An exploration of ambigrammatic sequences in narnaviruses

Joseph L. DeRisi[1,2], **Greg Huber**[1], **Amy Kistler**[1], **Hanna Retallack**[2], **Michael Wilkinson**[1,3], **David Yllanes**[1]

***For correspondence:**
joe@derisilab.ucsf.edu (JLD);
greg.huber@czbiohub.org (GH);
amy.kistler@czbiohub.org (AK);
hanna.retallack@ucsf.edu (HR);
m.wilkinson@open.ac.uk (MW);
david.yllanes@czbiohub.org (DY)

[1]Chan Zuckerberg Biohub, 499 Illinois Street, San Francisco, CA 94158, USA; [2]Department of Biochemistry and Biophysics, University of California, San Francisco, California, USA; [3]School of Mathematics and Statistics, The Open University, Walton Hall, Milton Keynes, MK7 6AA, England

Abstract Narnaviruses have been described as positive-sense RNA viruses with a remarkably simple genome of ~ 3 kb, encoding only a highly conserved RNA-dependent RNA polymerase (RdRp). Many narnaviruses, however, are 'ambigrammatic' and harbor an additional uninterrupted open reading frame (ORF) covering almost the entire length of the reverse complement strand. No function has been described for this ORF, yet the absence of stops is conserved across diverse narnaviruses, and in every case the codons in the reverse ORF and the RdRp are aligned. The > 3 kb ORF overlap on opposite strands, unprecedented among RNA viruses, motivates an exploration of the constraints imposed or alleviated by the codon alignment. Here, we show that only when the codon frames are aligned can all stop codons be eliminated from the reverse strand by synonymous single-nucleotide substitutions in the RdRp gene, suggesting a mechanism for *de novo* gene creation within a strongly conserved amino-acid sequence. It will be fascinating to explore what implications this coding strategy has for other aspects of narnavirus biology. Beyond narnaviruses, our rapidly expanding catalog of viral diversity may yet reveal additional examples of this broadly-extensible principle for ambigrammatic-sequence development.

Introduction

Narnaviruses (a contraction of *naked RNA viruses*) are RNA viruses with a seemingly simple genome (***Hillman and Cai, 2013***). The only manifestation of narnaviral infections documented to date is the presence of large concentrations of the viral RNA in the cytoplasm of the host cell, often detected as double-stranded RNA. These infections were first observed in cultured yeast (***Kadowaki and Halvorson, 1971***; ***Wesolowski and Wickner, 1984***). Subsequent metagenomic sequencing revealed narnaviruses in other fungi (***Osaki et al., 2016***), oomycetes (***Cai et al., 2012***), mosquitoes (***Cook et al., 2013***; ***Chandler et al., 2015***; ***Göertz et al., 2019***; ***Shi et al., 2017***), other arthropods (***Harvey et al., 2019***), algae (***Waldron et al., 2017***), trypanosomatids (***Grybchuk et al., 2018***; ***Akopyants et al., 2016***; ***Lye et al., 2016***), and potentially apicomplexans (***Charon et al., 2019***), although the precise host species is not always clear. The known examples of narnaviruses are approximately 3 kb in size and code for a single protein, an RNA-dependent RNA polymerase (RdRp), with the exception of two putative bipartite narnaviruses (***Lye et al., 2016***; ***Grybchuk et al., 2018***; ***Charon et al., 2019***).

A remarkable feature of some narnaviruses is the existence of an additional large open reading frame (ORF), defined as a region devoid of stop codons, which spans nearly the full length of the reverse complement sequence of the virus genome (***Firth, 2014***). We refer to these examples, with large reverse open reading frame (rORF) features, as *ambigrammatic narnaviruses.* Such large uninterrupted rORFs are very unlikely to occur by chance, since for a typical nucleotide sequence,

