustered using the same approach as for the reference genomes. A clustering into PCs was first achieved through an all-versus-all blastp using with ≤ 0.001 and bit score ≥ 50 or bit score ≥ 30 if both proteins are ≤ 70 amino acids. HMM profiles were constructed for the 5,142 PCs and these were compared all-versus-all using HHSearch, keeping hits with ≥90% probability and ≥50% coverage or ≥99% probability, ≥20% coverage and hit length of ≥100. This resulted in 4,008 protein families (iPFs).

The PCs were subsequently used for taxonomic classification of the inovirus sequences (see below), while iPFs were primarily used for functional affiliation. iPF families (iPFs).

functions were predicted based on the affiliation of iPF members against Pfam sequences (see below), while iPFs were primarily used for functional affiliation. iPF families (iPFs). 100. This resulted in 4,008 protein families (iPFs).

profiles were constructed for the 5,142 PCs and these were compared all-versus-all using HHSearch, keeping hits with ≥90% probability and ≥50% coverage or ≥99% probability, ≥20% coverage and hit length of ≥100. This resulted in 4,008 protein families (iPFs).

Gene-content-based clustering of inovirus genomes. A bipartite network was built in which genomes and PCs (as nodes) are connected by an edge when a predicted protein from the genome is a member of the PC. This network was then used to classify inovirus sequences as done previously for dsDNA viruses4. PCs were retrieved instead of all-versus-all hits, as they offer a higher subset of pl proteins (that is, tandem propehges) were excluded from this network-based classification as these could lead to improper connections between unrelated genomes. Singleton proteins were also excluded, and only PCs with at least 2 members were used to build the network. This network had a very low density (0.65%), reflecting the fact that most PCs were restricted to a minor fraction of the genomes. Nevertheless, this type of network can still be organized into meaningful groups through information theoretic approaches: here, sequence clusters were obtained through InfoMap, with default parameters and a two-level clustering (that is, groups can be associated with a group and a subgroup).

A summarized representation of the network was generated by displaying each subgroup (level 2) as a node with a size proportional to the number of species in the subgroup, and drawing an edge to a PC if ≥ 50% of the subgroup sequences encode this PC, except for the larger group (‘Protonoiviriidae:Subfamily_1’) where connections are drawn for PCs found in >25% of the sequences. The network was then visualized using Cytoscape5, with nodes from the same group (level 1) first gathered manually and with group automatically generated using Prefuse-directed layout (default spring strength of 200).

To evaluate the taxonomic rank to which these groups and subgroups would correspond, we calculated pairwise amino acid identity percentage of pl proteins for genomes (1) between groups and (2) within groups but between subgroups, using Sorensen’s Dice’s coefficient6. These were then compared to a comprehensive amino acid identity calculated with the same approach for established viral groups, namely, Caudovirales order using the terminase large subunit (TerL) as a marker protein, Microviridae using the major capsid protein (VP1) as a marker protein and Circoviridae using the replication initiation protein (Rep) as a marker protein (see Supplementary Notes).

Distribution of inovirus sequences by host and biome. The distribution of hosts for inovirus sequences was based on detections in IMG draft and complete genomes, that is, excluding all metagenome-derived detections but including detections in metagenome-assembled genomes (published draft genomes assembled from metagenomes). Host taxonomic classification was extracted from the IMG database. For visualization purposes, a set of 56 universal single-copy marker proteins45-47 was used to build phylogenetic trees for bacteria and archaea based on all available microbial genomes in IMG4 (genomes downloaded on 27 October 2017) and about 8,000 metagenome-assembled genomes from the Genome Taxonomy Database4 (downloaded on 18 October 2017). Marker proteins were identified with hmmssearch (version 3.1b2, hmmer.org) using a specific HMM for each of the markers. Genomes lacking a substantial proportion of marker proteins (≥28) or which had additional copies of ≥3 single-copy markers were removed from the data set.

