Supplementary information

for

Wide distribution of phage that infect freshwater SAR11 bacteria

Lin-Xing Chen[1], Yanlin Zhao[2], Katherine D. McMahon[3], Jiro F. Mori[4,11], Gerdhard L. Jessen[4,12], Tara Colenbrander Nelso[4], Lesley A. Warren[4,5], and Jillian F. Banfield[1,6,7,8,9,10*]

[1] Department of Earth and Planetary Sciences, University of California, Berkeley, CA 94720, USA

[2] Fujian Provincial Key Laboratory of Agroecological Processing and Safety Monitoring, College of Life Sciences, Fujian Agriculture and Forestry University, Fuzhou, Fujian, China

[3] Departments of Civil and Environmental Engineering, and Bacteriology, University of Wisconsin, Madison, WI 53705, USA

[4] Department of Civil and Mineral Engineering, University of Toronto, Toronto, Canada

[5] School of Geography and Earth Science, McMaster University, Hamilton, Canada

[6] Earth Sciences Division, Lawrence Berkeley National Laboratory, Berkeley, CA 94720, USA

[7] Department of Environmental Science, Policy, and Management, University of California, Berkeley, CA 94720, USA

[8] Chan Zuckerberg Biohub, San Francisco, CA, USA

[9] Innovative Genomics Institute at UC Berkeley, Berkeley, CA 94720, USA

[10] The University of Melbourne, Australia

[11] Present address: Graduate School of Nanobiosciences, Yokohama City University, Yokohama, Japan

[12] Present Address: Gerdhard L. Jessen, Instituto de Ciencias Marinas y Limnológicas, Facultad de Ciencias, Universidad Austral de Chile, Valdivia, Chile

[*]Corresponding author: Jillian F. Banfield

Email: jbanfield@berkeley.edu

Telephone: 510-316 4334

Address: McCone Hall, Berkeley, CA 94720

1. The absence of ribosomal protein L30 is a shared feature of all published SAR11 genomes

When evaluated the completeness using the 51 single copy genes that are universal in bacterial genomes, we found the *Fonsibacter* genome reported in this study had no ribosomal protein L30 **(**rpL30) gene. Further analyses indicated all SAR11 but one (AAA795-11P) genome lacked the rpL30 gene. For reference, we retrieved the two rpL30 genes in the Alphaproteobacteria bacterium casp-alpha1, a megabin that most close to SAR11 (but not SAR11; (1)). With these three rpL30 as queries, we searched NCBI using BLASTp for homologies. Phylogenetic analyses of the queries and their BLASTp hits indicated that, the one from AAA795-11P was clustered with those from Marinimicrobia, a candidatus group in the FCP superphylum (Fig. S2a). Moreover, the other three genes on the scaffold of AAA795-11P with rpL30, were also most close to those from Marinimicrobia (Fig. S2b). This result indicated that the rpL30 scaffold was misbinned into AAA795-11P. Thus, we concluded that the absence of rpL30 gene is a general feature of SAR11 genomes.

2. Alternatives for host cells lysis of SAR11 phages

No conventional lysozyme, key enzyme for an essential step in cell lysis of the virulence cycle, was detected for four phages in *HTVC019Pvirus* group I (including HTVC120P; Fig. 1c). However, a bacterial toxin which showed homology to phage lysozyme was present (Supplementary Table 3), and may perform this function (2).

For HTVC010P-related phages, only two of them contained a lysozyme (Fig. 2b). Next to the lysozyme within these two genomes, we detected a holin (GTA_holin_3TM; PF11351) (Fig. 2a), which was also found in all other phage genomes excluding the FFC draft genome likely due to incompleteness (Fig. 2b). Holin is thought to enable lysis by providing access to the peptidoglycan (3). In all but one of the phages without lysozyme, a gene encoding peptidase M15 (Peptidase_M15_3; PF08291) was detected next to the holin protein (Fig. 2b). Peptidase_M15_3 represents the C-terminal domain of zinc D-Ala-D-Ala carboxypeptidases from *Streptomyces* species (4), and related peptidase with peptidoglycan hydrolase activity has been documented (5). Moreover, the prediction of these peptidase M15 using SWISS-MODEL indicated they matched with the 1lbu.1.A template, which was a muramoyl-pentapeptide carboxypeptidase for bacterial cell wall degradation. Based on this information, we speculated that in phage without lysozyme, the holin and peptidase M15 work together for lysis.