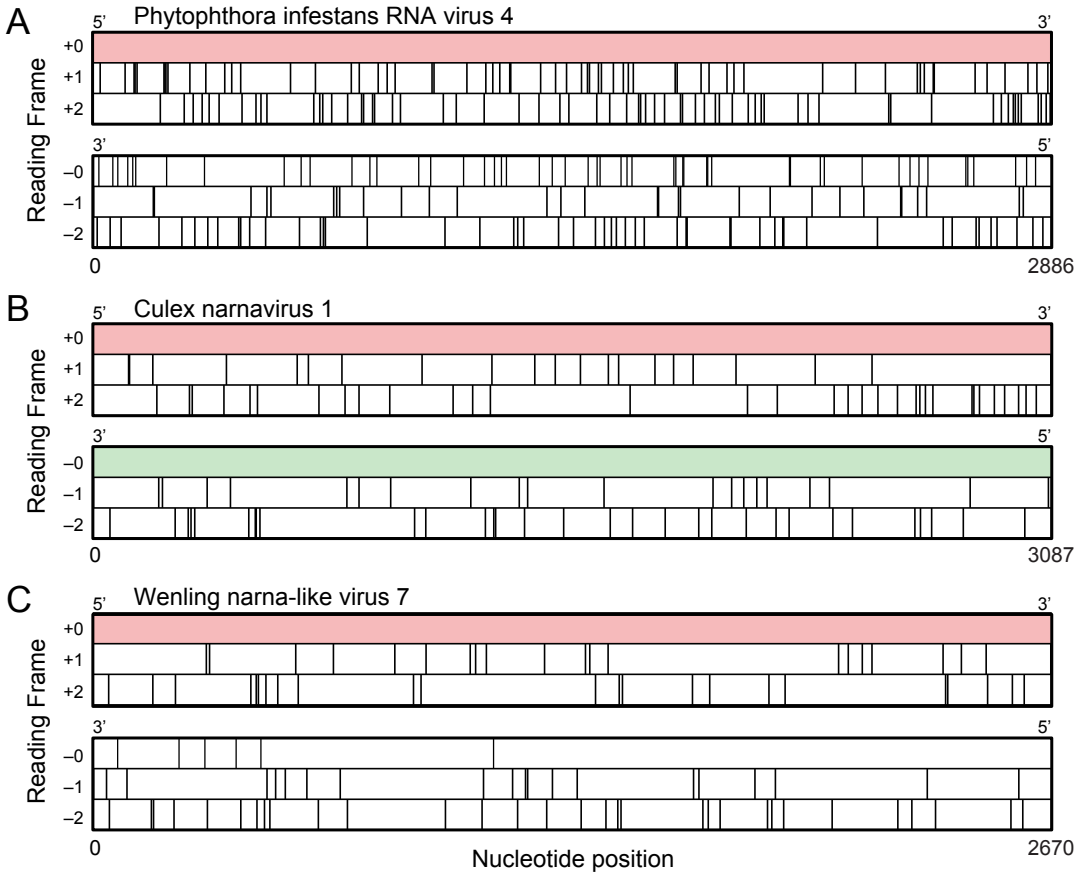

Figure 1. Ambigrammatic sequences in narnaviruses. Coding region for the RNA-dependent RNA polymerase (RdRp) of *Phytophthora infestans* RNA virus 4 (A), Culex narnavirus 1 (B), and Wenling narna-like virus 7 (C) in the reference +0 frame and all five other reading frames (see Figure 2 for our frame-labeling conventions). Stop codons in each frame are depicted as vertical lines. Large uninterrupted open reading frames (ORFs) are highlighted in color.

there are likely to be codons for which the reverse complement read is a stop codon. This is illustrated in Figure 1A, which shows a typical viral genome spanned by a single ORF and where all five alternative reading frames contain many stop codons. In contrast, Figure 1B shows the genetic sequence of Culex narnavirus 1, where one of the reverse reading frames also has an uninterrupted ORF spanning nearly the whole sequence. All known examples of ambigrammatic narnaviruses have the large reverse ORF in a reading frame where the codons are aligned with those in the forward direction (***Dinan et al., 2019***) (i.e., in "frame −0" in the conventions of Figure 2). Figure 1C considers an intriguing intermediate case, which will be discussed further below. The existence of these very long rORFs is a surprising observation which demands an exploration.

Interestingly, not all narnaviruses have an ambigrammatic genome (indeed, the example of Figure 1A is also in the *Narnaviridae* family). In agreement with previously reported observations (***Dinan et al., 2019***), overlaying the lengths of detectable rORFs on the narnavirus phylogeny shows that ambigrammatic sequences are present in at least two different clades. This finding, illustrated in Figure 3, indicates that the ambigrammatic feature is polyphetic and may have been gained and lost multiple times in the evolution of this viral family.

In this paper we explore how a large open reading frame might arise in the reverse complement sequence of the RNA. While there is an extensive literature on the statistics of codons in overlapping genes (discussed in ***Smith and Waterman, 1980***; ***Shukla, 2015***; ***Brandes and Linial, 2016***; ***Lèbre and Gascuel, 2017***) and on the evolution of overlapping genes in viruses (***Staden, 1984***; ***Krakauer, 2002***;

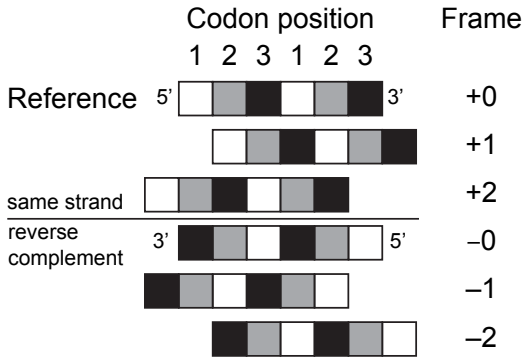

Figure 2. Labeling conventions used in this paper for reading frames.

Krakauer and Plotkin, 2002; ***Belshaw et al., 2007***; ***Rancurel et al., 2009***; ***Chirico et al., 2010***), the vast majority of these analyses handle cases of overlapping genes being translated in the same direction, with few papers considering antisense overlaps (see, e.g. ***Merino et al., 1994***). In contrast, narnaviruses are unique in demonstrating long uninterrupted reading frames in the reverse complement sequence of RNA. Moreover, most analyses of overlapping genes are complicated by the fact that, in the general case, the sequences of both genes can evolve. Analyzing cases where one of the overlapping genes is more strongly conserved is very difficult in general, and most earlier works treat the two overlapping genes symmetrically (***Smith and Waterman, 1980***). In this work we adopt a different approach and treat one of the genes as having a fixed amino acid sequence. This could represent a gene with a critical function such as viral polymerases which are often strongly conserved. We suggest that even under these very strong constraints it may be possible for a novel large rORF to arise, but only in *one* of the three possible reverse reading frames.[1]

We discuss how this mechanism may give rise to a narnaviral genome that is ambigrammatic, in the sense that it can code for two proteins, each translated from one of the two complementary strands of RNA. Finally, we note that ambigrammatic coding in other RNA viruses has not been observed to date, and make some speculative remarks about the potential role of the narnavirus rORF.

Results

Synonym Sudoku

Usually, an alternative reading frame for a gene is found to have many stop codons, so that it cannot, therefore, code for a polypeptide or protein of useful length. Let us assume that a nucleotide sequence already codes for a gene which has an essential function. In this case the sequence of amino acids should not be changed. The nucleotide sequence, on the other hand, can still be altered by replacing codons with synonyms (codons which code for the same amino acid). We can ask whether a sequence of single-nucleotide mutations can remove the stop codons that are expected to occur in alternative reading frames, while replacing codons with others that are synonyms in the original reading frame. For some base sequences this will always be possible, but we consider whether *all* of the stop codons in an alternative reading frame can be removed by single-nucleotide mutations in an *arbitrary* polypeptide sequence.

There are five possible alternative reading frames (two in the original strand and three in the reverse complement sequence, see Figure 2). Like a Sudoku puzzle, there is no alternative but to explore all five possibilities in turn. In our discussion, we frame our argument in terms of an RNA genome, so that the base pairings are A=U and G=C, and the stop codons are UAA, UAG, UGA. For

[1] The two alternative forward reading frames, though not directly relevant for experimental observations in narnaviruses, are also discussed for completeness.

