Submitted to Virus Evolution, 2021, pp. 1–9

doi: DOI HERE
Advance Access Publication Date: Day Month Year
Paper

PAPER

Polymorphism of Genetic Ambigrams

Gytis Dudas,[1,†] Greg Huber,[2,†] Michael Wilkinson[2,3,*,†] and David Yllanes[2,†]

[1]Gothenburg Global Biodiversity Centre, Carl Skottsbergs gata 22B, 413 19, Gothenburg, Sweden, [2]Chan Zuckerberg Biohub, 499 Illinois Street, San Francisco, CA 94158, USA and [3]School of Mathematics and Statistics, The Open University, Walton Hall, Milton Keynes, MK7 6AA, UK
[*]Corresponding author. michael.wilkinson@czbiohub.org [†]Authors in alphabetical order
FOR PUBLISHER ONLY Received on Date Month Year; revised on Date Month Year; accepted on Date Month Year

Abstract

Double synonyms in the genetic code can be used as a tool to test competing hypotheses regarding ambigrammatic narnavirus genomes. Applying the analysis to recent observations of *Culex narnavirus 1* and *Zhejiang mosquito virus 3* ambigrammatic viruses indicates that the open reading frame on the complementary strand of the segment coding for RNA-dependent RNA polymerase does *not* code for a functional protein. *Culex narnavirus 1* has been shown to possess a second segment, also ambigrammatic, termed 'Robin'. We find a comparable segment for *Zhejiang mosquito virus 3*, a moderately diverged relative of *Culex narnavirus 1*. Our analysis of Robin polymorphisms suggests that its reverse open reading frame also does not code for a protein. We make a hypothesis about its role.

Introduction

Of all the various types of viruses catalogued, narnaviruses rank among the simplest and most surprising (Cobián Güemes et al., 2016). Narnaviruses (a contraction of 'naked RNA virus') are examples of a minimal blueprint for a virus: no capsid, no envelope, no apparent assembly of any kind. The known narnavirus blueprint appeared for all intents and purposes to be a single gene, that which codes for an RNA-dependent RNA polymerase, abbreviated as RdRp, (Hillman and Cai, 2013). However, some narnaviruses have been found to have a genome with an open reading frame (i.e., a reading frame without stop codons) on the strand complementary to that coding for the RdRp gene, calling into question the general hypothesis of a one-gene blueprint (DeRisi et al., 2019; Dinan et al., 2020; Cepelewicz, 2020). This reverse open reading frame (rORF) has codon boundaries aligned with the forward reading frame. Because the genome can be translated in either direction, we say that these narnaviruses are *ambigrammatic*. The significance of an ambigrammatic genome is an open problem. In this paper we discuss how polymorphisms of sampled sequences can distinguish between competing hypotheses on the function and nature of ambigrammatic viral genomes. Our methods are applied to known ambigrammatic narnavirus genes and to the newly discovered ambigrammatic second segment of some narnaviruses, termed *Robin* (Batson et al., 2020).

Our discussion is based upon two rules about the genetic code and its relation to ambigrammatic sequences. Both of these *ambigram rules* are concerned with the availability of synonyms within the genetic code, which allow coding of the same amino acid with a different codon. The first rule states that for any sequence of amino acids coded by the forward strand, it is possible to use individual synonymous substitutions to remove all stop codons on the complementary strand (this result was discussed already in DeRisi et al., 2019). The second ambigram rule, described below, states that the genetic code contains double synonyms that allow polymorphisms, accessible by single-base mutations, even when the amino acids coded by both the forward and the complementary strands are fixed.