3. Potential reason for the low abundance of *Fonsibacter* phages in most Lake Mendota samples

Given the sampling strategy of Mendota Lake, that is, filtering microorganisms cells onto 0.2 µm pore-size filters (6), and assuming that the HTVC010P-related phages had a comparative capsid size as the Pelagiphage HTVC010P (50 nm in diameter (7)), we speculated that the obtained phages were primarily from the surface and/or inside the bacterial host cells. In this case, it is reasonable that the phages were with low relative abundance in most samples (Fig. 4c).

4. The sole example of CRISPR-Cas system in SAR11

To date, only one marine SAR11 genome (single-cell genome AAA240-E13) has been reported with a putative CRISPR locus (8). However, no other cas protein was identified near the locus excepting a cas4-like protein located on another scaffold. We tried but failed to link these two scaffolds or with any other scaffold in the genome bin, based on sequence overlap at the scaffold ends. Given the wide detection of cas4-like genes in archaeal, bacterial and phage genomes (9, 10), it remains unclear for the role of this only reported putative CRISPR-Cas system in SAR11.

5. Highly similar TerL genes of group 3 of HTVC010P-related phages found in marine and freshwater habitats.

In the main text, we described the detection of highly similar TerL genes related to the group 3 of HTVC010P-related phages represented by HTVC010P-related_33_76 (Fig. 5). Here we describe two study cases to show the details.

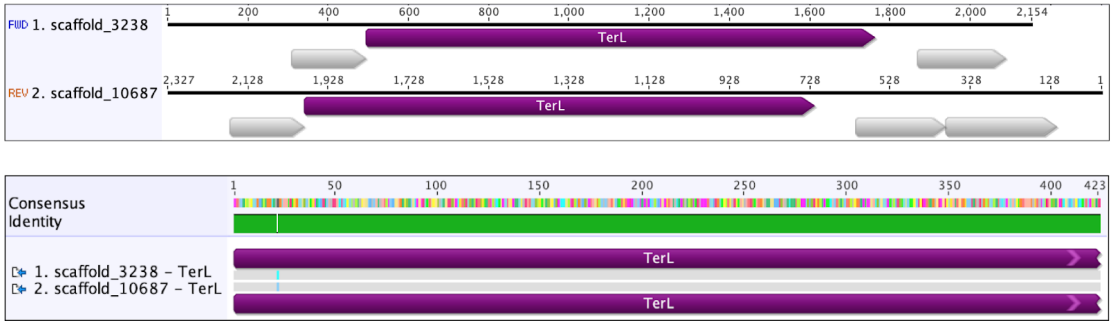
5.1 Case 1 - European eel may transition HTVC010P-related_33_76 and also marine SAR11

We detected one TerL in each of the two European eel related samples. The TerL were similar to that of HTVC010P-related_33_76 (80% and 82% amino acid similarity), and both of them were partial genes. To obtain full length TerL genes, we downloaded the raw reads from NCBI SRA under the accession number of SRR1586370 and SRR1586416 (11–13), which were sequenced with Illumina PE100 kit. Quality control was performed on the raw reads as described in the "Methods" section of the main text, followed by *de novo* assembly using idba_ud (parameters: --pre_correction --mink 20 --maxk 80 --step 20). The scaffolds were compared against the two incomplete TerL proteins using BLASTx, and the targeted scaffolds had complete TerL genes for both of them. The complete Terl shared high similarity with only one mismatch to that of HTVC010P-related_33_76 (Extended Data Fig. 1). We compared these two European eel related TerL proteins against all TerL proteins that we have identified in public databases, and found 12 of them had a minimum identity of 97% (up to 99.2%; Extended Data Fig. 2). These 12 TerL proteins were from Groves Creek Marsh (Skidaway Island, Georgia; sequences 3-8), White Oak River estuary (North Carolina, US; 9-13) and Delaware Coast (US; sequence 14).

