Virus Evolution, 2026, **12(1)**, veag030
https://doi.org/10.1093/ve/veag030
Advance Access Publication 23 May 2026
Research Article

Extreme GC3 codon bias in a novel brown seaweed virus results in pseudoambigrammatic characteristics

Rob J. Dekker[1], Wim C. de Leeuw[1], Han Rauwerda[1], Marina F. van Olst[1], Wim A. Ensink[1], Selina M. van Leeuwen[1], Job Cohen[2], Klaas Timmermans[2], Timo M. Breit[1,*], Martijs J. Jonker[1]

[1]RNA Biology Research Group, Swammerdam Institute for Life Sciences, Faculty of Science, University of Amsterdam, Science Park 904, 1098 XH Amsterdam, The Netherlands
[2]NIOZ Royal Netherlands Institute for Sea Research, Department of Estuarine and Delta Systems, Korringaweg 7, 4401 NT Yerseke, The Netherlands

*Corresponding author. RNA Biology research group, Swammerdam Institute for Life Sciences, Faculty of Science, University of Amsterdam, P.O. Box 1212, 1000 BE Amsterdam, The Netherlands. E-mail: t.m.breit@uva.nl

Abstract

Although viruses are often studied in relation to disease, they can also offer valuable insights into RNA functionality. Due to their vast range of hosts and high mutation rates, viruses can quickly adapt to changing environments. It is therefore likely that viruses exploit a broad range of RNA functionalities to develop strategies for host infection and propagation. Our aim is to discover RNA viruses with survival strategies that reveal unknown RNA characteristics and functionalities. Here, we report a new, unclassified negative-sense single-stranded RNA virus from the large brown seaweed *Saccharina latissima* that showed a unique genome characteristic. The novel virus, designated SalaUV-NL1, possesses a bipartite genome. The small segment (2120 nt) encodes a putative 328-aa protein and possibly an additional protein of ≥315 aa. The large segment (7065 nt) encodes a putative RNA-dependent RNA polymerase (RdRp) of 2296 amino acids on the anti-genomic strand, but strikingly contains a similarly sized and largely overlapping open reading frame (reverse ORF; rORF) on the genomic strand that encodes a putative protein of 2317 aa. This raises the possibility that SalaUV-NL1 represents an ambigrammatic virus. However, the RdRp ORF exhibits an extreme GC3 codon bias (98%), far exceeding the typical codon constraints of ambigrammatic viruses that mostly reflect the avoidance of stop codons on the reverse strand. Therefore, the rORF is likely incidental because it is the result of the GC3 codon bias on the RdRp-coding anti-genomic strand. We therefore designate SalaUV-NL1 as pseudoambigrammatic. Several other viruses with a similar pseudoambigrammatic signature, characterized by unusually high GC3 codon bias, were identified in GenBank, most of which lack a discernible rORF. The codon bias of SalaUV-NL1 surpasses even that of its host, *S. latissima*, although other hosts of pseudoambigrammatic viruses do not display elevated GC3 levels. Comparative analysis of multiple SalaUV-NL1 variants further revealed exceptionally high nucleotide substitution rates, particularly silent A/U → G/C transitions, reaching frequencies of up to 81%. This novel virus, therefore, exemplifies a previously unrecognized survival strategy among RNA viruses, the functional implications of which remain unclear.

Keywords **pseudoambigrammatic virus, GC3 codon bias, negative-sense single-stranded RNA virus (–ssRNA), *Saccharina latissima*RNA-dependent RNA polymerase (RdRp), Negarnaviricota**

Introduction

RNA viruses are frequently studied in the context of disease, but they also offer valuable insights into RNA functionality (Scholthof and Scholthof 2023). RNA is a versatile molecule that performs an exceptional set of cellular functions (Caprara and Nilsen 2000, Cech and Steitz 2014, Cao et al. 2024, Haseltine and Patarca 2024). Despite the many RNA functions and characteristics discovered over the past decades, new findings continue to emerge. To contribute to the understanding of RNA, we chose to study RNA viruses through observational research, as they are among the smallest operational units in nature in which RNA plays a dominant role. Viruses infect all sorts of life and display a remarkable array of strategies to enter their hosts and to propagate (International Committee on Taxonomy of Viruses (ICTV). 2023 Release n.d.). This is possible because RNA viruses have a high mutation rate, which, combined with natural selection, allows for quick adaptation to changes in their host environment (Wolf et al. 2018). These adaptations utilize the full capacity of known and unknown RNA functionalities (Firth and Brierley 2012, Newburn and White 2015, Ho et al. 2021). As Scholthof and Scholthof (Scholthof and Scholthof 2023) note: 'Anything you can think of happening to RNA, there is a plant virus somewhere doing exactly that'. The challenge remains to find RNA viruses with new characteristics and strategies

Received: October 28, 2025. **Revised:** May 12, 2026. **Accepted:** May 18, 2026

that will elucidate or hypothesize yet unknown RNA functions. To maximize the likelihood of such discoveries, we focused on host species that are (i) evolutionarily distant from well-studied mammals and plants, (ii) inhabit an environment fundamentally different from terrestrial ecosystems, (iii) remain relatively underexplored in terms of associated viruses, and (iv) are readily accessible: seaweeds (macroalgae) (Lachnit et al. 2016, Waldron et al. 2018, Chiba et al. 2020, Schroeder and McKeown 2021, Van Der Loos et al. 2023, Dekker et al. 2024, Zhao et al. 2024).

A common contemporary strategy for identifying novel non-pathogenic viruses is large-scale metatranscriptomics, often performed on pooled samples from diverse sources (Waldron et al. 2018, Batson et al. 2021, Van Der Loos et al. 2023). These studies typically aim to catalogue as many viral variants or novel species as possible, most often by detecting RNA-dependent RNA polymerase (RdRp) homology (Yin and Fischer 2008, Edgar et al. 2022). In contrast, our work takes a more reductionist approach (Dekker et al. 2024), focusing on the discovery of the complete genome of new RNA viruses, including hard-to-recognize segments, using creative bioinformatics analyses to facilitate their discovery and delineate their genomic characteristics.

In this study, we analysed several species of green and brown seaweed, both with and without possible disease phenotypes. Small RNA sequencing (sRNA-seq) was chosen for virus discovery because it captures the host antiviral RNA interference (RNAi) response, with virus-derived siRNAs providing indirect evidence of viral genome presence while also reflecting the underlying viral genome sequence. Compared to direct sequencing of viral transcripts, this approach can increase sensitivity for particular virus groups and sample types, while also providing stronger support that the virus is undergoing active replication (Kreuze et al. 2009, Llave 2010, Wu et al. 2010). Using this strategy, we identified a previously unknown RNA virus with a bipartite genome, characterized by two long, extensively overlapping open reading frames (ORFs) on opposite strands of the large segment, consistent with an ambigrammatic organization (Cook et al. 2013, DeRisi et al. 2019). However, further analysis revealed an exceptionally high GC3 codon bias in the viral genome, uncovering a previously unrecognized pseudoambigrammatic characteristic.

Materials and methods

Samples

The 24 seaweed samples analysed in experiment S1 have been described previously (Dekker et al. 2024). They included eight *Ulva lactuca* (sea lettuce; Chlorophyta, green algae) samples (S1–1 to S1–8), twelve *Saccharina latissima* (sugar kelp; Phaeophyceae, brown algae) samples (S1–9 to S1–16 and S1–26 to S1–28), and five *Alaria esculenta* (dabberlocks; Phaeophyceae, brown algae) samples (S1–21 to S1–25). These were supplemented with two additional *S. latissima* samples (S2–1 and S2–2). Samples originated from the Royal Netherlands Institute for Sea Research (NIOZ; Yerseke, southern Netherlands) and from an undisclosed commercial cultivation facility in the north of the Netherlands (Table 1).

RNA isolation

RNA was isolated using a CTAB extraction buffer as described previously (Sim et al. 2013) with a minor adaptation of method 5, using a final concentration of 2.5 M LiCl. The RNA concentration was measured using a NanoDrop ND-2000 (Thermo Fisher Scientific), and the RNA quality was assessed using the 2 200 TapeStation System with Agilent RNA ScreenTapes (Agilent Technologies).

RNA-sequencing

Barcoded small RNA-seq (sRNA-seq) and RNA-seq libraries were generated using the Small RNA-seq Library Prep Kit (Lexogen) and a TruSeq Stranded Total RNA with Ribo-Zero Plant kit (Illumina) using IDT for Illumina TruSeq RNA UD v2 indexes, respectively. The size distribution of the libraries with indexed adapters was assessed using a 2200 TapeStation System with Agilent D1000 ScreenTapes (Agilent Technologies). The sRNA-seq libraries were clustered and sequenced at 1x75 bp on a NextSeq 550 System using a NextSeq 500/550 High Output Kit v2.5 (75 cycles; Illumina). The two RNA-seq libraries were clustered and sequenced at 2x150 bp on a Novaseq X Plus System (Illumina).

RACE and amplicon sequencing

RACE was performed to obtain the complete 5′ and 3′ termini of the SalaUV-NL1–1-27 large segment. A total of 300 ng of total RNA was polyadenylated using *Escherichia coli* Poly(A) Polymerase (New England Biolabs). cDNA synthesis was then carried out with SuperScript IV Reverse Transcriptase (Thermo Fisher Scientific) and the oligo(dT)-adapter primer dT-B (GCGAGCACAGAATTAATACGACTCACTATAGG$(T)_{30}$VN), following the manufacturer's instructions. The resulting cDNA served as template for RACE using the following primer combinations: SalaUV-Rv (GAGTTCGAGGCGGAGGTG) + RACE-B (GCGAGCACAGAATTAATACGACTC) to amplify the 5′ end, and SalaUV-Fw (TTCTGCGCGTAGAACTCCTC) + RACE-B to amplify the 3′ end.

