Genome Integration and Reactivation of the Virophage Mavirus In the Marine Protozoan *Cafeteria roenbergensis*

Matthias G. Fischer[1]* and Thomas Hackl[1†]

[1]*Department of Biomolecular Mechanisms, Max Planck Institute for Medical Research, 69120 Heidelberg, Germany*

[†]*Present address: Massachusetts Institute of Technology, 15 Vassar Street, Cambridge, MA 02139, USA*

Endogenous viral elements are increasingly found in eukaryotic genomes, yet little is known about their origins, dynamics, or function. Here, we provide a compelling example of a DNA virus that readily integrates into a eukaryotic genome where it acts as an inducible antiviral defense system. We found that the virophage mavirus, a parasite of the giant virus CroV, integrates at multiple sites within the nuclear genome of the marine heterotrophic nanoflagellate *Cafeteria roenbergensis*. The endogenous mavirus is structurally and genetically similar to the eukaryotic Maverick/Polinton DNA transposons. Provirophage genes are activated by superinfection with CroV, which leads to the production of infectious mavirus particles. While provirophage-carrying cells are not directly protected from lysis by CroV, release of reactivated virophage particles promotes survival of other host populations. Our results corroborate the connection between mavirus and Maverick/Polinton elements and suggest that provirophages can defend natural protist populations against infection by giant viruses.

All viruses can potentially leave long-lasting imprints in cellular genomes. Some integrate into host genomes as part of their infection cycle; others lead exclusively lytic life styles, but can take advantage of rare stochastic events such as non-homologous DNA recombination or by exploiting helper functions from the cell and other viruses. To date, endogenous viral elements (EVEs) for all major groups of viruses have been identified in eukaryotic genomes[1]. Whereas viral insertions are often disadvantageous for the host, there are fascinating examples where EVEs have evolved a host benefit, such as coopted retroviruses in vertebrates[2–4], or the symbiotic relationship of polydnaviruses and parasitoid wasps[5]. However, for the majority of non-retroviral EVEs, neither host function nor the circumstances of their endogenization are known. Although most of the described EVEs are found in vertebrate genomes, their occurrence is not restricted to multicellular organisms. In particular the exogenous and endogenous viral spectrum of protists, which comprise the vast majority of eukaryotic diversity, remains largely untapped.

* Email: mfischer@mpimf-heidelberg.mpg.de

One of the biggest surprises in recent microbiological history was the discovery of protist-infecting giant viruses and their associated virophages. Giant viruses are double-stranded (ds) DNA viruses whose genomes can exceed 2000 kilobase pairs (kbp)[6,7]. They are members of the nucleocytoplasmic large DNA virus clade that includes the viral families *Ascoviridae, Asfarviridae, Iridoviridae, Marseilleviridae, Mimiviridae, Phycodnaviridae,* and *Poxviridae*[8,9], as well as the recently described pandoraviruses, pithoviruses, faustoviruses, and 'Mollivirus sibericum'[7,10–14]. Many giant viruses reproduce in cytoplasmic virion factories (VFs), where transcription, DNA replication, and particle assembly take place[15]. The presence of a viral transcription apparatus in these VFs permits the replication of so-called virophages, dsDNA viruses with 15-30 kbp genomes that parasitize giant viruses of the family *Mimiviridae*. Virophages are strictly dependent for their replication on a coinfecting giant virus[16,17]. The prediction that virophages are transcriptional parasites of giant viruses is based on regulatory signals shared by virophage and giant virus genes[17–19]. In addition, virophages encode their own morphogenesis and DNA replication genes and appear to be

autonomous for these processes. Coinfection with a giant virus and a virophage may result in decreased giant viral progeny and increased host survival rates.

Virophages are classified in the family *Lavidaviridae*[20] with currently three members: the amoeba-infecting Sputnik virus and Zamilon virus[16,21], and the Maverick-related virus (mavirus). Mavirus possesses a 19,063 bp circular dsDNA genome that is packaged inside a ≈75 nm wide icosahedral capsid[17]. The cellular host for mavirus is the heterotrophic nanoflagellate *Cafeteria roenbergensis*[22], a bacterivorous protist that is commonly found in marine environments and reproduces by binary fission. The viral host for mavirus is Cafeteria roenbergensis virus (CroV), a ≈700 kbp dsDNA virus with a 300 nm large capsid that lyses its host within 24 hours post infection (h p.i.)[23]. The mavirus genome codes for 20 proteins, seven of which have homologs among a group of mobile DNA elements called Mavericks or Polintons.

Maverick/Polinton elements (MPEs) are present in various eukaryotic lineages and stand out from other DNA transposons due to their size (15-20 kbp) and the viral nature of their genes[24–26]. All MPEs encode a retroviral integrase (rve-INT) and a protein-primed DNA polymerase B (pPolB); most elements also encode an FtsK-HerA-type genome packaging adenosine triphosphatase (ATPase), an adenovirus-like cysteine protease (PRO), and sporadically a superfamily 3 helicase (HEL). Two additional conserved MPE genes were recently identified as distant versions of the jelly-roll-fold minor and major capsid protein genes that are also encoded by virophages[26]. Whereas capsidless MPEs likely spread as transposons, the capsid-encoding MPEs can be considered endogenous viruses ("polintoviruses") and may in fact be the most broadly distributed family of EVEs among eukaryotes[27–29]. Conversely, mavirus-like virophages can be viewed as the infectious form of MPEs. Although the common evolutionary origin of MPEs and virophages is apparent, the directionality of this process is a matter of debate[17,28,30]. A central role in the virophage-MPE connection falls onto the integrases encoded by virophages. Sputnik carries a tyrosine recombinase and can integrate into the genome of the giant Lentille virus[31]. MPEs, mavirus, as well as a family of endogenous virophage-like elements found in the alga *Bigelowiella natans*[32] encode an rve-INT. Despite the widespread occurrence of integrase genes in virophages, it remains unclear under which conditions and how frequently these viruses are able to integrate into eukaryotic genomes.

This motivated us to test the endogenization potential of mavirus in its host *C. roenbergensis*. We show that mavirus readily integrates into the nuclear genome of *C. roenbergensis*, where it is vertically transmitted. We genetically characterized a host strain that carries more than eleven *de novo* mavirus integrations and demonstrate close structural similarity between the endogenous mavirus elements and MPEs. Our investigation reveals an inducible model system to study the integration and reactivation of a eukaryotic DNA virus. Furthermore, we show that provirophages can act as a kin-based defense system against giant viruses in protists.

Results

Isolation of a mavirus-positive host strain

We devised a straight-forward infection experiment to test whether mavirus was able to integrate *in vivo* into the nuclear genome of *C. roenbergensis* (Figure 1). For our experiments, we chose *C. roenbergensis* strain E4-10 (originally misclassified as *Bodo sp.*) that was isolated in the early 1990s from Pacific waters off the coast of Oregon, USA[33], and which had proven to be a productive host for CroV and mavirus[23]. To ensure that the host strain was genetically homogeneous, we performed single-cell dilutions and established clonal populations. This procedure was serially repeated two more times and one of the resulting clonal strains was selected for further experiments (Figure 1A). PCR testing of multiple mavirus target genes confirmed the absence of mavirus-specific sequences in this strain that we named E4-10P, being the parental strain for the study. The E4-10P strain was then either mock-infected with culture medium or infected with CroV at a multiplicity of infection (MOI) of 0.01 and with mavirus at an MOI of ≈1 (in contrast to CroV, no direct infectivity assay exists for mavirus and its MOIs are estimated from qPCR data). Under these conditions, mavirus inhibits CroV reproduction sufficiently to prevent complete lysis of the cell population (Figure S1). We then screened the surviving cells for host-integrated mavirus. After the survivor cells had been pelleted and washed

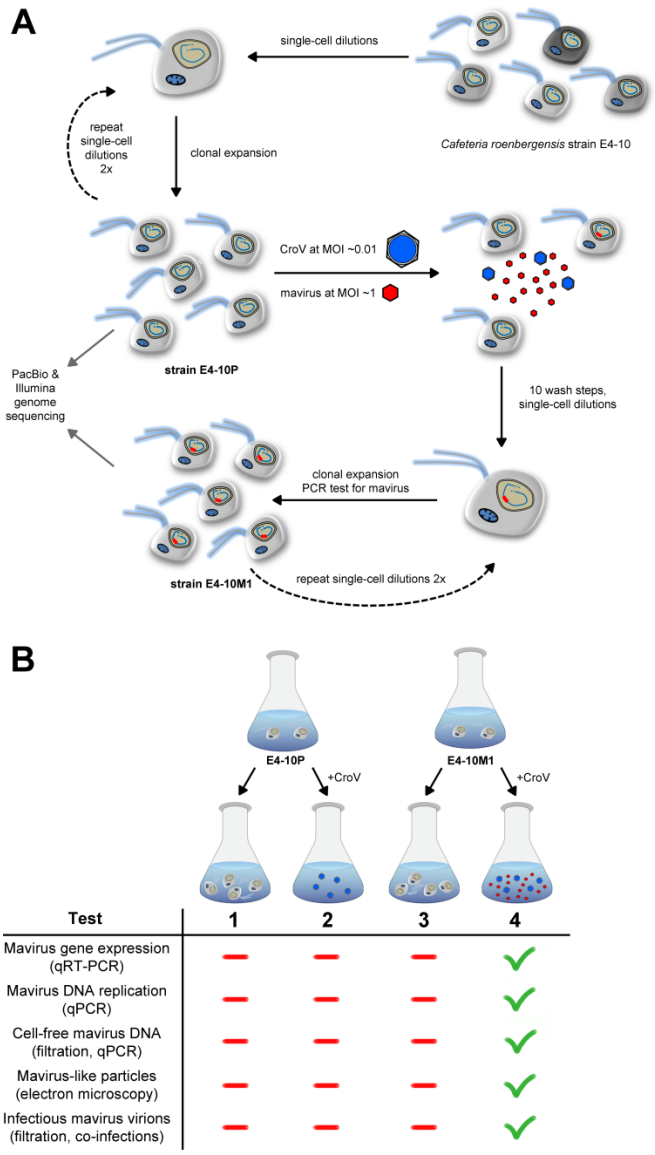

Figure 1: Experimental strategy to demonstrate integration and reactivation of mavirus. *(A) Strain E4-10 of the marine zooplankter* C. roenbergensis *was made clonal by repeatedly growing populations from single cells. The resulting strain E4-10P was infected with CroV and mavirus (see Figure S1). Surviving cells were washed free from residual virus particles and clonal strains were established. One of the strains that tested PCR-positive for mavirus was named E4-10M1 and further characterized. The genomes of E4-10P and E4-10M1 were sequenced on PacBio and Illumina MiSeq platforms to analyze mavirus integration sites. (B) Experiments conducted with E4-10P and E4-10M1 cells in this study.*

ten times, we performed three consecutive rounds of single-cell dilutions and extracted DNA from 66 of the resulting clonal strains. PCR analysis with mavirus-specific primers identified 21 (32%) mavirus-positive clonal populations. We chose the cell line with the highest qPCR signal for further analysis and named it E4-10M1 (for the first mavirus-positive strain). In order to confirm that the observed mavirus signal was associated with host cells and not caused by remaining free virus particles, we filtered the cell population through 5.0 µm, 0.45 µm, 0.22 µm or 0.1 µm pore-size syringe filters. *C. roenbergensis* cells are 5-10 µm in diameter and were retained by 5.0 µm and smaller pore-size filters, as confirmed by microscopy. In contrast, mavirus particles are ≈75 nm in diameter and will pass even through a 0.1 µm pore-size filter. DNA extracted from the filtrates was analyzed by qPCR with mavirus-specific primers. As a negative control, the same procedure was applied to E4-10P cells. A mavirus-positive signal was found only in the unfiltered E4-10M1 sample, strongly suggesting that mavirus DNA in host strain E4-10M1 was associated with cells and not with extracellular virions (Table 1).

