Science Advances
AAAS

advances.sciencemag.org/cgi/content/full/6/6/eaay5981/DC1

Supplementary Materials for

Expansion of known ssRNA phage genomes: From tens to over a thousand

J. Callanan, S. R. Stockdale, A. Shkoporov, L. A. Draper, R. P. Ross, C. Hill*

*Corresponding author. Email: c.hill@ucc.ie

Published 7 February 2020, *Sci. Adv.* **6**, eaay5981 (2020)
DOI: 10.1126/sciadv.aay5981

The PDF file includes:

Other Supplementary Material for this manuscript includes the following:

(available at advances.sciencemag.org/cgi/content/full/6/6/eaay5981/DC1)

Data S1 (.rar format). ssRNA phage finding hidden Markov model and associated sequences.
Data S2 (.rar format). Bioinformatic scripts used during data analysis.

Supplementary Text

Existing ssRNA phage sequences

Due to nomenclature/accession number disparities associated with sequences deposited to different databases/organisations, we graphically depicted all the ssRNA phage sequences we could identify for simplicity (Fig. S1). The latest International Committee on Taxonomy of Viruses (ICTV) report (*10*) was used to initially identify 57 *Leviviridae*, although not all had identifiable genomes within public sequence repositories (Fig. S1A). The single largest source of ssRNA phage sequences was from the recent metagenomic study of Krishnamurthy *et al*. (2016), where they identified 158 ssRNA phage sequences across invertebrate, vertebrate, sewage, aquatic and soil samples (Fig. S1B; (*21*)). By investigating the NCBI Taxonomy database, an additional 35 unique ssRNA phage genome sequences were identified (including those from the ICTV report).

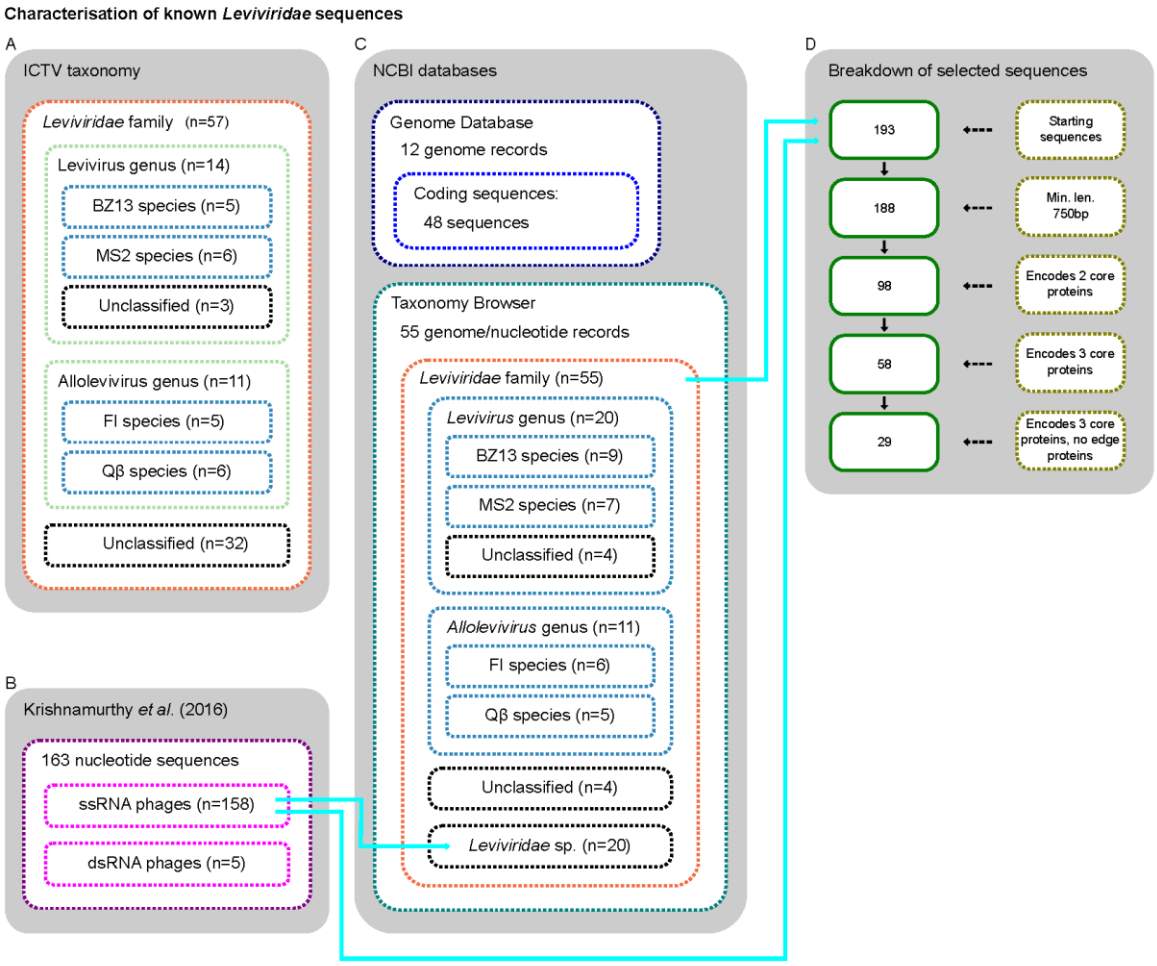


Fig. S1. Workflow depiction of known ssRNA phage sequences. Workflow depiction of known ssRNA phage sequences, outlining: **(A)** ICTV taxonomy, **(B)** Krishnamurthy *et al*. (2016), **(C)** NCBI Genome and Taxonomy database available sequences, and **(D)** the breakdown of the identifiable ssRNA phage sequences used in this study.