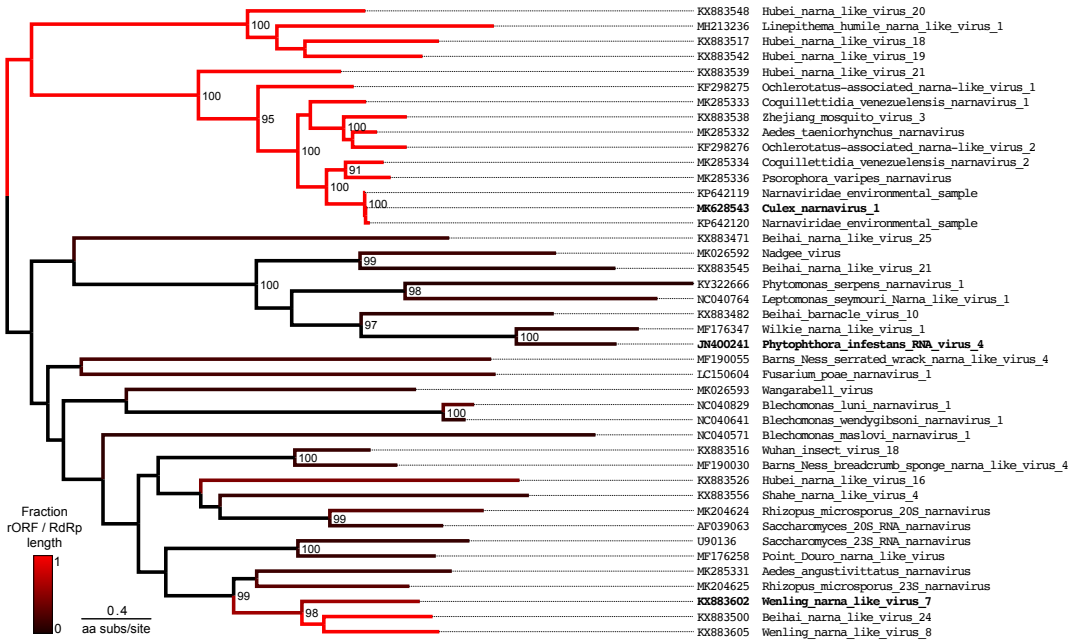

Figure 3. Maximum likelihood tree of amino-acid sequences for RNA-dependent RNA polymerase (RdRp) of 42 representative narnaviruses, identified by homology to the narnaviruses observed in culture, Culex narnavirus 1 and Phytophthora infestans virus 4 (NCBI Blastx, ***Altschul et al., 1990***). Unrooted tree shown with midpoint rooting for display. Branch coloring indicates the fraction of RdRp coding sequence overlapped by the longest open reading frame (defined as a region uninterrupted by stops) in the reverse complement aligned frame (-0 frame) for sequences at tips (see color bar, bottom left). The sequence names in bold correspond to those shown in Figure 1. Numbers at nodes indicate bootstrap values (shown when > 80). The branch length is given by the amino-acid substitutions per site, as illustrated by the scale bar.

the purposes of understanding ambigrammatic sequences observed in narnaviruses, we focus below on the three reverse complement reading frames; the two alternate forward reading frames are discussed in appendix 1.

Frame -0: aligned complementary read frames

First, consider the case where the reading frames of the forward and reverse complement sequences have their codons aligned (the -0 frame). In this context, stop codons UAA, UAG, UGA become, respectively, UUA, CUA, UCA in the $+0$ frame, encoding the amino acids Leu, Leu, Ser. Thus, only instances of leucine and serine in the $+0$ reading frame can result in stop codons in the -0 reading frame.

We should now consider if synonymous substitution of Leu or Ser codons in the $+0$ frame can remove the stop codons in the -0 frame. The synonyms of Leu are CU*, UUA, and UUG (where * means any nucleotide). The synonyms of Ser are UC*, AGU, and AGC. Hence, the Leu codon UUA can be transformed to UUG by a single substitution. Similarly, the Leu codon CUA can be transformed to CUU, CUG or CUC by single substitutions, while the Ser codon UCA is transformed to UCU, UCG or UCC by single substitutions. Thus, when frames are aligned, 7 types of single-nucleotide synonymous substitutions in the $+0$ frame are sufficient to remove stops in the reverse direction.

It is notable that ambigrammatic narnavirus strains have been isolated from a variety of different hosts, including those with genetic codes that contain only two stop codons, such as fungal mitochondria. Despite this, the narnavirus rORF sequences do not contain all three possible stop codons indicating that translation of the narnavirus rORF it is not limited to ribosomes which only recognize two stop codons.

Frames $-1, -2$: staggered reverse reading frames

The cases of staggered reverse complement reading frames are more complex, because a codon in the original $+0$ reading frame straddles two codons in each of these alternate reading frames. Let us first study the case of the -1 reading frame, where the codons of the reverse reading frame are shifted towards the $3'$ end (Figure 2). Consider the sequence CUA in the forward direction, with a shift of one base between frames, so we have, using | to denote triplet codon boundaries, $x, y, \ldots$ for unspecified bases and $\bar{x}, \bar{y}, \ldots$ for their pairing complementary bases:

$$\begin{array}{cccccccccccc} 5' & \ldots & x & y & \mathsf{C} & \mid & \mathsf{U} & \mathsf{A} & z & \mid & \ldots & 3' \quad +0 \text{ frame} \\ 3' & \ldots & \bar{x} & \bar{y} & \mid & \mathsf{G} & \mathsf{A} & \mathsf{U} & \mid & \bar{z} & \ldots & 5' \quad -1 \text{ frame} \end{array}$$

Note that the reversed read UAG is one of the stop codons we wish to avoid, which we could do by changing either UAz or xyC to a synonym. Let us start with UAz, which can only be Tyr (either UAU or UAC; the other two values of y correspond to stops). UAU and UAC are the only codons that translate to Tyr, so it is not possible to find a synonym of UAz that avoids the stop in the reverse sequence. The alternative is to find a synonym of xyC. Here we note that if transforming C to U is the only synonymous change, this will still yield a stop codon in the reverse read frame. The choices satisfying this condition are His, Asn, Asp, Ser. Exactly the same restrictions arise from considering the U|UA sequence, and no additional cases of non-removable stops arise from considering U|CA. Thus we find that the following combinations of four $+0$ codon pairs will prevent an ambigrammatic partner rORF in this frame: (His,Tyr), (Asn,Tyr), (Asp,Tyr), (Ser,Tyr).