The first of these rules addresses the 'how' of ambigrammatic genomes, by showing that stop codons on the complementary strand can be removed by single-point mutations, without altering the protein (in narnaviruses, the RdRp) coded in the forward direction. Here we argue that the second rule can help to resolve the 'why' of ambigrammatic genomes: the origin of ambigrammaticity itself. There are two distinct reasons why there might be an evolutionary advantage for a virus to evolve an ambigrammatic sequence. The first possibility is that the complementary strand might code for a functionally significant protein, for example, one that might interfere with host defence mechanisms. The second possibility is that the lack of stop codons on the complementary strand is significant, even if the

amino acid sequence that is coded is irrelevant. In particular, the lack of stop codons may promote the association between ribosomes and the complementary strand viral RNA (produced as part of its replication cycle). It is possible that a 'polysome' formed by a covering of ribosomes helps to shield the virus from degradation or from detection by cellular defence mechanisms (Cepelewicz, 2020; Retallack et al., 2020; Wilkinson et al., 2021). The second ambigram rule combined with data on the polymorphism of the virus genome can help distinguish whether the complementary strand codes for a functional protein. We shall argue that in the case of *Culex narnavirus 1* and *Zhejiang mosquito virus 3*, the evidence is in favour of this second hypothesis, namely that the open reading frame (ORF) on the complementary strand does not code for a functional protein.

After describing the genetic ambigram rules, we discuss how the existence of double synonyms can be used to assess whether the open reading frame on the complementary chain codes for functional protein. It is well known that, because RdRp is a highly-conserved gene, non-synonymous mutations are likely to be detrimental, so that most of the observed diversity consists of synonymous changes. Some of these synonymous mutations have the potential to be synonymous in the complementary strand. If the complementary strand also codes for a functional protein, we expect that doubly synonymous mutations will be favoured. In fact, there would be mutational 'hotspots' corresponding to the potential doubly-synonymous loci. We introduce two tests for whether the complementary strand is coding, based respectively on looking for mutational 'hotspots', and upon the mutational frequencies at loci which have double synonyms. We used these tests to analyse sequences for two different ambigrammatic narnaviruses: 46 RdRp segments of *Culex narnavirus 1* and 12 RdRp segments of *Zhejiang mosquito virus 3*, abbreviated to CNV and ZMV respectively. We find that neither of our tests supports the hypothesis that the translated sequence of the complementary strand of RdRp is under purifying selection. We also applied these tests to the second segment, termed *Robin*, which is found to be closely associated with this ambigrammatic narnavirus infection in mosquitos (Batson et al., 2020; Retallack et al., 2020). We also found that the complementary open reading frame of Robin does not appear to be under purifying selection. The discovery of Robin suggested that ambigrammatic companions may exist for other ambigrammatic viruses. Accordingly, we searched the assembled contigs of studies reporting the detection of ZMV, the only other ambigrammatic narnavirus observed multiple times in numerous locations, and discovered an ambigrammatic segment with similar properties to CNV Robin. Thus we consider four viral segments, denoted CNV-RdRp, CNV-Robin, ZMV-RdRp, ZMV-Robin. We shall report evidence that Robin does code for a protein in its forward direction, but that its complementary strand is non-coding. We find evidence that Robin segments are under detectable purifying selection. Figure 1 illustrates the phylogenetic relationship of CNV and ZMV, and ORF-wide dN/dS values of all their segments and coding directions (discussed in detail below).

Some careful consideration is required to reconcile our observations with results recently reported in Retallack et al. (2020), where it was shown that introducing mutations which are non-synonymous on the reverse open reading frame of *Culex narnavirus 1* can reduce the fitness of this virus. In the concluding section, we consider the interpretation of these observations, and discuss whether there may be implications for other viral families.

