Multiple origins of prokaryotic and eukaryotic single-stranded DNA viruses from bacterial and archaeal plasmids

Darius Kazlauskas, Arvind Varsani, Eugene V. Koonin & Mart Krupovic

Single-stranded (ss) DNA viruses are a major component of the earth virome. In particular, the circular, Rep-encoding ssDNA (CRESS-DNA) viruses show high diversity and abundance in various habitats. By combining sequence similarity network and phylogenetic analyses of the replication proteins (Rep) belonging to the HUH endonuclease superfamily, we show that the replication machinery of the CRESS-DNA viruses evolved, on three independent occasions, from the Reps of bacterial rolling circle-replicating plasmids. The CRESS-DNA viruses emerged via recombination between such plasmids and cDNA copies of capsid genes of eukaryotic positive-sense RNA viruses. Similarly, the rep genes of prokaryotic DNA viruses appear to have evolved from HUH endonuclease genes of various bacterial and archaeal plasmids. Our findings also suggest that eukaryotic polyomaviruses and papillomaviruses with dsDNA genomes have evolved via parvoviruses from CRESS-DNA viruses. Collectively, our results shed light on the complex evolutionary history of a major class of viruses revealing its polyphyletic origins.
Viruses with single-stranded (ss)DNA genomes represent a vast, highly diverse supergroup of medically, ecologically, and economically important pathogens infecting hosts from all three domains of cellular life. Although for years, ssDNA viruses have been thought to be relatively rare in the biosphere, recent metagenomics studies have increasingly revealed high abundance of these viruses in diverse environments. Currently, ssDNA viruses are classified into 13 families, 9 of which include (presumably) eukaryotic viruses, but many uncultivated ssDNA viruses remain unclassified. The majority of ssDNA viruses (9 families) have small circular genomes, which are known or predicted to be replicated by the rolling-circle mechanism. This mechanism of replication is initiated by the virus-encoded Rep protein of the HUH endonuclease superfamily, characterized by the signature HUH motif, in which two histidine residues are separated by a bulky hydrophobic residue. Informally, these viruses are often collectively referred to as circular, Rep-encoding ssDNA (CRESS-DNA) viruses. A variation on this theme is employed by members of the Parvoviridae family which have linear ssDNA genomes replicated by the rolling-hairpin mechanism initiated by a Rep protein homologous to those of the CRESS-DNA viruses. Members of the family Bidnaviridae apparently have evolved from paroviruses by replacing the HUH endonuclease domain with the DNA polymerase from pointoviruses.

The Rep proteins of ssDNA viruses of prokaryotes (bacteria and archaea) and eukaryotes display distinct domain organizations. In eukaryotic CRESS-DNA viruses, the endonuclease domain is fused to a superfamily 3 helicase (S3H) domain, which is responsible for unwinding of the double-stranded (ds) DNA replicative intermediate and, in some viruses, packaging of the viral genome into assembled empty capsids. By contrast, none of the bacterial or archaeal ssDNA viruses isolated to date encodes a Rep fused to a helicase domain. Instead, these viruses recruit a cellular helicase for the same function. Such dichotomy in the domain organization of the Rep proteins raises questions regarding the evolutionary relationship between ssDNA viruses infecting hosts from different cellular domains. Furthermore, HUH Reps are not restricted to ssDNA viruses, but are also functional in several groups of bacterial and archaeal dsDNA viruses, including certain members of the families Sphaero-olivoviridae, Rudiviridae, Corticoviridae, and Myoviridae. However, the implications of the presence of this gene for potential evolutionary links between these dsDNA viruses and ssDNA viruses remain unclear.

The HUH endonucleases are also encoded by diverse bacterial and archaeal as well as several eukaryotic plasmids and transposons, some of which have been shown experimentally to replicate and/or transpose via the rolling-circle mechanism. The homology between the endonuclease domains of the viral and bacterial plasmid Reps has been initially inferred from the conservation of 3 signature motifs and subsequently validated by structural analyses. Motif I, UUTU (U denotes hydrophobic residues), is thought to be involved in the recognition of the origin of replication. Motif II, HUH, is involved in the coordination of divalent metal ions, Mg$^{2+}$ or Mn$^{2+}$, which are essential for endonuclease activity at the origin of replication. Motif III (YxxK/YxxXY, where x is any amino acid) is involved in dsDNA cleavage and subsequent covalent attachment of the Rep through the catalytic tyrosine residue to the 5’ end of the cleaved product. The HUH endonucleases encoded by prokaryotic plasmids, viruses and transposons can have either two or one catalytic tyrosine residue in the motif III, whereas all known eukaryotic Rep-encoding viruses contain a single tyrosine residue. Notably, whereas most prokaryotic Reps consist of standalone endonuclease domains, some bacterial plasmids encode Reps with the domain organization similar to that characteristic of eukaryotic ssDNA viruses, that is, a nuclease-helicase fusion. For example, it has been shown that Reps encoded by plasmids of phytoplasma, plant-pathogenic bacteria, show the highest sequence similarity to Reps of plant-infecting geminiviruses. However, whether the similar domain organization is a result of convergent evolution or whether it alludes to a more recent common ancestry of the corresponding replicons remains unclear.