In the case of the -2 frame, the codons are shifted to the $5'$ end. Let us consider whether the stop codon in the reverse-read direction is UAA.

$$\begin{array}{cccccccccc} 5' & \ldots & \mid & x & \mathsf{U} & \mathsf{U} & \mid & \mathsf{A} & y & \ldots & 3' \quad +0 \text{ frame} \\ 3' & \ldots & \bar{x} & \mid & \mathsf{A} & \mathsf{A} & \mathsf{U} & \mid & \bar{y} & \ldots & 5' \quad -2 \text{ frame} \end{array}$$

Consider the possible synonym changes to the codons in the $+0$ frame that will remove the stop codon in the -2 frame. Changing the first base of the second codon (A) cannot give a synonym. Also, changing the second codon of the first base (U) never gives a synonym. Finally, changing the third base of the first codon always gives a synonym, but, if the first codon is UUU, there is only one synonym, UUC, which also gives a stop in the complementary chain. Similarly, if the first codon in the $+0$ frame is UUC, the stop codon UGA in the reverse read cannot be removed by synonym substitution. Considering the case of UC|A yields the same set of excluded combinations, and CU|A does not lead to further examples. We conclude that if the $+0$ frame contains Phe followed by Ile, Met, Thr, Asn, Ser or Arg, then there is a stop codon in the reverse-read frame which cannot be removed.

The potential of synonymous substitutions for generating a new ORF

In conclusion, the -0 reading frame is the only one of the three reverse reading frames where stop codons can always be removed by synonymous single-nucleotide mutations in the $+0$ original (conserved) amino-acid sequence. This is consistent with the observation that the forward and the reverse open reading frames in ambigrammatic narnaviruses have their codons aligned.

Thus, synonymous substitutions provide a possible route[2] to the generation of a new ORF. We should consider the extent to which this is an effective mechanism for creating new coding potential. In addition to removing stop codons in the -0 reading frame, replacing other codons of the original $+0$ ORF by synonyms can result in changes to the amino-acid sequence of the rORF, without changing the protein encoded in the $+0$ frame. It is these synonymous substitutions which allow scope for evolutionary adaptation of the new protein. If the changes in the $+0$ frame are strictly limited to synonym substitutions, the range of available proteins that can be produced by the new ORFs is quite constrained. We discuss how this can be quantified in Appendix 2.

[2] As discussed in Appendix 1, large ORFs could also be generated in the $+2$ (but not on the $+1$) reading frame by single-nucleotide synonymous substitutions, although this possibility has not been observed in narnaviruses.

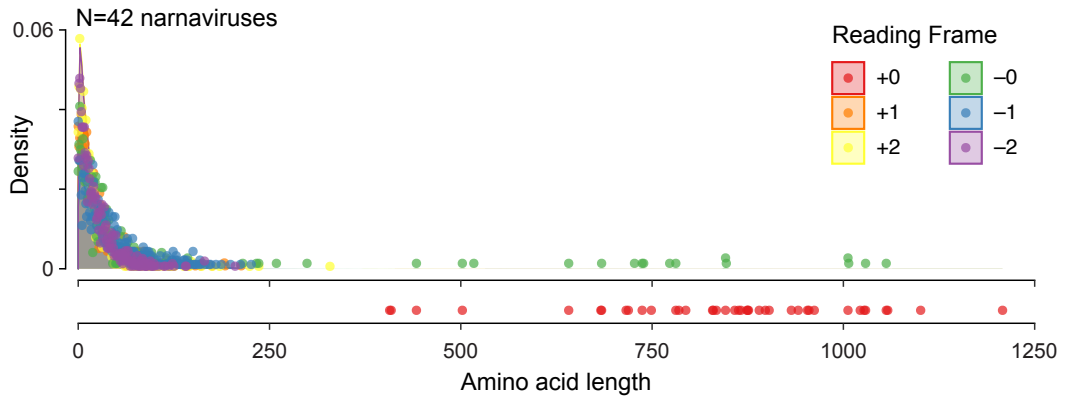

Figure 4. Probability distribution for ORF lengths in narnavirus-like sequences. Shading shows distribution of ORF lengths colored by reading frame after codon permutation test on RdRp coding sequences of 42 representative narnaviruses as in Figure 3. In brief, codons are randomly re-ordered and then ORF lengths in the 5 alternate frames are calculated (permutation methods as in ***Schlub et al., 2018***). Points give lengths of actual ORFs in reference sequences, colored according to reading frame, with the reference RdRp as +0 frame (red, below). Note that some annotated RdRp coding regions in the database may be fragments of the complete coding sequence.

A hypothesis about the evolution of narnaviruses

First we should consider a null hypothesis, that the rORF (spanning approximately 3 kb) is a chance occurrence. A priori this appears to be extremely unlikely: given that there are three stop codons out of 64 possibilities, we expect that the typical distance between stop codons will be approximately 60 base pairs. If the stop codons are randomly and independently scattered, with mean separation $\langle N \rangle$, the probability of a string of N bases containing no stops is expected to be $P(N) = \exp(-N/\langle N \rangle)$. If $\langle N \rangle = 60$, the probability of finding a sequence of 3 kb lacking a stop codon is approximately $\exp(-20) \approx 2 \times 10^{-9}$. The correct value of $\langle N \rangle$ depends upon the distribution of codons, so as a more refined check we examined the distribution of stop codons in each of the five possible alternative frames which arise from choosing a random permutation of the codon of the RdRp gene. Including both ambigrammatic and non-ambigrammatic narnaviruses to generate a null distribution of lengths of ORFs that may overlap the RdRp, we find that the expected probability of a long sequence without a stop codon is indeed well approximated by an exponential distribution, and that the expected probability of an ORF with the observed length is negligible (Figure 4). The scale length $\langle N \rangle$ varies from frame to frame, but is not greatly different from the simple estimate, $\langle N \rangle \approx 60$. Of course, even a highly unlikely feature can arise and become fixed in a population. However, there is sufficient variability in the RdRp sequence that stop codons would be expected to arise unless selected against. For instance, the average pairwise identity of the 11 sequences in the clade that includes Culex narnavirus 1 in Figure 3 is only 51 %. These analyses suggests that the chance occurrence and maintenance of a large rORF is highly unlikely, implying that it may offer some evolutionary advantage, broadly speaking.