To validate the (s)RNA-seq assemblies, the complete contig sequence of the large segment obtained from S1–27 was amplified in two overlapping fragments (F1 and F2). Reverse transcription was performed as described above except that the combined reverse primers (F1-Rv and F2-Rv) of both fragments were used. The resulting cDNA was then used in two PCR reactions: (i) for the 5′ half, primers F1-Fw (TCGACGCTGTAGTCCTCC) + F1-Rv (AGTTGAGGTTGCACGTGAAG); and (ii) for the 3′ half, primers F2-Fw (GCATCTCCTTCTCGTGCAAG) + F2-Rv (GCACCCGTTCCTGATGAAG). All amplicons were generated with Q5 High-Fidelity 2× Master Mix (New England Biolabs), purified with CleanNGS beads (CleanNA) at a 1.8× volume ratio, and size-validated using the Agilent 2200 TapeStation with D1000 ScreenTapes.

Purified PCR products were prepared for sequencing with the Native Barcoding Kit 24 V14 (SQK-NBD114.24; Oxford Nanopore Technologies) and run on a Flongle flow cell.

Bioinformatics: mapping and assembly

Sequencing reads were trimmed using trimmomatic v0.39 (Bolger et al. 2014) (parameters: LEADING:3; TRAILING:3; SLIDINGWINDOW:4:15; MINLEN:19). Mapping of the trimmed reads to the NCBI virus database was performed using Bowtie2 v2.4.1 (Langmead and Salzberg 2012). Contigs were assembled from sRNA-seq data using all trimmed reads as input for SPAdes v3.15.5 *De Novo* Assembler (Prjibelski et al. 2020) with parameter settings: only-assembler mode, coverage depth cutoff 10, and kmer length 17, 18, and 21. Assembly of contigs from RNA-seq data was performed with default settings. Scanning of contig sequences for potential RdRp-like proteins was performed using PalmScan (Hou et al. 2024) and LucaProt (Babaian and Edgar 2022).

Table 1 Number of sequencing reads that map to the SalaUV-NL1 virus genome in seaweed samples.

					Full virus name	*Saccharina latissima* unclassified virus NL1 isolate 1–27 segment L		*Saccharina latissima* unclassified virus NL1 isolate 1–27 segment S		
				Abbreviation		SalaUV-NL1–1-27 SL		SalaUV-NL1–1-27 SS		
Experiment set-up					**NCBI GenBank**	**PQ834486**		**PQ834489**		
Origin	**Name**	**Phenotype**	**Tissue**	**Type**	**Sample number**	**Number of reads**	**Coverage (%)**	**Number of reads**	**Coverage (%)**	**Total reads (M)**
NIOZ Dutch research facility (South NL)	*Ulva lactuca*	Healthy	Blade	Small RNA-seq	S1–1	7	1.0	8	5.0	21.7
					S1–2	2	0.5	7	4.7	24.2
					S1–3	14	3.3	4	2.7	26.0
					S1–4	8	1.3	7	4.7	19.5
					S1–5	4	1.0	4	2.6	23.9
					S1–6	9	1.6	3	3.9	21.0
					S1–7	4	0.8	5	4.4	19.3
					S1–8	2	0.6	5	1.8	19.6
	Saccharina latissima	Diseased	Blade w/o symptoms		S1–9	4619	99.2	3703	98.7	21.8
					S1–10	13	3.6	268	12.7	24.5
			Blade with bleaching spots		S1–11	6092	99.8	5333	98.8	22.8
					S1–12	7268	99.4	7342	99.8	22.3
					S1–13	19	3.9	373	7.5	22.4
					S1–14	8533	99.7	7869	98.8	23.6
					S1–15	5113	99.0	5116	99.8	24.0
					S1–16	8266	99.9	8259	98.7	21.0
		Healthy	Blade	RNA-seq	S2–1[b]	3921	99.9	5486	99.8	403.3
					S2–2[b]	937	99.8	1450	99.2	622.0
Dutch commercial facility (North NL)	*Alaria esculenta*	Healthy	Blade	Small RNA-seq	S1–21	5	1.1	8	3.8	15.0
					S1–22	5	0.8	10	3.8	23.8
					S1–23	8	1.6	6	2.5	22.8
					S1–24	11	1.6	15	4.8	40.5
					S1–25	13	3.9	15	5.1	44.5
England	*Saccharina latissima*	Unhealthy	Blade[a]		S1-26	1110	94.4	1878	98.3	18.6
Scotland		Healthy			S1–27	14509	100.0	6665	100.0	34.5
Germany			Sorus		S1–28	7	2.1	360	11.1	34.3

[a]Unhealthy appearance with dark spots [b]Note: these are not the same samples as S1–1 and S1–2

Bioinformatics: determining panhandle sequences

BLAST was used to align the last 25 nucleotides of the assembled sRNA-seq contigs to a database consisting of the first 25 nucleotides of the same set of contigs. A panhandle was called if the end of a sequence matches the start of the same contig in the reverse sense direction with an e-value <1e-8. An identical strategy was applied to identify panhandle sequences in other contigs in order to detect additional genome segments.

Bioinformatics: co-presence analysis

The trimmed reads of all samples were mapped using bowtie2 v.2.4.1 to all contigs (>200 nt) of each sample assembled by SPAdes by previously described method of sample S1–27. The read count for each contig in each sample was calculated using samtools idxstat v1.9 (Genome Research Ltd). This read count was normalized for each sample by dividing it by the total number of reads in the sample. Co-presence was estimated by calculating the Spearman correlation between the normalized presence of a contig over the sample and that of the sample S1–27 contigs of interest (NODE_1 and NODE_6). The 20 contigs with the highest correlation with the contig of interest were selected.

Bioinformatics: variant analysis

Assuming that the emergence of -ssRNA virus variants is primarily driven by mutations (Simon-Loriere and Holmes 2011, Peck and Lauring 2018), we analysed sequence variation by aligning the (s)RNA-seq reads from each SalaUV-NL1-positive sample containing >900 reads to the SalaUV-NL1–1-27 reference genome using Bowtie2 v2.4.1, followed by variant detection with FreeBayes v1.3.229 (Garrison and Marth 2012). Variants were further analysed in genomic context by incorporating detected variants with an allelic frequency >50% into the reference sequence with bcftools v1.17. This process of mapping, variant detection, and reference updating was repeated iteratively until no further changes were observed, to obtain sample-specific variant genome sequences. The variant genome sequences were compared using the edit distance (Levenshtein 1966) and subjected to ORF detection to analyse functional consequences of the mutations.

Bioinformatics: finding (pseudo)ambigrammatic viruses in GenBank

On 18 October 2024, a selection of virus sequences, with a non-human host, were downloaded from NCBI Virus (https://www.ncbi.nlm.nih.gov/labs/virus), which have a length ranging between 2500 and 29 600 nucleotides. To ensure the exclusion of artificial or modified sequences, any entries containing the terms 'Modified Microbial Nucleic Acid', 'vaccine', 'Patent', 'recombinant', or 'Expressions System' were removed. Subsequently, for each non-phage remaining sequence, the longest non-stop-ORF in an arbitrary frame without a stop codon was identified in both forward and reverse frames using R v4.3.2 (R Core Team 2023). The overlap between the longest forward and reverse ORFs (rORFs) was also calculated. Finally, for these ORFs, the GC content at the third codon position (GC3) and the overall GC content were determined.

Potential ambigrammatic virus sequences were identified based on two criteria: an overlap of >75 nucleotides between the ORFs encoded on the genomic and anti-genomic strands, and a total GC content of <70%. Potential pseudoambigrammatic virus sequences were identified based on two criteria: an anti-genomic strand ORF GC3 content over 90%, and a total GC content below 65%.

Bioinformatics: calculating the GC3 of host organisms

Protein-coding genes of the target organism were retrieved from NCBI using e-utils v15.6 with the query 'txid<tax_id>(Organism:exp) AND biomol_mrna(PROP)', where <tax_id> was replaced with the organism's NCBI taxonomy ID. The longest ORF for each sequence was identified using the ORFik R package v1.22.2 (Tjeldnes et al. 2021) in R v4.3.2. Sequences were then split into codons, and the last nucleotide of each codon was tabulated using standard R functions.

Bioinformatics: *S. latissima* transcriptome assembly and mapping

Paired-end RNA sequencing data was downloaded from the SRA (run ERR4387699). Quality control was performed with FastQC (Andrews 2010). Reads were trimmed with Trimmomatic (Bolger, Lohse, and Usadel 2014). High-quality reads were assembled with rnaSPAdes (Bushmanova et al. 2019). Using the ORFik R package (Tjeldnes et al. 2021) the longest ORF was determined in each contig. The resulting *S. latissima* transcriptome contigs can be found in Supplemental Document SD2.

Expression levels were calculated by mapping and counting the reads of sample S2–1 to the ORFs. The GC3 percentages and codon usages were calculated using standard R functions.

Results and discussion

A new unclassified virus species in *S. latissima*

To investigate the presence of unknown viruses with potential novel RNA-based survival strategies in edible brown and green algae, we analysed the virus-derived siRNA presence in eight *U. lactuca*, eleven *S. latissima*, and five *A. esculenta* samples, with and without phenotypic abnormalities, from an academic and a commercial research facility in the South and the North of the Netherlands, respectively (Table 1). Previously, we reported on the discovery of four new viruses in these *S. latissima* and *A. esculenta* samples (Dekker et al. 2024). Small RNA sequencing yielded ~641 million reads, resulting in an average of 49 million reads per sample. Additionally, we analysed two *S. latissima* samples with RNA-seq, which yielded on average 513 million reads per sample (Table 1; Supplemental Fig. SF1).