Here, we systematically explore the relationships among Rep-encoding DNA viruses, plasmids, and transposons from all three cellular domains. We identify 8 previously undescribed families of integrative plasmids that are widespread across different bacterial phyla and show that they have seeded the eukaryotic CRESS-DNA virosphere on at least 3 independent occasions. Similarly, the origins of bacterial and archaeal ssDNA viruses replicating by the rolling-circle mechanism can be traced to different families of prokaryotic plasmids, emphasizing tight evolutionary connections between viruses and capsid-less mobile genetic elements (MGE).

## Results

### Global network of the HUH replicons

To explore the evolutionary history of the HUH endonucleases, we collected a dataset of HUH endonucleases—the only protein encoded by all these replicons—representing each family of viruses, plasmids, and transposons associated with hosts across all three cellular domains. In this analysis, we did not consider Mob relaxases involved in plasmid conjugation. Enzymes in this family encompass circularly permuted conserved motifs which complicate their sequence-based comparison with the HUH endonucleases involved in DNA replication or transposition. The resulting dataset included 8764 sequences. These were grouped based on pairwise similarity, and clusters were identified using a convex clustering algorithm ($p$-value threshold of 1e−08) with CLANS. This analysis revealed 33 clusters which varied in size from 7 to 2711 sequences (Supplementary data 1). Following an inspection of the connectivity between clusters (Fig. 1), we defined 2 orphan clusters and 2 superclusters, which displayed either no or very few connections to each other (Supplementary data 1). Nevertheless, comparison of the available high-resolution structures for representatives of both orphan clusters and the 2 superclusters unequivocally confirm their common origin.

Orphan cluster 1 includes a single family of IS200/IS605 transposons which are widespread in bacteria and archaea. The HUH endonucleases of the IS200/IS605 insertion sequences have been extensively studied structurally and biochemically, resulting in a comprehensive understanding of their functions.

Although IS200/IS605 transposases have a structural fold common to that of other HUH endonucleases and contain all 3 signature motifs, they did not show appreciable sequence similarity to any other cluster of HUH endonucleases and thus remained disconnected from sequences in other clusters. Nevertheless, sequence diversity within the IS200/IS605 cluster is comparable to that within other clusters.

Orphan cluster 2 includes Rep proteins that are conserved in hyperthermophilic archaeal viruses of the family Rudiviridae. Structural studies of the Rep protein from the rudivirus SIRV1 revealed the canonical HUH endonuclease fold and biochemical characterization of the protein confirmed the expected nicking and joining activities in vitro. Like the IS200/IS605 transposases, the rudiviral Rep cluster does not connect to other HUH endonucleases, including homologs from other families of archaeal viruses and plasmids.

Conceivably, the uniqueness of the 2 orphan clusters is linked to the unusual transposition and replication mechanisms
employed by the respective elements. Indeed, IS200/IS605 insertion sequences transpose by a unique peel-and-paste mechanism\(^3\), whereas rudiviruses, unlike most other viruses and plasmids replicating by the rolling-circle mechanism, contain relatively large (~35 kb) linear dsDNA genomes with covalently closed termini\(^4\).

Supercluster 1 is by far the largest and most diverse HUH assemblage that includes 24 clusters (Supplementary data 1). Of these 24 clusters, 15 contain Reps from bona fide extrachromosomal plasmids of which 7 clusters also include Reps from diverse ssDNA (\textit{Microviridae}, \textit{Inoviridae}, and \textit{Pleolipoviridae}) and/or dsDNA (\textit{Myoviridae} and \textit{Corticoviridae}) viruses of bacteria and archaea. Three clusters consist of Reps encoded by microviruses of the subfamilies \textit{Gokushovirinae} and \textit{Bullavirinae}, and Xanthomonas inovirus Cf1 (family \textit{Inoviridae}), respectively. Notably, phiX174-like microviruses (\textit{Bullavirinae}) display similarity exclusively to microviruses of the subfamily \textit{Gokushovirinae}, indicative of the Rep monophyly in the two subfamilies of the \textit{Microviridae}, despite high sequence divergence. The bacterial IS91 (including \textit{ISCR} subfamily) and eukaryotic Helitron family transposons, respectively, form two distinct clusters. The two groups of transposons are not directly connected to each other, but are linked to distinct groups of bacterial and, in the case of IS91, archaeal plasmids, suggesting independent origins from bacterial extrachromosomal replicons. It has been previously suggested that helitrons might represent a missing link between eukaryotic CRESS-DNA viruses, namely, geminiviruses, and bacterial HUH replicons\(^5\) or that helitrons evolved from geminiviruses\(^6\). However, in our analysis, helitrons do not connect to any of the groups of CRESS-DNA viruses, suggesting independent evolutionary trajectories, consistent with the recent findings\(^7\).