While a total of 193 previously described unique ssRNA phage sequences were identified at our study's onset, we characterised how many of these represented complete or near-complete genomes (Fig. S1D). Using HMM 5-MC followed by manual curation, only 29 sequences fulfilled the requirement of encoding all three of the ssRNA phage core proteins (maturation protein, MP; coat protein, CP; RNA-dependent RNA polymerase, RdRp) without

their premature termination by the edge of a phage contig. Determining phage 'edge proteins' was performed by analysing the encoded start and stop codons of the MP and RdRp genes. Only proteins beginning with canonical or cognate start codons (AUG, GUG or UUG) or stop codons (UAG, UAA, or UGA) were considered full length.

Parameters for detecting ssRNA phages

A caveat with searches tools that provide users with an expected-value output is that E-values are dependent on the length of the query sequence, and also size of the searched database. Therefore, for consistent implementation across studies of different sizes, we report where possible hmmscan scores and not E-values. During the iterative development of our ssRNA phage detecting HMMs, we only considered hmmscan scores greater than 50 for continuation into the subsequent model (Fig. S2). However, during our final analysis and detection of ssRNA phages across metatranscriptome samples, we adopted a less stringent hmmscan score of 30.

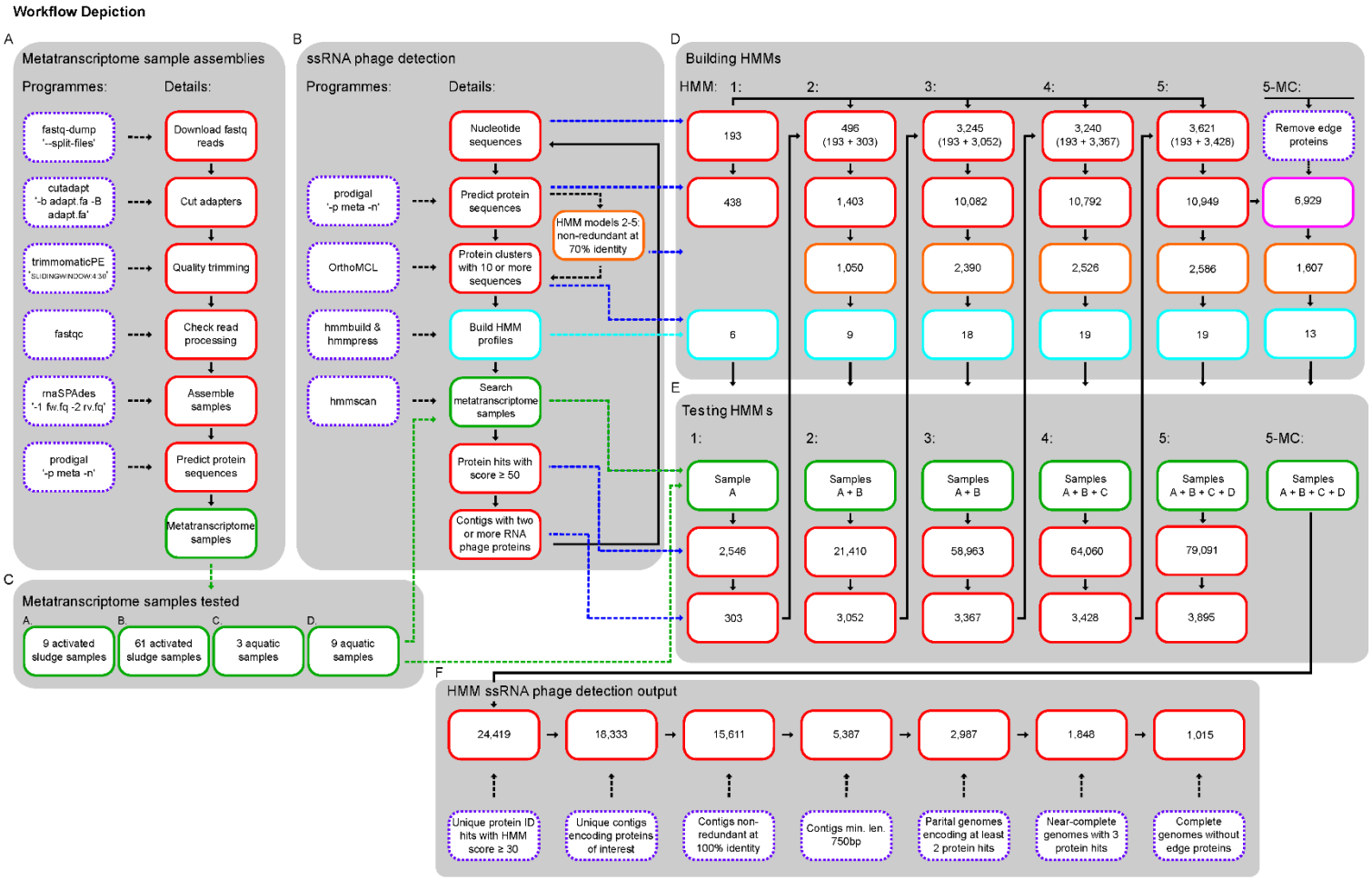


Fig. S2. Workflow depiction of the study pipeline. Workflow depiction of the study pipeline, outlining: **(A)** metatranscriptome sample assemblies, **(B)** the detection of ssRNA phages, **(C)** samples tested, and the breakdown of the **(D)** building, **(E)** testing, and **(F)** output of the HMM iterations.

The implementation of a strict, uncompromising set of parameters was decided early in our study as RdRp proteins are conserved across all RNA viruses. However, as RNA viruses have high nucleotide mutation rates (*60*, *61*), often little or no sequence similarity is observed between evenly closely related RNA viruses. Therefore, future studies could benefit from lowering the hmmscan score thresholds to find more diverse sequences, albeit at the risk of finding false positives.