It was hypothesized that the epidermal mucosa could work as a phage enrichment layer (14, 15), and which was documented using the European eel (*Anguilla anguilla*) as an animal model (12, 13). The European eel travel from

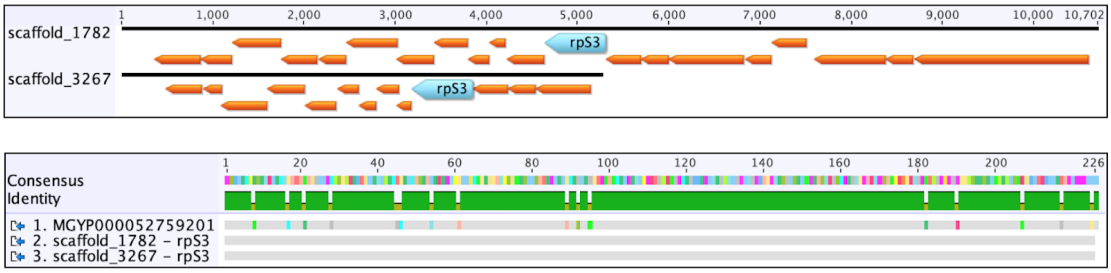
Europe to the East Coast of North America and back to Europe during their life, they usually spawn and lay eggs in the Sargasso Sea (16). Given the sampling site of these two European eel-related TerL proteins (sampled from Alfacada pond of Spain), and their high similarity to those detected in the East Coast of North America, we speculated that the European eel play a role in the transition of these phages into freshwater ecosystems in Europe. Moreover, the Sargasso Sea also plays a major role in the migration of the American eel and the American conger eel, we suspected if these eel species also have a similar role in the phage transition between the ocean and freshwater ecosystems in North America.

For the two European eel related samples, BLASTp search did not identify any proteins similar to *Fonsibacter* rpS3. We suspected if this is due to the low relative abundance of *Fonsibacter* in the corresponding communities. Upon this, we compared the quality-reads (see above) against all available SAR11 rpS3 nucleotide sequences with a minimum similarity of 80% using BLASTn (e-value threshold = 1e-10). As a result, no quality read showed the highest similarity to *Fonsibacter* rpS3 under these thresholds, we thus concluded that there was not *Fonsibacter* member in the sampled communities. However, we identified two identical rpS3 protein sequences assigned to the marine SAR11 subclade (scaffold_3267 and scaffold_1782; Extended Data Fig. 3). These proteins shared a 96% AA similarity to a Tara ocean protein (MGYP000052759201), via searching the Tara ocean database (https://www.ebi.ac.uk/metagenomics/sequence-search/search/phmmer)(no detailed geographic information of this sequence is available). Thus, we speculated that the European eel could also transition the marine SAR11 to freshwater ecosystems.



Extended Data Fig. 1. Upper panel: Genomic context of two European eel-related complete TerL (hypothetical proteins in gray). Bottom panel: Alignment of two European eel-related complete TerL proteins.

Extended Data Fig. 2. Upper panel: Similarity between European eel-related TerL proteins and those detected in global environments. Bottom panel: Alignment of European eel-related TerL proteins and relatives (> 97% similarity) from North America.