The remaining 5 clusters do not include any recognizable plasmid, viral or transposon sequences and thus are likely to represent new families of integrated MGE. Four of these groups are predominantly found in bacteria of the taxa Clostridiales, Actinobacteria, Neisseriales, and Bacteroidetes, respectively (labeled accordingly in Fig. 1), whereas the fifth group is specific to the candidate division MSBL1 (Mediterranean Sea Brine Lakes)\(^8\), a group of uncultured archaea found in different hypersaline environments. Most of the clusters display taxonomic uniformity at the domain level, i.e., clusters included either bacterial, or archaeal, or eukaryotic sequences (including the corresponding viruses and plasmids), suggesting that horizontal transfers of viruses or plasmids between host domains are infrequent. The two exceptions include the pUB110-like and IS91-like bacteria-dominated clusters, which include a handful of archaeal sequences. In the case of IS91 transposons, horizontal transfer from bacteria has been ascertained by phylogenetic analyses\(^9\). In addition, some of the clusters include sporadic sequences annotated as being eukaryotic; however, analysis of the corresponding contigs suggests that these are likely bacterial contaminants.

Of particular interest are the 7 clusters that include both viruses and plasmids. For instance, pEC316\_KPC-like cluster, besides plasmids, contains evolutionarily-unrelated viruses from 3 families, \textit{Myoviridae}, \textit{Corticoviridae}, and \textit{Inoviridae}, suggesting extensive horizontal spread of the \textit{rep} genes. Notably, Reps of inoviruses are distributed among 5 clusters. Given the scarcity of inoviral sequences in the pVT736-1-like and pUB110-like clusters, which include only Pseudomonas phage P3\(\beta\) and Propionibacterium phage B5, respectively, the directionality of gene transfer, from plasmids to the corresponding viruses,
Rps represent prophages. Furthermore, the pAS28-like cluster genomic neighborhoods suggests that only 6 of the remaining 60 whereas the rest come from bacterial genomes. Analysis of the Rep sequences, 2 of which are encoded by Pe194/pMV158-family plasmids, which contain only the nuclease domain, bacterial and algal SC2 Reps had the same nuclease domain organization as CRESS-DNA viruses. The same two-domain organization is also characteristic of the parovirus Reps. Thus, domain organization analysis corroborates the results of sequence clustering and further indicates that the bacterial SC2 Reps are more closely related to the Reps of eukaryotic viruses than to those from other prokaryotic plasmids and viruses.

We then sought to obtain additional information on the diversity and taxonomic distribution of the viral-like SC2 Reps that are encoded in bacterial genomes. Maximum likelihood phylogenetic analysis revealed 9 well-supported clades (Fig. 2a). Clustering and subsequent community detection analysis validated the 9 groups of bacterial Reps (Fig. 2b), where groups 1–3 correspond to the p4M-like cluster shown in Fig. 1, groups 4–8 to the pCPa-like cluster, and group 9 to the pPAPh2-like cluster. To emphasize their similarity to Reps of CRESS-DNA viruses, we refer to the 9 groups as pCRESS1 through pCRESS9. These groups displayed partially overlapping but distinct taxonomic distributions, covering several classes within 4 bacterial phyla (Supplementary Fig. 1 and Supplementary Table 1).

The majority of the Reps from pCRESS7 and pCRESS9 are encoded by extrachromosomal plasmids (Supplementary Table 1). By contrast, the vast majority (97.5%) of Reps found in other groups are encoded within mobile genetic elements site-specifically integrated into bacterial chromosomes (Supplementary Table 1; Fig. 2c; Supplementary Fig. 3; Supplementary Note 1). Notably, none of the elements encoded any homologs of currently known viral structural proteins (Supplementary Note 1). Collectively, these observations indicate that viral-like Reps in bacteria are encoded by diverse extrachromosomal and integrated plasmids.

Conserved features of bacterial and CRESS-DNA virus Reps. Sequence analysis showed that, despite considerable overall sequence divergence, Reps of pCRESS4 through 8 contain closely similar sequence motifs within the nuclease and helicase domains (Fig. 3), consistent with the results of the clustering and phylogenetic analyses (Fig. 2). In particular, these 5 pCRESS groups share a specific signature, YLxH (x, any amino acid) within motif III of the nuclease domain, which was not observed in Reps from pCRESS1–3 and 9 (Fig. 3). Thus, we refer to pCRESS4–8 collectively as the YLxH supergroup (rather than the pCPa-like cluster), to emphasize this shared feature. The YLxH signature was also conserved in Reps from the pE194/pMV158-like cluster, suggesting a closer evolutionary relationship between the two clusters, despite the fact that pE194/pMV158-like Reps lack the helicase domain. Also, pCRESS9 displays motifs similar to those of the P. pulchra plasmids and thus could be unified with these plasmids into a common assemblage. By contrast, pCRESS1, -2 and -3 (p4M-like cluster) display distinctive sets of motifs (Fig. 3; Supplementary Note 1).