We observed that the isolated ssRNA phage AP205 (NCBI accession NC_002700.2), which has previously been characterised (*35*), did not make it into our final curated dataset of 29 full length sequences. A detailed examination of phage AP205 highlighted that its CP is so diverse from other identified CP sequences it did not achieve an hmmscan score of 30 using HMM 5-MC. Therefore, as it did not fulfil our criteria for containing the three recognisable core proteins of ssRNA phages, our pipeline excluded it from the 1,044 sequences analysed in detail during this study. However, we investigated the CP of AP205 further and found it is similar to a potential CP encoded by another previously characterised ssRNA phage, ESE002 (NCBI accession KT462711.1;(*21*); with a BLASTp E-value 1.00e-5).

We found six additional putative CP sequences similar to AP205's encoded within the genomes of the 15,611 ssRNA phage sequences detected in this study. Five of the six hits demonstrate only weak sequence similarity to the CP of AP205 (BLASTp E-value ranges 9.00e-06 ≤ 1.00e-04), with a single closely related protein sequence (BLASTp E-value 1.00e-57). Due to our HMM development pipeline requiring a minimum of 10 similar protein sequences in order to generate a HMM, an additional ninth CP cluster was not generated that is specific for the third core protein associated with AP205, ESE002 and the aforementioned

6 additional ssRNA phage contigs. This demonstrates that there is clearly still undiscovered ssRNA phage diversity that additional studies, potentially from alternative environments, will no doubt capture.

Newly identified ssRNA phages

A search of the 82 metatranscriptome samples, generated from activated sludge and water environments (*62–65*), using HMM 5-MC, yielded a total of 24,419 ssRNA phage-associated proteins from distinct contigs using an hmmscan score of 30 (Fig. S2F). Of the proteins detected, there were 8,057 MP, 5,313 CP and 11,049 RdRp hits (Fig. 2A). This supports our observations whereby the CP is the most variable ssRNA phage protein separating into the most clusters, and the RdRp is the most conserved with only two clusters. Indeed, researchers have often used the RdRp as a phylogenetic marker for RNA viral diversity (*13*, *21*, *66*). Ergo, we defined the core genome of ssRNA phages as the MP, CP and RdRp, as these genes are encoded across all currently identified ssRNA phage genomes.

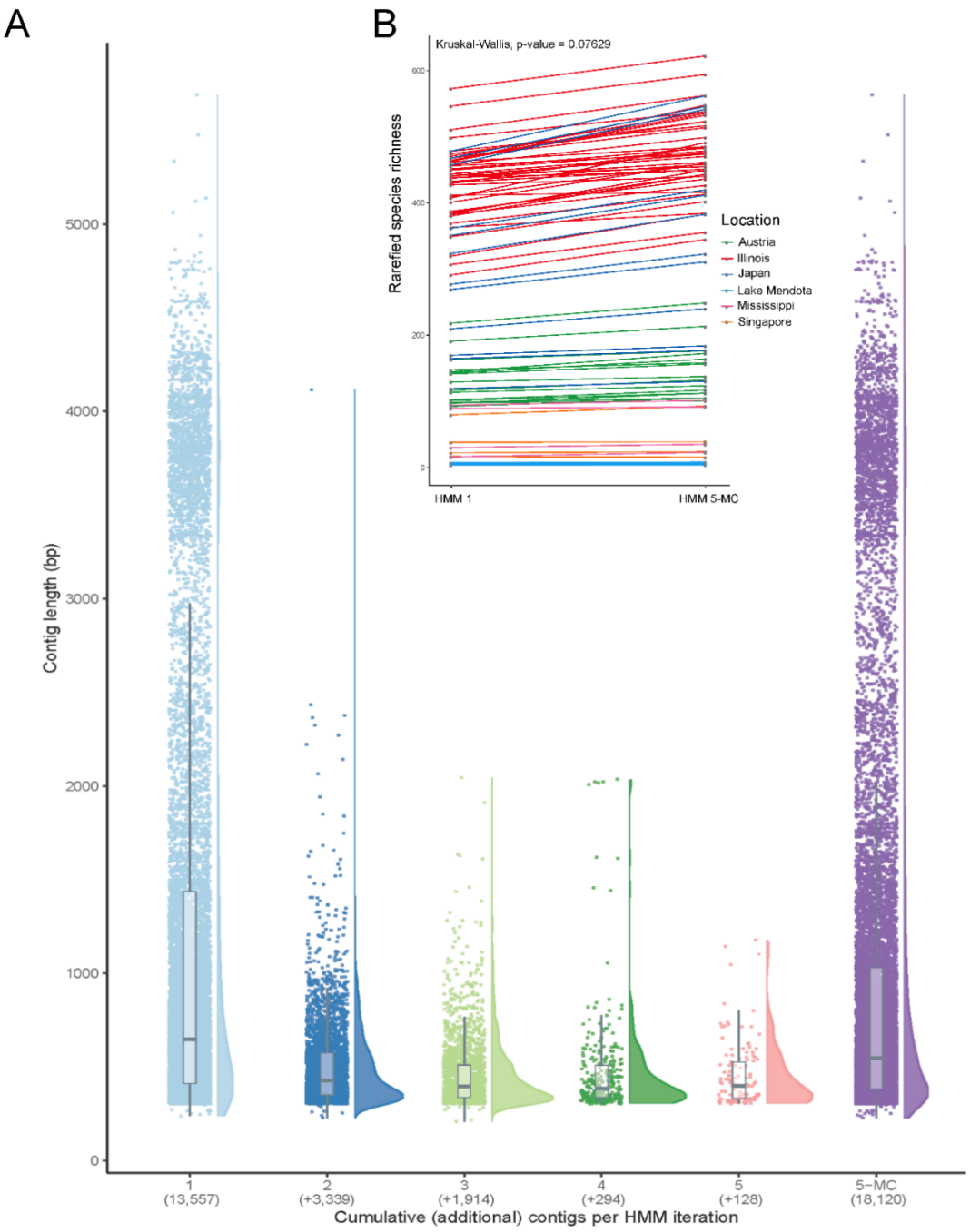


Fig. S3. Identification of ssRNA phage contigs within 82 metatranscriptome samples. **(A)** HMMs 1-5 were iteratively improved to detect additional contigs within the same samples (redundant contig hits indicated within brackets). The final manually curated HMM, 5-MC, was developed using only complete proteins after removing 'edge-proteins'. **(B)** Paired-sample rarefied species richness comparisons using HMMs 1 and 5-MC.