De novo assembly of the sRNA-seq dataset (Supplementary Document SD1) yielded a contig of particular interest in sample S1–27, of which the continuity and accuracy were verified using amplicon sequencing. Extending this contig using nanopore sequencing of 5′ and 3' RACE amplicons resulted in the reconstruction of a full-length viral genome of 7065 nucleotides. This sequence contains two oppositely oriented, extensively overlapping ORFs: a forward ORF (fORF) on the anti-genomic/positive strand encoding a 2296 amino-acid protein, and a rORF on the genomic/negative strand, encoding a putative 2317 amino-acid protein (Fig. 1a). This genomic organization is characteristic of ambigrammatic viruses (Cook et al. 2013, DeRisi et al. 2019). RdRp analysis (Edgar et al. 2022) (https://serratus.io) revealed an RdRp domain in the protein encoded by the fORF, indicated by the presence of a conserved RdRp palmprint (palmID; Fig. 1d). This analysis further suggested that the virus may belong to

the phylum *Negarnaviricota* (realm *Riboviria*, kingdom *Orthornavirae*) (Kuhn et al. 2023), a group of viruses characterized by negative-sense single-stranded RNA (-ssRNA) genomes. Additional support was provided by a distinctive feature of the genome: reverse-complement similarity within the first 54 nt, forming a panhandle RNA structure with a characteristic protruding cytosine at position 11 of the 5′ end (Fig. 1c) (Ren et al. 2020, Malet et al. 2023). Such panhandle structures are often present in Negarnaviricota viruses, for instance, in bunyaviruses. Furthermore, the size distribution of the virus-associated sRNA-sequencing reads shows a peak at 21 nucleotides, indicating the presence of siRNAs (Fig. 1h). These 21-mer reads are distributed across the entire virus sequence (Fig. 1e). The presumable siRNAs also shown a bias for their 5′ first nucleotide, being uracil (84) and 3′ final nucleotide being cytosine (42) ies (Waldron et al. 2018, Kim 2008).

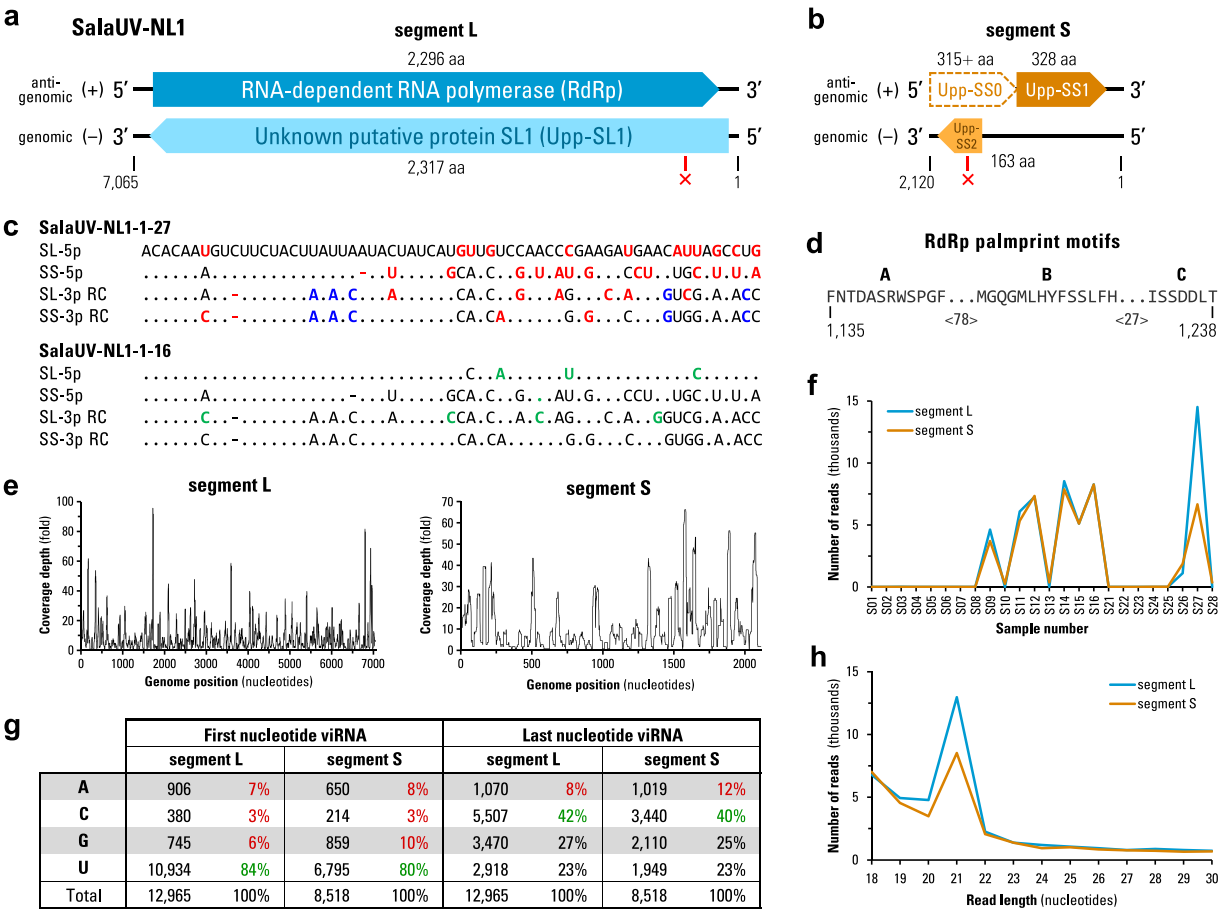
BLAST analysis against GenBank revealed that the genomic RNA and predicted protein sequences were most similar to 'Barns Ness serrated wrack bunya/phlebo-like virus 1 (BNswBPV1), isolate C7, segment L' (MF190051.1) (Waldron et al. 2018). The RNA genome shared 68.2% nucleotide sequence identity with 34% query coverage (E-value = 0.0), whereas the encoded RdRp shared 34.8% amino acid sequence identity with 94% query coverage (E-value = 0.0; Supplemental Table ST1). The rORF-encoded protein on the genomic strand produced only one notable BLASTp hit in the GenBank non-redundant protein database, to an ambigrammatic virus protein annotated as an RdRp (QJW70364.1; 36.3% identity, 33% query coverage, E-value = 6×10^{-22}). However, further inspection indicated that this hit likely corresponded to the rORF-encoded protein rather than the RdRp, because the fORF-derived protein from the same accession showed similarity to >100 known RdRp sequences, with up to 37.3% identity over 98% query coverage. Analysis of the amino acid composition showed that the putative protein encoded by the rORF is severely biased, with ten of the 20 amino acid types occurring at frequencies below 1% (Supplemental Table ST2). When the sRNA-seq and RNA-seq reads from all samples were mapped to this newly identified virus sequence, >900 reads corresponding to the virus were detected in ten samples from both facilities (Table 1 and Fig. 1f). Given that BNswBPV1, despite its relative dissimilarity, is classified merely as bunya/phlebo-like, and that the new virus was detected exclusively in *S. latissima* samples (Table 1), we assigned the name *S. latissima* unclassified virus Netherlands 1 (SalaUV-NL1) (King et al. 2012).

Discovery of a second, small RNA segment in SalaUV-NL1

Another characteristic of *Negarnaviricota* viruses is their multisegmented RNA genome (Kuhn et al. 2023). Given the relatively large size of the new virus sequence (7 065 nucleotides), as well as the presence of an RdRp domain, we assumed it to be a large segment (SalaUV-NL1 SL). Next, we set out to find possible other, smaller genomic RNA segments of the new SalaUV-NL1 virus. In the absence of a reference virus sequence, we used the 5′ and 3′ panhandle motifs to search for matching contig ends in sample S27, which contained the highest number of reads for the new virus. This yielded an RNA contig of 2120 nucleotides with almost identical panhandle sequences for the first 15 nucleotides at both ends (Fig. 1c). We applied an alternative approach by mapping all 21-mer sRNA-seq reads of all samples to all RNA contigs of sample S1–27 (Supplemental Document SD1) and identifying contigs that showed a read abundance distribution over the samples that was almost identical to SalaUV-NL1 SL (Table 1, Fig. 1f, and Supplemental Table ST3). Interestingly, the contig with the best read-distribution correlation was the same one identified using the panhandle-sequence approach. Furthermore, sRNA-seq reads associated with this contig show the same 21-mer peak in their size distribution (Fig. 1h) and the same prevalence for the first and last nucleotide as SalaUV-NL1 SL (Fig. 1g). Hence, we concluded that, considering its size of 2 120 nucleotides (Fig. 1b), this contig likely represents the small segment of the new virus (SalaUV-NL1 SS). As we were unable to identify any other contigs or sRNA-seq reads with obvious similarity to the four observed panhandle sequences (Fig. 1c) of the new virus and found no contigs with conclusive co-presence across these samples, we assumed that the virus has no additional genome segments. Given that for some viruses, panhandle sequences have been reported for the large and medium genomic segments but not for the small segment (Ren et al. 2020, Malet et al. 2023), we cannot exclude the possibility that additional segments of SalaUV-NL1 exist. Also, we were unable to identify a coat protein by similarity searches. SalaUV-NL1 SS contained two ORFs: one fORF on the anti-genomic strand encoding a putative protein of 328 amino acids (Upp-SS1) and an rORF on the genomic strand encoding a putative protein of 163 amino acids (Upp-SS2; Fig. 1b). Upp-SS2 almost completely lacks six amino acids (Cys, Gln, Gly, His, Ile, and Tyr; Supplemental Table ST2). Both proteins showed no convincing similarity to any GenBank virus protein sequence.