Computational analysis confirms mavirus integration

In order to assess whether the mavirus genome was integrated in the nuclear genome of strain E4-10M1, or whether it persisted extra-chromosomally, we sequenced genomic DNA from strains E4-10P and E4-10M1 on Illumina MiSeq and Pacific BioSciences (PacBio) RS II platforms. The PacBio reads were error-corrected with the trimmed paired-end MiSeq reads, and we created hybrid assemblies for each strain from the paired-end MiSeq and high accuracy PacBio reads. The read data suggested that *C. roenbergensis* has a diploid genome (Figure S2), which obstructed the direct assembly of mavirus-containing contigs because integration at a specific site occurred at only one of the two alleles, thus introducing a structural ambiguity. The assembly software resolved this ambiguity by either producing separate contigs or by parsimoniously ignoring the mavirus-containing alleles altogether. We therefore scanned the E4-10M1 genome assembly indirectly for integrated mavirus sequences by aligning corrected PacBio reads to the mavirus

Condition	Host strain	Targets per mL	Unfiltered	5.0 µm filtrate (CN)	0.45 µm filtrate (PES)	0.22 µm filtrate (PES)	0.1 µm filtrate (PVDF)
Uninfected	P	Cells	(1.53±0.18)E+06	(9.20±4.20)E+03	BDL	BDL	BDL
		Mavirus	BDL	BDL	BDL	BDL	BDL
		CroV	BDL	BDL	BDL	BDL	BDL
	M1	Cells	(1.39±0.14)E+06	(5.50±0.00)E+03	BDL	BDL	BDL
		Mavirus	(4.03±0.61)E+06	(1.04±0.46)E+04	BDL	BDL	BDL
		CroV	BDL	BDL	BDL	BDL	BDL
CroV-infected	P	Cells	BDL	BDL	BDL	BDL	BDL
		Mavirus	BDL	BDL	BDL	BDL	BDL
		CroV	(5.40±0.87)E+07	(1.35±0.10)E+06	(2.34±1.00)E+06	(3.11±0.47)E+04	(9.55±4.00)E+04
	M1	Cells	BDL	BDL	BDL	BDL	BDL
		Mavirus	(1.89±0.17)E+09	(1.25±0.05)E+09	(9.90±0.53)E+08	(8.97±0.71)E+08	(4.76±0.59)E+08
		CroV	(1.41±0.12)E+08	(3.03±0.15)E+07	(1.00±0.03)E+07	(1.48±0.48)E+04	(1.18±0.20)E+04
Mavirus-spiked	P	Cells	(1.83±0.19)E+06	(2.29±0.64)E+04	BDL	BDL	BDL
		Mavirus	(1.22±0.02)E+09	(7.37±4.67)E+08	(7.64±4.52)E+08	(9.74±0.72)E+08	(7.23±0.14)E+08
		CroV	BDL	BDL	BDL	BDL	BDL
Mechanical lysis	M1	Cells	(1.83±1.41)E+04	BDL	BDL	BDL	BDL
		Mavirus	(1.24±0.05)E+07	(1.15±0.04)E+07	(1.16±0.02)E+07	(1.17±0.05)E+07	(1.17±0.10)E+07
		CroV	BDL	BDL	BDL	BDL	BDL

Table 1: Cell and virus concentrations in different size fractions of mock-infected and CroV-infected E4-10P and E4-10M1 populations.
Uninfected host cultures or CroV-infected cultures after cell lysis were passed through syringe filters of various nominal pore sizes. As controls, E4-10P cells were spiked with mavirus particles immediately prior to filtration, and uninfected E4-10M1 cells were mechanically lysed by sonication and then filtered. DNA extracted from identical volumes of each filtrate was used as template in qPCR assays with mavirus- and CroV-specific primers. Cell concentrations were determined by microscopy counts. Shown are the average values of three independent experiments with error bars representing ± SD. C. roenbergensis *cells are 5-10 µm in diameter, CroV particles are 300 nm in diameter, mavirus particles are 75 nm in diameter. BDL, below detection limit (≈1E+03 cells or viruses per mL); CN, cellulose nitrate; PES, polyethersulfone; PVDF, polyvinylidene difluoride.*

reference genome, extracting those reads, and assembling them into contigs. The longest resulting contig was 30,556 bp in length and contained a 19,055 bp sequence that was 100% identical to the 19,063 bp mavirus reference genome (GenBank accession HQ712116). The 8 bp that were missing from the 3' end of the reference genome were found directly adjacent to the 5' end of the integrated mavirus genome, indicating that our initial prediction for the genome linearization site was off by 8 bp[17]. In contrast to the reference mavirus genome, the endogenous virus genome was flanked on either side by 615/616 bp-long TIRs that were 99.7% identical to each other. The longer TIRs result in a total length of 20,190 bp for the endogenous mavirus genome, compared to 19,063 bp for the reference genome. Although most mavirus integrations had 615/616 bp-long TIRs that started with six guanine residues, some contained two additional Gs. The TIR length thus appears to differ slightly for each integration event. The host sequence directly adjacent to the provirophage genome featured target site duplications (TSDs) that were mostly 6 bp, in some cases 5 bp, long. The TSD sequences differed between integration sites with no obvious consensus motif. By recruiting reads to the flanking regions of mavirus TIRs, we found 11 well-supported integration sites in the E4-10M1 genome (Table 2). The actual number may be slightly higher due to integrations in repetitive genomic regions which we could not resolve.

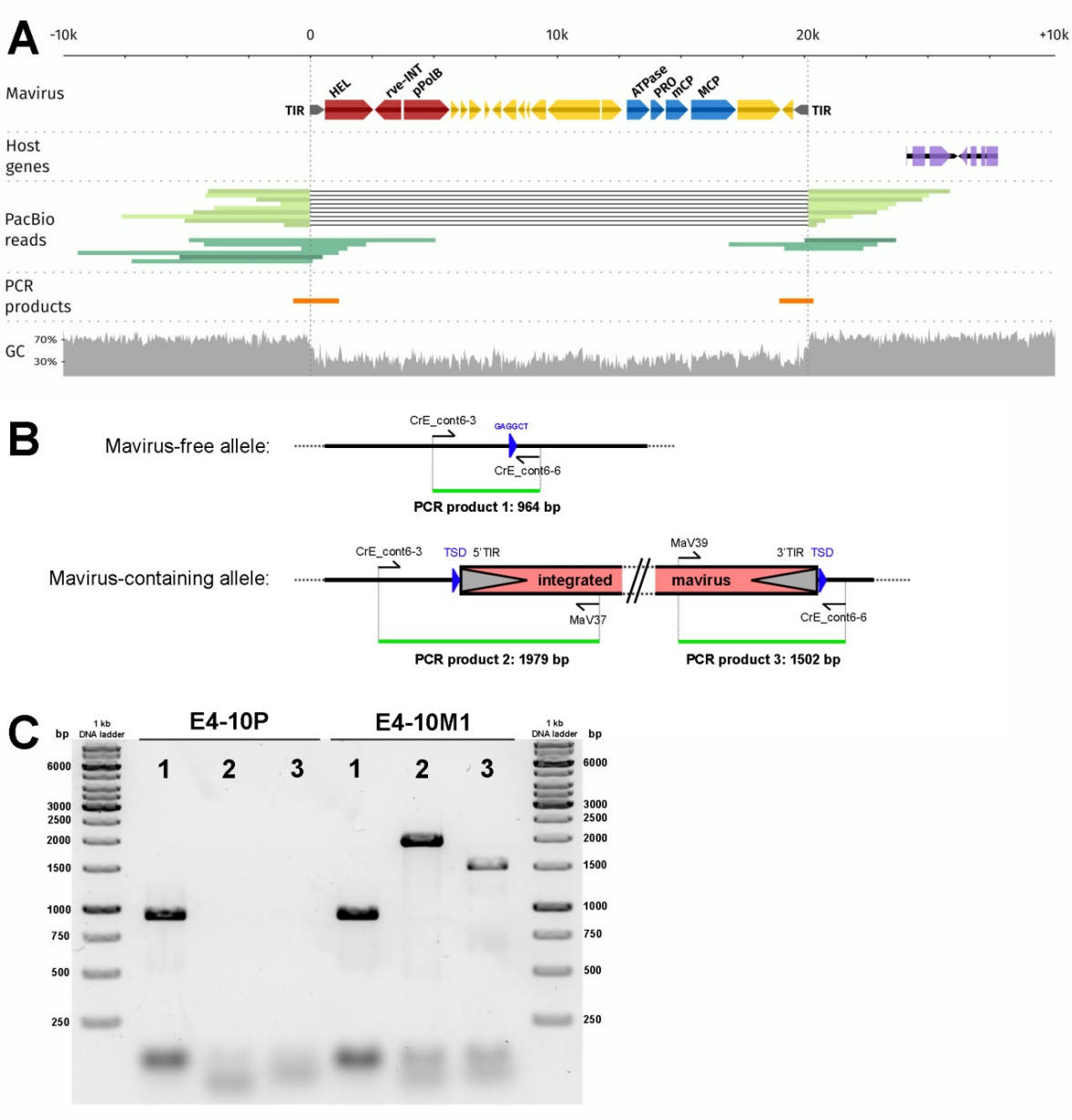

Figure 2: Characterization of a mavirus integration site in* C. roenbergensis *strain E4-10M1.
(A) Overview of the integration site. The ≈20 kbp long mavirus genome is flanked by 10 kbp of host sequence on either side (part of a 208 kbp-long contig). Mavirus genes for replication and integration are shown in red, morphogenetic genes are shown in blue, other genes are shown in yellow and terminal inverted repeats (TIR) are indicated in grey. The exon structure of two adjacent host gene models (function unknown) is shown in purple. PacBio reads covering the integration site are shown in green. Reads that span the integration site and contain only host sequence are shown in light green, whereas reads that cross the virus-host junction are shown in dark green. The two read populations represent mavirus-free and mavirus-containing alleles of the diploid host genome, respectively (see also Figure S2). PCR products spanning the host-virus junction are shown in orange. The GC content plot is based on a 30 bp sliding window. (B) Schematic representation of the host genomic region in A) illustrating the PCR primer binding sites and expected PCR products that were used to confirm the integration site. (C) Gel image of the PCR products obtained from E4-10P and E4-10M1 genomic templates with primers spanning the integration site. The lanes are labelled according to the primer combinations and products shown in B).

Site #	Contig length (bp)	Position of integration site within contig (bp)	TSD	# of terminal C/G nucleotides in the TIR
1	223481	121402	CGGAA	7
2	161079	145387	TGACAC	7
3	358887	10602	ATTTC	7
4	58090	3551	CAAACT	6
5*	208205	118064	GAGGCT	6
6	116230	100054	CGACA	7
7	14206	376	TGTCAA	6
8	209165	76864	CTGTG	7
9	403211	1650	CACCTC	7
10	19689	16490	CCACAC	7
11	8714	3914	TGCAC	7

Table 2: Details on the 11 bioinformatically well-supported mavirus integration sites in* C. roenbergensis *strain E4-10M1.
The integration site described in detail in Figure 2 is marked with an asterisk.

One of the integration sites was characterized in further detail and its reconstruction is shown in Figure 2. Aligning PacBio reads to any of the 11 well-supported integration sites consistently resulted in two distinct read populations, one that connected the regions directly flanking the integration site without the viral insertion, and another population that spanned the viral-cellular junctions of the integration site (Figure 2A). Mavirus integrations thus occurred at only one of two homologous sites in the diploid flagellate genome, resulting in the E4-10M1 strain being heterozygous for each mavirus provirophage. The heterozygous condition was confirmed by PCR analysis (Figure 2B). The large difference in GC content between host genome (70% GC) and mavirus genome (30% GC, Figure 2A) required careful optimization of PCR primers and annealing temperatures for products that were part host and part virus. The optimized PCR yielded products for both mavirus-free and mavirus-containing alleles in strain E4-10M1, whereas only products for the mavirus-free allele were obtained with E4-10P template DNA (Figure 2C).

CroV infection induces gene expression and genome replication of endogenous mavirus

To test if the endogenous mavirus genes were expressed, we analyzed selected transcripts by quantitative reverse transcription PCR (qRT-PCR). Because mavirus gene promoters are highly similar to the late gene promoter motif in CroV[17], we considered the possibility that gene expression of the endogenous mavirus might be facilitated by the presence of CroV. Therefore, we isolated total RNA from mock-infected and CroV-infected E4-10P and E4-10M1 cells at 0 and 24 h p.i. The RNA was DNase-treated and converted into cDNA using a mix of random and oligo(dT) primers. Using gene-specific primers (Table S1), we then quantified in the cDNA pool five mavirus genes (*MV03* [pPolB], *MV15* [ATPase], *MV16* [PRO], *MV17* [mCP], *MV18* [MCP]), and three CroV target genes: the isoleucyl-tRNA synthetase (IleRS) gene *crov505*, the DNA polymerase B (PolB) gene *crov497*, and the MCP gene *crov342*. These CroV genes are classified as early, intermediate, and late, respectively, because their transcripts were first detected at 0 h, 3 h, and 6 h p.i. in a DNA microarray study[23]. As a host control, we used the *C. roenbergensis* aspartyl-tRNA synthetase gene (AspRS, EST:MMETSP0942-20120912|8440_1[34]).