A total of 24,419 predicted proteins were associated with 18,333 ssRNA phage sequences, of which 15,611 were non-identical (made non-redundant at 100% identity to remove duplicates). Parameters for minimum length and number of core proteins within a genome resulted in; 5,387 contigs with a minimum length of 750bp, 2,987 contigs that were over 750bp and encoded at least two core proteins, and 1,848 contigs were of sufficient length and encoded all three core proteins. In order to analyse near-complete ssRNA phage genomes, we selected sequences that encoded full-length versions of the three core proteins (i.e. no 'edge proteins'). This yielded 1,015 near-complete novel ssRNA phage genomes (near-complete including genomes with potentially incomplete non-coding genome termini). When considering near-complete and partial genomes, our study increases the number of known ssRNA phage sequences approximately 60-fold.

Genome architecture of ssRNA phages

We predicted ssRNA phage encoded proteins using Prodigal, which was designed for the annotation of bacterial genomes. It is a well reported feature that the compact genomes of ssRNA phages use diverse and atypical mechanisms for the control and production of additional functional proteins. These mechanisms include the formation of RNA secondary structures (*67*, *68*), translational frame-shifting (*69*), and encoding protein sequences within the boundaries of another larger gene sequence (*70*). Therefore, new tools are needed to fully predict coding sequences within not just ssRNA phages, but across all phages.

Of the non-core proteins predicted within ssRNA phage genomes, three specific locations were frequently noted (Fig. 2D). Hypothetical proteins could exist before the MP, after the CP, or following the RdRp. We termed the locations the Alpha position (preceding the MP and closest to the genomes' 5' termini), the Beta position (ensuing the CP), and the Gamma

position (following the RdRp at the genomes' 3' termini). Open reading frames (ORFs) located at the Alpha and Beta positions have previously been shown to encode a lysin protein in several isolated ssRNA phages, such as AP205 (Alpha), and MS2, PP7 and PRR1 (Beta). On 20 instances, there were two hypothetical proteins situated before the MP (termed Alpha 1 and Alpha 2, the former closest to the genomes' termini). However, mapping of the occurrence of hypothetical ORFs on the phylogram of ssRNA phages showed no specific clustering of the Alpha 2 hypothetical protein, but several clades of related sequences encoding a hypothetical in the Alpha 1 position (Fig. S4B). The conservation of an ORF at this Alpha 1 position suggests these hypothetical genes indeed have a biological function.