To reduce redundancy and to enable a representative taxon sampling, DNA-directed DNA polymerase subunit α160kd (COG0086) was identified using hmmssearch (hmmer 3.1b2) and the HMM of COG0086 (ref. 48). Protein hits were then extracted and clustered with cd-hit at 65% sequence similarity, resulting in 99 archaeal and 837 bacterial clusters. Genomes with the greatest number of markers were selected as cluster representatives. For every marker protein, alignments were built with MAFFT7 (v7.294b) and subsequently trimmed with BMGE (v1.12) using BLOSUM30 (ref. 49). Single-protein alignments were then concatenated, resulting in an alignment of 11,220 sites for the archaea and 16,562 sites for the bacteria. Maximum-likelihood phylogenies were inferred with FastTree2 (v2.1.9 SSE3, OpenMPI)70 using the options: -spr 4 -mclac 2 -slownm -lg.

A distribution of inovirus sequences across biomes was obtained by compiling ecosystems and sampling location of all metagenomes where at least one inovirus sequence was detected. This information was extracted from the GOLD database6, and the map was generated using the BaseMap functions from the matplotlib python library71.

Estimation of inovirus prevalence and co-infection patterns. Prevalence and co-infection patterns were evaluated from the set of sequences identified in complete and draft microbial genomes from the IMG database, that is, excluding detections from metagenome assemblies. To control for the presence of near-identical genome regions, the IMG filtered all PICR members against Pfam HMMs (v30) and made use of HHpred8 to further classify host genomes based on pairwise ANI (cut-offs: 95% nucleotide identity on ≥95% alignment fraction). Prevalence was calculated at the host genus rank as the number of genomes with one or more inovirus sequence detected. Co-occurrence of inoviruses was evaluated based on the detections of distinct species in the same host genomes. Finally, we evaluated the rate of bacteria and archaea co-infected by an inovirus and a member of the Caudovirales order, the group of dsDNA viruses including most of the characterized bacteriophages (both lytic and temperate) as well as several archaeoviruses. To identify Caudovirales infections, we used the gene encoding the terminase large subunit as a marker gene, and searched the same genomes from the IMG database for hits to the PFAM domains terminalase_1, terminalase_3, terminalase_6 and terminase_GpA (hmmssearch, score ≥ 30).

Phylogenetic trees of inovirus sequences. Phylogenies of inovirus sequences were based on multiple alignment of pl protein sequences. To obtain informative multiple alignments of all pl families of all pI proteins and used to identify the nearest neighbours of sequences of interests. For sequences detected in archaeal genomes, an additional 10 most closely related sequences with ≤ 0.001, bit score ≥ 50 and a blast hit covering ≥50% of the query sequence were recruited for each archaea-associated sequence to help populate the tree. A similar approach was used for the tree based on the integrase genes from archaea-associated inoviruses: the protein sequences for the three integrase genes were compared with the NCBI nr database with blastp (bit score ≥ 50, ≤ 0.001) to gather their closest neighbours across archaeal and bacterial genomes.

Results data sets were first filtered for partial sequences as follows: the average sequence length was calculated excluding the top and bottom 10%, and all sequences shorter than half of this average were excluded. These protein sequences were next aligned with MUSCLE (v3.8.151)41, automatically trimmed with trimAL (v1.4.rev15)42 (option gappyout), and trees were constructed using IQ-TREE (v1.5.5) with an automatic detection of optimal model43 and displayed using iTOL44. The optimal substitution model, selected based on the Bayesian information criterion, was VT + F + R55 + G (model.autoselect:shortest.autosequences) for the archaea inoviruses and LG + R4 for the intece phylogeny of archaeal inoviruses. Annotated trees are available at http://itol.embl.de/shared/Siroyx (project ‘Inovirus’).

Functional affiliation of iPFs. An automatic functional affiliation of all iPFs was provided through COG assignments16. These were then compared to PFAM (data extracted from the IMG). To refine these annotations for functions of interest, namely, replication initiation proteins, integration proteins, DNA methylases and toxin–antitoxin systems, individual iPF alignments were submitted to the HHPred website72, and the alignments were visually inspected for conserved residues and/or motifs (Supplementary Table 5, motifs extracted from refs. 73 and the PFAM database30 (ref. 73)).