As shown in Figure 3A, this gene showed slightly higher expression levels at 0 h than at 24 h, potentially because the cells at 0 h were growing exponentially whereas at 24 h, the cells had already reached stationary phase (Supplemental Spreadsheet). Expression of the CroV IleRS, PolB and MCP genes could be clearly detected at 24 h p.i. in the CroV-infected cultures and was comparable between E4-10P and E4-10M1 strains. The mavirus genes in E4-10M1 cells were quiescent under normal conditions and also immediately after inoculation with CroV. In contrast, all of the five tested mavirus genes were expressed at 24 h in the CroV-infected E4-10M1 strain, with the MCP (*MV18*) gene reaching the highest expression level (Figure 3A). The quantification cycle (*C*q) values of the mavirus RT(-) controls for the CroV-infected E4-10M1 cultures at 24 h p.i. were on average 19.2 cycles above those of the respective RT(+) samples (Figure 3A), indicating a low level of genomic DNA contamination. Our qRT-PCR data thus strongly suggest that CroV infection induces expression of the endogenous mavirus genes in E4-10M1 cells.
Addition of the protein biosynthesis inhibitor cycloheximide (CHX) or the DNA polymerase inhibitor aphidicolin (APH) effectively inhibited host cell growth and CroV DNA replication (Supplemental Spreadsheet). CHX treatment inhibited expression of the intermediate PolB gene *crov497* and the late MCP gene *crov342* (Figure 3B); thus a functional host translation system is required for the expression of these genes. However, cDNA of the IleRS gene *crov505* was detected at 24 h p.i. in the presence of CHX, demonstrating that protein biosynthesis is expendable for the expression of this early CroV gene. This finding is in line with the presence of a viral transcription apparatus in the CroV virion[35] and strongly suggests that CroV, like other strictly cytoplasmic large DNA viruses such as mimi- and poxviruses, express their early genes with a pre-packaged viral transcriptase. In the presence of APH, all three CroV genes were expressed at low levels, with cDNA of the late MCP gene being barely detectable.
Crucially, treatment with CHX or APH also inhibited mavirus gene expression in CroV-infected E4-10M1 cells (Figure 3B). Only the *MV03* pPolB gene was weakly expressed with APH treatment. These results indicate that *de novo* protein synthesis and CroV DNA replication are prerequisites for provirophage gene induction.

In the next set of infection experiments, we examined whether CroV infection would induce DNA replication of the integrated mavirus genome. Again, triplicate cultures of E4-10P and E4-10M1 were either mock-infected or infected with CroV, and DNA from each sample was extracted daily over the course of one week to measure the copy numbers of CroV and mavirus genomes by qPCR. We tested several target amplicons for each virus and found them to yield comparable results. Hence we chose a 125 bp region of the *MV18* MCP gene as a proxy for mavirus genome quantification. Similarly, a 128 bp amplicon of the *crov283* VV D11-like transcription factor gene was used to quantify CroV genome copies. Host cell density was assessed by staining the cells with Lugol's acid iodine solution and counting them on a hemocytometer. As shown in Figure 4A, no virus DNA was found in mock-infected E4-10P cells, whereas a latent mavirus signal was present in mock-infected E4-10M1 cells.
When E4-10P cells were infected with CroV, the cell numbers started to decline after 1-2 days and CroV genome copies increased ≈10,000fold. CroV-infected E4-10M1 cells also started to lyse after 24 h p.i. and CroV genome replication was comparable to that in the E4-10P strain. The titers of infectious CroV in the E4-10P and E4-10M1 lysates were measured to ≈5E+07 per ml. In contrast to the CroV-infected E4-10P cells as well as to the uninfected E4-10M1 cells, a sharp increase in the mavirus signal was observed in the CroV-infected E4-10M1 strain. Within 48 h p.i., the mavirus signal in the CroV-infected E4-10M1 cultures was ≈500 times higher than in the uninfected E4-10M1 cultures. The increase in mavirus signal coincided with the increase in CroV genome copies and the decrease of host cell numbers. These results demonstrate that CroV infection of E4-10M1 cells induces genome replication of mavirus provirophages.

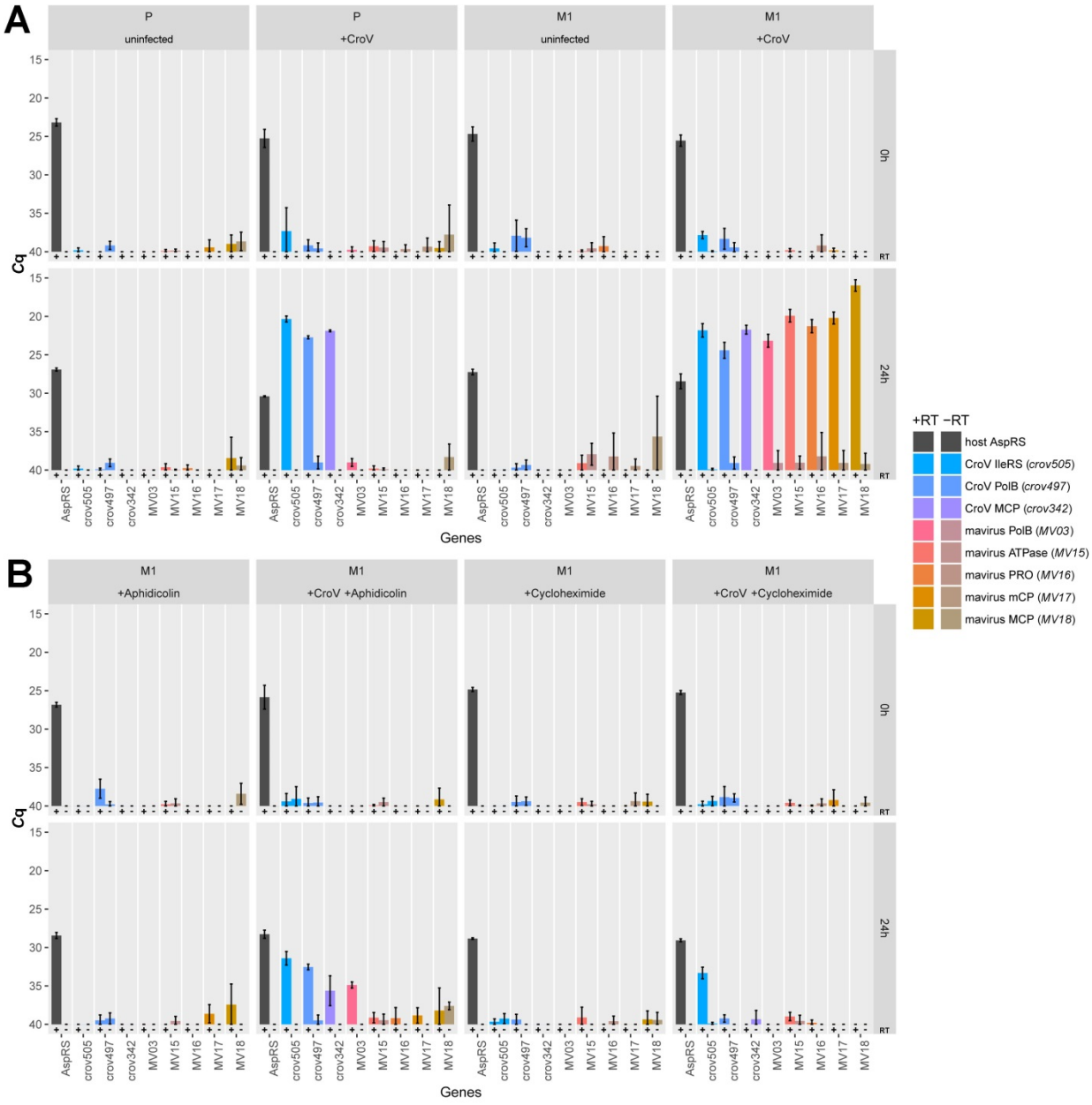

Figure 3: Gene expression analysis of the endogenous mavirus genome.
(A + B) Selected cellular and viral transcripts isolated from differently treated C. roenbergensis *cultures were quantified by qRT-PCR. Shown are the average quantification cycle (Cq) values of three independent experiments with error bars representing ± SD. The following genes were assayed: host* AspRS, C. roenbergensis *E4-10 aspartyl-tRNA synthetase;* crov342, *CroV major capsid protein;* crov497, *CroV DNA polymerase B;* crov505, *CroV isoleucyl-tRNA synthetase;* MV03, *mavirus pPolB;* MV15, *mavirus ATPase;* MV16, *mavirus PRO;* MV17, *mavirus mCP;* MV18, *mavirus MCP. Cq values of the control reactions without reverse transcriptase (-RT) are shown in darker shades directly to the right of the respective +RT results. Accession numbers are listed in Table S1. See also Supplemental Spreadsheet. (A) Gene expression in mock-infected or CroV-infected E4-10P and E4-10M1 cultures at 0 and 24 h p.i. (B) Gene expression in mock-infected or CroV-infected, aphidicolin- or cycloheximide-treated E4-10M1 cultures at 0 and 24 h p.i.*

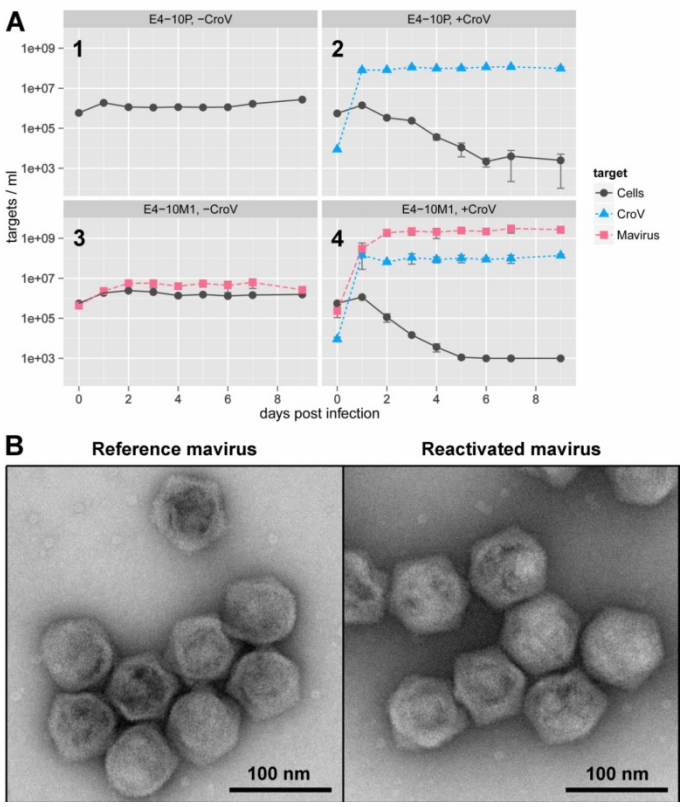

Figure 4: CroV infection induces replication and particle production of the endogenous mavirus.
(A) C. roenbergensis *strains E4-10P and E4-10M1 were mock-infected (1,3) or infected with CroV (2,4) and cells and viruses were monitored for 9 days. Cell densities are based on microscopy counts, viral numbers are derived from qPCR data assaying short amplicons of the mavirus* MV18 *gene and the* crov283 *gene, respectively. The detection limit for both methods was ≈1E+03/mL. Data were pooled from three independent experiments and error bars represent ± SD. See also Supplemental Spreadsheet. (B) Negative-stain electron micrograph of virus particles from the CroV-infected E4-10M1 strain, compared to reference mavirus particles. Both samples were purified on CsCl density gradients. See also Figures S3, S4.*

CroV induces the production of infectious mavirus particles

We then investigated in which size fraction the replicated mavirus genomes were present in the culture medium of CroV-lysed E4-10M1 cultures. At 3 days (d) p.i., samples of CroV-infected E4-10P and E4-10M1 cultures were filtered through 5.0 µm, 0.45 µm, 0.22 µm, or 0.1 µm pore-size filters. DNA extracted from each filtrate was analyzed by qPCR with CroV- and mavirus-specific primers. Only supernatants from CroV-infected E4-10M1 cultures tested positive for mavirus in all filtrates (Table 1). A control lysate created by sonication of uninfected E4-10M1 cells, as well as an uninfected E4-10P culture that was mixed with mavirus particles immediately prior to filtration, also contained mavirus DNA in all filtrates. This suggested that mavirus DNA was cell-associated in uninfected E4-10M1 cultures, whereas supernatants from CroV-infected E4-10M1 cultures contained extracellular mavirus DNA. We then examined the lysates for mavirus-like capsids using electron microscopy. A comparison of 0.1 µm-filtered supernatants sampled at 3 d p.i. from the four infection experiments shown in Figure 4A revealed that only the CroV-infected E4-10M1 sample contained virus-like particles similar to mavirus capsids (Figure S3). To acquire more biological material, we repeated the infection experiments on a larger scale (3 l) and, after the CroV-infected cultures had lysed, we concentrated the 0.22 µm pre-filtered culture supernatants 200-fold by tangential flow filtration (100 kDa cutoff) and analyzed the concentrates by isopycnic CsCl density gradient ultracentrifugation. A reference mavirus preparation was run in parallel as a positive control and yielded a band at a density of ≈1.29 g/ml CsCl (Figure S4A). Only the concentrate of the CroV-infected E4-10M1 sample also displayed a band at this density. The visible bands as well as material from equivalent positions in the other CsCl gradients were extracted from the gradients and imaged by electron microscopy (Figure S4B). The icosahedral particles produced by the CroV-infected E4-10M1 cells were indistinguishable from reference mavirus particles (Figure 4B) and DNA extracted from these particles tested PCR-positive for mavirus (Figure S4C), corroborating the conclusion that CroV infection induces capsid