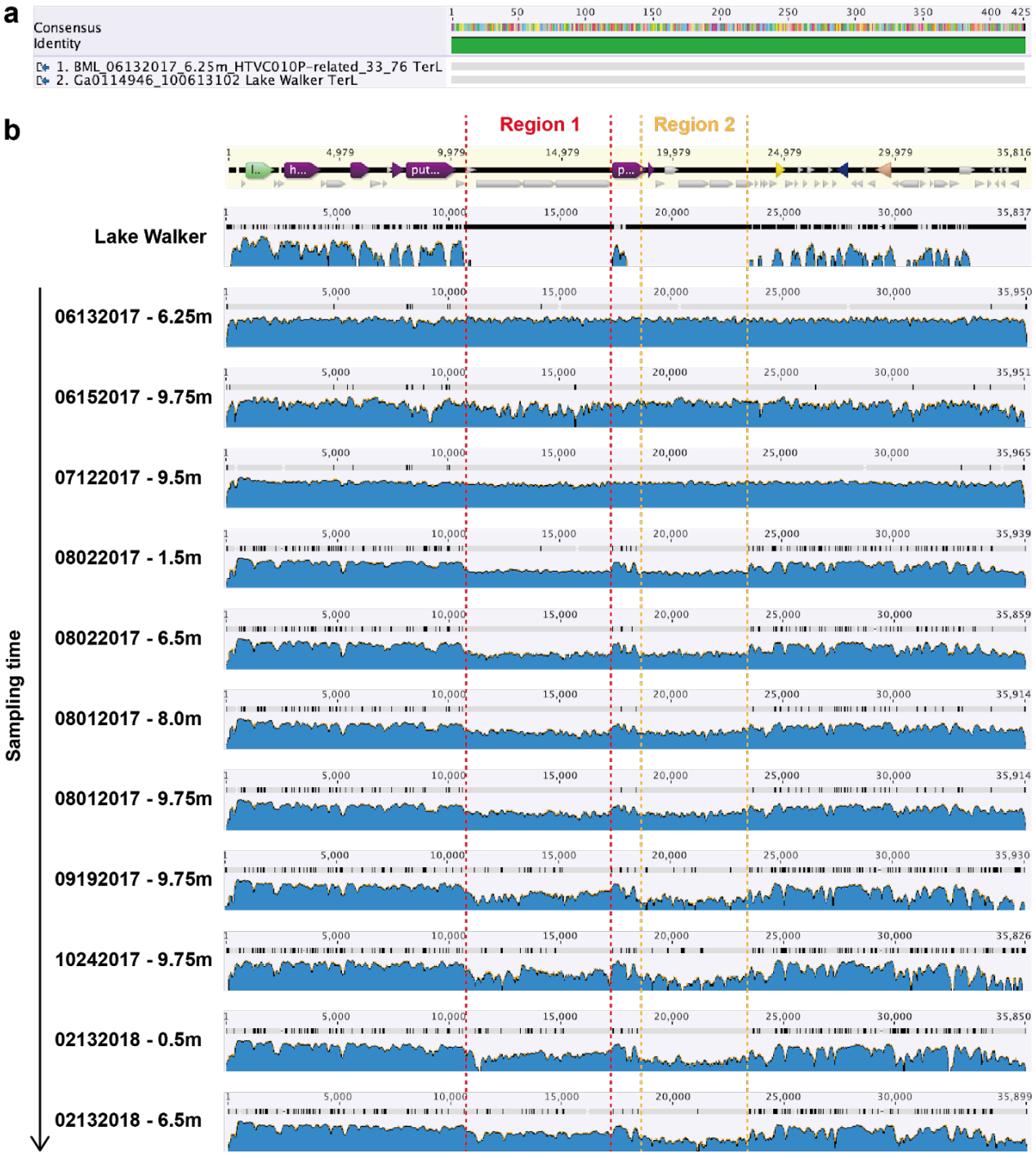


Extended Data Fig. 3. Upper panel: Genomic context of the SAR11 rpS3. Bottom panel: Alignment of the two identical SAR11 rpS3 proteins with one from Tara ocean project database.

5.2 Case 2 - Identical HTVC010P-related_33_76 TerL found in a sediment sample of Lake Walker

The Lake Walker had a high concentration of total dissolved solids (TDS) due to the lower water level, which resulted from the overuse of water in the River Walker, the only input of Lake Walker (https://www.walkerbasin.org/history-of-walker-lake). As a result, the high TDS can no longer support the native fish and wildlife populations. For example, the Lahontan cutthroat trout have not been observed since year 2010 when TDS reached 20,000 mg/L. For reference, a TDS range of 8,000 - 12,000 mg/L is optimal for lake health conditions, typical seawater TDS is 40,000 mg/L, brackish TDS is around 10,000 mg/L.