SalaUV-NL1 variants

High viral read abundance (>900 reads) for the newly discovered virus SalaUV-NL1 was observed in ten samples, comprising eight from the NIOZ research facility and two from the commercial North NL facility (Table 1). The read coverage for both SalaUV-NL1 segments from the eight samples of the NIOZ facility averaged 99.6% (Supplemental Table ST4). To assess potential nucleotide variation across all ten samples, we aligned the (s)RNA-seq reads from each sample separately against the SalaUV-NL1–1-27 reference genome derived from sample S1–27. The nucleotide variations identified in each individual sample were then used to revise the original reference sequence independently for that sample, resulting in the identification of eight additional SalaUV-NL1 variants (Supplemental Document SD3). Comparison of NIOZ facility sample isolates 1–16 and 2–1 showed 99.8% nucleotide identity across the aligned RNA genome positions and 99.9% amino acid identity across the aligned protein residue positions. In contrast, comparison of the NIOZ facility isolates with the North NL facility isolate 1–27 showed lower sequence identity: 97.2% nucleotide identity across the aligned RNA genome positions and 96.0% amino acid identity across the aligned protein residue positions (Supplemental Table ST4). These differences are clearly illustrated by the edit distance matrixes of the SalaUV-NL1 variant sequences (Supplemental Fig. SF2). Notably, a C-to-U substitution (100% alternative allele frequency) at position 526 on the genomic strand of the large segment of the SalaUV-NL1 1–16 isolate creates a premature stop codon in the rORF, shortening the predicted protein from 2317 to 165 amino acids, without altering the length of the RdRp fORF (Fig. 1a). This suggests that the full-length Upp-SL1 protein encoded by the rORF of isolates 1–27 and 2–1 is possibly non-essential for the virus (DeRisi et al. 2019). Interestingly, we also discovered a premature stop codon at position 1760 (100% alternative allele frequency) in the rORF coding for Upp-SS2 on the small segment of isolates 1–16 and 2–1 (Fig. 1b), truncating this small putative protein from 163 to 64 aa. These results show that



g

	First nucleotide viRNA				Last nucleotide viRNA			
	segment L		segment S		segment L		segment S	
A	906	7%	650	8%	1,070	8%	1,019	12%
C	380	3%	214	3%	5,507	42%	3,440	40%
G	745	6%	859	10%	3,470	27%	2,110	25%
U	10,934	84%	6,795	80%	2,918	23%	1,949	23%
Total	12,965	100%	8,518	100%	12,965	100%	8,518	100%

Figure 1 Schematic representation of *S. latissima* unclassified virus NL1 (SalaUV-NL1) and characteristics of related siRNAs. a) Schematic representation of the SalaUV-NL1 isolate 1–27 large segment. RdRp encoded by the fORF on the anti-genomic strand and the unknown putative protein SL1 (Upp-SL1) encoded by the rORF on the genomic strand are shown. The stop codon found in SalaUV-NL1 isolate 1–16 is marked with a 'x'. b) Schematic representation of the SalaUV isolate 1–27 small segment. The unknown putative protein SS1 (Upp-SS1) encoded by the fORF on the anti-genomic strand and the unknown putative protein SS2 (Upp-SS2) encoded by the rORF on the genomic strand are shown. The Upp-SS0 fORF on the anti-genomic strand (dashed box) lacks a start codon but has a stop codon (UAG) immediately preceding the Upp-SS1 fORF start codon. Using cap-snatching, thereby including a start codon from a nascent host RNA, translation of the Upp-SS0 fORF might be achieved, resulting in a 315+ amino acid protein. The stop codon found in SalaUV isolates 1–16 and 2–1 is marked with a 'x'. c) Similarities between the 63 nt 5′ and 3′ terminus (panhandle) sequence motifs of the large (SL) and small (SS) RNA segments of SalaUV isolates 1–27 and 1–16. To facilitate interpretation of the sequence complementarity, the RC of the 3′ terminal sequences is presented. Nucleotides that differ from the preferred nucleotide are highlighted in red or blue if they are similar in the 3p sequences. Nucleotides that differ between SalaUV isolates 1–27 and 1–16 are highlighted in green. d) A, B, and C motifs of the RdRp palmprint are shown, with the nucleotide distances between motifs indicated in angle brackets (<#>). e) Coverage plots of the sRNA-seq reads mapping to SalaUV for the large and small segment. f) The total numbers of sRNA-seq reads that are related to the SalaUV large (blue) and small segment (orange) are shown for each sample. g) Summary of nucleotide occurrence at the ends of SalaUV-related sRNA-seq reads (viRNA) from all samples. h) Read size distribution plot for SalaUV-related sRNA-seq reads for the large (blue) and small (orange) segments.

the full-length rORF-derived proteins, as annotated in variant 1–27, are not strictly conserved as single continuous proteins. This suggests that the full-length rORF protein products do not seem essential; however, we cannot exclude the possibility that functional proteins can still arise through translational stop codon readthrough or from alternative translation initiation sites.

The SalaUV-NL1 host

The genomic sequence of the virus SalaUV-NL1 was obtained through metagenomics, which complicates the identification of its true host. Although *S. latissima* is the most likely host, it remains possible that the virus instead infects another organism associated with the seaweed, such as a fungus. To resolve such uncertainty, metagenomic datasets are often examined for genomic signatures of other organisms. Previous studies (Waldron et al. 2018, Batson et al. 2021) have proposed that detecting an additional pathogen can point to that organism as the possible source of the virus. In our case, however, the samples were analysed individually rather than as pooled material, making a co-occurrence analysis across the sRNA-seq dataset a more appropriate approach. Accordingly, we searched for contigs that exhibited the same presence pattern across the ten samples in which SalaUV-NL1 was detected at a level of at least 900 reads. As expected, the SalaUV-NL1 contigs were the highest co-presence contigs. Of the remaining top 10 co-presence contigs per SalaUV-NL1 segment, 16 showed similarity to seaweed sequences in GenBank (Supplemental Table ST3). This outcome strongly suggests that *S. latissima* is the host of the SalaUV-NL1 virus. Remarkably, the regions that were consistently

found were two adjacent *Saccharina japonica* U-linked SDR genomic sequences (Luthringer et al. 2015, Du et al. 2022) suggesting that this region may be involved in seaweed virus infection.

Ambigrammatic viruses

In ambigrammatic viruses, the RdRp fORF avoids the triplets UUA, CUA, and UCA, which are the reverse complements (RCs) of the three canonical stop codons (Cook et al. 2013, DeRisi et al. 2019). This constraint prevents the formation of stop codons in the phase-0 (in-frame) rORF on the reverse strand. A start codon near the 5′ end of the reverse strand creates an uninterrupted rORF that is comparable in length to the RdRp fORF, making the virus ambigrammatic. There have been several reports of ambigrammatic viruses (Cook et al. 2013, DeRisi et al. 2019, Dinan et al. 2020, Batson et al. 2021, Abbo et al. 2023, Zhang et al. 2023, Cheng et al. 2024). To get an overview of known ambigrammatic viruses, we analysed all available GenBank virus sequences, excluding those annotated 'human', for largely overlapping ORFs on both strands. About 114 out of 221 000 virus sequences showed an overlap between the fORF and the rORF of >75 nucleotides, while having a GC content <70% (Supplemental Table ST5), many (40) being Culex narnavirus (Göertz et al. 2019). For these ambigrammatic viruses, keeping the rORF intact appears to be important (Dudas et al. 2021, Retallack et al. 2021, Wilkinson et al. 2021). This finding calls the ambigrammatic nature of SalaUV NL1 into question, because the virus appears not to depend on the putative protein encoded by the rORF, as evidenced by the NIOZ facility isolate carrying a severely truncating stop codon (Fig. 1a) and further supported by the near-complete absence of approximately half of the amino acid types (Supplemental Table ST2). The closest related virus sequence (BNswBPV) likewise lacks a definitive protein-coding rORF, but in this case, the absence results from a missing start codon in an otherwise uninterrupted ORF. We therefore investigated the origin of the ostensible ambigrammatic nature of SalaUV-NL1.

Pseudoambigrammatic characteristics of SalaUV-NL1

Codon usage analysis revealed not only the absence of reverse-complement stop codons from the SalaUV-NL1 SL RdRp fORF, but also a striking bias towards G and C nucleotides at the third codon position (1% A, 1% U, 47% C, 51% G; Table 2 and Supplemental Table ST6). Hence, the fORF (RdRp) of this virus has a third codon position GC content (GC3) of over 98%, which is also true for the fORF of BNswBPV1 (Table 2). The most extreme GenBank example of a virus sequence with such a high GC3 is 'Trichoderma gamsii negative-stranded virus 1' (Table 2), with only one codon ending with a U and two with an A, resulting in an fORF of 2 268 codons and a non-stop rORF of 2 300 codons with no start codon at all. Thus, the observed absence of stop codons in the phase-0 reading frame on the genomic strand of SalaUV-NL1 SL, appears to merely be a byproduct of the near-complete avoidance of A and U nucleotides at the third codon positions of the RdRp fORF (Supplemental Table ST6), rather than an adaptation to generate a functional protein from the rORF. We therefore propose to characterize our SalaUV-NL1 virus as pseudoambigrammatic.