Given this, it is desirable to speculate how the rORF observed in narnaviruses may have arisen. The virus could have originally existed as a sequence which just coded for the RdRp, with stop codons in the complementary reading frame. It could have evolved by gradually removing the stop codons, and at the same time making other synonym changes in the $+0$ frame, as the coding sequence lengthens. These mutations could result in progressively longer rORFs, selected by their capacity to increase the fitness of the virus.

The Wenling-7 narnavirus (Figure 1C), which has a single stop codon in the middle of a large rORF, is intriguing as a potential transitional form in which the rORF is either being gained or lost. If it is being gained, the mechanism in the previous paragraph indicates that the rORF would increase in length incrementally. If the rORF is being lost (for instance, if transfer to exploit a new host

removes its advantage), we expect one or more stops randomly scattered in a large ORF. Because this latter picture is more consistent with the Wenling-7 genome, we speculate that this strain is in the process of losing the large rORF feature. With the presently available data, however, we cannot conclusively support this intuition. Filling out the phylogeny with more sequences could allow us to clarify the potential direction of change.

At this time there is very limited information as to whether the rORF can increase the fitness of the narnavirus. There may be unusual mechanisms in which uninterrupted rORFs alter the translation or protein processing machinery in the cell to the virus's advantage, even if no protein is made or if the amino acid sequence is not important for its function. For instance, RNA forms of several viruses are thought to be subject to nonsense-mediated decay (***LeBlanc and Beemon, 2004***; ***Balistreri et al., 2014***; ***Fontaine et al., 2018***). The narnavirus rORF may be another strategy to increase viral-RNA stability through antagonism of RNA decay pathways or other methods.

Alternatively, if we postulate that the rORF does indeed code for a functional protein, we might speculate on what can be learned from its sequence. A search of the Protein Data Bank (PDB) and NCBI's translated sequence database (NCBI nr) revealed no sequences with significant identity to the translated rORF. Explorations of secondary structure and other protein features were similarly uninformative, but do not rule out the possibility for a functional protein (see Appendix 3). Perhaps a protein translated from the rORF could enable the narnavirus to evade host-cell defenses, allow for movement of the virus between cells, enhance replication by complexing with the genome and RdRp, be required for replication in a particular host species, interact with additional viral elements, or have any of a number of other functions that have not yet been described for this family of viruses.

Discussion

Motivated by observations of ambigrammatic sequences in narnaviruses, we have shown that an existing ORF can give rise to a large uninterrupted ORF in the reverse complement sequence by synonymous substitutions that preserve the amino-acid sequence of a conserved forward ORF and remove stop codons from the rORF. We find that this mechanism for making ambigrammatic genes only works when the forward and reverse read frames are aligned. These findings are consistent with the observed alignment of overlapping $+0$ RdRp ORFs and -0 rORFs among many naturally occurring narnavirus sequences.

Any function for the narnavirus rORF remains an intriguing mystery, as does the machinery and processes that may be involved in translating complementary strands of the same RNA sequence. To our knowledge, there are no biologically validated examples of overlapping genes in the reverse complement orientation among RNA-only viruses (***Belshaw et al., 2007***; ***Rancurel et al., 2009***; ***Sabath et al., 2012***). While some such overlaps have been predicted (***Schlub et al., 2018***), they exhibit neither the overlap length nor the conservation across related strains that is seen among narnaviruses. Indeed, the narnavirus overlap is the longest yet observed among RNA viruses (see, e.g., ***Brandes and Linial, 2016***). As our observations about the structure of the genetic code are extensible beyond this family, it will be interesting to see whether the explosion in metagenomic sequencing data will reveal more ambigrammatic viruses.

By virtue of packing, replication, and transmission requirements, viral genomes display a myriad of diverse innovations that provoke us to consider what is possible at the extremes of sequence evolution. Here, the ambigrammatic genes found in some narnaviruses are one such innovation, and their existence likely points to new biology that may be equally as fascinating.

Acknowledgments

The authors thank John Pak for illuminating discussions on the protein-coding potential of the rORF and Timothy Schlub and Edward Holmes for generously sharing the code used to produce Figure 1. MW thanks the Chan Zuckerberg Biohub for its hospitality. Author contributions: MW, DY and GH

conceived of the proposal and performed reading frame analysis. HR analyzed sequencing data that stimulated this investigation, and generated the figures. JLD and AK provided critical input on viral biology. HR, MW and DY wrote the manuscript with input from all authors.

References

Akopyants NS, Lye LF, Dobson DE, Lukeš J, Beverley SM. A Narnavirus in the Trypanosomatid Protist Plant Pathogen Phytomonas serpens. Genome Announc. 2016 Aug; 4(4):e00711–16. https://mra.asm.org/content/4/4/e00711-16, doi: 10.1128/genomeA.00711-16.

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J Mol Biol. 1990 Oct; 215(3):403–410. http://www.sciencedirect.com/science/article/pii/S0022283605803602, doi: 10.1016/S0022-2836(05)80360-2.

Balistreri G, Horvath P, Schweingruber C, Zünd D, McInerney G, Merits A, Mühlemann O, Azzalin C, Helenius A. The host nonsense-mediated mRNA decay pathway restricts Mammalian RNA virus replication. Cell Host Microbe. 2014 Sep; 16(3):403–411. doi: 10.1016/j.chom.2014.08.007.

Belshaw R, Pybus OG, Rambaut A. The evolution of genome compression and genomic novelty in RNA viruses. Genome Res. 2007; 17:1496–1504. http://www.genome.org/cgi/doi/10.1101/gr.6305707, doi: 10.1101/gr.6305707.