Origin of the SF3 helicase domain. Sequence analyses suggest that the SF3 helicase domain-containing plasmid Reps, especially those from pCRESS2, pCRESS3, and pCRESS9, and P. pulchra, are closely related to the Reps of CRESS-DNA viruses. However, the directionality of evolution, i.e., whether plasmid Reps evolved from those of CRESS-DNA viruses or vice versa, is not obvious. Although it is tempting to take the absence of the helicase domain in the pE194/pMV158-like cluster as an indication that this group is ancestral to the helicase-containing Reps, it cannot be ruled out that the helicase domain was lost by these plasmids. Thus, we set out to investigate the provenance of the SF3 helicase domain in the plasmid and viral Reps. Sensitive sequence searches with HMMER against the nr30 database showed that the helicase domain appears obvious. Furthermore, many inoviruses do not encode HUH endonucleases, but rather encode replication initiators of an evolutionarily unrelated superfamily, Rep_trans (Pfam id: PF02486)45, which also abounds in bacterial plasmids.40, whereas inoviruses of the genus Vespertiliovirus lack Reps and instead replicate by transposition using IS3 and IS30 family transposases derived from the corresponding insertion sequences46. Collectively, these observations indicate that the replication modules of inoviruses have been exchanged with distantly related and even non-homologous replication modules from various plasmid and transposon families. Similarly, archaeal pleiopivirus are split between two clusters corresponding to different families of archaeal plasmids, pGRB1-like and pTP2-like, respectively, suggesting that exchange of replication-associated genes is common in bacterial and archaeal viruses with small, plasmid-sized genomes. In some cases, it is difficult to ascertain the viral versus plasmid membership of Reps encoded in cellular chromosomes because both types of MGE can integrate into the host genomes. For example, the XacF1-like cluster includes 62 Rep sequences, 2 of which are encoded by filamentous phages, whereas the rest come from bacterial genomes. Analysis of the genomic neighborhoods suggests that only 6 of the remaining 60 Reps represent prophages. Furthermore, the pAS28-like cluster includes one plasmid, pAS28 (ref. 47); however, related Reps have been previously identified in prophages48, but not in characterized viruses, giving the erroneous impression that the pAS28-like Rep is plasmid-exclusive. To further characterize the evolutionary relationships between Reps encoded by different types of MGE, we constructed maximum likelihood phylogenetic trees for the 7 clusters that included Reps from both viruses and plasmids (Supplementary Fig. 2a–g). The results of phylogenetic analyses suggest horizontal transfer of the rep genes between plasmids and viruses, with viral sequences typically being nested among plasmid-encoded homologs.

Supercluster 2 (SC2) consists of 7 clusters (Supplementary Table 1) which include all known classified and unclassified eukaryotic CRESS-DNA viruses, paroviruses, a cluster of plasmids from the red alga Pyropia pulchra49, and 4 clusters containing bacterial Rep sequences. The vast majority of the bacterial Reps in the pCPa-like and p4M-like clusters are encoded in bacterial genomes rather than in plasmids and have not been previously characterized. In our network, the CRESS-DNA viruses are connected to pCPa-like, p4M-like, pPAPh2-like and P. pulchra-like clusters, whereas the pE194/pMV158-like cluster does not form direct connections to the CRESS-DNA viruses, but rather encode replication initiators of filamentous paroviruses and derived endo- nuclease (Parvoviridae) cluster, including paroviruses and derived endogenous viruses integrated in various eukaryotic genomes, is loosely connected directly to the CRESS-DNA viruses, suggesting that paroviruses with linear ssDNA genomes share common ancestry with CRESS-DNA viruses which, by definition, have circular genomes. Intrigued by the seemingly close evolutionary connection between eukaryotic CRESS-DNA viruses and bacterial and algal Reps, we investigated these relationships in greater detail, as reported in the following sections.

The diversity of viral-like Reps in bacterial genomes. To investigate the extent of similarity between the Reps of eukaryotic CRESS-DNA viruses and non-viral replicons from SC2, we compared their domain organizations. With the exception of pE194/pMV158-family plasmids, which contain only the nuclease domain, bacterial and algal SC2 Reps had the same nuclease-helicase domain organization as CRESS-DNA viruses. The same two-domain organization is also characteristic of the parovirus Reps. Thus, domain organization analysis corroborates the results of sequence clustering and further indicates that the bacterial SC2 Reps are more closely related to the Reps of eukaryotic viruses than to those from other prokaryotic plasmids and viruses.
domains of plasmid and CRESS-DNA viral Reps are most closely related to those of eukaryotic positive-sense RNA viruses (order *Picornavirales* and family *Caliciviridae*) as well as the AAA+ ATPase superfamily\(^50,51\). In this analysis, we also included the SF3 sequences of parvoviruses, polyomaviruses, and papillomaviruses that are thought to be evolutionarily related to CRESS-DNA viruses\(^2,25\). Several groups of more distant SF3 helicases from viruses with large dsDNA genomes\(^52\) were disregarded. Due to the high sequence divergence and relatively short length, phylogenetic analyses of the SF3 helicase domains were not informative, resulting in star-shaped tree topologies, irrespective of the evolutionary models or taxonomic sampling used. However, clustering analysis based on pairwise similarities provided insights into the relationships between the different ATPase families (Fig. 4a). In particular, the close relationship between the SF3 helicase domains of bacterial Reps and CRESS-DNA viruses was clearly supported. Both groups connect to the RNA viruses, but only bacterial Reps, particularly those of the YLxH supergroup, show connections to AAA+ superfamily ATPases, namely, bacterial helicase loader DnaC and, to a lesser extent, DnaA and Cdc48-like ATPases (Fig. 4a). The closer similarity between the YLxH supergroup and bacterial AAA+ ATPases is supported by comparison of the catalytic motifs which revealed several shared derived characters, to the exclusion of other groups (Supplementary Fig. 4). At the same clustering threshold, neither eukaryotic DNA nor RNA viruses linked to any group of ATPases other than those from bacterial plasmids. The SF3 helicases of parvoviruses linked to those of CRESS-DNA viruses, consistent with the analysis of full-length Rep sequences (Fig. 1). Papillomaviruses and polyomaviruses formed 2 clusters which connected to each other and to parvoviruses.