One TerL from a Lake Warker sediment sample (collected on Nov 2nd of 2013; TDS = 19,000 mg/L) was identical to that of HTVC010P-related_33_76 (Extended Data Fig. 4a). We downloaded the raw reads from NCBI SRA (SRR5747948, generator of this unpublished data: Duane Moser), and performed quality control (as described in the "Methods" section of the main text). However, the relatively low abundance of this Lake Walker phage inhibited attempt for genome reconstruction, instead, we mapped the quality-reads to the HTVC010P-related_33_76 genome with 98% similarity. The result showed the high similarity of all the protein-coding genes with predicted function (Fig. 2), while not for some large genes encoding hypothetical protein (Regions 1 and 2; Extended Data Fig. 4b). By mapping EPL quality reads to the genome of HTVC010P-related_33_76 with 98% similarity, we detected similar coverage profiles as observed for the Lake Walker sample that, most of the EPL samples had a lower coverage in Regions 1 and 2. This observation likely indicated the occurrence of a close phage without the genes in Regions 1 and 2.



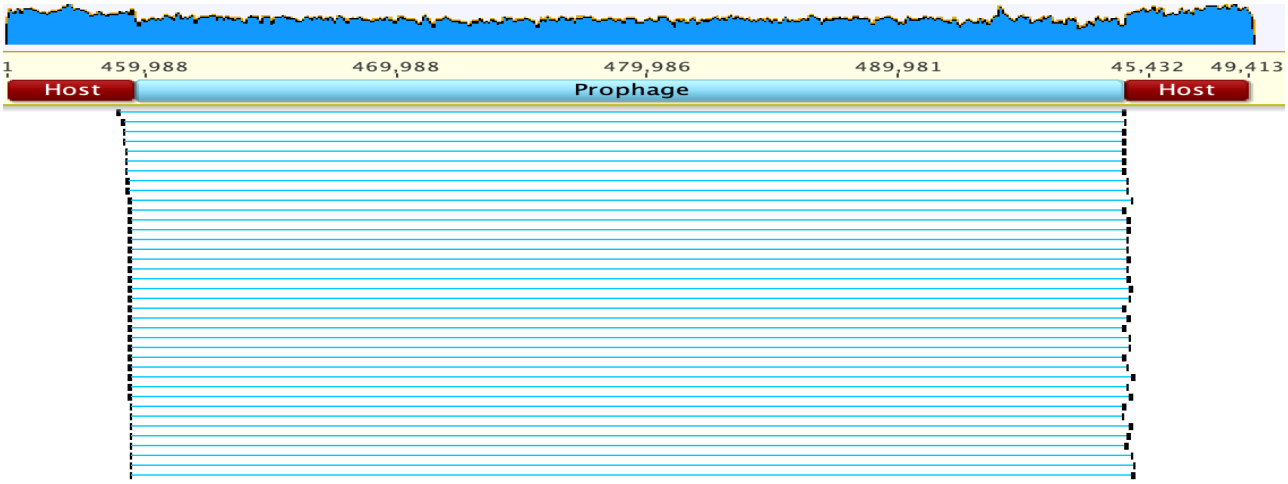
Extended Data Fig. 4. Comparative analyses related phages from EPL (HTVC010P-related_33_76) and Lake Walker sediment. (a) Identical TerL of HTVC010P-related_33_76 and the related phage from Lake Walker. (b) mapping of paired-end reads from the Lake Walker sediment sample, and different sampling points (and depth) to the genome of HTVC010P-related_33_76 (with 98% read similarity to the phage genome).

6. A guide to obtaining complete phage genome from prophage-containing scaffold

Some phages could perform the lysogenic life strategy by inserting the genome into their bacterial hosts, which are prophages. Once we obtain a bacterial host scaffold with prophage, it generally indicates that most of this phage population perform lysogenic life strategy (otherwise the phage genome will be in a separated scaffold, not in the host scaffold). However, in the same community or other related communities (collected at different time points or different locations within the same region), may exist the same host cells without the prophage. Thus, we may be able to confirm the exact recombination site of phage genome into the host genome, and also the true length of the phage genome. Here we show a detailed guide of how to obtain complete phage genome from the prophage genome in the host.