To identify toxin–antitoxin protein partners, all inovirus sequences were screened for co-occurring genes including an iPF annotated as toxin and/or antitoxin, and the list of putative pairs was next manually curated (Supplementary Table 5). This enabled the identification of putative toxin proteins detected as single sequence hits and only observed in the pI pl proteins of inoviruses (see Supplementary Table 5).

Finally, putative structural proteins and DNA-interacting proteins were specifically searched for. Putative structural proteins were predicted as described above for the isolate reference genomes, that is, as sequences of 30–90 amino acids, after in silico removal of signal peptide, if detected, and displaying 1 or 2 mismatches. These were then compared to PFAM (data extracted from the IMG). To refine these annotations for functions of interest, namely, replication initiation proteins, integration proteins, DNA methylases and toxin–antitoxin systems, individual iPF alignments were submitted to the HHPred website72, and the alignments were visually inspected for conserved residues and/or motifs (Supplementary Table 5, motifs extracted from refs. 73 and the PFAM database30 (ref. 73)).
explored to identify the ones with a clear associated CRISPR–Cas system and to affiliate these systems to the different types described. Only spacers for which a gene could be identified in a region of ±10 kb were retained. The CRISPR–Cas system affiliations were based on the set of criteria already identified around the spacer and performed following the guidelines from ref. 7.7

For host genomes with a self-targeting spacer, additional (that is, non-inovirus) prophages were detected using virusSorter2.9. The number of distinct prophages was also estimated using the detection of large terminus subunits (hmmsearch against PFAM database, score ≥ 30). Putative Acr and anti-CRISPR-associated (Aca) prophages were detected through homology to previously described Acr systems80 (blastp, e ≤ 0.001 and score ≥ 30). Putative Acr and Aca proteins were identified by searching for HTH-domain-containing proteins identified based on HTH domains in the SMART database (see above) in inovirus sequences displaying a match to a CRISPR spacer extracted from the same host genome.

Microscopy and PCR investigation of a predicted provirus in M. profundus

MobM. M. profundus strain MobM cells were grown in anaerobic DSMZ medium 479 at 37 °C with 5 mM methanol added as a methanogenic substrate instead of trimethyamine16. After 35 h of growth, anaerobic motilis YC C was added to the culture at a final concentration of 1.0 µg ml−1 to induce the provirus. Samples were collected before and 4 h after induction and were filtered with 0.22-µm pore size polyethersulfone filters (Millipore, Fisher Scientific) to obtain a ‘cellular’ (≥0.22 µm) and a ‘viral’ (<0.22 µm) fraction.

The four types of samples (with or without induction, cellular and viral fractions) were prepared and imaged at the Molecular and Cellular Imaging Center, Indiana University, Bloomington, OH, USA. An equal volume of 2x fixative (6% glutaraldehyde and 2% paraformaldehyde in 0.1 M potassium phosphate buffer pH 7.2) was added directly to the culture post-induction. Of the medium, 30 µl was applied to a formvar and carbon-coated copper grid for 5 min, blotted and then stained with 2% uranyl acetate for 1 min. Samples were examined with a Hitachi H7500 electron microscope and imaged with the SIA-L12C (16 megapixel) digital camera.

PCRs were initially run for induced and non-induced samples on both size fractions with three pairs of primers: one internal to the predicted provirus (P primers), one spanning the insertion site (I primers) and one spanning the junction of the predicted excised circular genome (C primers). The reactions were conducted for 35 cycles with denaturation, annealing and extension cycles of 0.5, 0.5 and 1 min at 95°C, 52.0 and 72.0 °C, respectively. For C primers, numerous nonspecific amplification products were obtained with these conditions, and another set of PCRs was conducted with higher annealing temperatures of 56.5 °C and 57.5 °C, both in triplicates. The PCR product was then cleaned to remove polymerase, free dNTPs and primers (Zymo Research) and subsequently used as templates for Sanger sequencing. The resulting chromatograms were analysed using the R2 packages sangerseq2,3, sangeranalyse4 and readr5. The extracted primary sequences were aligned to the MobM genome using blastn6 and MUSCLE7,8, and the alignment was visualized with Jalview9.