Brandes N, Linial M. Gene overlapping and size constraints in the viral world. Biol Direct. 2016 May; 11(1):26. https://doi.org/10.1186/s13062-016-0128-3, doi: 10.1186/s13062-016-0128-3.

Cai G, Myers K, Fry WE, Hillman BI. A member of the virus family Narnaviridae from the plant pathogenic oomycete Phytophthora infestans. Archives of Virology. 2012 Jan; 157(1):165–169. https://doi.org/10.1007/s00705-011-1126-5, doi: 10.1007/s00705-011-1126-5.

Chandler JA, Liu RM, Bennett SN. RNA shotgun metagenomic sequencing of northern California (USA) mosquitoes uncovers viruses, bacteria, and fungi. Front Microbiol. 2015; 6. https://www.frontiersin.org/articles/10.3389/fmicb.2015.00185/full, doi: 10.3389/fmicb.2015.00185.

Charon J, Grigg MJ, Eden JS, Piera KA, William T, Rose K, Davenport MP, Anstey NM, Holmes EC. Matryoshka RNA virus 1: a novel RNA virus associated with Plasmodium parasites in human malaria. bioRxiv. 2019 May; p. 624403. https://www.biorxiv.org/content/10.1101/624403v1, doi: 10.1101/624403.

Chirico N, Vianelli A, Belshaw R. Why genes overlap in viruses. Proc Biol Sci. 2010 Dec; 277(1701):3809–3817. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2992710/, doi: 10.1098/rspb.2010.1052.

Cook S, Chung BYW, Bass D, Moureau G, Tang S, McAlister E, Culverwell CL, Glücksman E, Wang H, Brown TDK, Gould EA, Harbach RE, Lamballerie Xd, Firth AE. Novel Virus Discovery and Genome Reconstruction from Field RNA Samples Reveals Highly Divergent Viruses in Dipteran Hosts. PLOS ONE. 2013 Nov; 8(11):e80720. https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0080720, doi: 10.1371/journal.pone.0080720.

Dinan AM, Lukhovitskaya NI, Olendraite I, Firth AE. A case for a reverse-frame coding sequence in a group of positive-sense RNA viruses. bioRxiv. 2019 Jun; http://biorxiv.org/lookup/doi/10.1101/664342, doi: 10.1101/664342.

Drozdetskiy A, Cole C, Procter J, Barton GJ. JPred4: a protein secondary structure prediction server. Nucleic Acids Res. 2015 Jul; 43(W1):W389–W394. https://academic.oup.com/nar/article/43/W1/W389/2467870, doi: 10.1093/nar/gkv332.

Firth AE. Mapping overlapping functional elements embedded within the protein-coding regions of RNA viruses. Nucleic Acids Res. 2014 Nov; 42(20):12425–12439. https://academic.oup.com/nar/article/42/20/12425/2903286, doi: 10.1093/nar/gku981.

Fontaine KA, Leon KE, Khalid MM, Tomar S, Jimenez-Morales D, Dunlap M, Kaye JA, Shah PS, Finkbeiner S, Krogan NJ, Ott M. The Cellular NMD Pathway Restricts Zika Virus Infection and Is Targeted by the Viral Capsid Protein. mBio. 2018; 9(6). doi: 10.1128/mBio.02126-18.

Göertz GP, Miesen P, Overheul GJ, van Rij RP, van Oers MM, Pijlman GP. Mosquito small RNA responses to West Nile and insect-specific virus infections in Aedes and Culex mosquito cells. Viruses. 2019 Mar; 11(3):271. https://www.mdpi.com/1999-4915/11/3/271, doi: 10.3390/v11030271.

Grybchuk D, Akopyants NS, Kostygov AY, Konovalovas A, Lye LF, Dobson DE, Zangger H, Fasel N, Butenko A, Frolov AO, Votýpka J, d'Avila Levy CM, Kulich P, Moravcová J, Plevka P, Rogozin IB, Serva S, Lukeš J, Beverley SM, Yurchenko V. Viral discovery and diversity in trypanosomatid protozoa with a focus on relatives of the human parasite Leishmania. PNAS. 2018 Jan; 115(3):E506–E515. https://www.pnas.org/content/115/3/E506, doi: 10.1073/pnas.1717806115.

Harvey E, Rose K, Eden JS, Lawrence A, Doggett SL, Holmes EC. Identification of diverse arthropod associated viruses in native Australian fleas. Virology. 2019 Sep; 535:189–199. http://www.sciencedirect.com/science/article/pii/S0042682219301850, doi: 10.1016/j.virol.2019.07.010.

Hillman BI, Cai G. The family *Narnaviridae*: simplest of RNA viruses. In: Ghabrial SA, editor. *Mycoviruses*, vol. 86 of Advances in Virus Research; 2013.p. 149–176. doi: 10.1016/B978-0-12-394315-6.00006-4.

Kadowaki K, Halvorson HO. Appearance of a New Species of Ribonucleic Acid During Sporulation in Saccharomyces cerevisiae. J Bacteriol. 1971; 105(3):826–830. https://jb.asm.org/content/105/3/826.

Krakauer DC. Evolutionary principles of genomic compression. Comments on Theoretical Biology. 2002; 7:215–236.

Krakauer DC, Plotkin JB. Redundancy, antiredundancy, and the robustness of genomes. PNAS. 2002; 99(3):1405–1409. doi: 10.1073/pnas.032668599.

LeBlanc JJ, Beemon KL. Unspliced Rous Sarcoma Virus Genomic RNAs Are Translated and Subjected to Nonsense-Mediated mRNA Decay before Packaging. J Virol. 2004 May; 78(10):5139–5146. https://jvi.asm.org/content/78/10/5139, doi: 10.1128/JVI.78.10.5139-5146.2004.

Lèbre S, Gascuel O. The combinatorics of overlapping genes. J Theor Biol. 2017 Feb; 415:90–101. http://www.sciencedirect.com/science/article/pii/S0022519316303150, doi: 10.1016/j.jtbi.2016.09.018.