Step 1: when phage-specific proteins are identified in a scaffold of a bacterial/archaeal genome bin, paired-end reads should be mapped to this scaffold (using bowtie or similar tools), to confirm if there is a complete prophage genome in this scaffold (it means the prophage genome is in the middle of the scaffold; bacterial/archaeal genes + phage genes + bacterial/archaeal genes).

Step 2: check the mapping profile to see if there are paired-end reads spanning the potential prophage region (Extended Data Fig. 5). If existed, these paired-end reads mapped on the host scaffold are from host cells without the prophage, and these in the prophage are from lytic phage cells. In this example, there are only two paired-end reads in the prophage region.



Extended Data Fig. 5. Long-distance spanning paired reads showing the location of prophage in host genome scaffold.

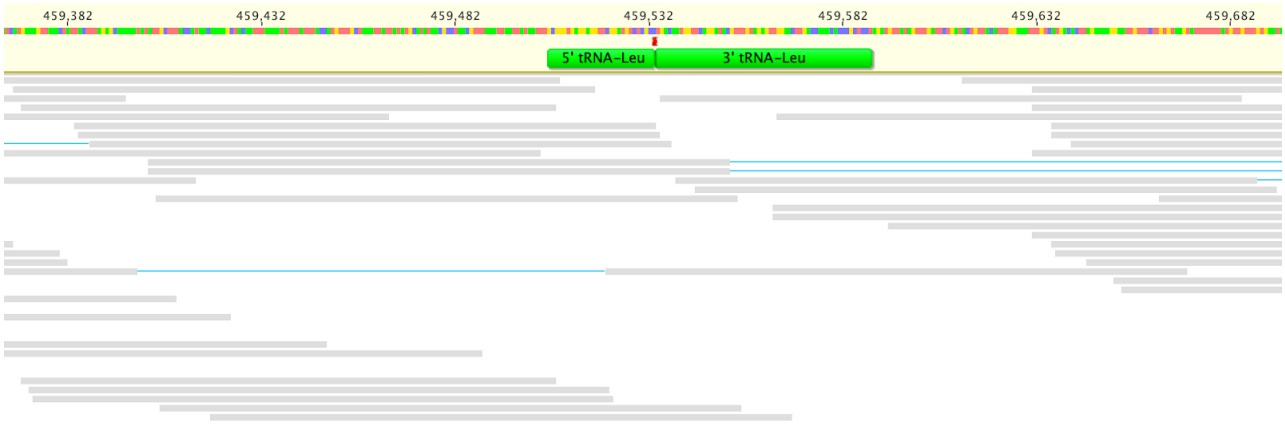
Step 3: the key step is to find some reads (multiple should be better) that with part of them perfectly matched to the scaffold, while the other part could not be matched, also the unmatched part from these partially-mapped reads could be aligned perfectly (Extended Data Fig. 6). In this example, "CTGGTATCGAAAGATGTGAAGGTTCAAGTCCTTTCTCCCGCACCATGTAGAATATGAAAGTA" (see the red zoom-in). Note that these reads were from the host cells without prophage. Upon this, the unmatched part must be found near the right end of the prophage on the scaffold (see the blue zoom-in). If this is true, then the recombination site could be determined as shown in Extended Data Fig. 6.

Extended Data Fig. 6. Upper panel: Overview of an example for a host scaffold with prophage. Middle panel: How to determine the left edge of prophage into host genome. Bottom panel: How to determine the right edge of prophage into host genome

Generally, we could do this by starting from the left end, or from the right end. However, there is a small difference if started from the right end, as we could see the unmatched part is not exactly at the end of the prophage (blue zoom-in box; Extended Data Fig. 6), this is because the phage shared a small part of DNA with the host (Extended Data Fig. 7), that is "AGTCTTAGGAA" ('core sequence'), which is part of the host tRNA-Leu sequence, and also the recombination site. This recombination site sequence is with only one mismatch from that of Pelagiphage HTVC025P ("GTCTTAGGAAC", (17)).