Experimental characterization of hypothetical proteins from self-targeted Pseudomonas inoviruses. Hypothetical proteins predicted on inovirus prophages, which were (1) found in Pseudomonas genomes, (2) predicted to be targeted by at least one CRISPR spacer from the same genome, and (3) for which no acr locus could be identified anywhere else in the same genome, were selected for further functional characterization. The ten candidate genes were first codon optimized for expression in P. aeruginosa using an empirically derived codon usage table. Codon optimization and vendor defined synthesis constraints removal were performed before testing.

P. aeruginosa strains (PA14: p-lac I-C CRISPR–Cas, PA14 and 4386) were cultured on LB agar or liquid media at 37 °C. The pHERD30T plasmids were electroporated into P. aeruginosa strains, and LB was supplemented with 50 µg ml−1 gentamicin to maintain the pHERD30T plasmid. Phages DMS3macrIF1 by streaking out cells from a solid plate infection and screening for colonies resistant to superinfection by DMS3macrIF1. Lysogeny was confirmed by prophage induction. The same plasmid transformation approach was then used to test the effect of inovirus self-targeting on host cell viability.

Quantification and statistical analysis. Sequence similarity searches were conducted with thresholds of E-value ≤ 0.001 and bit score ≥ 30 or 50, the former being used mainly for short proteins. The different classifiers (random forest, conditional random forest and generalized linear model) used to identify inovirus sequences were evaluated using a tenfold cross-validation approach. For all boxplots, the lower and upper hinges correspond to the first and third quartiles, respectively, and the whiskers extend no further than ±1.5 times the interquartile range.

Reporting Summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

The following files are available at https://genome.jgi.doe.gov/petal/Inovirus/Inovirus_home.html: GB_files_inoviruses.zip: GenBank files of all representative genomes for each inovirus species; Ref_PCs_inoviruses.zip: PCs from the references (raw fasta, alignment fasta and hmm profile); iPS_inoviruses.zip: protein families from the extended inovirus data set (raw fasta, alignment fasta and hmm profile); MobM_C_primer_amphlicon.fasta: multiple sequence alignment of the C primer products with the Methanoblobus MobM genome (NZ_FOU1000007), confirming that C primer products span the excised genome. Accession numbers of all inovirus sequences used as reference are listed in Supplementary Table 1. Accession numbers of all genomes and metagenomes mined, including detailed information for each (meta)genome in which some inovirus sequences were detected are available in Supplementary Table 2. Finally, the list of all inovirus genome accession numbers, along with taxonomic and environmental distribution information, is provided in Supplementary Table 3.

Code availability

The set of scripts and models used to detect inovirus sequences is available at https://bitbucket.org/srouxjgi/inovirus/src/master/Inovirus_detector/.

Received: 12 February 2019; Accepted: 5 June 2019; Published online: 22 July 2019.