This pattern of connectivity suggests a specific vector of evolution and appears to be best compatible with the following scenario. The SF3 helicase domain of bacterial plasmids evolved from a bacterial DnaC-like ATPase; this helicase domain was appended to the nuclease domain of Reps of *pE194/pMV158*-like plasmids yielding the ancestor of the YLxH supergroup; bacterial plasmid Reps were passed on to the CRESS-DNA viruses; the SF3 helicase of RNA viruses was horizontally acquired either from bacterial plasmids or, more likely, from eukaryotic CRESS-DNA viruses; CRESS-DNA viruses have spawned parvoviruses which in turn gave rise to papillomaviruses and polyomaviruses (Fig. 4b). The alternative scenario, under which SF3 helicases of eukaryotic RNA viruses gave rise to the universal bacterial DnaC and DnaA proteins, through bacterial plasmids, appears non-parsimonious and extremely unlikely. Indeed, DnaA is ubiquitous and essential in bacteria\(^50,51\), so the capture of the helicase from a plasmid would have to occur at the very origin of the bacterial domain of
with respect to their nuclease and helicase domains, to avoid CRESS-DNA viruses which were previously judged to be chimeric. We used a dataset of SC2 Reps, excluding Reps of plasmids and CRESS-DNA viruses. For phylogenetic analyses, we sought an adequate outgroup for the phylogeny of Reps from bacterial plasmids. This assemblage includes geminiviruses and genomoviruses joining pCRESS9 plasmids of phytoplasma, and the other one includes CRESSV6 and P. pulchra plasmids. Notably, P. pulchra plasmids appear to emerge directly from within the CRESSV6 diversity, with the closest relationship to the CRESSV6 subclade of viruses sequenced from wastewater samples. The relationship between geminiviruses/genomoviruses and pCRESS9 plasmids is not resolved in the phylogeny. However, clustering analyses strongly suggest that Reps of pCRESS9 plasmids evolved from geminiviruses-genomoviruses (Figs. 1 and 4). Consistent with this scenario, phytoplasmal pCRESS7 and pCRESS9 plasmids, despite encoding phylogenetically distinct Reps, share the gene content, namely, the copy number control protein, PRK06752-like SSB protein and conserved hypothetical protein (Supplementary Fig. 3g, i). Furthermore, geminiviruses and CRESSV6 encode homologous capsid proteins suggesting that they evolved from a common viral ancestor rather than converged from two groups of plasmids by capturing homologous capsid protein genes. Clade 2 includes bacterial Reps of pCRESS1–3 and, as a sister group, CRESS-DNA viruses of the families Nanoviridae/Alphasatellitidae, Smacoviridae, and Circoviridae as well as unclassified CRESSV1 through CRESSV5, whereas GasCSV-like viruses are nested within bacterial pCRESS2.

The robustness of the PhyML tree was validated by additional analyses (Supplementary Note 1), including (i) maximum likelihood phylogenetic analyses using RAxML and IQ-TREE, with alternative branch support methods (Figure S5); (ii) phylogenetic reconstruction using the 20-profile mixture model (Figure S5); (iii) alternative branch support methods; and (iv) maximum likelihood phylogenetic analyses using RAxML and IQ-TREE, with alternative branch support methods (Figure S5).
(iii) statistical analysis of the unconstrained and 3 constrained tree topologies (Supplementary Table 2). Collectively, these results indicate that the obtained tree topology is highly robust and is likely to accurately reflect the evolutionary history of Reps encoded by CRESS-DNA viruses and plasmids.

Notably, analysis of the conserved motifs (Fig. 3) suggests a specific association between the virus Reps in clade 1 and bacterial pCRESS3 (rather than pCRESS1–3 collectively), implying that the phylogenetic placement might be affected by ancient recombination events. Furthermore, bacilladnaviruses were omitted from the global phylogenetic tree because their Reps displayed unstable position in the phylogeny depending on the taxon sampling (Supplementary Fig. 6), possibly, due to the small number of available sequences, their high divergence and potential chimerism. Regardless, phylogenetic analysis strongly suggests that the majority of CRESS-DNA viruses, including circoviruses, smacoviruses, nanoviruses, and CRESSV1–5, evolved from a common ancestor with bacterial Reps of pCRESS1–3, whereas the uncultivated GasCSV-like viruses emerge directly from the bacterial pCRESS2 Reps (Fig. 5).