Underscoring the biological relevance of the unusually high GC3 content in the fORF, the terminal 5′ and 3′ regions of the SalaUV-NL1 large segment exhibit substantially lower GC3 values of 21% and 25%, respectively (Supplemental Fig. SF3). Moreover, the small segment of SalaUV-NL1 also exhibits a high GC3 content of 98.8% (Table 2), as well as the low GC3 percentages (13% and 20%) at its 5′ and 3′ ends, respectively. This is remarkable since there appears to be only a relatively small fORF (Upp-SS1) present in this segment (Fig. 1b). However, the small segment also contains a 315-codon fORF that lacks a canonical start and spans from the beginning of the anti-genomic strand to the stop codon immediately upstream of the start codon of the fORF encoding Upp-SS1. Given that SalaUV-NL1 is a negative-strand virus, it is possible that this segment uses cap-snatching, thereby including a start codon from a nascent host RNA, to achieve translation of the Upp-SS0 fORF (Fig. 1b) (Ho et al. 2020), which would explain the presence of the high GC3 percentage in this virus segment. Nonetheless, we observed that the high-GC3 region in both viral segments continued past the boundaries of the proposed fORFs. This is intriguing, as it raises questions about whether the GC3 pattern is truly a result of just translation processes.

Other pseudoambigrammatic viruses

Pseudoambigrammatic viruses can be readily detected in databases, since they are characterized by a distinctly elevated GC3 percentage despite having an average overall GC content. Importantly, elevated overall GC content may also account for the absence of reverse-complement stop codons, given that each contains at least one adenine and one uracil. We therefore queried the GenBank nucleotide database for viral sequences characterized by elevated GC3 content (>90%) in combination with near-average overall GC content (<65%). The database search identified 20 viral sequences of this type, 18 of which belong to Negarnaviricota, a phylum of negative-sense single-stranded viruses typically characterized by segmented genomes (Table 2 and Supplemental Table ST7). The elevated GC3 percentage is likely to produce a codon bias, which can be quantified using the *effective number of codons* (ENC) (Wright 1990). On average, the ORFs of pseudoambigrammatic viruses displayed an ENC of 27 (Supplemental Table ST7), whereas true ambigrammatic sequences showed a higher average ENC of 49 (Supplemental Table ST5), indicating that the former exhibit a much stronger codon bias, with limited codon diversity, while the latter use a broader and more balanced set of synonymous codons. A strong correlation was observed between the codon usage of the SalaUV-NL1 SL RdRp ORF and that of RdRp ORFs from other pseudoambigrammatic viruses, in agreement with their pronounced GC3 bias (Supplemental Table ST8). Among the 21 viral sequences analysed, only a single A-ending codon, AUA, was substantially represented (>10 occurrences) in 10 sequences (48%). Interestingly, the GC3 bias extended to the stop codons, with 16 of 18 (89%) being UAG. Furthermore, protein sequence comparison showed that, apart from a few exceptions, the pseudoambigrammatic virus proteins share little sequence identity (Supplemental Table ST1). By contrast, the amino acid distribution of the SalaUV-NL1 SL RdRp protein showed a strong correlation with that of RdRp proteins from the other pseudoambigrammatic viruses (Supplemental Table ST9). The lowest correlation observed among pseudoambigrammatic proteins was 79% (QXN75452.1), which is still markedly higher than the 43% observed for the RdRp protein of a true ambigrammatic virus (QBR53296.1).

Origin of codon usage bias of SalaUV-NL1

High GC3 codon usage bias is known to occur in highly expressed genes (Quax et al. 2015, Parvathy et al. 2022, Wint et al. 2022).

Table 2 Summary of GenBank pseudoambigrammatic virus sequences with extreme GC3 codon bias.

Description[a]	Genome type	Virus type	Ambigrammatic[b]	Size (nt/aa)			ORF		Host	
				Seq	ORF	Protein	GC3%	GC%	Species	GC3%
Saccharina latissima unclassified virus NL1, Segment L[c]	(−) RNA	Unclassified	+/− ambi	7065	6954	2296	98.1	58	*Saccharina latissima*	63.1
Saccharina latissima unclassified virus NL1, Segment S[c]	(−) RNA	Unclassified	-	2129	987	328	98.8	61	*Saccharina latissima*	63.1
Barns Ness serrated wrack bunya/phlebo-like virus 1	RNA	Narnaviridae	-	6977	6774	2270	99.6	60	*Fucus serratus*	58.3
Botrytis cinerea negative-stranded RNA virus 6	(−) RNA	Negarnaviricota	-	7071	6777	2258	99.9	55	*Botrytis cinerea*	46.1
Erysiphe necator associated neg-stranded RNA virus 11	(−) RNA	Negarnaviricota	-	6883	6823	2265	98.4	54	*Erysiphe necator*	38.1
Erysiphe necator associated neg-stranded RNA virus 13	(−) RNA	Negarnaviricota	-	6961	6837	2278	99.6	54	*Erysiphe necator*	38.1
Erysiphe necator associated neg-stranded RNA virus 14	(−) RNA	Negarnaviricota	-	6866	6759	2252	99.1	54	*Erysiphe necator*	38.1
Erysiphe necator associated neg-stranded RNA virus 15	(−) RNA	Negarnaviricota	-	7116	6861	2286	92.9	53	*Erysiphe necator*	38.1
Erysiphe necator associated neg-stranded RNA virus 16	(−) RNA	Negarnaviricota	ambi	6957	6807	2268	99.8	56	*Erysiphe necator*	38.1
Freshwater macrophyte assoc. bunya-like virus 1[d]	RNA	Bunyavirales	-	6786	6615	2204	99.6	54	Various	-
Heterobasidion irregulare neg-stranded RNA virus 1[c]	(−) RNA	Negarnaviricota	-	7043	6873	2290	98.6	55	*Heterobasidion irregulare*	62.3
MAG: Baoding Narna tick virus 1	RNA	Narnaviridae	-	2700	2649	900	97.1	54	*Haemaphysalis longicornis*	64.1
MAG: Grapevine-associated mycobunya-like virus 1	RNA	Bunyavirales	-	2697	2643	880	99.7	53	*Vitis vinifera*	46.7
MAG: Grapevine-associated mycobunya-like virus 3[d]	RNA	Bunyavirales	-	2582	2542	847	99.8	55	*Vitis vinifera*	46.7
MAG: Grapevine-associated mycobunya-like virus 4	RNA	Bunyavirales	-	5663	5529	1842	99.1	53	*Vitis vinifera*	46.7
MAG: *Trichoderma gamsii* negative-stranded virus 1	(−) RNA	Mononegavirales	(ambi)	6902	6804	2267	99.9	56	*Trichoderma gamsii*	58.0
MAG: *Trichoderma harzianum* negative-stranded virus 1	(−) RNA	Mononegavirales	-	7061	6810	2269	99.6	54	*Trichoderma harzianum*	57.9
Plasmopara viticola lesion ass. Myco-bunyavirales-like virus 2	RNA	Bunyavirales	-	6881	6798	2265	98.2	54	*Plasmopara viticola*	56.4
Plasmopara viticola lesion ass. Myco-bunyavirales-like virus	RNA	Bunyavirales	-	5020	4917	1639	99.9	54	*Plasmopara viticola*	56.4
Rhizoctonia solani bunya/phlebo-like virus 1	RNA	Bunyavirales	(ambi)	7804	7543	2513	97.4	55	*Rhophitulus solani*	53.2
Schistocephalus solidus bunya-like virus	RNA	Bunyavirales	-	6533	6465	2154	93.0	56	*Schistocephalus solidus*	-
Suillus luteus-associated bunya-like virus 5	RNA	Bunyavirales	-	6952	6804	2267	99.9	56	*Suillus luteus*	54.3
Culex narnavirus 1	RNA	Narnaviridae	ambi	3324	3090	1029	66.1	63	*Culex tarsalis*	62.2

[a]Corresponding GenBank accession numbers are in Supplemental Table ST7 [b](ambi) = small ambigrammatic ORF [c]Sequence contains 5′ and 3′ panhandle motifs [d]Partial sequence