References

1. Rakonjac, J., Bennett, N. J., Spagnuolo, J., Gagic, D. & Russel, M. filamentous bacteriophages: biology, phage display and nanotechnology applications. Curr. Issues Mol. Biol. 13, 51–76 (2011).
2. Faquet, C. M. The diversity of single stranded DNA. Viral Divers. 7, 38–46 (2004).
3. Marvin, D. A., Symmons, M. F. & Strauss, S. K. Structure and assembly of M13 viruses. Microbiol. Mol. Biol. Rev. 55, 1–49 (1991).
4. Fauquet, C. M. The diversity of single stranded DNA. Viral Divers. 7, 38–46 (2004).
5. Marvin, D. A., Symmons, M. F. & Strauss, S. K. Structure and assembly of filamentous bacteriophages. Prog. Biophys. Mol. Biol. 114, 80–122 (2014).
6. Bradbury, A. R. M. & Marks, J. D. Antibodies from phage antibody libraries. J. Immunol. Methods 290, 29–49 (2004).
7. Nam, K. T. et al. Stamped microbattery electrodes based on self-assembled phage-mimetic nanoparticles. Nat. Nanotechnol. 13, 290–295 (2018).
8. Hu, Z. & Sun, W. Drug delivery vectors based on filamentous bacteriophages and phage-mimetic nanoparticles. Drug Deliv. 24, 1898–1908 (2017).
9. Henry, K. A., Arabiti-Ghahroudi, M. & Scott, J. K. Beyond phage display: non–E. coli applications of the filamentous bacteriophage as a vaccine carrier, therapeutic biologic, and bioconjugation scaffold. Front. Microbiol. 6, 755 (2015).
10. Ilyina, T. S. filamentous bacteriophages and their role in the virulence and evolution of pathogenic bacteria. Mol. Genet. Microbiol. Virol. 30, 1–9 (2015).
11. Shi, Y., Li, J. W. & Turner, P. E. Evolution of viralism from phage genes in experimental virus populations. Evolution 72, 707–712 (2018).
12. Sweere, J. M. et al. Bacteriophage trigger anti-viral immunity and prevent clearance of bacterial infection. Science 363, eaat9691 (2019).
42. Delcher, A. L., Bratke, K. A., Powers, E. C. & Salzberg, S. L. Identifying bacterial genes and endosymbiont DNA with Glimmer. Bioinformatics 23, 673–679 (2007).
43. Edgar, R. C. MUSCLE: a multiple sequence alignment method with reduced time and space complexity. BMC Bioinformatics 5, 113 (2004).
44. Remmert, M., Biegert, A., Hauser, A. & Soding, J. HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment. Nat. Methods 9, 173–177 (2011).
45. Lu, L., Niu, B., Zhu, Z., Wu, S. & Li, W. CD-HIT: accelerated for clustering the next-generation sequencing data. Bioinformatics 28, 3150–3152 (2012).
46. Eddy, S. R. Accelerated profile HMM searches. PLoS Comput. Biol. 7, e1001295 (2011).
47. Petersen, T. N., Brunak, S., Von Heijne, G. & Nielsen, H. SignalP 4.0: discriminating signal peptides from transmembrane regions. Nat. Methods 8, 775–780 (2011).
48. Krogh, A., Larsson, B., von Heijne, G. & Sonnhammer, E. L. S. Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. J. Mol. Biol. 305, 567–580 (2001).
49. Camacho, C. et al. BLAST+: architecture and applications. BMC bioinformatics 10, 421 (2009).
50. Liaw, A. & Wiener, M. Classification and regression by randomForest. R News 2, 18–22 (2002).
51. Strobl, C., Boulesteix, A. L., Kneib, T., Augustin, T. & Zeileis, A. Conditional variable importance for random forests. BMC Bioinformatics 9, 307 (2008).
52. Friedman, N., Hastie, T. & Tibshirani, R. Regularization paths for Cox’s proportional hazards model via coordinate descent. J. Stat. Softw. 39, 1–13 (2011).
53. Wickham, H. ggplot2: Elegant Graphics for Data Analysis (Springer, 2016).