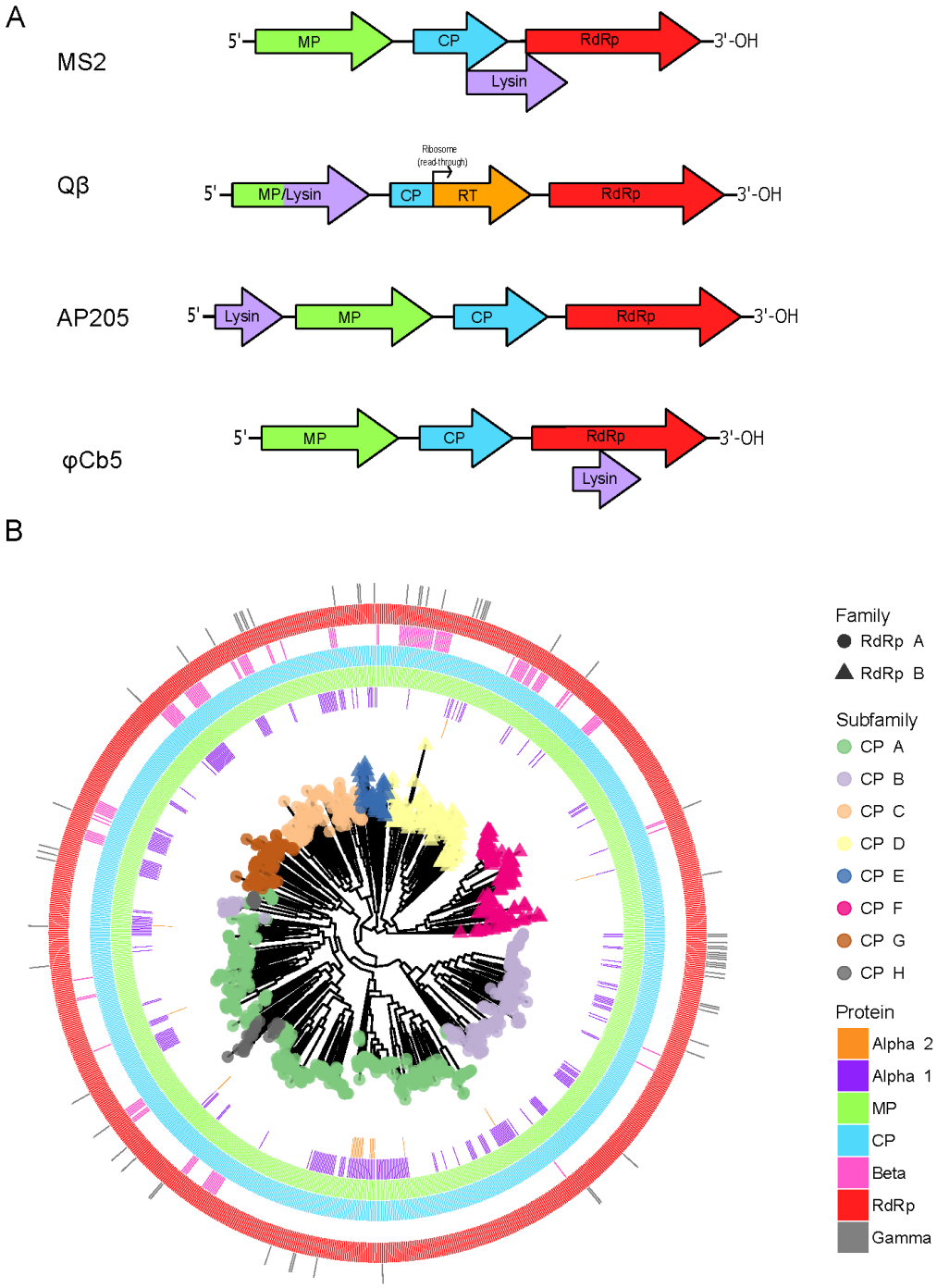


Fig. S4. Genome architecture of ssRNA phages. (A) Genome architecture described for previously known ssRNA phages. Notably, the position and method of translation of the lysin protein is the most variable between sequences. Due to the overlapping, compact nature of MS2 and φCb5 lysin, current computational approaches did not detect these coding

sequences. **(B)** The position of predicted hypothetical proteins mapped onto the suggested phylogeny of ssRNA phages. Specific clades are observed which share hypothetical proteins in the same genomic architectural position.

Following further investigation, we discovered the ORFs predicted to occur after the RdRp had weak sequence similarity to the RdRp clusters (hmmscan score < 30) and may have arisen due to insertion of a premature stop codon during sequencing assembly. However, RNA phages have previously been shown to bypass stop codons as a mechanism to regulate the translational frequency of CP (*28*). We investigated if related ssRNA phages all encode a hypothetical protein downstream of the RdRp, but observed no specific clades of ssRNA phage with hypotheticals in the Gamma position (Fig. S4B), suggesting these ORFs are not conserved and likely a computational artefact. However, only biochemical investigations will completely determine if these hypotheticals are indeed functional proteins, such as the lysin, or an alternative viral replication control mechanism.

Genome architecture examples of previously described ssRNA phages highlight the diverse methods for the production of their associated lysin proteins (Fig. S4A). For phage MS2, its encoded lysin overlaps with the 3'-end of the CP and 5'-edge of the RdRp. During manual curation of MS2-encoded non-core proteins, no lysin was predicted using Prodigal. Therefore, atypical approaches to identify coding sequences will be required to automate the detection of non-core proteins within ssRNA phages. Nonetheless, we already note clades of related ssRNA phages encoding hypothetical proteins in the same Alpha and Beta positions (Fig. S4B).

A total of 506 hypothetical proteins were predicted across the 1,105 full length ssRNA phage genomes. These proteins were clustered using CD-HIT at a 70% sequence identity threshold using a word length of 5 (version 4.6, (*71*)). A total of 29 clusters, representing 262 protein sequences, were generated using CD-HIT that contained 5 or more sequences. All 262 sequences were queried locally against the prokaryotic virus orthologous groups (pVOGs) profile hidden Markov model database (*72*). All sequences associated with the 29 CD-HIT clusters were also aligned using MUSCLE and the alignment queried against a PDB database (version 23 Feb) using the online MPI HHpred Bioinformatics Toolkit (*73*). Only a single CD-HIT cluster, containing 5 protein sequences all from previously identified phages (Qß and closely related phages), were found as similar to previously characterised lysins of ssRNA phages. Both the local pVOG and online HHpred queries identified the lysin (see Supplementary Data).

Phylogenetic assessment of ssRNA phages

To validate previous taxonomic cut-offs against the aforementioned 29 known ssRNA phages with complete genomes, we performed pairwise amino acid identity (AAI) comparisons of the RdRp sequences (Fig. S5). It was found that the current cut-offs for current ssRNA genera and species equated to 50% and 80% pairwise AAIs, respectively. Applying these cut-offs in a bottom-up approach to classifying the 1,044 ssRNA phages (the 1,015 of this study and the 29 previously identified), we predict 331 species and 247 genera (Fig. S6). As these species and genera taxa were defined independently, and not in a hierarchical fashion, we subsequently verified that no members of a single species were detected across two or more genera.