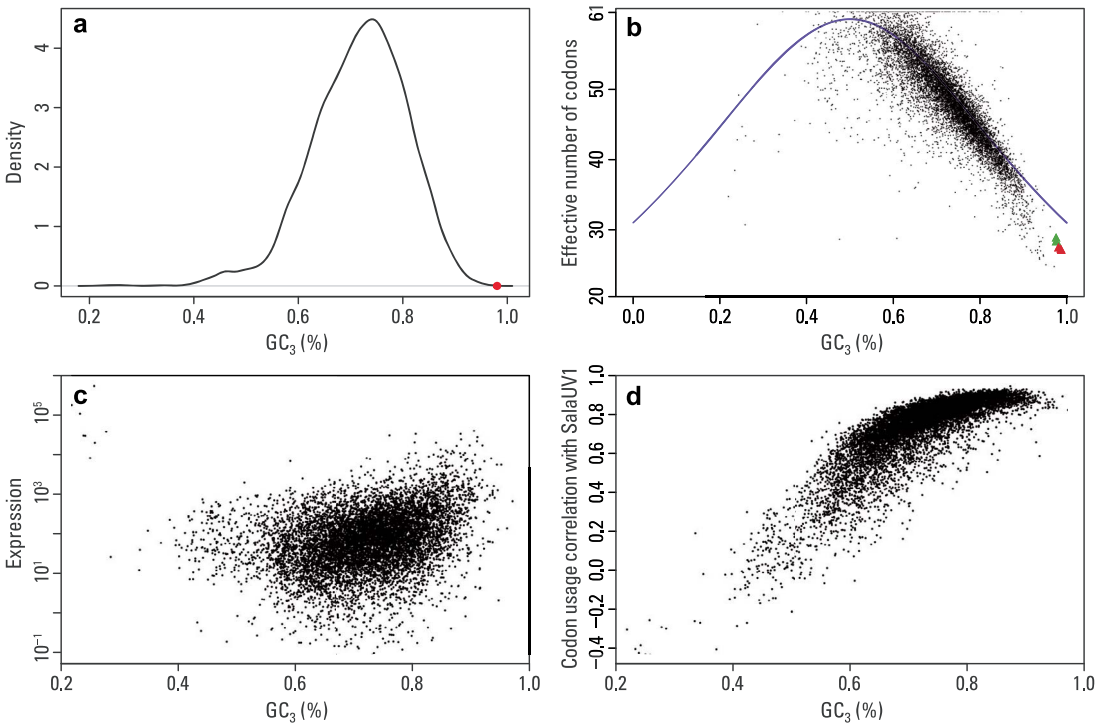


Figure 2 Relationships between GC3 content, codon-usage, and expression of *S. latissima* transcript ORFs. For each transcript in the *de novo* assembled *S. latissima* transcriptome, the longest ORF was identified and its GC3 content computed. a) Density distribution of the GC3 content of all *S. latissima* ORFs. The dot marks the GC3 content of the SalaUV large segment ORF (RdRp), indicating its position relative to the transcriptome-wide distribution. b) Relationship between *S. latissima* ORF GC3 content and the ENC for each ORF (black points). ENC is a single-number summary of synonymous codon-usage bias (20 = maximal bias, only one codon used per amino acid; 61 = no bias, all synonymous codons are used equally). The blue curve represents the expected neutrality relationship in the absence of codon usage bias, i.e. when codon usage is determined by GC3 composition only. Red and green triangles indicate the SalaUV large segment and small segment ORFs, respectively. c) *Saccharina latissima* ORF GC3 content *versus* ORF expression in sample S2–1; expression is the k-mer coverage reported by the rnaSPAdes transcriptome assembler. d) *Saccharina latissima* ORF GC3 content *versus* the correlation between that ORF's codon-usage profile and that of the SalaUV large segment RdRp ORF.

However, most of the *S. latissima* transcriptome shows a relatively high GC3 codon bias (Luthringer et al. 2015) (average 63.1%; Fig. 2a/b and Table 2) that does not seem to correlate with gene expression (Fig. 2c). Even so, the GC3 content in SalaUV-NL1 (98%) is higher than that of almost all *S. Latissima* transcripts (Fig. 2a). Additionally, there is an obvious correlation between the GC3 percentage and the GC3 codon bias of SalaUV-NL1 (Fig. 2d). Together, these findings suggest that the GC3 codon bias observed in SalaUV-NL1 may originate from the virus mimicking the codon usage of the majority of genes of its host organism. To investigate whether this holds true for other pseudoambigrammatic viruses, we examined the GC3 codon bias in the host organisms from other pseudoambigrammatic viruses (Supplemental Table ST7). The GC3 content in these organisms ranged from 38 to 64% (Table 2), indicating no relevant relationship between the GC3 codon bias of pseudoambigrammatic viruses and the GC3 codon usage of their host organisms. Nonetheless, the position of the SalaUV-NL1 ENC values well below the neutrality line in the GC3 content *versus* ENC plot (Fig. 2b) demonstrates that its GC3 codon usage is non-random and shaped by selection pressure (Zhou et al. 2015).

Characterizing the sequence differences between SalaUV-NL1 variants

In this study, we identified three distinct SalaUV-NL1 variants (1–16, 1–27, and 2–1). Given the pronounced GC3 bias, we investigated whether sequence variation among these viruses was associated with specific codon base positions within the ORFs. Comparative analysis identified 164 nucleotide differences within the L segment fORF (RdRp) and 93 across the entire segment S, including 39 within its fORF (Upp-SS1; Supplemental Table ST10). Variation at the RNA sequence level may result in protein sequence changes, such as missense mutations, or in synonymous (silent) mutations that do not alter the protein sequence. Our analysis revealed a pronounced difference among the SalaUV-NL1 variants in the frequency of A/U-to-G/C substitutions at the third codon position, where such changes give rise to synonymous mutations (Table 3). While most other differences across the ORFs of the large and small segments involved only 0.04% to 3.5% of nucleotides, the third-base A/U-to-G/C substitutions between variants 1–27 and 1–16 ranged from 41% to 80%. A similar pattern was observed between variants 2–1 and 1–16, with substitutions ranging from 38% to 50% (Table 3). These findings suggest a selective pressure favouring the replacement of the remaining A/U nucleotides at third codon positions with G/C, consistent with the strong GC3 bias in which nearly all other codons already terminate in G or C. This is reflected in the greater number of unused codons in the L segment of variant 1–16 (22) compared with variant 1–27 (14) (Supplemental Table ST6). An exception was noted: in the L segment of variant 1–27, 24 codons end with A as the third base. Of these, all 8 non-AUA codons were replaced by G/C-ending codons, while only 2 of the 16 AUA (Ile) codons changed to AUC (Ile), compared with the L segment of variant 1–16. AUA is noteworthy because it encodes isoleucine, the only amino acid

Table 3 Sequence changes between SalaUV-NL1 variants related to codon position[a].

Variant	Mutation		SalaUV-NL1 SL 1–27			SalaUV-NL1 SS 1–27 minus ends			SalaUV-NL1 Upp-SS1 1–27		
			1st base	2nd base	3rd base	1st base	2nd base	3rd base	1st base	2nd base	3rd base
1–16	SILENT	A/U > G/C	1.2% (15)	–	62.2% (28)	0.3% (1)	–	41.2% (7)	0.6% (1)	–	80.0% (4)
		G/C > A/U	1.3% (13)	–	0.6% (13)	0.3% (1)	–	0.4% (3)	0.6% (1)	–	0.6% (2)
		A < > U	–	–	–	–	–	–	–	–	–
		G < > C	–	–	1.2% (26)	–	–	1.2% (8)	–	–	0.6% (2)
	MISSENSE	A/U > G/C	0.5% (6)	0.4% (6)	2,2% (1)	2.8% (10)	3.2% (14)	3.5% (1)	3.5% (6)	2.5% (5)	–
		G/C > A/U	1.3% (13)	0.9% (7)	–	3.4% (11)	3.1% (8)	–	1.9% (3)	2.4% (3)	–
		A < > U	–	0.2% (2)	–	0.3% (1)	–	–	–	–	–
		G < > C	0.1% (1)	0.1% (1)	0.4% (10)	–	0.4% (1)	0.7% (5)	–	–	0.9% (3)
	Mutation		SalaUV-NL1 SL 2–1			SalaUV-NL1 SS 2–1 minus ends			SalaUV-NL1 Upp-SS1 2–1		
			1st base	2nd base	3rd base	1st base	2nd base	3rd base	1st base	2nd base	3rd base
	SILENT	A/U > G/C	0.4% (4)	–	38.7% (12)	–	–	41.2% (7)	–	–	50.0% (2)
		G/C > A/U	0.4% (5)	–	0.4% (10)	0.3% (1)	–	0.4% (3)	0.6% (1)	–	0.3% (1)
		A < > U	–	–	–	–	–	–	–	–	–
		G < > C	–	–	–	–	–	–	–	–	–
	MISSENSE	A/U > G/C	0.3% (3)	0.1% (2)	–	0.8% (3)	0.7% (3)	5.9% (1)	–	–	–
		G/C > A/U	0.3% (4)	0.2% (2)	–	1.2% (4)	0.8% (2)	–	–	–	–
		A < > U	–	–	–	0.3% (1)	–	–	–	–	–
		G < > C	0.1% (1)	–	0.04% (1)	–	0.4% (1)	0.3% (2)	–	–	–
Variant	All bases		SalaUV-NL1 SL			SalaUV-NL1 SS minus ends			SalaUV-NL1 Upp-SS1		
			1st base	2nd base	3rd base	1st base	2nd base	3rd base	1st base	2nd base	3rd base
1–27	A/U		1 296	1 495	45	360	432	17	172	203	5
	G/C		1 001	802	2 252	328	256	671	157	126	324
2–1	A/U		1 299	1 496	31	359	427	17	168	201	4
	G/C		998	801	2 266	329	261	671	161	128	325
1–16	A/U		1 301	1 496	29	361	426	12	169	201	3
	G/C		996	801	2 268	327	262	676	160	128	326

[a] Genbank accessions SalaUV-NL1 variants: 1–16 SL (PQ834485); 1–27 SL (PQ834486); 2–1 SL (PQ834487); 1–16 SS (PQ834488); 1–27 SS (PQ834489); 2–1 SS (PQ834490).

represented by exactly three codons. This finding suggests that the selective pressure influencing codon usage operates primarily at the level of tRNA availability or efficiency, rather than at the protein level.

Concluding remarks

In our quest to discover new RNA viruses with yet unknown RNA functionalities in seaweed, we encountered a novel multisegmented RNA virus in *S. latissima* samples. The RNA genome of the virus, named SalaUV-NL1, consists of a large segment and a small segment, and shares limited similarity with any known viral sequence, yet possesses remarkable genome characteristics. At first glance, the genome architecture was consistent with an ambigrammatic virus. In these viruses, stop codons are depleted not only from the fORF (RdRp) but also from the phase-0 reading frame on the reverse strand, thereby creating a long rORF that could, in principle, encode a protein comparable in length to the fORF product (Cook et al. 2013, DeRisi et al. 2019). However, the large segment of the new virus showed an unusually strong codon bias in the fORF encoding a large protein that includes an RdRp domain. Specifically, this fORF preferentially uses codons that do not end in A or U, implying a high GC content at third codon positions (GC3). Because the phase-0 rORF shares the same triplet boundaries as the fORF, this GC3 nucleotide-composition bias is mirrored on the genomic strand, but instead occurs at first codon positions (GC1). Since all three canonical stop codons begin with U, the depletion of U at these relevant positions on the genomic strand makes stop codons infrequent. Consequently, a long, uninterrupted putative rORF arises that is similar in length to the fORF (RdRp), even if this pattern is driven by nucleotide composition rather than selection for a functional ambigrammatic coding strategy. Therefore, in contrast to ambigrammatic viruses, we do not consider the rORF to encode a protein that is essential for viral replication or persistence (Dudas et al. 2021, Retallack et al. 2021, Wilkinson et al. 2021). This interpretation is supported by our observation that viral variants identified at the NIOZ facility harboured a premature stop codon that yielded a markedly truncated rORF, reducing its predicted length from 2 317 to 165 amino acids. We hypothesize that the rORF arises as an incidental consequence of the pronounced GC3 bias in the fORF and propose the term pseudoambigrammatic to describe this property, i.e. the apparent presence of an additional ORF on the complementary strand that is primarily attributable to compositional constraints on the RdRp coding strand rather than to selection for a functional gene product on the genomic strand. A GenBank query revealed the existence of several other pseudoambigrammatic viruses in different host species (Table 2). Even if pseudoambigrammatic viruses may not encode a functional rORF protein, this does not rule out that the ribosome protection hypothesis (DeRisi et al. 2019, Dudas et al. 2021, Retallack et al. 2021, Wilkinson et al. 2021) or other factors (Plotkin and Kudla 2011, Cardinale et al. 2013) could still be relevant to them.