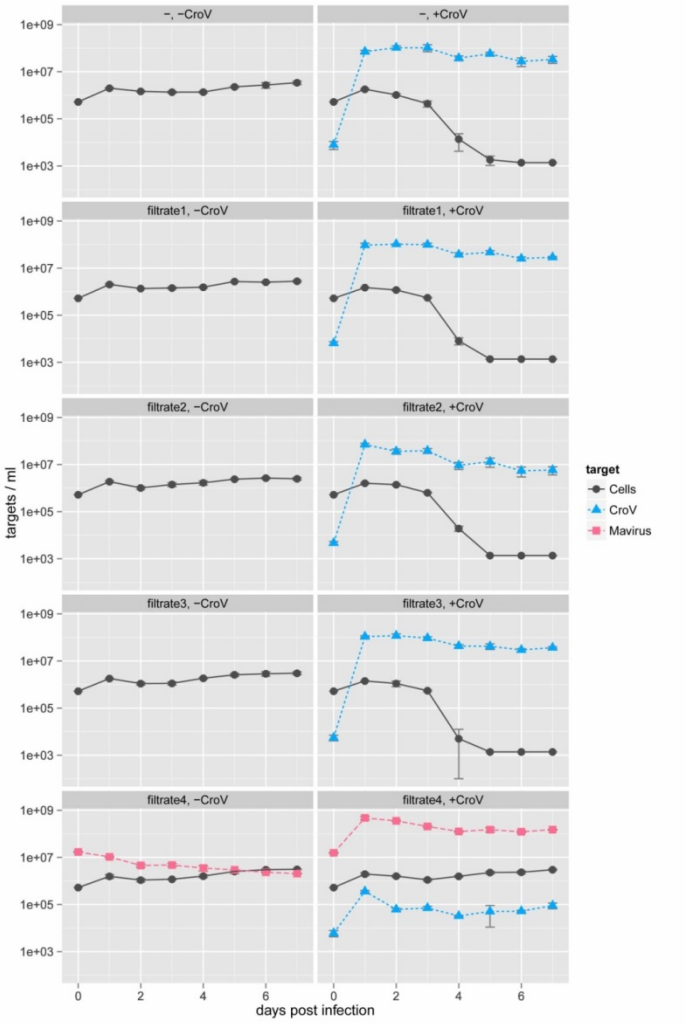

Figure 5: Reactivated mavirus protects host cells from CroV-induced lysis.
C. roenbergensis *strain E4-10P was mock-infected (left column) or CroV-infected (right column) and simultaneously inoculated with 0.02% (v/v) of 0.1 µm filtrates sampled at 3 d p.i. of the infection experiments shown in Figure 4A. Cell densities are based on microscopy counts, viral numbers are derived from qPCR data assaying short amplicons of the mavirus* MV18 *gene and the* crov283 *gene, respectively. The detection limit for both methods was ≈1E+03/mL. Data were pooled from three independent experiments and error bars represent ± SD. Only filtrate 4, derived from the CroV-infected E4-10M1 culture, contained infectious mavirus that was able to suppress CroV replication and protect the host cell population. See also Figure S5 and Supplemental Spreadsheet.*

formation of the endogenous mavirus. To examine whether the mavirus-like particles from the CroV infected E4-10M1 cultures were infectious, we used the 0.1 µm filtrates from the infection experiment in Figure 4A to inoculate E4-10P cultures that were simultaneously infected with CroV. If any of these filtrates contained infectious mavirus particles, coinoculation of cells with CroV should result in mavirus genome replication, which can be detected by qPCR. As shown in Figure 5, only filtrate 4 (from CroV-infected E4-10M1 cells) contained mavirus DNA. When filtrate 4 was added to mock-infected E4-10P cells, the mavirus signal slowly declined, indicating DNA degradation. In the presence of CroV and filtrate 4, however, the mavirus signal increased ≈50-fold within 24 h. We conclude from these results that CroV induces the production of infectious mavirus particles in strain E4-10M1.

Reactivated mavirus inhibits CroV propagation and protects host populations from lysis by CroV.

The reactivated mavirus not only replicated, it also suppressed CroV genome replication by 2-3 orders of magnitude, which resulted in survival of the host cell population (Figure 5). None of the other filtrates elicited a similar effect, thus only the CroV-infected E4-10M1 culture produced an agent that inhibited CroV replication and led to host survival. Although the most parsimonious explanation posits this agent to be reactivated mavirus, it would in principle be possible that a 0.1 µm filterable agent of non-viral nature, released by the CroV-infected E4-10M1 strain, might be responsible for the observed protective effect. We therefore performed an additional control experiment, in which we irradiated filtrate 4 with 500 J/m^2 (50 mW sec/cm^2) of ultraviolet (UV) light (λ=254 nm) and repeated the relevant infection experiments as shown in Figure 5. UV treatment is an effective means to inactivate viruses and other microorganisms[36]. Similar to the previous infections, co-inoculation of E4-10P cells with CroV and non-irradiated filtrate 4 inhibited CroV replication and prevented extensive lysis of the host cell population. UV treatment of lysate 4 abrogated the CroV-inhibitory and host-protective effects, and the mavirus signal decreased ~100-fold over the course of the experiment (Figure S5), corroborating the assumed role of mavirus as the causative agent.

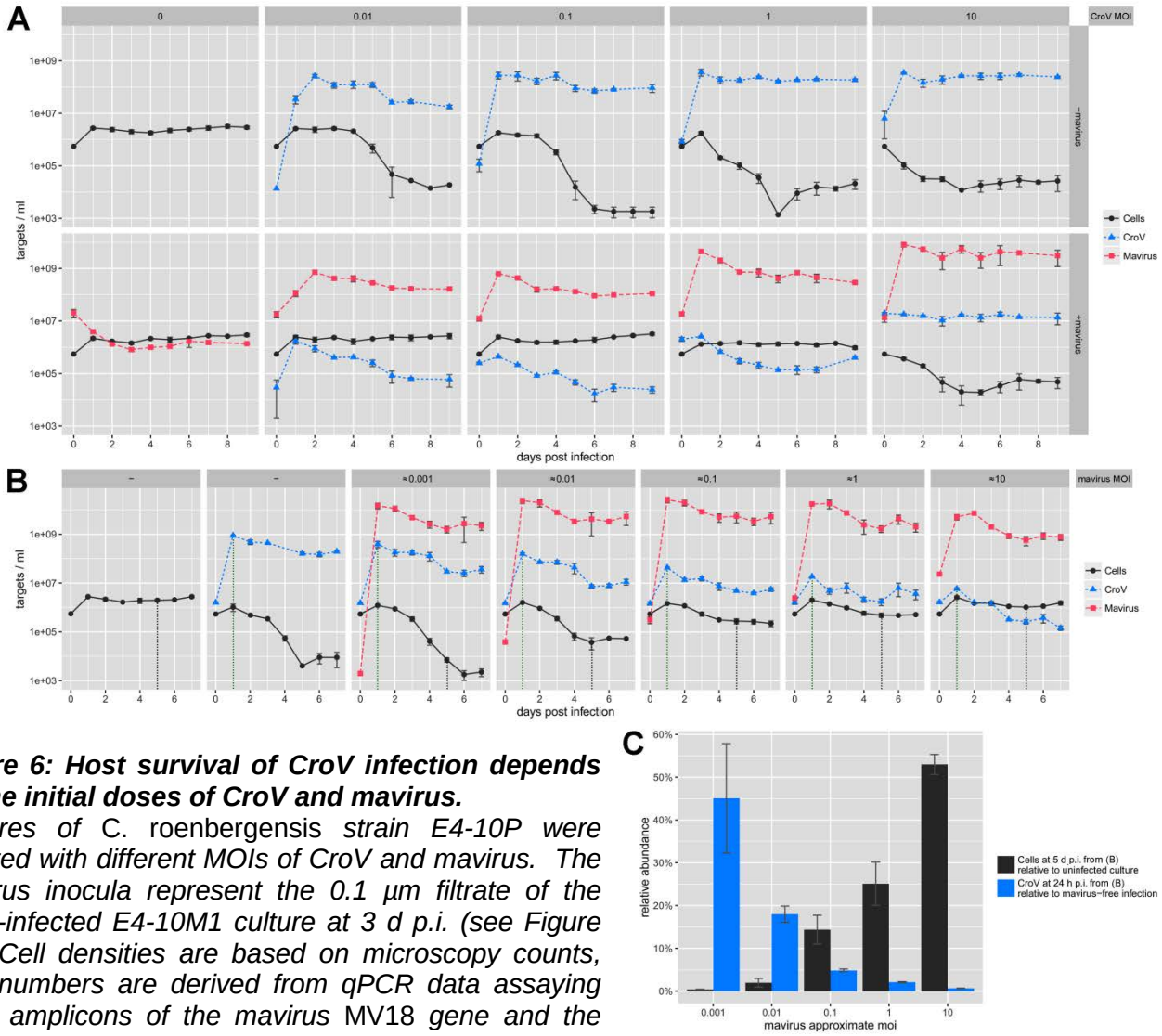

Figure 6: Host survival of CroV infection depends on the initial doses of CroV and mavirus.
Cultures of C. roenbergensis *strain E4-10P were infected with different MOIs of CroV and mavirus. The mavirus inocula represent the 0.1 µm filtrate of the CroV-infected E4-10M1 culture at 3 d p.i. (see Figure 4A). Cell densities are based on microscopy counts, viral numbers are derived from qPCR data assaying short amplicons of the mavirus* MV18 *gene and the* crov283 *gene, respectively. The detection limit for both methods was ≈1E+03/mL. Data were pooled from three independent experiments and error bars represent ± SD. See also Supplemental Spreadsheet. (A) Infection experiments with different MOIs of CroV in the absence (upper row) or presence (lower row) of mavirus at an MOI of ≈10. (B) Infection experiments with C. roenbergensis strain E4-10P similar to (A), but with a fixed CroV MOI of 1 and increasing MOIs of mavirus. The leftmost panel shows the mock-infected control, the second panel from the left shows the mavirus-free, CroV-infected control. Vertical dotted lines mark the reference points for the analysis in (C). (C) Summary of the influence of mavirus MOI on CroV DNA replication and host cell survival from the infection experiments shown in (B). Black columns show the host cell densities at 5 d p.i. with increasing mavirus MOIs relative to mock-infected 5 d p.i. cultures. Blue columns show the CroV genome copy concentration at 24 h p.i. with increasing mavirus MOIs relative to mavirus-free 24 h p.i. CroV infections.*

To gain more insight in the virophage-virus-host dynamics, we infected E4-10P cells with different MOIs of CroV and of reactivated mavirus. Figure 6A shows a series of infections with CroV MOIs varying from 0.01 to 10, either without added mavirus (upper panels) or in the presence of mavirus at MOI ≈10 (lower panels). The number of virions that each cell receives at a given MOI follows a Poisson distribution, therefore the percentage of infected cells at an MOI of 1 is 63%,

and an MOI of 10 is needed to ensure that >99.99% of cells are infected. With every cell infected with mavirus, host populations survived an infection with CroV at MOIs of 0.01 to 1 (Figure 6A). Although CroV did not replicate at MOI 10 in the presence of mavirus, the cells still lysed (96.5% decline after 5 days). These data indicate that nearly every cell infected with CroV is destined to lyse, irrespective of mavirus, and that mavirus rather prevents the spread of CroV by inhibiting its replication. We then infected E4-10P cells with a CroV MOI of 1 and mavirus MOIs ranging from ≈0.001 to ≈10. As shown in Figure 6B, these experiments revealed a clear dose-response relationship between the mavirus inoculum on one hand and host survival and inhibition of CroV DNA replication on the other hand. Even low MOIs of mavirus significantly inhibited CroV. At 24 h p.i. and a mavirus MOI of ≈0.001, CroV DNA replication was reduced to 45% of the level observed in a mavirus-free CroV infection; a mavirus MOI of ≈0.01 reduced CroV DNA replication by 82%, and a mavirus MOI of ≈0.1 resulted in a 95% reduction of CroV DNA replication (Figure 6C). Likewise, host cell survival at a CroV MOI of 1 improved with increasing mavirus MOIs. At 5 d p.i. and a mavirus MOI of ≈0.1, host cell density reached 15% of the density of an uninfected 5 d p.i. culture; it increased to 25% with mavirus at MOI ≈1, and about half of the cell density of an uninfected culture was reached at a mavirus MOI of ≈10 (Figure 6C).

Combined, these results lead us to conclude that host-integrated mavirus genomes are transcriptionally silent under normal conditions, and that CroV infection triggers gene expression, genome replication, and virion synthesis of the endogenous mavirus. Although the provirophage-carrying cells were not directly protected from CroV-induced lysis, reactivated mavirus particles were able to inhibit CroV in subsequent coinfection experiments, which resulted in dose-dependent survival of the host population.