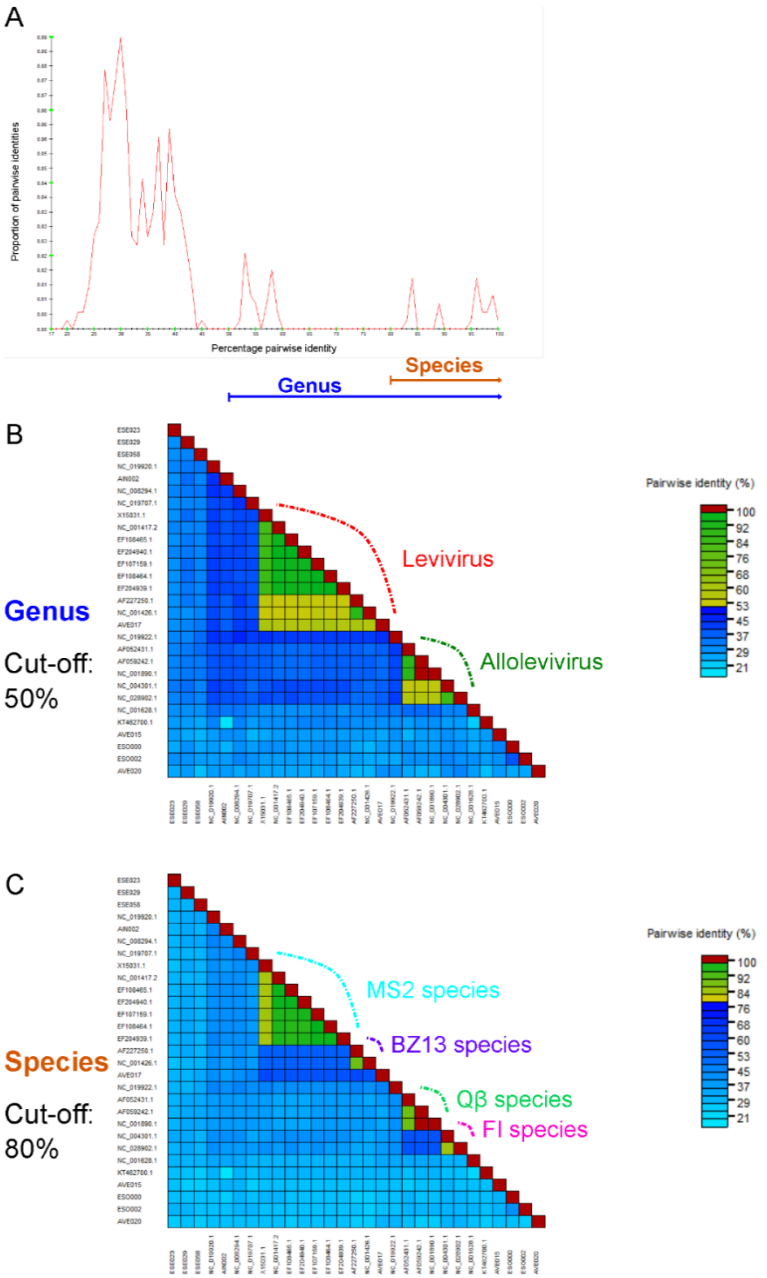


Fig. S5. Taxonomic cutoff values for ssRNA phage genera and species. (A) Pairwise RdRp amino acid identity (AAI) comparisons. Four currently recognised species had AAI ≥80%, while the two genera had AAI ≥50%. Visualisation of AAI at **(B)** genera level, and **(C)** species level with taxonomic groups depicted.

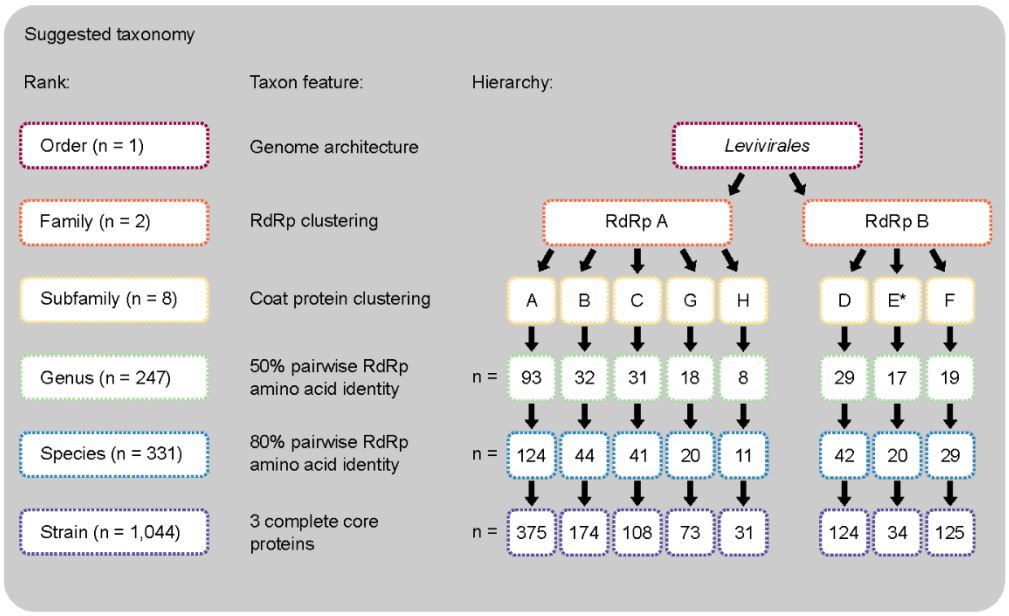


Fig. S6. Potential taxonomic restructuring for ssRNA phages. An outline of defining features for taxonomic ranks and their numerical breakdown is detailed for all 1,044 near-complete ssRNA phage genomes. Asterisk denotes the AVE006 outlier.

For higher taxonomic divisions, where little or no nucleotide or amino acid sequence similarities are observed, we adopted the graph-based clustering of the most variable core protein, the CP, as a feature to distinguish potential subfamilies. Once more, sequences were assessed to ensure no genera or species were found across multiple subfamilies. The most conserved core protein, the RdRp, was subsequently used to distinguish the most distant relationships between ssRNA phages at a potential family taxonomic rank.

A single phage, AVE006, which was identified in a previous study (*21*), did not adhere to the taxonomic defining features outlined in this study (Fig. S6, indicated by asterisk). The CP of AVE006 is most similar to cluster E (hmmscan score 36.4), while its replicase is most similar

to RdRp cluster B (hmmscan score 229.6). All other phages with CP cluster E group by RdRp cluster A. However, AVE006's RdRp is also very similar to RdRp cluster A (hmmscan score 205.2). In agreement with this observation is the position of AVE006 in our phylogenetic analysis (Fig. 3A, highlighted with green arrow head), where AVE006 is situated on the border between suggested ssRNA phage families.