Given that it is unlikely that the GC3 codon bias observed in the fORF (RdRp) is driven by codon usage constraints imposed by the overlapping rORF, a more plausible explanation is that this bias reflects adaptation to the host's tRNA pool. Specifically, the preference for G and C-ending codons at the third synonymous position may be influenced by the relative abundance and availability of corresponding tRNAs in the host, thereby enhancing translational efficiency. This is supported by the observations that the fORF of almost all pseudoambigrammatic viruses uses the only G ending stop codon (UAG) in their fORF, and that biased GC3 codon usage often occurs in highly expressed genes. Another argument to consider the rORF irrelevant is that the virus seems to mimic the codon bias of highly expressed genes. Since these genes only produce a sense mRNA strand, the antisense strand has no influence on the creation of the GC3 high codon bias. However, interpreting the virus's pseudoambigrammatic character is further complicated by the fact that the strong GC3 bias was not confined to the relatively short fORF on the small segment, but instead extended across nearly the entire remainder of the segment. Only the extreme 5′ and 3′ termini showed the opposite pattern, with elevated AT3 bias, similar to the ends of the large segment. It might be that there is an unrecognized ORF in this RNA sequence, possibly created by so-called cap/ATG snatching (Ho et al. 2020), as the glutamic acid-rich region in the GC3-rich frame bears some resemblance to that of the latency-associated nuclear antigen (LANA) in herpesvirus(Verma and Robertson 2003). However, the high GC3 content in both SalaUV-NL1 virus segments extended well outside the identified fORFs. Thus, if translational selection is not the primary driver of the near-maximal GC3 in this virus, the observed GC3 bias may reflect selection for RNA stability, analogous to stability-driven codon usage patterns reported for human mRNA (Hia et al. 2019).

Another puzzling finding emerged when comparing variants from the two facilities: there was a high proportion of A/U ↔ G/C transitions at third codon positions, yielding predominantly synonymous codon differences among the SalaUV-NL1 SL variants. This increased the GC3 content from 98.0% in variant 1–27 to 98.7% in variant 1–16. In total, there were 45 codons in variant 1–27 with A/U at the third position, compared to only 29 in variant 1–16. Although the relationship between the Scottish (1–27) and Dutch (1–16) samples is unknown, it is striking that 28 of 45 (62%) third-position A/U bases in variant 1–27 are G/C in variant 1–16, all yielding synonymous codons. In contrast, the reverse pattern is rare: only 0.6% show the reciprocal change (G/C to A/U) at the third position, and missense G/C bases occur in just 0.2%.

The uneven distribution of genomic differences is evident in the fact that 16 of the 24 (67%) codons in variant 1–27 that end with an A at the third position are AUA codons, while the remaining eight are replaced by G/C-ending codons in variant 1–16. The AUA codon is particularly noteworthy, as it shares its first two nucleotides with the canonical start codon, AUG. The mechanisms driving these specific genomic changes remain unknown. To resolve this unusual pattern, it will be necessary to collect additional SalaUV-NL1 variants and perform comparative genomic analyses.

Acknowledgements

This research was directly and indirectly funded by the Swammerdam Institute for Life Sciences of the University of Amsterdam.

Supplementary material

Supplementary material is available at *VEVOLU Journal* online.

Conflicts of interest

None declared.

Funding

None declared.

Data availability

The raw sequence reads have been deposited in the NCBI Sequence Read Archive under BioProject accession numbers PRJNA1089059 and PRJNA1195823. The following virus genome sequences have been deposited in NCBI GenBank: SalaUV-NL1 1–16 SL (PQ834485); SalaUV-NL1 1–27 SL (PQ834486); SalaUV-NL1 2–1 SL (PQ834487); SalaUV-NL1 1–16 SS (PQ834488); SalaUV-NL1 1–27 SS (PQ834489); SalaUV-NL1 2–1 SS (PQ834490).

References

Abbo SR, De Almeida JPP, Olmo RP *et al.* The virome of the invasive Asian bush mosquito *Aedes japonicus* in Europe. *Virus. Evolution* 2023;**9**:vead041. https://doi.org/10.1093/ve/vead041

Andrews S. *FastQC: A Quality Control Tool for High Throughput Sequence Data*. Cambridge, UK: Babraham Bioinformatics, Babraham Institute; 2010. Available from: http://www.bioinformatics.babraham.ac.uk/projects/fastqc/

Babaian A, Edgar R. Ribovirus classification by a polymerase barcode sequence. *PeerJ* 2022;**10**:e14055. https://doi.org/10.7717/peerj.14055

Batson J, Dudas G, Haas-Stapleton E *et al.* Single mosquito metatranscriptomics identifies vectors, emerging pathogens and reservoirs in one assay. *eLife* 2021;**10**:e68353. https://doi.org/10.7554/eLife.68353

Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* 2014;**30**:2114–20. https://doi.org/10.1093/bioinformatics/btu170

Bushmanova E, Antipov D, Lapidus A *et al.* rnaSPAdes: a *de novo* transcriptome assembler and its application to RNA-Seq data. *GigaScience* 2019;**8**:giz100. https://doi.org/10.1093/gigascience/giz100

Cao X, Zhang Y, Ding Y *et al.* Identification of RNA structures and their roles in RNA functions. *Nat Rev Mol Cell Biol* 2024;**25**:784–801. https://doi.org/10.1038/s41580-024-00748-6

Caprara MG, Nilsen TW. RNA: versatility in form and function. *Nat Struct Biol* 2000;**7**:831–3. https://doi.org/10.1038/82816

Cardinale D, DeRosa K, Duffy S. Base composition and translational selection are insufficient to explain codon usage bias in plant viruses. *Viruses* 2013;**5**:162–81. https://doi.org/10.3390/v5010162

Cech TR, Steitz JA. The noncoding RNA revolution—trashing old rules to forge new ones. *Cell* 2014;**157**:77–94. https://doi.org/10.1016/j.cell.2014.03.008

Cheng X, Wu X, Fang R. The minus strand of positive-sense RNA viruses encodes small proteins. *Trends Microbiol* 2024;**32**:6–7. https://doi.org/10.1016/j.tim.2023.10.012

Chiba Y, Tomaru Y, Shimabukuro H *et al.* Viral RNA genomes identified from marine macroalgae and a diatom. *Microbes Environ* 2020;**35**:n/a. https://doi.org/10.1264/jsme2.ME20016

Cook S, Chung BYW, Bass D *et al.* Novel virus discovery and genome reconstruction from field RNA samples reveals highly divergent viruses in dipteran hosts. *PLoS One* 2013;**8**:e80720. https://doi.org/10.1371/journal.pone.0080720

Dekker RJ, De Leeuw WC, Van Olst M *et al.* Discovery of novel RNA viruses in commercially relevant seaweeds *Alaria esculenta* and *Saccharina latissima*. bioRxiv 2024.05.22.594653. 2024. https://doi.org/10.1101/2024.05.22.594653

DeRisi JL, Huber G, Kistler A *et al.* An exploration of ambigrammatic sequences in narnaviruses. *Sci Rep* 2019;**9**:17982. https://doi.org/10.1038/s41598-019-54181-3

Dinan AM, Lukhovitskaya NI, Olendraite I *et al.* A case for a negative-strand coding sequence in a group of positive-sense RNA viruses. *Virus Evol* 2020;**6**:veaa007. https://doi.org/10.1093/ve/veaa007

Du Y, Liu PF, Li Z *et al.* Discerning the putative U and V chromosomes of *Saccharina japonica* (Phaeophyta) by cytogenetic mapping of sex-linked molecular markers. *Front Mar Sci* 2022;**9**:821603. https://doi.org/10.3389/fmars.2022.821603

Dudas G, Huber G, Wilkinson M *et al.* Polymorphism of genetic ambigrams. *Virus. Evolution* 2021;**7**:veab038. https://doi.org/10.1093/ve/veab038

Edgar RC, Taylor B, Lin V *et al.* Petabase-scale sequence alignment catalyses viral discovery. *Nature* 2022;**602**:142–7. https://doi.org/10.1038/s41586-021-04332-2

Firth AE, Brierley I. Non-canonical translation in RNA viruses. *J Gen Virol* 2012;**93**:1385–409. https://doi.org/10.1099/vir.0.042499-0

Garrison E, Marth G. Haplotype-based variant detection from short-read sequencing, arXiv:1207.3907. Preprint, arXiv. 2012. https://doi.org/10.48550/arXiv.1207.3907

Göertz GP, Miesen P, Overheul GJ *et al.* Mosquito small RNA responses to West Nile and insect-specific virus infections in aedes and culex mosquito cells. *Viruses* 2019;**11**:271. https://doi.org/10.3390/v11030271

Haseltine WA, Patarca R. The RNA revolution in the central molecular biology dogma evolution. *IJMS* 2024;**25**:12695. https://doi.org/10.3390/ijms252312695

Hia F, Yang SF, Shichino Y *et al.* Codon bias confers stability to human mRNA s. *EMBO Rep* 2019;**20**:e48220. https://doi.org/10.15252/embr.201948220

Ho JSY, Angel M, Ma Y *et al.* Hybrid gene origination creates human-virus chimeric proteins during infection. *Cell* 2020;**181**:1502–1517.e23. https://doi.org/10.1016/j.cell.2020.05.035

Ho JSY, Zhu Z, Marazzi I. Unconventional viral gene expression mechanisms as therapeutic targets. *Nature* 2021;**593**:362–71. https://doi.org/10.1038/s41586-021-03511-5

Hou X, He Y, Fang P *et al.* Using artificial intelligence to document the hidden RNA virosphere. *Cell* 2024;**187**:6929–6942.e16. https://doi.org/10.1016/j.cell.2024.09.027

International Committee on Taxonomy of Viruses (ICTV). 2023 Release. International Committee on Taxonomy of Viruses (ICTV). 2023 release. n.d. ictv.global/taxonomy/.