Discussion

Here, we explored the evolutionary relationships among different classes of bacterial, archaeal, and eukaryotic replicons encoding HUH endonucleases (Reps). Our analysis revealed widespread exchange of rep genes among bacterial and archaeal viruses and plasmids, with some of the Rep clusters being particularly promiscuous, as in the case of pEC316_KPC-like Reps which are encoded not only in the corresponding plasmid but also in evolutionarily unrelated bacteriophages from 3 different families. Conversely, Reps of filamentous bacteriophages (family Inoviridae) fall into 5 distinct HUH clusters, indicating that, in this
virus family, replication modules are readily exchangeable, presumably, for those better suited in particular hosts. Thus, the genome replication module of inoviruses shows extreme promiscuity and cannot serve as a phylogenetic marker, consistent with a recent analysis of 10,000 inovirus genomes\textsuperscript{15}, so that the family is held together by the shared morphogenetic module. By contrast, the rep is the only gene conserved in all CRESS-DNA viruses and can serve as a vertically transmitted character against which various evolutionary events associated with the diversification of this virus class are mapped. Ultimately, however, no single gene or even functional module can fully represent the evolution of a given virus group. Instead, a more "holistic" approach is needed, where the provenance of all or most virus genes is deciphered.

Our present analysis pinpoints the origins of the replication modules of CRESS-DNA viruses. We identified 9 groups of bacterial Reps which share the nuclease-helicase domain organization with CRESS-DNA viruses. These bacterial Reps are encoded by previously unknown plasmids integrated into the genomes of diverse bacteria. By tracing the evolution of the helicase domain, we inferred the likely vector of evolution, namely, from plasmid Reps to the Reps of CRESS-DNA viruses. Although the Reps of CRESS-DNA viruses are generally considered to be monophyletic\textsuperscript{17}, our analysis shows that this might not be the case sensu stricto. Instead, the CRESS-DNA virus diversity has likely been seeded on 3 independent occasions from different groups of bacterial plasmids at different stages of evolution (Fig. 6). Conversely and contrary to the previous conclusion\textsuperscript{33}, our results also indicate that CRESS-DNA viruses have given rise to (at least) 2 groups of plasmids in red alga and phytopathogenic phytoplasma, respectively. Thus, transitions between the virus and plasmid states appear to be bidirectional.
CRESSV1 assemblage including circoviruses, smacoviruses, nanoviruses and early stages of eukaryotic evolution, as in the case of the viral corresponding plasmids. Some of these events could have occurred at the inception of life involving symbiotic or parasitic bacterial donors of the corresponding genes from RNA viruses has occurred in eukaryotic cells, possibly involving endosymbiosis or parasitic bacterial donors of the corresponding plasmids. Some of these events could have occurred at the early stages of eukaryotic evolution, as in the case of the viral assemblage including circoviruses, smacoviruses, nanoviruses and CRESSV1–5. By contrast, GasCSV-like viruses probably emerged relatively recently. Given the close relationship between pCRESS2 and GasCSV-like viruses, the latter group might infect bacteria rather than eukaryotes. Alternatively, the transition from a plasmid to a CRESS-DNA virus ancestor could have occurred once and was followed by replacement of the rep genes with counterparts from other plasmids, resulting in the 3 contemporary lineages of CRESS-DNA viruses. However, given that neither Rep nor capsid proteins appear to be orthologous in the 3 virus groups, this alternative scenario cannot be substantiated at this point. Regardless, it is clear that Rep and capsid protein genes have been repeatedly exchanged with distantly related homologs from other viruses, even in the more recent history of CRESS-DNA viruses. Thus, these viruses, most likely, evolved from ssDNA viruses but their evolution involved a drastic change in both the genome DNA structure and the replication mechanism such that the HUH domain switched from an enzymatic to a structural role. Clustering analysis of the SF3 helicase domains suggests that both polyomaviruses and papillomaviruses are homologous to the HUH endonuclease domains of CRESS-DNA viruses and paroviruses. Thus, these viruses, most likely, evolved from ssDNA viruses but their evolution involved a drastic change in both the genome DNA structure and the replication mechanism such that the HUH domain switched from an enzymatic to a structural role. Clustering analysis of the SF3 helicase domains suggests that both polyomaviruses and papillomaviruses are homologous to the HUH endonuclease domains of CRESS-DNA viruses and paroviruses. The current classification of CRESS-DNA viruses largely relies on the phylogeny of the Rep proteins. The ever-growing diversity of sequenced CRESS-DNA virus genomes calls for revision of the taxonomy of this class of viruses. Our analysis reveals two larger groupings of CRESS-DNA viruses, each including several families/clades, which could be equivalent to new orders, whereas all CRESS-DNA viruses could be unified at a yet higher taxonomic level. Finally, the membership of paroviruses,
polymaviruses, and papillomaviruses in this assemblage could be also considered, at the highest taxonomic level. Indeed, it is not unprecedented that the same taxon contains viruses with different nucleic acids types. For instance, the order Ortervirales includes reverse-transcribing viruses with RNA and DNA genomes89, whereas members of the family Pleolipoviridae have either ssDNA or dsDNA genomes90. Notably, the ICTV has recently announced that taxonomic ranks above the order level are now officially accepted91, opening the door for the formal unification of the whole spectrum of evolutionarily related CRESS-DNA viruses.