54. R Core Team R: A Language and Environment for Statistical Computing (R Foundation for Statistical Computing, 2018).
55. Adrienensn, E. M., Krupovic, M. & Knezevic, P. Taxonomy of propharyc viruses: 2016 update from the ICTV bacterial and archaeal viruses subcommittee. Arch. Virol. 162, 1153–1157 (2017).
56. Kurtz, S. et al. Versatile and open software for comparing large genomes. Genome Biol. 5, R12 (2004).
57. Alva, V., Nam, S.-Z., Söding, J. & Lupas, A. N. The MPI bioinformatics toolkit as an integrative platform for advanced protein sequence and structure analysis. Nucleic Acids Res. 44, W410–W415 (2016).
58. Demchak, B. et al. Cytoscape: the network visualization tool for GenoSpace workflows. F1000Res. 3, 151 (2014).
59. Muhire, B. M., Varanasi, A. & Martin, D. P. SDT: a virus classification tool based on pairwise sequence alignment and identity calculation. PLoS ONE 9, e108277 (2014).
60. Elof-Fadros, E. A. et al. Global metagenomic survey reveals a new bacterial candidate phylum in geothermal springs. Nat. Commun. 7, 10476 (2016).
61. Yu, B. et al. Microbiocidal-based mim-metagenomes enables discovery of novel microbial lineages from complex environmental samples. eLife 6, e26580 (2017).
62. Parks, D. H. et al. Recovery of nearly 8,000 metagenome-assembled genomes substantially expands the tree of life. Nat. Microbiol. 2, 1533–1542 (2017).
63. Tatusov, R. L. The COG database: a tool for genome-scale analysis of protein families. Nucleic Acids Res. 28, 33–36 (2000).
64. Katoh, K. & Standley, D. M. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol. Biol. Evol. 30, 772–780 (2013).
65. Crisculo, A. & Gribaldo, S. BMGE (block mapping and gathering with entropy): a new software for selection of phylogenetic informative regions from multiple sequence alignments. BMC Evol. Biol. 10, 210 (2010).
66. Price, M. N., Dehal, P. S. & Arkin, A. P. FastTree 2—approximately maximum-likelihood trees from multiple sequence alignments. BMC Evol. Biol. 10, 210 (2010).
67. Price, M. N., Dehal, P. S. & Arkin, A. P. FastTree 2—approximately maximum-likelihood trees for large alignments. PLoS ONE 5, e9490 (2010).
68. Mulkherjee, S. et al. Genomes Online Database (GOLD) v6: data updates and feature enhancements. Nucleic Acids Res. 45, D446–D456 (2017).
69. Hunter, I. D. Matplotlib: a 2D graphics environment. Comput. Sci. Eng. 9, 90–95 (2007).
70. Capella-Gutiérrez, S., Silla-Martínez, J. M. & Gabaldón, T. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. Bioinformatics 25, 1972–1973 (2009).
71. Nguyen, L. T., Schmidt, H. A., Von Haeseler, A. & Minh, B. Q. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. Mol. Biol. Evol. 32, 268–274 (2015).
72. Letunic, I. & Bork, P. Interactive Tree of Life (iTOL) v3: an online tool for the display and annotation of phylogenetic and other trees. Nucleic Acids Res. 44, W242–W245 (2016).
73. Krupovic, M. Networks of evolutionary interactions underlying the polyphyletic origin of ssDNA viruses. Curr. Opin. Virol. 3, 578–586 (2013).
74. Carr, S. B., Phillips, S. E. V. & Thomas, C. D. Structures of replication initiation proteins from staphylococcal antibiotic resistance plasmids reveal protein asymmetry and flexibility are necessary for replication. Nucleic Acids Res. 44, 2417–2428 (2016).
74. Finn, R. D. et al. The Pfam protein families database: towards a more sustainable future. *Nucleic Acids Res.* **44**, D279–D285 (2016).