Discussion

Our genome analysis and infection experiments with the marine protozoan *C. roenbergensis* reveal a biphasic life cycle for the virophage mavirus, with lytic and latent modes of infection. Based on the high percentage (32%) of mavirus-associated survivor cells we observed after a single coinfection of *C. roenbergensis* strain E4-10P with CroV and mavirus, we can confidently state that mavirus integrations occur frequently under laboratory conditions. We predict that integration events may be similarly frequent in natural environments, although frequencies may vary depending on the host and virophage strain. This strategy is astonishing, given that genome integration is not a mandatory step of the mavirus infection cycle, and may rather be a means to an end for long-term survival of the virophage. In addition to a susceptible host cell, a virophage needs a permissive giant virus to enable virophage gene expression, and it needs to coinfect the host within a narrow time window. Although widespread, neither cellular nor viral hosts for virophages are particularly abundant in marine habitats. Thus, if virophages depended solely on horizontal transmission and if virophage decay rates were higher than the rate with which the next giant virus-infected host cell can be encountered, the virophage could easily go extinct. By integrating into a host genome, however, the virophage can persist over prolonged periods of time until a suitable giant virus induces the lytic cycle of the virophage. In this regard, the mavirus system displays parallels to previous observations on adeno-associated virus (AAV). In the absence of a helper virus (adenovirus or herpesvirus), AAV integrates into a host chromosome[37,38]. This ensures the persistence of the AAV genome until more favorable conditions for AAV arise. Upon superinfection of a proviral cell line with a helper virus, the AAV genome is excised and resumes its extrachromosomal replication cycle[39]. A similar example is found in the bacterial domain, where the provirus form of the integrative plasmid P4 can be mobilized by the helper bacteriophage P2[40]. However, AAV genome integration is site-specific and the viral DNA typically integrates as a concatemer, which we did not observe for mavirus. Another difference is that the AAV genome integrates via homologous or non-homologous recombination and does not possess an integrase. Mavirus, on the other hand, encodes an rve-type integrase that is assumed to catalyze the integration reaction. Homologous integrases are found in retroviruses, retroelements, MPEs, and the *Tetrahymena*-specific Tlr elements. The C-terminus of the mavirus rve-INT encodes a chromatin organization modifier (chromo) domain that could direct the nucleoprotein integration

complex to heterochromatic regions, akin to genome integration of chromoviruses[41]. However, with the currently small number of integration sites available, it remains unclear whether mavirus integration is site-specific. Upon integration, rve-INTs create short duplicated sequences (TSDs) that are 5-6 bp long in the case of mavirus and MPEs[24,25]. Another feature shared with MPEs are the 615/616 bp long TIRs that border the endogenous mavirus genomes. In MPEs, TIR length typically ranges from 400 to 700 bp[24], but may vary from 100 bp to more than 1 kbp[25]. Our findings strengthen the close relationship between MPEs and mavirus: in addition to similar TIRs and TSDs, they share several homologous genes (HEL, rve-INT, pPolB, ATPase, PRO, mCP, MCP), they have similar length (15-30 kbp) and an overlapping host range (eukaryotes, including many protist lineages). MPEs were initially classified as transposable elements based on computational analysis, in particular the presence of an integrase, TIRs and TSDs. However, there is no experimental data to demonstrate that MPEs are capable of transposition, and the recent identification of capsid genes in many MPEs rather puts these elements in the category of endogenous viruses[26,28,29]. Whether MPEs are capable of active replication and whether they can be encapsidated under certain conditions remains to be seen. The Cafeteria-CroV-mavirus triad provides an excellent model system to study virophage integration, replication, and potential excision/transposition events, which may help to clarify the true nature of MPEs.

Our data suggest that CroV infection leads to cell lysis regardless of mavirus (Figure 6A). The most plausible explanation for the occurrence of survivor cells with multiple *de novo* mavirus integrations is thus that several mavirus particles infected and integrated in the same CroV-free cell. This hypothesis is supported by the CroV-independent mavirus entry mode via endocytosis[17]. Less likely alternatives are that mavirus integration is a post-replicative process (which would require CroV), or that endogenous mavirus genomes spread horizontally via replicative transposition. The latter scenario would imply that provirophages are not stable over time and can multiply within the host genome, akin to transposable elements. Our sequence data do not support the hypothesis that mavirus provirophages are genetically mobile, at least not during the estimated few hundred cell divisions that have passed between isolation and genome sequencing of the E4-10M1 strain.

In contrast to the endogenous *B. natans* virophages[32], the endogenous mavirus genes were not transcriptionally active under normal conditions, although we cannot exclude activation under specific, CroV-independent circumstances. Our qRT-PCR data show that the provirophage genes were highly expressed upon CroV infection of the E4-10M1 strain. Based on this observation and the similarity of transcriptional signals between virophages and their giant viruses, we propose that a CroV-encoded transcription factor (TF) may be responsible for provirophage activation (Figure 7). The conserved sequence motif in the promoter region of all 20 mavirus genes is highly similar to the late promoter motif of CroV[17] and a CroV late TF that specifically binds this motif would be the most probable candidate for the mavirus-inducing agent. Late TFs of large DNA viruses are synthesized *de novo* during infection; hence translational inhibition by CHX should prevent mavirus gene induction by CroV. Our experimental results confirm these predictions (Figure 3B). Furthermore, viral DNA replication is often a prerequisite for the onset of late phase, thus APH treatment should inhibit the expression of late CroV genes and prevent provirophage induction. With the exception of *MV03*, mavirus gene expression was suppressed by APH, further corroborating the late TF hypothesis. While we cannot explain the marginal *MV03* gene activity in the presence of APH (which was ≈3000fold lower than in APH-free CroV-infected E4-10M1 cells), weak gene expression was also observed for the CroV late gene *crov342* (MCP). It is thus possible that very low levels of the late TF were still synthesized in the APH-treated cultures. The gene for the late TF has not been identified yet, nor is it known how the TF activates mavirus transcription. Whereas the TF and the associated CroV-encoded DNA-dependent RNA polymerase complex could easily access a mavirus genome located in the cytoplasmic VF, the endogenous mavirus genome is separated from the VF by the nuclear envelope. The TF would thus have to gain access to the nucleus, where it could interact with a cellular RNA polymerase complex to initiate mavirus gene transcription. Alternatively, the entire CroV-encoded transcription machinery could

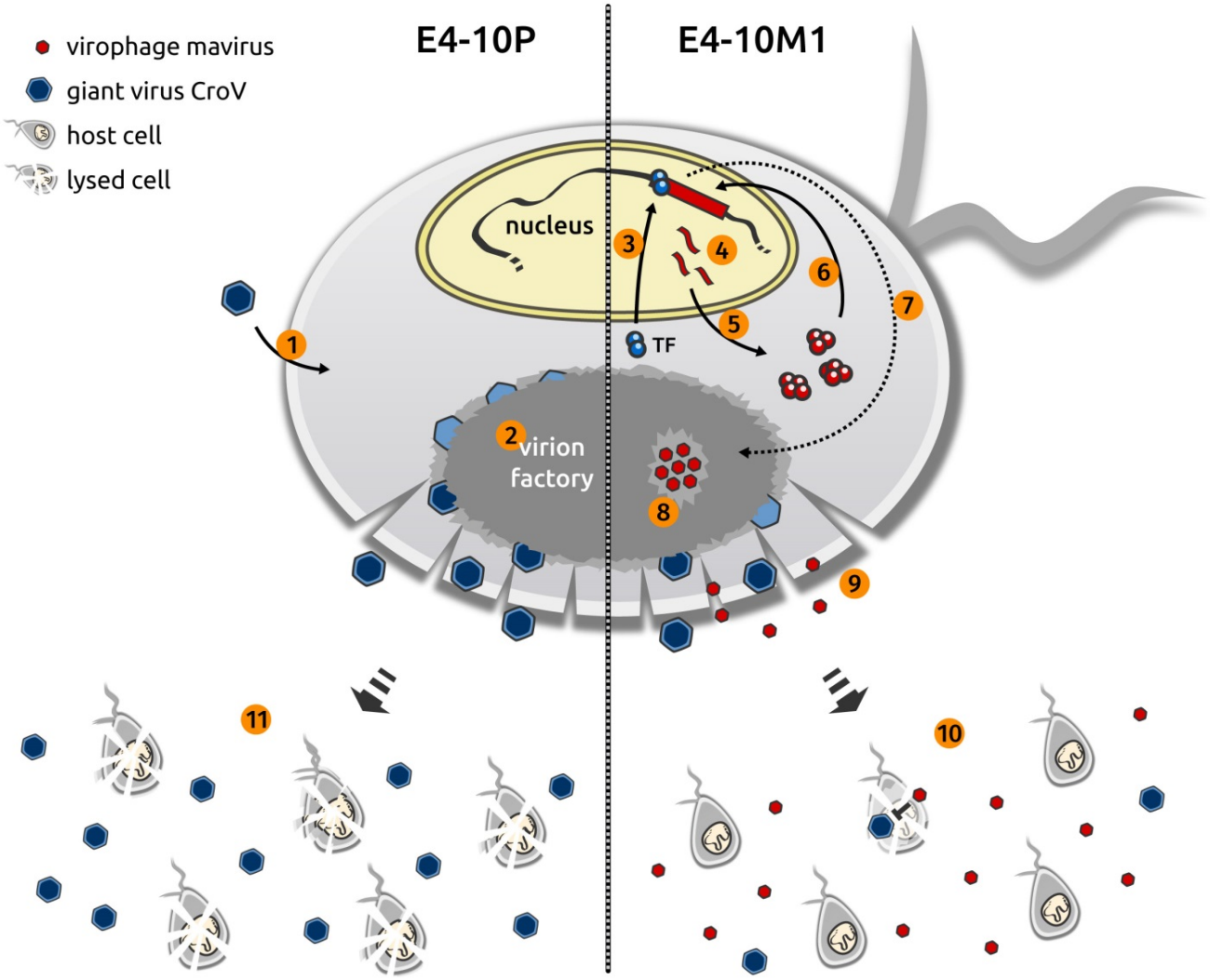

Figure 7: Hypothesis for CroV-induced reactivation of endogenous mavirus.
Shown is a schematic C. roenbergensis *cell displaying select events of a CroV infection cycle in strains E4-10P (left) and E4-10M1 (right). Following CroV entry (1), the virion factory forms in the cytoplasm. At the onset of late phase, a CroV-encoded transcription factor (TF) recognizing the late CroV promoter motif is synthesized (2). In E4-10M1 cells, the late TF enters the nucleus (3), binds the mavirus promoter sequences and activates gene expression of the provirophage (4). Mavirus-specific transcripts are exported and translated (5) and some of the mavirus proteins return to the nucleus to excise or replicate the provirophage genome (6). The mavirus genome then translocates to the CroV factory (7), where genome replication, particle assembly, and genome packaging occur (8). Cell lysis releases the newly synthesized CroV and mavirus particles (9) and the reactivated virophages inhibit further spread of the CroV infection in coinfected cells, leading to enhanced survival of the host population (10). In contrast, CroV infection of an E4-10P cell does not induce a virophage response and CroV continues to infect other host populations (11).*

locate to the nucleus. After mRNA synthesis, the mavirus transcripts must be translated and, at a minimum, the rve-INT or pPolB protein would need to return to the nucleus to initiate excision or replication of the endogenous mavirus genome. Finally, the excised or replicated mavirus genome would have to locate to the cytoplasm for virion assembly and genome packaging within the VF (Figure 7). It is also possible that reactivation occurs only during very late stages of infection, at which the integrity of the nuclear membrane may become compromised. Based on these considerations, it appears likely that the mavirus response to CroV infection is slower for the provirophage than for an exogenous mavirus during a coinfection. This would explain why CroV gene expression and DNA replication were not impaired in infected E4-10M1 cultures (Figures 3A, 4A, Supplemental Spreadsheet). In contrast, CroV replication was severely impaired in CroV/mavirus-coinfected E4-10P cultures, which led to increased survival of the host cell population (Figures 5, 6).

The beneficial action of provirophages on their cellular host has been proposed before[17,32] and we have now confirmed it experimentally, albeit with an interesting twist. Our results show that provirophage-mediated protection of flagellate cells against CroV infection occurs at the population level, rather than at the infected-cell level. The infection experiments with different CroV MOIs (Figure 6A) suggest that mavirus has no effect on the fate of a CroV-infected cell, but that it can inhibit the release of new CroV virions from a coinfected cell. The CroV-induced reactivation of endogenous mavirus is thus an altruistic response in which the initially infected cells lyse and release infectious mavirus particles into the surrounding medium. Mavirus then halts the spread of CroV in subsequent coinfections, thereby protecting any remaining uninfected host cells (Figures 5-7). We propose that virophage integration and reactivation play an ecologically important role in regulating virus-mediated mortality of natural protist populations. However, the magnitude of this effect cannot be extrapolated from our results. These experiments were done under defined laboratory conditions, with a single host strain, a single virus strain, high nutrient availability, and in the absence of other pathogens and predators of *Cafeteria*. Crucially, we conducted our experiments under well-mixed conditions, which is not representative of natural habitats where genetic diversity, micro-niches and fluctuations of biotic and abiotic factors may influence the fate of an individual cell. On the other hand, even if reactivation of provirophages in response to giant virus infection leads to only a slightly increased survival rate in certain groups of heterotrophic nanoflagellates, then provirophage protection is likely to have a significant effect on the ecology of these unicellular eukaryotes.