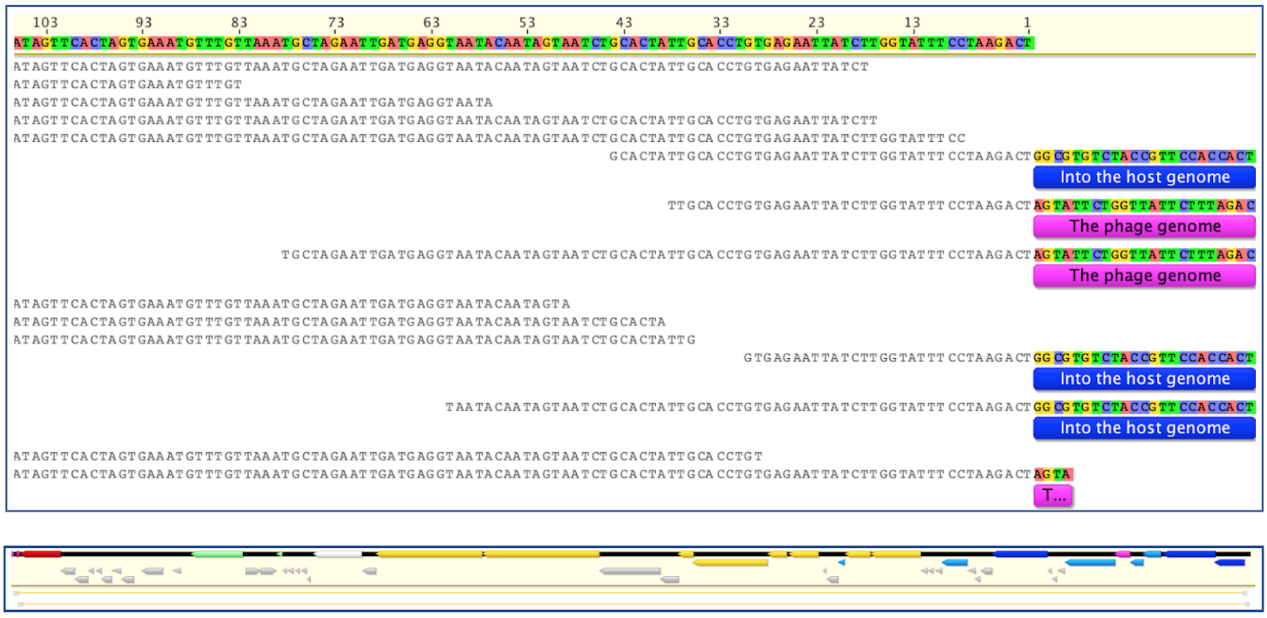
Extended Data Fig. 7. The 'core sequence' at the insertion site of prophage into host genome. The attB, aatP, attL and attR locations are shown and the recombination site was identified by sequence alignment.

Step 4: delete the prophage from the host genome scaffold based on the recombination site determined in step 3, and mapped quality reads to the host scaffold without prophage, to confirm the insertion location. Paired reads mapped to this region confirmed the insertion position (Extended Data Fig. 8).



Extended Data Fig. 8. Reads mapping to the host scaffold with prophage deleted. The red stripe shows the recombination site of the phage into the host genome, which split the host tRNA-Leu into two fragments.

Step 5: map paired-end reads to the prophage genome to confirm if it is circular. If the prophage could be circularized, there should be reads that cover both end of the prophage genome. As shown below, we found three such reads (indicated in pink) (Extended Data Fig. 9). Also, we found two paired reads that flank both ends of the prophage genome. This information suggests the existence of free living particles of this phage.



Extended Data Fig. 9. Upper panel: Reads mapping shows the prophage genome is circular. Bottom panel: Reads mapping shows end-to-end flanking paired reads.

It should be noted that under some circumstances the recombination site may not be determined, and/or the prophage could not be circularized into a complete phage genome, for example, (1) when all the host cells are with the prophage, (2) when the sequencing coverage is too low and there are not enough reads spanning the recombination sites, (3) when the prophage has several variants that share high sequence similarity, sometimes it will be impossible to determine.