75. Kelley, L. A., Mezulis, S., Yates, C., Wass, M. & Sternberg, M. The Phyre2 web portal for protein modelling, prediction, and analysis. *Nat. Protoc.* **10**, 845–858 (2015).

76. Letunic, I. SMART 4.0: towards genomic data integration. *Nucleic Acids Res.* **32**, 142D–144D (2004).

77. Makarova, K. S. et al. An updated evolutionary classification of CRISPR–Cas systems. *Nat. Rev. Microbiol.* **13**, 722–736 (2015).

78. Mochimaru, H. et al. *Methanobobus profundi* sp. nov., a methylothetic methanogen isolated from deep subsurface sediments in a natural gas field. *Int. J. Syst. Evol. Microbiol.* **59**, 714–718 (2009).

79. Hill, J. T. et al. Poly peak parser: method and software for identification of unknown indels using sanger sequencing of polymerase chain reaction products. *Dev. Dyn.* **43**, 1632–1636 (2014).

80. Lanfear, R. sangeranalyseR: a suite of functions for the analysis of Sanger sequence data in R v.1.20.0 (2015).

81. Wickham, H., Hester, J. & Francois, R. readr: read rectangular text data v.1.3.1 (2017).

82. Waterhouse, A. M., Procter, J. B., Martin, D. M. A., Clamp, M. & Barton, G. J. Jalview version 2—a multiple sequence alignment editor and analysis workbench. *Bioinformatics* **25**, 1189–1191 (2009).

83. Oberortner, E., Cheng, J. F., Hillson, N. J. & Deutsch, S. Streamlining the design-to-build transition with build-optimization software tools. *ACS Synth. Biol.* **6**, 485–496 (2017).

84. Qiu, D., Damron, F. H., Mima, T., Schweizer, H. P. & Yu, H. D. PRAD-based shuttle vectors for functional analysis of toxic and highly regulated genes in *Pseudomonas* and *Burkholderia* spp. and other bacteria. *Appl. Environ. Microbiol.* **74**, 7422–7426 (2008).

Acknowledgements

The MobM strain was provided by D. J. Ferguson, Miami University, Oxford, OH, USA. Its genome was sequenced and assembled by the US Department of Energy Joint Genome Institute through a Community Science Program initiative to K.C.W. (CSP no. 1777). T. Meulam at the Molecular and Cellular Imaging Center, Ohio State University, Wooster, OH, USA, performed the transmission electron microscopy of MobM samples. We gratefully acknowledge the contributions of many principal investigators who sent extracted DNA for isolate genome and metagenome sequencing as part of the Department of Energy Joint Genome Institute Community Science Program, and allowed us to include in our study the inovirus sequences detected in these publicly available data sets regardless of publication status (the complete list of data sets in which inovirus sequences were detected including principal investigators is available in Supplementary Table 2). This work was conducted by the US Department of Energy Joint Genome Institute, a Department of Energy Office of Science User Facility; under contract no. DE-AC02–05CH11231 and used resources of the National Energy Research Scientific Computing Center, which is supported by the Office of Science of the US Department of Energy under contract no. DE-AC02–05CH11231. R.A.D. and K.C.W. were partially supported by funding from the National Sciences Foundation Dimensions of Biodiversity Award (1342701). M.K. was supported by l’Agence Nationale de la Recherche (France) project ENVIRA (ANR-17-CE15-0005-01). The Bondy-Denomy lab (A.L.B. and J.B.-D.) is supported by the UCSF Program for Breakthrough in Biomedical Research, funded in part by the Sandler Foundation, the NIH Office of the Director (DP5-OD021344) and NIGMS (R01GM127489). Research of P.B.M.C. was funded by the US Department of Energy award DE-AC02–05CH11231. Funding for A.S. was provided by the National Science Foundation grant EAR 1331940 (the Eel River Critical Zone Observatory).

Author contributions

S.R., M.K. and E.A.E.-F. conceived the study. R.A.D. and A.L.B. designed the archaeal inovirus induction and functional characterization of inovirus genes in *Pseudomonas* experiments, respectively. A.S. and P.B.M.C. contributed unpublished metagenomic data. S.R. and M.K. performed the data and metadata curation. S.R. developed the computational tools. S.N. and F.S. contributed additional computational analyses. R.A.D., J.-F.C. and A.L.B. performed the experiments. S.R., M.K., R.A.D., A.L.B., S.N., F.S., J.-F.C., N.N.I., J.-B.-D., A.V., N.C.K. and E.A.E.-F. designed and wrote the manuscript. All authors reviewed and corrected the final manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41564-019-0510-x.

Reprints and permissions information is available at www.nature.com/reprints.

Correspondence and requests for materials should be addressed to S.R. or E.A.E.-F.