Materials and methods

Host and virus strains

C. roenbergensis strain E4-10 was isolated from coastal waters near Yaquina Bay, OR, as described previously (Gonzalez & Suttle 1993). The cell suspension culture has since been continuously passaged approximately every 4 weeks in f/2 enriched natural or artificial seawater medium supplemented with 1-3 autoclaved wheat grains per 10 mL to stimulate bacterial growth. For culture experiments, cells were grown in f/2 enriched artificial seawater medium supplemented with 0.05% (w/v) Bacto™ yeast extract (Becton, Dickinson and Company, Germany). For f/2 artificial seawater medium, the following sterile stock solutions were prepared: 75 g/L $NaNO_3$; 5 g/L NaH_2PO_4; 1000x trace metal solution containing 4.36 g/L Na_2EDTA x 2 H_2O, 3.15 g/L $FeCl_3$ x 6 H_2O, 0.01 g/L $CuSO_2$ x 5 H_2O, 0.18 g/L $MnCl_2$ x 4 H_2O, 0.006 g/L Na_2MoO_4 x 2 H_2O, 0.022 g/L $ZnSO_4$ x 7 H_2O, 0.01 g/L $CoCl_2$ x 6 H_2O; and a 50,000x vitamin solution containing 5 g/L thiamine-HCl, 25 mg/L biotin, and 25 mg/L cyano-cobalamine. The vitamin solution was stored at -20°C, the other solutions at room temperature. To prepare 1 L of f/2 artificial seawater medium, 33 g of Red Sea Salt (Red Sea Meersalz, www.aquaristikshop.com) were dissolved in ultra-pure water (ELGA, Veolia Water Technologies, Germany), then 1 mL each of the 75 g/L $NaNO_3$, 5 g/L NaH_2PO_4, and 1000x trace metal solutions as well as 20 µL of the 50,000x vitamin solution were added. After autoclaving, the medium was 0.22 µm filtered and stored at 4°C. Cultures were grown in flat-bottom 125 mL or 250 mL polycarbonate Erlenmeyer flasks (VWR, Germany) at 23°C in the dark.

The viruses used for infection experiments were Cafeteria roenbergensis virus (CroV) strain BV-PW1[23,42] and mavirus strain Spezl[17].

Viral infectivity assays

The infectivity of CroV was measured by end-point dilution assays and the statistical method by Reed and Muench[43] was used to determine the 50% end point. The resulting cell culture infectious dose at which 50% of the cultures lysed ($CCID_{50}$) was in good agreement with counts of SYBR-stained CroV particles by epifluorescent microscopy and also with gene copy numbers derived by quantitative PCR (qPCR). End-point dilution assays were carried out in 96-well plates with 200 µl of 1E+06 cells/mL exponentially growing host cells in f/2 medium + 0.05% (w/v) yeast extract per well. Each row (12 wells) was inoculated with a different dilution of CroV suspension (10 µL/well). Dilutions ranged from 1E-02 to 1E-09. The plates were stored at 23°C in the dark and analyzed after 6 days for cell lysis by microscopy. For mavirus, end-point dilution assays could not be employed because, in contrast to CroV, a productive mavirus infection does not result in cell lysis or cytopathic effects. Epifluorescence microscopy-based particle counts were too unreliable due to the small size of mavirus particles and the presence of bacteriophages in the host cultures. We therefore relied on a qPCR assay to quantify the number of mavirus major capsid gene copies, which can be seen as an upper approximation of the number of infectious mavirus particles. The actual titer of infectious virions is likely to be lower than the qPCR estimates because of free (non-encapsidated) mavirus DNA and an unknown proportion of non-infectious yet genome-containing particles.

Infection experiments

Typically, host cell suspension cultures were diluted daily to a cell density of 1-5E+05 cells/mL with f/2 medium containing 0.05% (w/v) yeast extract, until the desired culture volumes were reached. On the day of infection, when the cells had reached a density of >1.0E+06 cells/mL, the cultures were diluted with f/2 medium containing 0.05% (w/v) yeast extract to a cell density of 5-7E+05 cells/mL. Depending on the experiment, aliquots of 20 mL or 50 mL were dispensed in 125 mL or 250 mL polycarbonate flat-base Erlenmeyer flasks (Corning, Germany; through VWR International) and inoculated with virus-containing lysate or virus-free f/2 medium (for mock infections). The CroV inoculum varied between different infection experiments (see Supplemental Spreadsheet), according to the desired MOI and the titer of the CroV working stock, which was stored at 4°C and replaced every few months. For instance, the 50 mL infection experiments shown in Figures 4A, 5, and S5 received 100 µl of a CroV lysate with a $CCID_{50}$ of approximately 5E+06/mL per flask. Mock-infected cultures received an equal volume of f/2 medium. For testing culture supernatant from previous infection experiments for mavirus activity, 1 mL of the appropriate 0.1 µm-filtered lysate were added to the flask immediately prior to the CroV inoculum. Cultures were incubated at 23°C in the dark. Cell concentrations were measured by staining a 10 µL aliquot of the suspension culture with 1 µL of Lugol's Acid Iodine solution and counting the cells on a hemocytometer (Neubauer Improved Counting Chamber, VWR Germany). This method does not distinguish between live and dead cells and will also include cells that are already dead but have not lysed yet. Aliquots (200 µl) for DNA extraction were taken at appropriate time points and were immediately frozen and stored at -20°C until further processing. All infections, except the ones shown in Figure S1, were carried out in triplicates.

Isolation of *C. roenbergensis* strains E4-10P and E4-10M1

C. roenbergensis strain E4-10 was made clonal by repeated single-cell dilutions. Each well of a 96-well plate was filled with 200 µl of f/2 medium containing 0.01% (w/v) yeast extract. Then 1 µl of an E4-10 culture diluted to 300 cells/ml were added to each well, so that on average every 3rd well received one cell. After 6 days at 23°C, wells were inspected for cell growth and positive samples were transferred to 20 ml of f/2 medium containing 0.05% (w/v) yeast extract. This procedure was repeated serially two more times. DNA from the final isolate was extracted and tested by qPCR to confirm the absence of mavirus. 20 ml cultures of the resulting E4-10P strain at 5E+05 cells/ml in f/2 medium containing 0.05% (w/v) yeast extract were then either mock-infected, infected with CroV at MOI=0.01, or coinfected with CroV (MOI=0.01) and mavirus (MOI≈1). Eight days post infection (see Figure S1), the surviving cells from the coinfection were pelleted by centrifugation (5 min at 7,000 x g,

23°C), the pellets were resuspended in 50 ml f/2 medium and the centrifugation/dilution procedure was repeated 9 more times. The washed cells were then subjected to three consecutive rounds of single-cell dilution as described above. DNA was extracted from the resulting 66 clonal strains and tested by qPCR with mavirus-specific primers. The strain with the highest qPCR signal was named E4-10M1.

Filtration assay

Host strains E4-10P and E4-10M1 were either infected with CroV or mock-infected with f/2 medium. At 5 days p.i., when the CroV-infected cells had lysed, aliquots from the four different samples were passed through syringe filters of different nominal pore sizes, ranging from 5.0 µm to 0.1 µm, and DNA was extracted from 200 µl of each filtrate as well as from 200 µl of the unfiltered samples. The following syringe filters were used: 0.1 µm pore-size PVDF Millex (Millipore Merck, Ireland), 0.22 µm pore-size PES (TPP, Switzerland), 0.45 µm pore-size PES (TPP, Switzerland), 5.0 µm pore-size CN-S Whatman (Fisher Scientific GmbH, Germany). E4-10M1 cells were mechanically lysed by sonication with a Branson Sonifier 250 equipped with a microtip, duty cycle 50%, output setting 2. Two milliliter aliquots of an E4-10M1 suspension culture containing 1.4E+06 cells/mL were sonicated for 2x 30 sec with 30 sec incubation on ice in between. As a positive control, an E4-10P suspension culture was mixed with 0.1 µm-filtered reactivated mavirus to yield a final flagellate concentration of 1.8E+06 cells/ml. The sonicated and positive control samples were then filtered and processed as described above.

DNA extraction and quantitative PCR

We used qPCR with the SYBR-related EvaGreen™ dye to quantify viral DNA target sequences. Genomic DNA (gDNA) was extracted from 200 µL of suspension culture with the DNeasy 96 Blood & Tissue Kit (Qiagen, Hilden, Germany) following the manufacturer's instructions for DNA purification of total DNA from cultured cells, with a single elution step in 100 µL of double-distilled (dd) H_2O and storage at -20°C. DNA concentrations in the eluted samples typically ranged from 1 to 10 ng/µL, as measured on a NanoDrop 2000c spectrophotometer (Thermo Scientific, Germany). One microliter of gDNA was used as template in a 20 µL qPCR reaction containing 10 µL of 2X Fast-Plus EvaGreen® Master Mix with low ROX dye (Biotium, Inc. via VWR, Germany), 10 pmol of each forward and reverse primer (see Table S1), and 8.8 µL of ddH_2O. No-template controls (NTC) contained ddH_2O instead of gDNA. Each qPCR reaction (sample, NTC, or standard) was carried out in technical duplicates, with individual replicates differing in their quantification cycles (*C*q) by about 0.5% on average (0.49% ± 0.43%, n=200). The limit of detection for this assay was ≈10 copies, which equates to ≈5000 copies per mL of suspension culture. The *C*q values of the NTC controls were consistently below the limit of detection. Thermal cycling was carried out in a Stratagene Mx3005P qPCR system (Agilent Technologies, Germany) with the following settings: 95°C for 5 min, 40 cycles of 95°C for 10 s followed by 60°C for 25 s and 72°C for 25 s, a single cycle of 72°C for 5 min, and a final dissociation curve was recorded from 50°C to 95°C. qPCR results were analyzed using MxPro™ qPCR software v4.10 (Stratagene, La Jolla, USA). The threshold fluorescence was set using the amplication-based option of MxPro™ software. During PCR optimization, qPCR products were analyzed by agarose gel electrophoresis for correct product length and absence of unspecific products. During sample analysis, dissociation curves were used to monitor product specificity. Standard curves were calculated from a 10-fold dilution series that ranged from 10^1 to 10^8 molecules of a linearized pEX-A plasmid (Eurofins Genomics, Germany) carrying the fragment of the *MV18* MCP gene (GenBank Accession No: ADZ16417) that was amplified by primers Spezl-qPCR-5 and Spezl-qPCR-6 (Table S1) for mavirus quantification, or gDNA extracted from a known amount of CroV particles, the concentration of which had been determined by epifluorescence microscopy. To directly compare the two different kinds of template DNA used for virus quantification, the linearized plasmid also contained the target sequence for the *crov283* gene (GenBank Accession No: ADO67316.1) that is amplified by primers CroV-qPCR-9 and CroV-qPCR-10 and used as an approximation for CroV genome copies. The resulting standard curves and *C*q values of the plasmid and gDNA templates were highly similar to each other (Figure S6),

which implies that the quantification of mavirus using a plasmid-encoded target sequence is a valid approach. Owing to the large number of samples to be analyzed and to ensure plate-to-plate consistency, one full set of standard dilutions was recorded on the first plate of each primer set, and only two of the eight standard dilutions were repeated on consecutive plates. Using these two repeated standard dilutions as calibrators with the Multiple Experiment Analysis feature of MxPro™ software, the parameters of the full standard curve (fluorescence threshold and standard curve equation) were applied to all subsequent analyses. For mavirus quantification with primers Spezl-qPCR-5 and Spezl-qPCR-6, the R^2 value for the standard curve was 0.996, the amplification efficiency was 109.7%, and the standard curve equation was Y=-3.109*log(x)+33.89. For CroV quantification with primers CroV-qPCR-9 and CroV-qPCR-10, the R^2 value for the standard curve was 1.000, the amplification efficiency was 103.0%, and the standard curve equation was Y=-3.253*log(x)+34.77.