The interactions of ssRNA phages within microbiomes

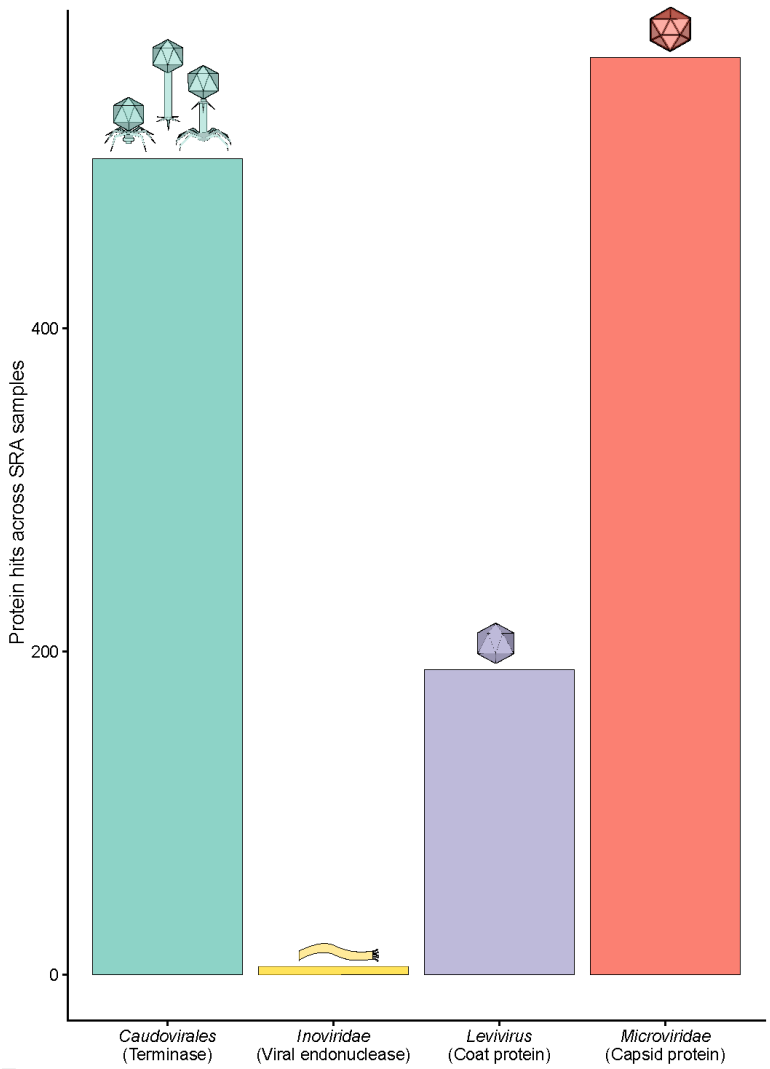
In recent years, there has been a greater effort to understand environmental microbes and their impact on various biogeochemical and nutrient cycles. This includes efforts to better understand the role phages play in shaping microbial community structures and metabolic pathways. For instance, while phages are capable of infecting and killing their microbial hosts, they have also been shown within aquatic environments to augment the photosynthetic capacity of cyanobacteria (*74*). As ssRNA phages have been overlooked within the majority of microbiomes until now, their ecological importance remains to be fully elucidated.

Using our expanded repertoire of ssRNA phage sequences, we queried the bacterial RefSeq database (release 89) and, not surprisingly, found no evidence for ssRNA phage lysogens within bacterial genomes. Therefore, alternative approaches may be required to assign phage-host pairs. We proceeded to characterize the association of ssRNA with alternative community members by searching for evidence of phages co-existing and potentially co-infecting bacteria. In order to perform like-for-like comparisons of ssRNA phages with *Caudovirales*, *Inoviridae* and *Microviridae*, we built profile hidden Markov models (HMMs) using the currently available Pfam (version 32.0) protein families; PF01819, PF04466, PF11726, and PF02305, respectively. As hits against ssRNA phages could potentially occur against either a native *ex vivo* virion or an actively transcribed genome, it is not accurate to

perform direct comparisons against phage genomes composed of DNA. Nonetheless, there is clear evidence for transcription of caudoviral terminases and microviral capsids within activated sludge and aquatic environments (Fig. S7A). Only low levels of *Inoviridae* transcription were observed within the assembled metatranscriptome samples of this study. However, this limited detection may be the result of a poor inovirus-detecting HMM, as it was recently shown they are far more prevalent within environmental samples than previously appreciated (*75*). As ssRNA phages were detected alongside replicating ssDNA and dsDNA phages of differing morphologies, this highlights the complex challenges faced at understanding all facets of a microbiome's viral constituents.

We also looked for evidence of CRISPR spacers against ssRNA phages that could implicate potential host bacteria. Using an in-house pipeline, we built a database of all potential CRISPR spacers predicted using 'pilercr' (version 1.06; (*76*)). A total of 37,095 CRISPR spacers, between 20-75bp in length, were predicted from the 82 metatranscriptome sample assemblies. The BLASTn was adapted for short sequences ("-evalue 1 -word_size 7 -gapopen 10 -gapextend 2 -penalty -1 -dust no"). No CRISPR spacers were found to target the 15,611 new ssRNA phages identified within this study. However, when the predicted CRISPR spacers were queried locally against the viral RefSeq database (version 89), they perfectly matched sequences observed in *Staphylococcus*, *Bacillus*, *Synechococcus* and *Streptococcus*-infecting phages (Fig. S7B).

A



B

CRISPR spacer ID	Viral RefSeq hit	Spacer length (bp)	% Identity
SRR6049586_34554_length_594_cov_4.930902_g32347_i0_16659	Staphylococcus phage SpaA1	29	100
SRR6049586_34554_length_594_cov_4.930902_g32347_i0_16659	Bacillus phage phi4J1	29	100
SRR6049586_34554_length_594_cov_4.930902_g32347_i0_16659	Bacillus phage Waukesha92	29	100
SRR6254353_131237_length_313_cov_1.625000_g128065_i0_20879	Synechococcus phage S-RSM4	28	100
SRR6960509_109618_length_418_cov_6.208696_g104094_i0_25281	Streptococcus phage Str-PAP-1	33	100

Fig. S7. Analysis of microbial community complexity. (A) Detection of diverse phages with differing morphology, infection strategies and encoded using alternative genetic material. The search was performed using HMMs built from Pfam protein families of

caudoviral terminase (PF04466), microviral capsid protein (PF02305), inoviral viral endonuclease (PF11726), and *Levivirus* coat protein (PF01819), against the 82 metatranscriptomic samples.