Although the Reps of prokaryotic viruses of the families Microviridae, Inoviridae, and Sphaerolipoviridae lack the helicase domain, their HUH nucleocases show clear affinities with those from different groups of plasmids, suggesting routes of evolution parallel to those of the CRESS-DNA viruses. Furthermore, the evolution of inoviruses appears to have involved multiple replacements of the rep gene with those from various plasmids. The results presented here shed light on the origin of a major part of the virosphere, the ssDNA viruses replicating via the rolling-circle mechanism and, in particular, CRESS-DNA viruses. Arguably, evolution of the ssDNA viruses is the most compelling manifestation of the previously noted general trend in virus evolution, namely, tight evolutionary connections between viruses and capsid-less MGEs92,93.

**Methods** Databases. Homologs of the HUH endonucleases were retrieved by running searches against protein sequence databases filtered to 50 and 90% sequence identity (UniRef50 and UniRef90, respectively) which were downloaded from http://www.uniprot.org. Search for bacterial homologs of CRESS-DNA virus Reps was performed against nr90 (NCBI’s nr database (ftp://ftp.ncbi.nlm.nih.gov/blast/db/)) filtered to 90% identity. To detect remote sequence similarity, we used sequence profile databases which included profiles from PDB (www.pdb.org), SCOP94, Pfam95, and CDD96. For query profile generation nr70 database was used.

**Sequence searches and clustering.** Homologs of the HUH superfamily endonuclease domains for each representative Rep sequence were obtained by performing three jackhammer72 iterations against the UniRef50 database. Representative Reps were selected as queries for homology searches based on exhaustive review of literature on the HUH superfamily97,98. In addition, for HUH groups with less than 10 homologs in UniRef50, we repeated searches against UniRef90 database. The resultant set of sequences was grouped using a convex clustering method (program kmeans++) to define based on the presence of direct repeats corresponding to their full genomes assembled from virome reads. The integrated plasmids were identified by thorough analysis of genomic neighborhoods of the Rep-encoding genes. The precise borders of integration were defined based on the presence of direct repeats corresponding to attachment sites. The repeats were searched for using Unipro UGENE99. Genes of integrated plasmids were annotated based on the HHsearch searches72. Genome maps were compared and visualized using Easyfig with tBLASTx option80.

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

**Data availability** The authors declare that the data supporting the findings of this study are available within the paper and its supplementary information files. Accession numbers for all proteins analyzed in this study as well as alignments used to generate the trees shown in Figs. 2 and 5 are included in the Supplementary Data 1, 2, and 3.

Received: 6 October 2018 Accepted: 10 July 2019

**References**

1. Zhao, L., Rosario, K., Breitbart, M. & Duffy, S. Eukaryotic circular rep-encoding single-stranded DNA (CRESS DNA) viruses: ubiquitous viruses with small genomes and a diverse host range. Adv. Virus Res 103, 71–133 (2019).

2. Kropovic, M. Networks of evolutionary interactions underlying the polyphyletic origin of ssDNA viruses. Curr. Opin. Virol. 3, 578–586 (2013).

3. Wang, H. et al. Plasma virome of cattle from forest region revealed diverse small circular ssDNA viral genomes. Virol. J. 15, 11 (2018).

4. Chow, C. E. & Suttle, C. A. Biogeography of viruses in the sea. Annu Rev. Virol. 2, 41–66 (2015).

5. Labonte, J. M. & Suttle, C. A. Previously unknown and highly divergent ssDNA viruses populate the oceans. ISME J. 7, 2169–2177 (2013).

6. Roux, S., Kropovic, M., Poulet, A., Debroux, O. & Enault, F. Evolution and diversity of the Microviridae viral family through a collection of 81 new complete genomes assembled from virome reads. PLoS ONE 7, e40418 (2012).

7. Wang, Y. et al. The feline virome of red-crowned cranes. Arch. Virol. 164, 3–16 (2019).

**Multiple sequence alignments and phylogenetic analysis.** To construct multiple sequence alignments for phylogenetic analysis we used MAFFT90 and TrimAl91. MAFFT options G-INS-i and L-INS-i and TrimAl gap thresholds 0.05 and 0.15 were used to generate alignments for Figs. 2 and 5, respectively. The resulting alignments covered both HUH and SF3 (where available) domains and contained 743 and 508 positions, respectively. Both alignments can be found in the Supplementary Data 2 and 3. Phylogenetic trees were calculated with PhyML92 using automatic model selection and aBayes branch support. Substitution models VT G + I + F (VT, amino acid replacement matrix; G, gamma shape parameter: estimated (1.864); I, proportion of invariable sites: estimated (0.005); F, equilibrium frequencies: empirical) and LG + G (LG, amino acid replacement matrix; G, estimated by the program) substitution models were selected for phylogenetic analyses shown in Figs. 2 and 5, respectively. Additional trees were constructed using IQ-TREE v1.6.8 (ref. 79) with Ultrafast Bootstrap Approximation branch support90, and RAXML with non-parametric bootstrapping93. Mixture model tree was constructed with IQ-TREE79 using model parameters (LG + C20 + F + G) and ultrafast bootstrapping (with 1000 replicates). Alignment and guide tree (parameters “-s” and “-f”), respectively, were the same as in Fig. 5. Highly diverged sequences forming long branches were removed before constructing final trees. Bacilladnaiviridae viruses were also removed, because their position was not stable in trees with different sequence sampling (Supplementary figure 6). Phylogenetic trees are available from the authors upon request. The trees shown in Figs. 2, 3, 5 and 56 can be found in the Supplementary Data 4 to 10.
21. Krupovic, M. & Koonin, E. V. Evolution of eukaryotic single-stranded DNA viruses.