PCR verification of an example mavirus integration site

The mavirus integration site shown in Figure 2 was verified by PCR analysis and Sanger sequencing of the PCR products. Due to the difficulty of obtaining PCR products that were part host sequence with 70% GC content and part mavirus sequence with 30% GC content, primers were designed manually and several primers had to be tested under various PCR cycling conditions before the predicted products could be obtained. Primer sequences are listed in Table S1. PCR amplifications were performed using 2 ng of genomic DNA template from strain E4-10P or E4-10M1 in a 25 µl reaction mix containing 5 µl Q5® Reaction Buffer (NEB, Germany), 0.5 U of Q5® High-Fidelity DNA Polymerase (NEB, Germany), 0.2 mM dNTPs and 0.5 µM of each primer. In addition, the PCR mixes to amplify the empty integration site with primers CrE_cont6-3 and CrE_cont6-6 (amplifying only host sequence with 70% GC content) contained 5 µl of Q5 High GC Enhancer solution. The PCRs were carried out in a TGradient thermocycler (Biometra, Germany) with the following cycling conditions: 30 s denaturation at 98°C; 35 cycles of 10 s denaturation at 98°C, 30 s annealing at 68°C (for primer pair CrE_cont6-3 & MaV37) or 69°C (for primer pairs MaV39 & CrE_cont6-6 and CrE_cont6-3 & CrE_cont6-6) and 1 min extension at 72°C; and a final 2 min extension at 72°C. For product analysis, 5 µl of each reaction were mixed with loading dye and pipetted on a 1% (w/v) agarose gel supplemented with GelRed. The marker lanes contained 0.5 µg of GeneRuler™ 1 kb DNA Ladder (Fermentas, Thermo-Fisher Scientific, USA). The gel was electrophoresed for 2 h at 70 V and visualized on a ChemiDoc™ MP Imaging System (BioRad, Germany).

Cycling conditions for the PCR shown in Figure S4C were: 45 s denaturation at 98°C; 35 cycles of 10 s denaturation at 98°C, 30 s annealing at 58°C (primer pairs MaV21F & MaV21R) and 1 min extension at 72°C; and a final 2 min extension at 72°C.

RNA extraction and quantitative reverse-transcriptase PCR

Triplicate 50 mL cultures of strains E4-10P and E4-10M1 at an initial cell density of 6E+05 cells/mL were either mock-infected with f/2 medium or infected with CroV at an approximate MOI of 0.2. Aphidicolin-treated cultures were supplemented with 125 µl of a 2 mg/ml aphidicolin solution in DMSO (Sigma-Aldrich, Germany) for a final concentration of 5 µg/ml. Cycloheximide-treated cultures were supplemented with 37.5 µl of a 66.6 mg/ml cycloheximide solution in DMSO (Sigma-Aldrich, Germany) for a final concentration of 50 µg/ml. Cultures were incubated at 23°C. For extraction of total RNA, 1 mL aliquots were taken from each culture at 0 h p.i. and 24 h p.i. and centrifuged for 5 min at 10,000 x g, 21°C. The supernatants were discarded and the cell pellets were immediately flash-frozen in N_2(l) and stored at -80°C until further use. RNA extraction was performed with the Qiagen RNeasy® Mini Kit following the protocol for purification of total RNA from animal cells using spin technology. Cells were disrupted with QIAshredder homogenizer spin columns and an on-column DNase I digest was performed with the Qiagen RNase-Free DNase Set. RNA was eluted in 30 µl of 60°C warm RNase-free molecular biology grade water. The RNA was then treated with 1 µl TURBO DNase (2 U/µl) for 1 h at 37°C according to the manufacturer's instructions (Ambion via ThermoFisher Scientific, Germany). RNA samples

were analyzed for quantity and integrity on a Fragment Analyzer™ capillary gel electrophoresis system (Advanced Analytical, USA) with the DNF-471 Standard Sensitivity RNA Analysis Kit. Six microliters of each RNA sample were then reverse transcribed into cDNA using the Qiagen QuantiTect® Reverse Transcription Kit according to the manufacturer's instructions. This protocol included an additional DNase treatment step and the reverse transcription reaction using a mix of random hexamers and oligo(dT) primers. Control reactions to test for gDNA contamination were done for all samples by omitting reverse transcriptase from the reaction mix. The cDNA was diluted twofold with RNase-free H_2O and analyzed by qPCR with gene-specific primers. The qPCR reagents and conditions were the same as described above for genomic DNA qPCR. For data presentation purposes, any qPCR reactions that yielded no *C*q value after 40 PCR cycles were treated as *C*q=40. The no-template controls had an average *C*q value of 39.16 with a standard deviation of 2.20.

Concentration, purification, and electron microscopy of reactivated mavirus particles

Five hundred milliliter cultures of strains E4-10P and E4-10M1 at 5E+05 cells/mL in 3 L polycarbonate Fernbach flasks were either mock-infected with f/2 medium or infected with CroV at an MOI of 0.02. Six replicates were prepared for a total volume of 3 L per condition (E4-10P or E4-10M1, mock-infected or CroV-infected). At 3 d p.i., the cultures were centrifuged for 40 min at 7000 x g and 4°C (F9 rotor, Sorvall Lynx centrifuge) and the supernatants were filtered on ice through a 0.2 µm PES Vivaflow 200 tangential flow filtration (TFF) unit (Sartorius via VWR, Germany). The filtrates were then concentrated on ice with a 100,000 MWCO PES Vivaflow 200 TFF unit to a final volume of ≈15 mL. The concentrates were passed through a 0.1 µm pore-size PVDF Millex syringe filter (Millipore Merck, Ireland) and analyzed on 1.1-1.5 g/mL continuous CsCl gradients. The CsCl gradients were prepared by underlayering 6.5 mL of 1.1 g/mL CsCl solution in 10 mM Tris-HCl, pH 8.0, 2 mM $MgCl_2$ with an equal volume of 1.5 g/mL CsCl solution in 10 mM Tris-HCl, pH 8.0, 2 mM $MgCl_2$ in a SW40 Ultra-Clear™ centrifuge tube (Beckman Coulter, Germany). Tubes were capped and continuous gradients were generated on a Gradient Master (BioComp Instruments, Canada) with the following settings: tilt angle 81.5°, speed 35 rpm, duration 75 sec. After replacing 3.9 mL of solution from the top of the gradients with 4 mL of concentrated culture supernatants, the gradients were centrifuged for 24 h, 205,000 x g, 18°C using a SW40 rotor (Beckman Coulter, Germany) in a Beckman Optima™ ultracentrifuge. Bands in the gradients were visualized by illumination with an LED light source from the top of the gradient. One milliliter of gradient material from the mavirus band material (or equivalent positions of gradients were no such band was visible) were extracted with a syringe by puncturing the centrifuge tube with a 21G needle. The extracted band material was dialyzed for 24 h at 4°C in 3 mL dialysis cassettes (Pierce, 20 kDa cutoff) against 1 L of 10 mM Tris-HCl, pH 8.0, 2 mM $MgCl_2$. After dialysis, each sample was diluted to 4 ml with 10 mM Tris-Cl, pH 8.0, 2 mM $MgCl_2$ and centrifuged in Ultra-Clear™ tubes (Beckman Coulter, Germany) in a SW60 rotor for 1 h, 100,000 x g, 18°C. The supernatant was discarded and the pellets were softened overnight at 4°C in 50 µl of 10 mM Tris-Cl, pH 8.0, 2 mM $MgCl_2$ and then resuspended by pipetting. Aliquots (≈3 µL) of the concentrated samples were incubated for 2 min on Formvar/Carbon coated 75 mesh Cu grids (Plano GmbH, Germany) that had been hydrophilized by glow discharge. Grids were rinsed with ddH_2O, stained for 90 sec with 1% uranyl acetate, and imaged on a Tecnai T20 electron microscope (FEI, USA) with an acceleration voltage of 200 kV.

UV treatment of reactivated mavirus particles

A Stratalinker® UV crosslinker 2400 (Stratagene) was used for irradiation of virus samples with UV-C (λ=254 nm) light. Five hundred microliter drops of 0.1 µm-filtered reactivated mavirus suspension were pipetted on Parafilm and irradiated with a single dose of 500 J/m^2 of UV-C light. The dose was monitored with a VLX 3W radiometer (Vilber-Lourmat). The irradiated virus suspension was then kept in the dark to prevent eventual light-induced DNA repair. Infection experiments were carried out as described above and cultures were incubated in the dark for the entire duration of the experiment. Samples for DNA extraction and qPCR analysis were taken and processed as described above.

MiSeq and PacBio genome sequencing

Genomic DNA from 1E+09 cells each of the clonal *C. roenbergensis* strains E4-10P and E4-10M1 was isolated using the Qiagen Blood & Cell Culture DNA Midi Kit. The genomes were sequenced on an Illumina MiSeq platform (Illumina Inc., San Diego, USA) using the MiSeq reagent kit v3 at 2 x 300 bp read length configuration. The E4-10P genome was sequenced by GATC Biotech AG (Constance, Germany) with the standard MiSeq protocol. The E4-10M1 genome was prepared and sequenced at the Max Planck Genome Centre (Cologne, Germany) with NEBNext® High-Fidelity 2X PCR Master Mix chemistry and a reduced number of enrichment PCR cycles (six) in order to reduce AT-bias. The total output was 6.8 Gbp and 4.5 Gbp for E4-10P and E4-10M1, respectively. Overall sequencing quality was assessed with FastQC v0.11.3. Reads were trimmed for low quality bases and adapter contamination using Trimmomatic v0.32 [44] and customized parameters (minimum phred score 20 in a 10 bp window, minimum length 75 bp, Illumina TruSeq3 reference adapter) resulting in 5.0 Gbp and 2.9 Gbp high quality paired-end sequences, respectively. We also sequenced genomic DNA of strains E4-10P and E4-10M1 on a Pacific Biosciences RS II platform (two SMRT cells each, Max Planck Genome Centre Cologne, Germany), which resulted in 0.52 Gbp and 1.30 Gbp of raw reads, respectively. The reads were extracted from the raw data files with DEXTRACTOR rev-844cc20 and general quality was assessed with FastQC v0.11.3.

Read correction and assembly

Proovread v2.12[45] was used for hybrid correction of the PacBio reads with the respective trimmed MiSeq read sets. Correction generated 423 Mbp (N50: 5994 bp) and 741 Mbp (N50: 7328 bp) of high accuracy long reads for E4-10P and E4-10M1, respectively. Reads were assembled into contigs with SPAdes v3.5.0[46] using the dipspades.sh module. Trimmed MiSeq reads were provided as paired-end libraries and corrected PacBio reads as single-end libraries. To account for structurally diverging sister chromosomes caused by asexual reproduction, the –expect-rearrangements flag was set. Assembly metrics were assessed with QUAST v2.3[47]. The E4-10P data set was assembled into 326 consensus contigs of at least 1000 bp, with a total assembly length of 40.3 Mbp and an N50 of 290 kbp. The E4-10M1 genome was assembled into 463 consensus contigs longer than 1000 bp, with a total assembly length of 31.4 Mbp and an N50 of 177 kbp.

Proovread.cfg

```
#-- SI: proovread.cfg ----------------------#
'seq-filter' => {
   '--trim-win' =>  "10,1",
   '--min-length' => 500,
},
'sr-sampling' => {
   DEF => 0, # no sampling - entire sr-file
},
```

Reference-guided assembly of the integrated mavirus genome

The E4-10M1 genome assembly was scanned for mavirus integration sites with blastn [NCBI BLAST v2.2.29+[48]]. The search returned one partial hit with 7000 bp and a few small hits with less than 600 bp alignment length. Additionally, partial hits were visualized and analyzed in context of the assembly graph structure using Bandage v0.4.2[49]. A full-length assembly of the potentially integrated mavirus genome sequence from the E4-10M1 set was generated through a reference guided assembly approach: Corrected PacBio reads of the E4-10M1 strain were aligned to the mavirus reference genome with blastn and strict settings (-evalue 10e-10 -perc_identity 96). Matching reads longer than 1000 bp were extracted and assembled with SPAdes v3.5.0[46] with the --only-assembler flag set.

Detection/analysis of integration sites

Mavirus integration sites in the host genome were detected indirectly by identification of reads covering the junctions between a location in the *C. roenbergensis* genome and the terminal region of mavirus. In preparation, paired E4-10M1 MiSeq reads were merged with FLASh v1.2.11[50] into longer single-end fragments to maximize the chances for unambiguous hits in subsequent mappings. The merged fragments as well as the corrected E4-10M1 PacBio reads were aligned to the revised TIR region of the mavirus genome with bwa mem [BWA v0.7.10-r984-dirty[51]] and samtools

v1.1[52]. Fragments with a minimum alignment length of 30 bp and a minimum overlap of 10 bp at the TIR 5' prime were identified and extracted with a custom script. Due to the total length of 615/616 bp for the TIR, no merged MiSeq fragment spanned the entire TIR, and hence, no information about the strand-orientation of the mavirus core genome could be inferred from the MiSeq data. A read subset containing orientation information was generated by aligning extracted TIR-matching PacBio reads to the full mavirus genome and extracting end overlapping reads with a minimum alignment length of 650 bp. These reads spanned the entire TIR and extended into one side of the core region by at least 34 bp, thus yielding information about the orientation of the integrated element. The extracted mavirus end-overlapping MiSeq and PacBio reads were mapped with bwa mem onto the E4-10M1 genome assembly. Mapping locations of the reads were considered potential integration sites and have been further analyzed manually in a JBrowse[53] genome browser instance, previously set up for the *C. roenbergensis* genome assemblies.