Kim VN. Sorting out small RNAs. *Cell* 2008;**133**:25–6. https://doi.org/10.1016/j.cell.2008.03.015

King AMQ, Adams MJ, Carstens EB *et al.* Virus taxonomy: classification and nomenclature of viruses. In: King AMQ, Adams MJ, Carstens EB, Lefkowitz EJ (eds.), *Ninth Report of the International Committee on Taxonomy of Viruses*. San Diego: Elsevier, 2012.

Kreuze JF, Perez A, Untiveros M *et al.* Complete viral genome sequence and discovery of novel viruses by deep sequencing of small RNAs: a generic method for diagnosis, discovery and sequencing of viruses. *Virology* 2009;**388**:1–7. https://doi.org/10.1016/j.virol.2009.03.024

Kuhn JH, Abe J, Adkins S *et al.* Annual (2023) taxonomic update of RNA-directed RNA polymerase-encoding negative-sense RNA viruses (realm Riboviria: Kingdom Orthornavirae: Phylum Negarnaviricota). *J Gen Virol* 2023;**104**:001864. https://doi.org/10.1099/jgv.0.001864

Lachnit T, Thomas T, Steinberg P. Expanding our understanding of the seaweed Holobiont: RNA viruses of the red alga Delisea pulchra. *Front Microbiol* 2016;**6**:6. https://doi.org/10.3389/fmicb.2015.01489

Langmead B, Salzberg SL. Fast gapped-read alignment with bowtie 2. *Nat Methods* 2012;**9**:357–9. https://doi.org/10.1038/nmeth.1923

Levenshtein VI. Binary codes capable of correcting deletions, insertions, and reversals. *Soviet Physics-Doklady* 1966;**10**:707–10.

Llave C. Virus-derived small interfering RNAs at the core of plant–virus interactions. *Trends Plant Sci* 2010;**15**:701–7. https://doi.org/10.1016/j.tplants.2010.09.001

Luthringer R, Lipinska AP, Roze D *et al.* The Pseudoautosomal regions of the U/V sex chromosomes of the Brown alga Ectocarpus exhibit unusual features. *Mol Biol Evol* 2015;**32**:2973–85. https://doi.org/10.1093/molbev/msv173

Malet H, Williams HM, Cusack S *et al.* The mechanism of genome replication and transcription in bunyaviruses. *PLoS Pathog* 2023;**19**:e1011060. https://doi.org/10.1371/journal.ppat.1011060

Newburn LR, White KA. Cis-acting RNA elements in positive-strand RNA plant virus genomes. *Virology* 2015;**479-480**:434–43. https://doi.org/10.1016/j.virol.2015.02.032

Parvathy ST, Udayasuriyan V, Bhadana V. Codon usage bias. *Mol Biol Rep* 2022;**49**:539–65. https://doi.org/10.1007/s11033-021-06749-4

Peck KM, Lauring AS. Complexities of viral mutation rates. *J Virol* 2018;**92**:10.1128/jvi.01031-17. https://doi.org/10.1128/jvi.01031-17

Plotkin JB, Kudla G. Synonymous but not the same: the causes and consequences of codon bias. *Nat Rev Genet* 2011;**12**:32–42. https://doi.org/10.1038/nrg2899

Prjibelski A, Antipov D, Meleshko D *et al.* Using SPAdes De novo assembler. *Curr Protoc Bioinformatics* 2020;**70**:e102. https://doi.org/10.1002/cpbi.102

Quax TEF, Claassens NJ, Söll D *et al.* Codon bias as a means to fine-tune gene expression. *Mol Cell* 2015;**59**:149–61. https://doi.org/10.1016/j.molcel.2015.05.035

R Core Team. *R: A Language and Environment for Statistical Computing, Version 4.3.2*. Vienna, Austria: R Foundation for Statistical Computing, 2023. https://www.R-project.org/

Ren F, Zhou M, Deng F *et al.* Combinatorial minigenome Systems for Emerging Banyangviruses Reveal Viral Reassortment Potential and Importance of a protruding nucleotide in genome "panhandle" for promoter activity and reassortment. *Front Microbiol* 2020;**11**:599. https://doi.org/10.3389/fmicb.2020.00599

Retallack H, Popova KD, Laurie MT *et al.* Persistence of Ambigrammatic Narnaviruses requires translation of the reverse open reading frame. *J Virol* 2021;**95**:e0010921–1. https://doi.org/10.1128/JVI.00109-21

Scholthof HB, Scholthof KBG. Plant virology: an RNA treasure trove. *Trends Plant Sci* 2023;**28**:1277–89. https://doi.org/10.1016/j.tplants.2023.06.019

Schroeder DC, McKeown DA. Viruses of seaweeds. In: Hurst CJ (ed.), *Studies in Viral Ecology*2nd edn. Hoboken (NJ): Wiley Online Library, 2021.

Sim MC, Ho CL, Phang SM. A simple and effective method for RNA isolation and cDNA library construction from the brown seaweed Sargassum polycystum (Fucales, Phaeophyceae). *J Appl Phycol* 2013;**25**:1277–85. https://doi.org/10.1007/s10811-013-9980-z

Simon-Loriere E, Holmes EC. Why do RNA viruses recombine? *Nat Rev Microbiol* 2011;**9**:617–26. https://doi.org/10.1038/nrmicro2614

Tjeldnes H, Labun K, Torres Cleuren Y *et al.* ORFik: A comprehensive R toolkit for the analysis of translation. *BMC Bioinformatics* 2021;**22**:336. https://doi.org/10.1186/s12859-021-04254-w

Van Der Loos LM, De Coninck L, Zell R *et al.* Highly divergent CRESS DNA and picorna-like viruses associated with bleached thalli of the green seaweed *Ulva*. *Microbiol Spectrum* 2023;**11**:e00255–23. https://doi.org/10.1128/spectrum.00255-23

Verma SC, Robertson ES. ORF73 of herpesvirus saimiri strain C488 tethers the viral genome to metaphase chromosomes and binds to *cis* -acting DNA sequences in the terminal repeats. *J Virol* 2003;**77**:12494–506. https://doi.org/10.1128/JVI.77.23.12494-12506.2003

Waldron FM, Stone GN, Obbard DJ. Metagenomic sequencing suggests a diversity of RNA interference-like responses to viruses across multicellular eukaryotes. *PLoS Genet* 2018;**14**:e1007533. https://doi.org/10.1371/journal.pgen.1007533

Wilkinson M, Yllanes D, Huber G. Polysomally protected viruses. *Phys Biol* 2021;**18**:046009. https://doi.org/10.1088/1478-3975/abf5b5

Wint R, Salamov A, Grigoriev IV. Kingdom-wide analysis of fungal protein-coding and tRNA genes reveals conserved patterns of adaptive evolution. *Mol Biol Evol* 2022;**39**:msab372. https://doi.org/10.1093/molbev/msab372

Wolf YI, Kazlauskas D, Iranzo J *et al.* Origins and evolution of the global RNA Virome. *mBio* 2018;**9**:e02329–18. https://doi.org/10.1128/mBio.02329-18

Wright F. The 'effective number of codons' used in a gene. *Gene* 1990;**87**:23–9. https://doi.org/10.1016/0378-1119(90)90491-9

Wu Q, Luo Y, Lu R *et al.* Virus discovery by deep sequencing and assembly of virus-derived small silencing RNAs. *Proc Natl Acad Sci* 2010;**107**:1606–11. https://doi.org/10.1073/pnas.0911353107

Yin Y, Fischer D. Identification and investigation of ORFans in the viral world. *BMC Genomics* 2008;**9**:24. https://doi.org/10.1186/1471-2164-9-24

Zhang Y, Liang X, Zhao M *et al.* A novel ambigrammatic mycovirus, PsV5, works hand in glove with wheat stripe rust fungus to facilitate infection. *Plant Commun* 2023;**4**:100505. https://doi.org/10.1016/j.xplc.2022.100505

Zhao J, Nair S, Zhang Z *et al.* Macroalgal virosphere assists with host–microbiome equilibrium regulation and affects prokaryotes in surrounding marine environments. *ISME J* 2024;**18**:wrae083. https://doi.org/10.1093/ismejo/wrae083

Zhou H, Yan B, Chen S *et al.* Evolutionary characterization of Tembusu virus infection through identification of codon usage patterns. *Infect Genet Evol* 2015;**35**:27–33. https://doi.org/10.1016/j.meegid.2015.07.024

Virus Evolution, 2026, **12**, veag030
https://doi.org/10.1093/ve/veag030
Research Article