24. Scott, J. F., Eisenberg, S., Bertsch, L. L. & Kornberg, A. A mechanism of duplex DNA replication revealed by enzymatic studies of phage phi X174: catalytic NTP-binding domains encoded by genomes of small DNA and RNA viruses.

17. Rosario, K., Duffy, S. & Breitbart, M. A high density of ssDNA phages from the Family Microviridae detected within the gut of a protochordate model organism (Ciona robusta). Viruses 10, E404 (2018).

27. Forterre, P., Krupovic, M., Raymann, K. & Soler, N. Plasmids from hyperthermophilic archaea of the order Thermococcales from South Africa and New Zealand.

19. Koonin, E. V. & Ilyina, T. V. Computer-assisted dissection of rolling circle viruses of the Bidnaviridae family from genes of four other groups of widely different viruses.

30. Wawrzyniak, P., Plucienniczak, G. & Bartosik, D. The different faces of microviruses. Proc Natl Acad Sci USA 98, 8714–8719 (2001).

37. Filee, J., Siguier, P. & Chandler, M. Insertion sequence diversity in archaea. PeerJ 4, e121 (2016).

38. He, S. et al. The IS200/IS605 Family and replication-associated proteins of uncultured single-stranded DNA viruses. Front Microbiol 2, 11 (2011).

39. Prangishvili, D., Koonin, E. V. & Krupovic, M. Genomics and biology of Rudiviruses, a model for the study of virus-host interactions in Archaea.

45. Wang, Y. et al. Rolling-circle replication initiation protein of haloarchaeal phanerophages SR1: its selective binding to bacterial transposases of the IS91 family insertion sequences. J. Gen. Virol. 99, 416–421 (2018).

46. Forterre, P. Single-stranded DNA viruses employ a variety of mechanisms for integration into host organisms. Ann. N. Y Acad. Sci. 1341, 41–53 (2015).

47. Kato, J. et al. Development of a genetic transformation system for an alga-archaeal symbiont of a metazoan. Appl Environ. Microbiol 64, 2061–2064 (1998).

48. Krupovic, M. & Bamford, D. H. Putative prophages related to lytic tailless viruses. Proc. Natl Acad. Sci. USA 105, 20079–20084 (2008).

53. Varsani, A. & Krupovic, M. Sequence-based taxonomic framework for the virus genome: implications for the Rep mechanism and viral replication. Arch. Virol. 155, 251–276 (2010).

54. Dayaram, A. et al. Novel ssDNA virus recovered from estuarine Mollusc (Amphithoe crenata) whose replication associated protein (Rep) shares similarities with Rep-like sequences of bacterial origin. J. Gen. Virol. 94, 1104–1110 (2013).

55. Diemer, G. S. & Stedman, K. M. A novel virus genome discovered in an alien virus community on Antarctica. Arch. Virol. 155, 2371–2382 (2010).

63. Varsani, A. & Krupovic, M. Sequence-based taxonomic framework for the virus genome: implications for the Rep mechanism and viral replication. Arch. Virol. 155, 251–276 (2010).

64. Oke, M. et al. A dimeric Rep protein initiates replication of a linear archael virus genome: implications for the Rep mechanism and viral replication. J. Virol. 85, 925–931 (2011).

65. Filee, J., Siguier, P. & Chandler, M. Insertion sequence diversity in archaea. Microbiol Mol. Biol. Rev. 71, 121–157 (2007).

66. He, S. et al. The IS200/IS605 Family and “Peel and Paste” Single-stranded DNA Transposition Mechanism. Microbiol. Spectr. 3, MDA2-0039-2014 (2015).

67. Krupovic, M., Ghabrial, S. A., Jiang, D. & Varsani, A. Genomoviridae: a new family of widespread single-stranded DNA viruses. Arch. Virol. 161, 2633–2643 (2016).

68. Varsani, A. & Krupovic, M. Sequence-based taxonomic framework for the classification of uncultured single-stranded DNA viruses of the family Genomoviridae. Virus Evol. 3, vewe025 (2016).

69. Diez-Villaescusa, C. & Rodriguez-Valera, F. CRISPR analysis suggests that small circular single-stranded DNA smacoviruses infect Archaea instead of humans. Nat. Commun. 10, 294 (2019).

70. Hickman, A. B., Ronning, D. R., Kotin, R. M. & Dyda, F. Structural unity among viral origin binding proteins: crystal structure of the nuclease domain of adenovirus-associated virus Rep. Mol. Cell 10, 327–337 (2002).

71. Krupovic, M., Ghabrial, S. A., Jiang, D. & Varsani, A. Genomoviridae: a new family of widespread single-stranded DNA viruses. Arch. Virol. 161, 2633–2643 (2016).

72. Varsani, A. & Krupovic, M. Sequence-based taxonomic framework for the classification of uncultured single-stranded DNA viruses of the family Genomoviridae. Virus Evol. 3, vewe025 (2016).