Reconstruction of a mavirus integration site

Direct assembly of an integrated mavirus genome into the host genome was prevented by the diploid state of the *C. roenbergensis* genome and by the repetitive nature of the multiple mavirus integrations, which could not be properly resolved in assembly graph structures. Therefore, we manually reconstructed a contig comprising a mavirus integration site from the previously obtained integration site coordinate information and read evidence available in the MiSeq and PacBio data sets. For the reconstruction, we chose the predicted integration site at nucleotide position 118,064 on contig 5 (length: 208,205 bp). To validate the reconstructed sequence, MiSeq and corrected PacBio reads were mapped back against the artificial contig with bwa mem. Genomic features were annotated by mapping previously obtained host genome annotations (maker v2.31.8)[54] and mavirus gene annotations (PROKKA v1.11 with custom mavirus database[55]) onto the new contig. Annotations were mapped with a custom script based on UCSC annotation lift-over strategies (LiftOver_Howto, Minimal_Steps_For_LiftOver) utilizing Kenttools v302[56]. Visualization of the annotated contig was generated with bio2svg v0.6.0.

Ploidy assessment based on k-mer coverage frequency distribution

19-mer counts of the raw *C. roenbergensis* E4-10P Illumina MiSeq read data set were calculated with jellyfish v2.2.4[57] in canonical representation and plotted with custom R scripts. Peak positions in Figure S2 were identified manually.

Accession Numbers

C. roenbergensis strains E4-10P and E4-10M1 have been deposited in the Roscoff Culture Collection (strain numbers RCC 4624 and RCC 4625, respectively). The GenBank accession number for the reconstructed mavirus integration site of *C. roenbergensis* strain E4-10M1 is KU052222.

Author Contributions

M.G.F. conceived the study, designed and performed experiments, collected and analyzed data and wrote the manuscript. T.H. corrected and assembled sequence data, and analyzed, interpreted and visualized data.

Acknowledgements

This research was supported by the Max Planck Society. We are grateful to C. Suttle for access to host and virus strains, and to the Roscoff team for maintaining and distributing protist strains. We thank K. Barenhoff, K.-A. Seifert, and K. Fenzl for technical assistance, U. Mersdorf for electron microscopy expertise, C. Roome for IT support, L. Czaja and the Max Planck Genome Centre in Cologne for bioinformatic assistance, S. Higgins and E.V. Koonin for helpful suggestions, K. Haslinger and J. Reinstein for critical reading of the manuscript, and I. Schlichting for mentoring and continued support. The authors declare no conflict of interest.

References

1. Aiewsakun, P. & Katzourakis, A. Endogenous viruses: Connecting recent and ancient viral evolution. *Virology* **479-480,** 26–37 (2015).
2. Jern, P. & Coffin, J. M. Effects of Retroviruses on Host Genome Function. *Annu. Rev. Genet.* **42,** 709–732 (2008).
3. Chuong, E. B., Elde, N. C. & Feschotte, C. Regulatory evolution of innate immunity through co-option of endogenous retroviruses. *Science* **351,** 1083–7 (2016).
4. Zeng, M. *et al.* MAVS, cGAS, and endogenous retroviruses in T-independent B cell responses. *Science* **346,** 1486–92 (2014).
5. Herniou, E. A. *et al.* When parasitic wasps hijacked viruses: genomic and functional evolution of polydnaviruses. *Philos. Trans. R. Soc. B Biol. Sci.* **368,** 20130051–20130051 (2013).
6. Raoult, D. *et al.* The 1.2-megabase genome sequence of Mimivirus. *Science* **306,** 1344–50 (2004).
7. Philippe, N. *et al.* Pandoraviruses: Amoeba Viruses with Genomes Up to 2.5 Mb Reaching That of Parasitic Eukaryotes. *Science* **341,** 281–286 (2013).
8. Iyer, L. M., Aravind, L. & Koonin, E. V. Common origin of four diverse families of large eukaryotic DNA viruses. *J. Virol.* **75,** 11720–34 (2001).
9. Iyer, L. M., Balaji, S., Koonin, E. V & Aravind, L. Evolutionary genomics of nucleo-cytoplasmic large DNA viruses. *Virus Res.* **117,** 156–184 (2006).
10. Abergel, C., Legendre, M. & Claverie, J.-M. The rapidly expanding universe of giant viruses: Mimivirus, Pandoravirus, Pithovirus and Mollivirus. *FEMS Microbiol. Rev.* **39,** 779–796 (2015).
11. Legendre, M. *et al.* In-depth study of Mollivirus sibericum , a new 30,000-y-old giant virus infecting Acanthamoeba. *Proc Natl Acad Sci U S A* 201510795 (2015). doi:10.1073/pnas.1510795112
12. Legendre, M. *et al.* Thirty-thousand-year-old distant relative of giant icosahedral DNA viruses with a pandoravirus morphology. *Proc Natl Acad Sci U S A* **111,** 4274–9 (2014).
13. Reteno, D. G. *et al.* Faustovirus, an Asfarvirus-Related New Lineage of Giant Viruses Infecting Amoebae. *J Virol* **89,** 6585–6594 (2015).
14. Fischer, M. G. Giant viruses come of age. *Curr. Opin. Microbiol.* **31,** 50–7 (2016).
15. Mutsafi, Y., Fridmann-Sirkis, Y., Milrot, E., Hevroni, L. & Minsky, A. Infection cycles of large DNA viruses: Emerging themes and underlying questions. *Virology* **467,** 3–14 (2014).
16. La Scola, B. *et al.* The virophage as a unique parasite of the giant mimivirus. *Nature* **455,** 100–4 (2008).
17. Fischer, M. G. & Suttle, C. A. A Virophage at the Origin of Large DNA Transposons. *Science* **332,** 231–234 (2011).
18. Claverie, J.-M. & Abergel, C. Mimivirus and its virophage. *Annu. Rev. Genet.* **43,** 49–66 (2009).
19. Legendre, M. *et al.* mRNA deep sequencing reveals 75 new genes and a complex transcriptional landscape in Mimivirus. *Genome Res.* **20,** 664–74 (2010).
20. Krupovic, M., Kuhn, J. H. & Fischer, M. G. A classification system for virophages and satellite viruses. *Arch. Virol.* **161,** 233–47 (2016).
21. Gaia, M. *et al.* Zamilon, a novel virophage with mimiviridae host specificity. *PLoS One* **9,** e94923 (2014).
22. Fenchel, T. & Patterson, D. J. *Cafeteria roenbergensis* nov. gen., nov. sp., a heterotrophic microflagellate from marine plankton. *Mar. Microb. Food Webs* **3,** 9–19 (1988).
23. Fischer, M. G., Allen, M. J., Wilson, W. H. & Suttle, C. A. Giant virus with a remarkable complement of genes infects marine

zooplankton. *Proc. Natl. Acad. Sci.* **107,** 19508–19513 (2010).

24. Pritham, E. J., Putliwala, T. & Feschotte, C. Mavericks, a novel class of giant transposable elements widespread in eukaryotes and related to DNA viruses. *Gene* **390,** 3–17 (2007).

25. Kapitonov, V. V & Jurka, J. Self-synthesizing DNA transposons in eukaryotes. *Proc. Natl. Acad. Sci. U. S. A.* **103,** 4540–5 (2006).

26. Krupovic, M., Bamford, D. H. & Koonin, E. V. Conservation of major and minor jelly-roll capsid proteins in Polinton (Maverick) transposons suggests that they are bona fide viruses. *Biol. Direct* **9,** 6 (2014).

27. Krupovic, M. & Koonin, E. V. Self-synthesizing transposons: unexpected key players in the evolution of viruses and defense systems. *Curr. Opin. Microbiol.* **31,** 25–33 (2016).

28. Krupovic, M. & Koonin, E. V. Polintons: a hotbed of eukaryotic virus, transposon and plasmid evolution. *Nat. Rev. Microbiol.* **13,** 105–115 (2014).

29. Yutin, N., Shevchenko, S., Kapitonov, V., Krupovic, M. & Koonin, E. V. A novel group of diverse Polinton-like viruses discovered by metagenome analysis. *BMC Biol.* **13,** 95 (2015).

30. Yutin, N., Raoult, D. & Koonin, E. V. Virophages, polintons, and transpovirons: a complex evolutionary network of diverse selfish genetic elements with different reproduction strategies. *Virol. J.* **10,** 158 (2013).

31. Desnues, C. *et al.* Provirophages and transpovirons as the diverse mobilome of giant viruses. *Proc. Natl. Acad. Sci. U. S. A.* **109,** 18078–83 (2012).

32. Blanc, G., Gallot-Lavallée, L. & Maumus, F. Provirophages in the Bigelowiella genome bear testimony to past encounters with giant viruses. *Proc. Natl. Acad. Sci. U. S. A.* **112,** E5318–26 (2015).

33. Gonzalez, J. M. & Suttle, C. A. Grazing by Marine Nanoflagellates on Viruses and Virus-Sized Particles - Ingestion and Digestion. *Mar. Ecol. Prog. Ser.* **94,** 1–10 (1993).

34. Keeling, P. J. *et al.* The Marine Microbial Eukaryote Transcriptome Sequencing Project (MMETSP): Illuminating the Functional Diversity of Eukaryotic Life in the Oceans through Transcriptome Sequencing. *PLoS Biol.* **12,** e1001889 (2014).

35. Fischer, M. G., Kelly, I., Foster, L. J. & Suttle, C. a. The virion of Cafeteria roenbergensis virus (CroV) contains a complex suite of proteins for transcription and DNA repair. *Virology* **466-467,** 82–94 (2014).

36. Hijnen, W. a M., Beerendonk, E. F. & Medema, G. J. Inactivation credit of UV radiation for viruses, bacteria and protozoan (oo)cysts in water: a review. *Water Res* **40,** 3–22 (2006).

37. Berns, K. I., Pinkerton, T. C., Thomas, G. F. & Hoggan, M. D. Detection of adeno-associated virus (AAV)-specific nucleotide sequences in DNA isolated from latently infected Detroit 6 cells. *Virology* **68,** 556–60 (1975).

38. Cheung, A. K., Hoggan, M. D., Hauswirth, W. W. & Berns, K. I. Integration of the adeno-associated virus genome into cellular DNA in latently infected human Detroit 6 cells. *J. Virol.* **33,** 739–48 (1980).

39. Samulski, R. J., Berns, K. I., Tan, M. & Muzyczka, N. Cloning of adeno-associated virus into pBR322: rescue of intact virus from the recombinant plasmid in human cells. *Proc Natl Acad Sci USA* **79,** 2077–81 (1982).

40. Christie, G. E. & Dokland, T. Pirates of the Caudovirales. *Virology* **434,** 210–21 (2012).

41. Gao, X., Hou, Y., Ebina, H., Levin, H. L. & Voytas, D. F. Chromodomains direct integration of retrotransposons to heterochromatin. *Genome Res* **18,** 359–69 (2008).

42. Garza, D. R. & Suttle, C. A. Large double-stranded DNA viruses which cause the lysis

of a marine heterotrophic nanoflagellate (Bodo sp) occur in natural marine viral communities. *Aquat. Microb. Ecol.* **9,** 203–210 (1995).

43. Reed, L. & Muench, H. A simple method of estimating fifty per cent end points. *Am. J. Hyg.* **27,** 493–7 (1938).

44. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30,** 2114–20 (2014).

45. Hackl, T., Hedrich, R., Schultz, J. & Förster, F. proovread: large-scale high-accuracy PacBio correction through iterative short read consensus. *Bioinformatics* **30,** 1–8 (2014).

46. Bankevich, A. *et al.* SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *J. Comput. Biol.* **19,** 455–77 (2012).

47. Gurevich, A., Saveliev, V., Vyahhi, N. & Tesler, G. QUAST: quality assessment tool for genome assemblies. *Bioinformatics* **29,** 1072–5 (2013).

48. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *J. Mol. Biol.* **215,** 403–10 (1990).

49. Wick, R. R., Schultz, M. B., Zobel, J. & Holt, K. E. Bandage: interactive visualization of de novo genome assemblies. *Bioinformatics* **31,** 3350–2 (2015).

50. Magoč, T. & Salzberg, S. L. FLASH: fast length adjustment of short reads to improve genome assemblies. *Bioinformatics* **27,** 2957–63 (2011).

51. Li, H. Toward better understanding of artifacts in variant calling from high-coverage samples. *Bioinformatics* **30,** 2843–51 (2014).

52. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25,** 2078–9 (2009).

53. Skinner, M. E., Uzilov, A. V, Stein, L. D., Mungall, C. J. & Holmes, I. H. JBrowse: A next-generation genome browser. *Genome Res.* **19,** 1630–1638 (2009).

54. Campbell, M. S., Holt, C., Moore, B. & Yandell, M. in *Current Protocols in Bioinformatics* **48,** 4.11.1–4.11.39 (John Wiley & Sons, Inc., 2014).

55. Seemann, T. Prokka: rapid prokaryotic genome annotation. *Bioinformatics* **30,** 2068–9 (2014).

56. Kent, W. J. The Human Genome Browser at UCSC. *Genome Res.* **12,** 996–1006 (2002).

57. Marçais, G. & Kingsford, C. A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinformatics* **27,** 764–70 (2011).