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

This is a U.S. government work and not under copyright protection in the U.S.; foreign copyright protection may apply 2019

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article’s Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article’s Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit http://creativecommons.org/licenses/by/4.0/.
Statistics
For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

- n/a
- A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
- The exact sample size (n) for each experimental group/condition, given as a discrete number and unit of measurement
- The statistical test(s) used AND whether they are one- or two-sided
- A description of all covariates tested
- A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
- A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
- For null hypothesis testing, the test statistic (e.g. F, t, r) with confidence intervals, effect sizes, degrees of freedom and P value noted

Give P values as exact values whenever suitable.
For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
Estimates of effect sizes (e.g. Cohen’s d, Pearson’s r), indicating how they were calculated

Software and code
Policy information about availability of computer code

Data collection
Data were collected using a custom set of scripts specifically designed to identify inovirus genomes. These are available at https://github.com/simroux/Inovirus/tree/master/Inovirus_detector

Data analysis
For specific reference inoviruses, genes were predicted de novo using Glimmer v3. Sequence similarity searches were conducted using blast+ v2.7.1, hmpmr [online at https://toolkit.tuebingen.mpg.de/#/tools/hmpmr] and hmschear v2.0.15. Sequences were clustered using InfoMap 0.18.25, mummer 3.0, SDT 1.0, and cd-hit 4.7. Viral sequences (non-inoviruses) were automatically detected using VirSorter v1.0.5. Signal peptide and transmembrane domains were predicted using SignalP 4.1 and TMHMM 2.0c. Trees were built using FastTree2 and IQ-Tree 1.5.5, based on alignments computed with muscle 3.8 or MAFFT v7.294b and automatically trimmed with trimAL v1.4 or BMGE v1.12. Alignments were manually inspected using Jalview v10.0.2. Statistical analyses, sanger sequenced reads interpretation, and automatic classifier design were conducted in R 3.4.3, using the following packages: randomForest, party, glmnet, sangerseqR, sangeranalyseR, and readR. Secondary structure of putative inovirus major capsid proteins were predicted using Phyre v2.0. Figures were generated with R 3.4.3 using the ggplot2 package, Cytoscape v3.6.1, ITOL v4.4.1, and python v3.6.2 using matplotlib v2.0.2.
Constraints in sequences to be synthesized were automatically identified and adjusted using BOOST v1.3.3.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data
Policy information about availability of data
All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:
- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

The following data are available at https://genome.jgi.doe.gov/portal/Inovirus/Inovirus.home.html:
Gb_files_inoviruses.zip: GenBank files of all representative genomes for each inovirus species.
Ref_PCs_inoviruses.zip: Protein clusters from the references (raw fasta, alignment fasta, hmm profile).
iPFs_inoviruses.zip: Protein families from extended inovirus dataset (raw fasta, alignment fasta, hmm profile).
MobM_C_primer_amplicon.fasta: Multiple sequence alignment of the C primer products with Methanologus MobM genome (NZ_FOUJ01000007) confirming that C primer products span the junction of the excised genome.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☐ Life sciences    ☐ Behavioural & social sciences    ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

| Sample size | No sample size calculation was performed, as the largest collection of publicly available data possible was mined. |
|-------------|--------------------------------------------------------------------------------------------------|
| Data exclusions | No data were excluded. |
| Replication | None of the findings was found to be impossible to replicate. This includes PCR amplification of the putative archaeal inovirus provirus, which was repeated either two or three times with similar results (see Supplementary Fig. 11), and the superinfection experiments which were conducted twice and produced similar results. |
| Randomization | None of the analyses involved allocation of samples to different groups. |
| Blinding | None of the analyses required blind investigation since the study does not involve a treatment vs control trial (with the exception of "obvious" negative controls such as "no template" PCR). |

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your research. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

| Materials & experimental systems | Methods |
|----------------------------------|---------|
| n/a | Involved in the study | n/a | Involved in the study |
| ☒ Antibodies | ☐ ChiP-seq | ☒ | Flow cytometry |
| ☒ Eukaryotic cell lines | ☒ MRI-based neuroimaging | ☒ | Palaeontology |
| ☒ Palaeontology | | ☒ | Animals and other organisms |
| ☒ Animals and other organisms | | ☒ | Human research participants |
| ☒ Human research participants | | ☒ | Clinical data |
| ☒ Clinical data | | | |

nature research | reporting summary

October 2018