Lye LF, Akopyants NS, Dobson DE, Beverley SM. A Narnavirus-Like Element from the Trypanosomatid Protozoan Parasite Leptomonas seymouri. Genome Announc. 2016 Aug; 4(4):e00713–16. https://mra.asm.org/content/4/4/e00713-16, doi: 10.1128/genomeA.00713-16.

Merino E, Balbás P, Puente JL, Bolívar F. Antisense overlapping open reading frames in genes from bacteria to humans. Nucleic Acids Res. 1994 May; 22(10):1903–1908. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC308092/.

Osaki H, Sasaki A, Nomiyama K, Tomioka K. Multiple virus infection in a single strain of *Fusarium poae* shown by deep sequencing. Virus Genes. 2016 Dec; 52(6):835–847. doi: 10.1007/s11262-016-1379-x.

Rancurel C, Khosravi M, Dunker AK, Romero PR, Karlin D. Overlapping Genes Produce Proteins with Unusual Sequence Properties and Offer Insight into De Novo Protein Creation. J Virol. 2009 Oct; 83(20):10719–10736. https://jvi.asm.org/content/83/20/10719, doi: 10.1128/JVI.00595-09.

Sabath N, Wagner A, Karlin D. Evolution of Viral Proteins Originated De Novo by Overprinting. Mol Biol Evol. 2012 Dec; 29(12):3767–3780. https://academic.oup.com/mbe/article/29/12/3767/1006311, doi: 10.1093/molbev/mss179.

Schlub TE, Buchmann JP, Holmes EC. A simple method to detect candidate overlapping genes in viruses using singl genome sequences. Mol Biol Evol. 2018 Oct; 35(10):2572–2581. https://academic.oup.com/mbe/article/35/10/2572/5067730, doi: 10.1093/molbev/msy155.

Shi M, Neville P, Nicholson J, Eden JS, Imrie A, Holmes EC. High-Resolution Metatranscriptomics Reveals the Ecological Dynamics of Mosquito-Associated RNA Viruses in Western Australia. J Virol. 2017; 91(17). doi: 10.1128/JVI.00680-17.

Shukla A. Analysis of overlapping reading frames in viral genomes. PhD thesis, University of Lübeck; 2015.

Smith TF, Waterman MS. Protein Constraints induced by Multiframe Encoding. Mathematical Biosciences. 1980; 49:17–26.

Staden R. Measurements of the effects that coding for a protein has on a DNA sequence and their use for finding genes. Nucleic Acids Res. 1984 Jan; 12(1 Pt 2):551–567. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC321070/.

Tretyachenko V, Vymětal J, Bednárová L, Kopecký V, Hofbauerová K, Jindrová H, Hubálek M, Souček R, Konvalinka J, Vondrášek J, Hlouchová K. Random protein sequences can form defined secondary structures and are well-tolerated *in vivo*. Sci Rep. 2017; 7:15449. doi: 10.1038/s41598-017-15635-8.

Waldron FM, Stone GN, Obbard DJ. Metagenomic sequencing suggests arthropod antiviral RNA interference is highly derived, and identifies novel viral small RNAs in a mollusc and a brown alga. bioRxiv. 2017; https://www.biorxiv.org/content/early/2017/07/21/166488, doi: 10.1101/166488.

Wesolowski M, Wickner RB. Two new double-stranded RNA molecules showing non-mendelian inheritance and heat inducibility in Saccharomyces cerevisiae. Mol Cell Bio. 1984 Jan; 4(1):181–187. doi: 10.1128/mcb.4.1.181.

Appendix 1

Alternative forward reading frames

In the main text we focused on the possibility of ORFs in the reverse strand, because that was the situation relevant for narnaviruses. However, since the arguments presented in this paper are just based on the genetic code and not specific to these viruses, it is worth considering the alternative reading frames in the forward direction as well.

Frame $+2$

Consider first the case where the new gene is read in the forward direction and left shifted by a single base (frame $+2$ in the convention of Figure 2). For example, if the codons of the $+0$ sequence are

$$\begin{array}{llllllllll} 5' & \ldots & \mid x & y & \mathrm{U} & \mid \mathrm{A} & \mathrm{A} & z & \mid \ldots & 3' \quad +0 \text{ frame} \\ 5' & \ldots & \bar{x} & \bar{y} & \mid \mathrm{U} & \mathrm{A} & \mathrm{A} & \mid \bar{z} & \ldots & 3' \quad +2 \text{ frame} \end{array}$$

so that there is a stop codon in the new reading frame.

In this case, replacing xyU with xyw, $w \neq$ U, removes the stop codon in the new frame. In seven cases (Leu, Pro, Arg, Thr, Val, Ala, Gly) any of the four possibilities for w gives a synonym. For another six cases (Phe, Tyr, Cys, His, Asn, Asp), a synonym substitution is also possible but $w =$C is the only option. In the case of Ser, UCU allows three changes but AGU allows only AGC. For Ile there are two possible changes and the remaining amino acids can never have U in the final position. Hence, a synonym for the first codon that removes the stop can always be found.

Because, when the first base is A, changing either the first or the second base always changes the amino acid, we cannot construct a synonym by single-nucleotide mutation of the second codon, and there are no other cases to consider. The same reasoning holds true if AA in the second codon is replaced by AG or GA (the other two possible combinations that would yield a stop in the $+2$ frame). So, in the case of a left shift, synonym substitution is possible.

Frame $+1$

In the case of a right shift (frame $+1$), not all stop codons can be removed by a synonym transformation. For example if the codons in the $+0$ frame are

$$\begin{array}{llllllllll} 5' & \ldots & \mid x & \mathrm{U} & \mathrm{A} & \mid \mathrm{A} & y & z & \mid \ldots & 3' \quad +0 \text{ frame} \\ 5' & \ldots & \bar{x} & \mid \mathrm{U} & \mathrm{A} & \mathrm{A} & \mid \bar{y} & \bar{z} & \ldots & 3' \quad +1 \text{ frame} \end{array}$$

then we want to remove the stop codon UAA by a single-base substitution which gives a synonym, for any choice of x. If $x =$ U, then we have to find a replacement for the second U or the first A that still codes for Leu. The only possible synonym of Leu that involves changing the second or third letter is UUG, but this is unsatisfactory because UGA is another stop codon. Similarly, it is impossible to obtain a synonym by changing the A in the second codon. The same reasoning holds if we consider the UAG or UGA stop codons. So the Leu codon UUA followed by a codon beginning with either A or G results in a non-removable stop. Similarly, the Leu codon UUG followed by a codon beginning with A results in a stop in the new frame that cannot be removed by a single-nucleotide substitution (although it can be removed by two substitutions, for example UUG → CUG → CUC).