With the availability of multiple ssRNA phage sequences, we were able to investigate regions within ssRNA phage genomes under evolutionary selective pressure. These hotspots are often involved in phage-host interactions. As the MP of ssRNA phages is the host receptor-binding protein, we focused our attention on this specific protein. When we investigated 15 ssRNA phage's MP protein sequences of cluster A, which varied in their BLASTp similarities (sharing 92 to 52% identity), we found three regions across these MPs with high variability which had also been highlighted in the case of a ssRNA phage AP205 (Fig. S8A; (*35*)). Through protein homology modelling with PyMOL (version 2.2.2) using PDB model 5TC1 (*77*), it was found these three regions, when folded, formed the beta-sheet domain involved in host-binding, as found in a previous study focused on Qß by Gorzelnik *et al* (Fig. S8B; (*78*)). In addition, the specific MP variable region is on the exposed virion surface (Fig. S8C), with the more conserved alpha-helical domain in contact with CP subunits and the viral ssRNA genome (Fig. S8D).

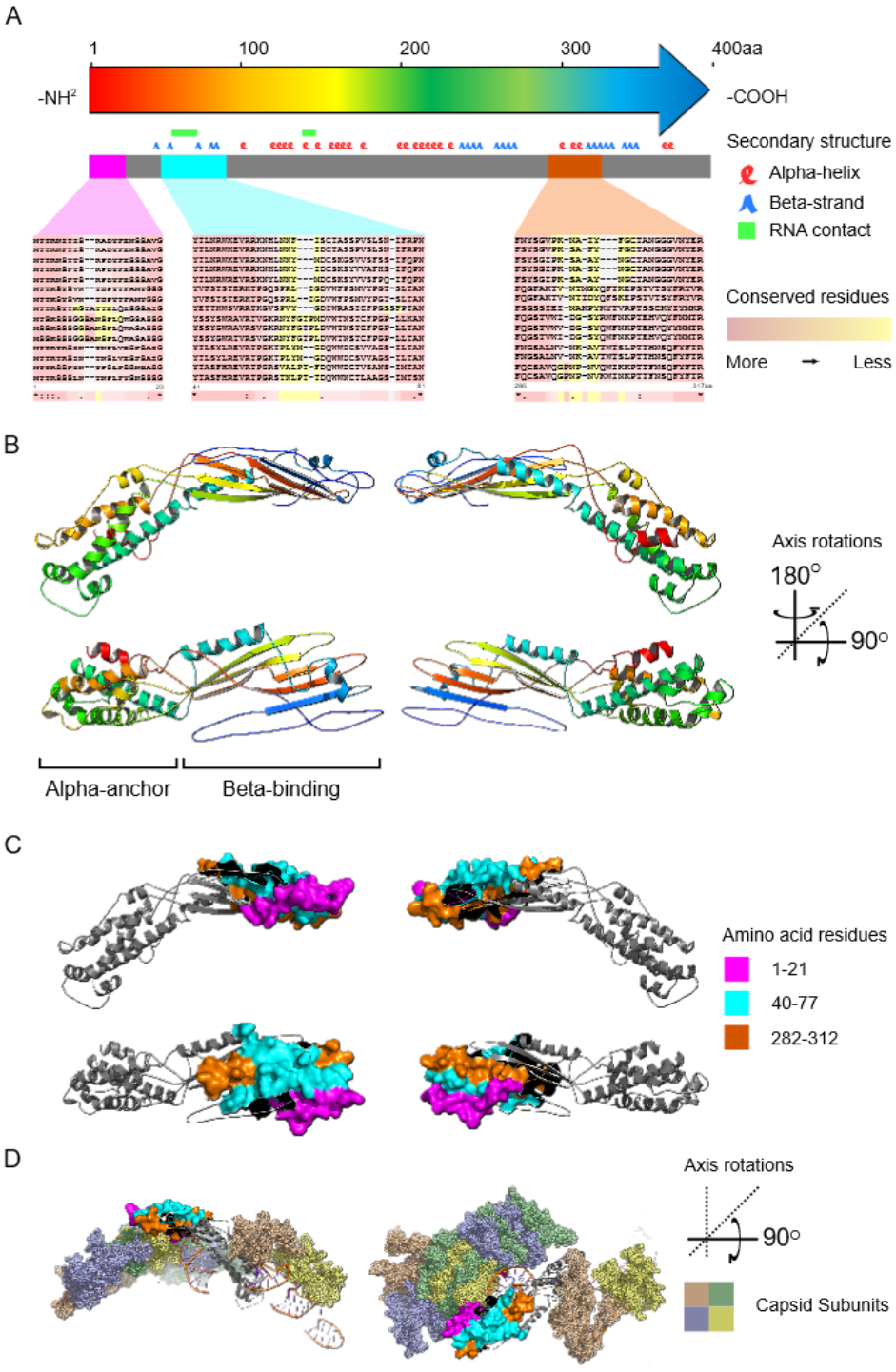


Fig. S8. Structural investigation of ssRNA phage–host interactions. (A) A cartoon representation of the MP of ssRNA phages, with amino acid (aa) length, predicted secondary structure, RNA contact sites, and variable regions highlighted. **(B)** The MP has an alpha-helical and beta-sheet domain involved in anchoring the protein within the phage virion, and

binding host bacteria, respectively. **(C)** The three variable regions of the MP are in close proximity in the folded MP protein. The exposed variable surface area of the beta-sheet domain is displayed, and uses consistent colours to panel A. The corresponding panel images displayed in B and C are identical, but the individual panels are rotated 180˚ around the y-axis, and 90˚ around the x-axis. **(D)** The MP incorporated in a partial reconstruction of an ssRNA phage virion, emphasising the variable regions with respect to CP subunits and the ssRNA genome.

Metatranscriptomic samples

This work involved publically available datasets that were generated by the work of several other lab groups, including; the Liu, Woyke, McMahon, Crump and Gin groups. While there work has been referenced, where possible, we also wish to extend our sincerest gratitude to each of these groups for making their sequences available through the JGI website.