We conclude that it is possible for new ORFs, transcribed in the forward direction, to overlap a perfectly conserved gene. If we exclude cases requiring more than one substitution, this can only happen in the $+2$ frame.

Appendix 2

Quantifying the evolutionary space for the companion gene

In a standard gene, a sequence of N codons allows for $\mathcal{N} = 20^N$ different combinations of amino acids. In the case where a new gene overlaps an existing gene which is perfectly conserved, we must confine our choices to the set of amino acids which correspond to synonyms in the original gene. If the codon for position j allows n_j different synonyms, the number of possible amino-acid sequences is

$$\mathcal{N} = n_1 n_2 \cdots n_N \ .$$

This number grows rapidly with the length of the sequence:

$$\mathcal{N} \approx \mathrm{e}^{sN}$$

where s describes the amount of freedom that the polypeptide sequence has (in physics, it would be termed an *entropy*). For unconstrained evolution we have an entropy per codon equal to

$$s_0 = \ln 20 \ .$$

This is a measure of the range of possible proteins that can be constructed in the original gene.

In the cases that we analyze, a new coding sequence in an alternate frame is constrained by the requirement that the amino-acid sequence of the original gene is conserved, so that codons must be chosen from the set of synonyms. The number of possible sequences for the new gene is much smaller, but still grows exponentially with the length of the sequence. Because there are 20 amino acids and 64 codons, there are approximately three possible choices for each codon, so that the entropy of the new gene is (approximately) $s_1 \approx \ln 3$. Let us consider this more precisely for the case where the new gene is evolving in the -0 frame (that is, reverse complement sequence, with codons aligned). In this case the entropy of the new sequence is

$$s_1 = \sum P_i \ln n_i$$

where P_i is the fraction of amino acids of type i, and n_i is the number of distinct amino acids which can arise from the reverse complement of each synonym. For example, if the amino acid of the original gene is Ser, there are six possible synonyms (UC*, AGU, AGC). The complementary codons, (UGA, CGA, AGA, GGA, ACU, GCU) represent, respectively Stop, Arg, Arg, Gly, Thr, Ala, so that for i = Ser, the number of distinct codons, excluding Stop, is $n_{\mathrm{Ser}} = 4$. The corresponding numbers for all of the amino acids are $n_{\mathrm{Phe}} = 2$, $n_{\mathrm{Leu}} = 4$, $n_{\mathrm{Ile}} = 4$, $n_{\mathrm{Val}} = 4$, $n_{\mathrm{Ser}} = 4$, $n_{\mathrm{Pro}} = 3$, $n_{\mathrm{Thr}} = 4$, $n_{\mathrm{Ala}} = 4$, $n_{\mathrm{Tyr}} = 2$, $n_{\mathrm{His}} = 1$, $n_{\mathrm{Gln}} = 1$, $n_{\mathrm{Asn}} = 2$, $n_{\mathrm{Lys}} = 2$, $n_{\mathrm{Asp}} = 2$, $n_{\mathrm{Glu}} = 2$, $n_{\mathrm{Cys}} = 2$, $n_{\mathrm{Trp}} = 1$, $n_{\mathrm{Arg}} = 4$, $n_{\mathrm{Gly}} = 4$. Using the codon frequencies P_i determined from the Culex narnavirus 1 sequence, we find $s_1 \approx 1.104$, which is remarkably close to the value $s_1 \approx \ln 3 \approx 1.099$ estimated in the previous paragraph.

Because s_1 is significantly smaller than s_0, the set of possible sequences which can be coded by the new gene is quite restricted. In particular, in a sequence of length N the ratio of the number of possible amino-acid sequences between a standard gene ($\mathcal{N}_0$) and the constrained one ($\mathcal{N}_1$) is

$$\frac{\mathcal{N}_0}{\mathcal{N}_1} \approx \mathrm{e}^{(s_0 - s_1)N} \approx 6.63^N \ .$$

For $N \sim 1000$ (as in narnaviruses), this ratio would be $\sim 10^{820}$. The constraint is eased if we allow changes in amino acids of the original protein as well as synonyms. It is expected that

the actual evolution of the virus genome will represent a compromise between conserving the function of the original protein and allowing scope for evolution of the new protein.

Appendix 3

Exploration of the possible secondary structure in the rORF

The translated rORFs of ambigrammatic narnaviruses are predicted to have median α-helical and β-strand contents of 22 % and 12 % respectively (calculated using JPred4, ***Drozdetskiy et al., 2015***). This degree of secondary structure is consistent with a structured (or folded) protein, but a significant presence of secondary structure can also be observed in random amino-acid sequences (***Tretyachenko et al., 2017***). We further note that the isoelectric point (PI) of the RdRp is high (median 10.4, range 7.7–11.6 for sequences >2 kb in Figure 3), due to a high frequency of Arg, a basic amino acid (median 9.9 %, range 6–13.3 %). Notice that this does not necessarily lead to a high concentration of Arg in the rORF.[a] The translated rORFs also have high PIs (median 10.2, range 8.3–11.4), which does not seem to be dictated by the amino-acid composition in the RdRp, and yet is similar to PIs calculated for translations of the -0 frame for non-ambigrammatic narnaviruses (median 10.1, range 6.8–12.7). Basic residues can be involved in binding negatively-charged nucleic acids, but without experimental information we can only speculate on the role for a putative protein translated from the rORF.

[a]The codons for Arg are CG*, AGA and AGG, none of which leads to an Arg in the -0 frame.