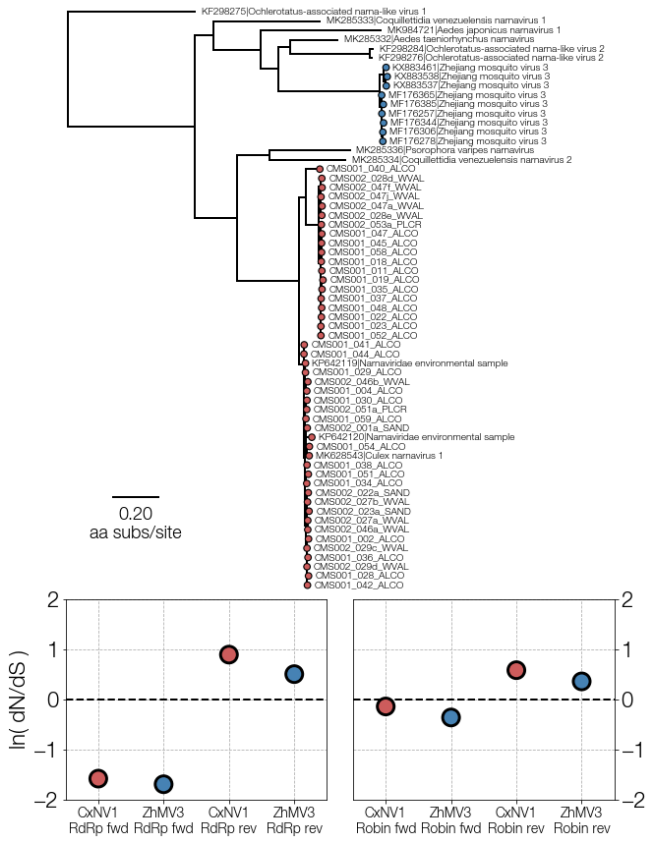

Fig. 1. **a** A maximum-likelihood tree illustrating the relationship between CNV (*Culex narnavirus 1*) (red) and ZMV (*Zhejiang mosquito virus 3*) (blue). **b** ORF-wide dN/dS values for forward and reverse directions of RdRp and Robin segments for both viruses.

There are many examples of overlapping viral genes with staggered reading frames: this was first clearly described in Barrell et al. (1976), and has been reviewed in Chirico et al. (2010). Recent work by Nelson, Ardern and Wei (Nelson et al., 2020) discusses how these can be identified. Our investigations indicate that the ambigrammatic ORFs discussed in this work are a different phenomenon, because they are non-coding. Our approach to analysing the ambigrammatic sequences is quite distinct from the rather complex machinery proposed in Nelson et al. (2020), because it emphasises the role of double synonyms as an unambiguous discriminant of the role of the ambigrammatic sequences.

Ambigram rules and their significance

We start by describing the two genetic ambigram rules.

Rule 1 *All complementary-strand stops are removable*

Consider the reading frame on the complementary strand that has its codons aligned with those on the forward strand. Every codon on the forward strand corresponds to a complementary-strand codon read in the reverse direction. The rule states that any stop codon on the complementary strand can be removed by a single-point mutation which leaves the amino acid specified by the forward-read codon unchanged.

This result is demonstrated by the following argument, as discussed in DeRisi et al. (2019). Reversing the read direction and taking the pairing complement, the stop codons UAA, UAG, UGA in the standard genetic code become, respectively, UUA, CUA, UCA, for which the amino acids are Leu, Leu, Ser. It is only instances of leucine and serine in the forward sequence that can result in stop codons in the reverse read. The synonyms of Leu are CUN, UUA, UUG (where N means any base). The synonyms of Ser are UCN, AGU, AGC. The undesirable 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. And the Ser codon UCA is transformed to UCU, UCG or UCC by single substitutions. We conclude that every stop codon on the reverse reading frame can be removed by a synonymous, single site nucleotide mutation.

Furthermore, it is found that complementary-strand stops cannot always be removed by synonymous substitutions in the other two read frames for the complementary strand (each case requires a separate and somewhat involved argument, also given in DeRisi et al., 2019). As a consequence of these two arguments, we need discuss only the complementary read frame with aligned codons.

Rule 2 *There exist double synonyms*

Most synonymous mutations of the forward strand produce a non-synonymous change in the complementary strand, but the genetic code does include a number of double synonyms, where the reverse complement of a synonymous mutation is also a synonym. For example codon AGG (Arg) can become CGG (Arg) via a synonymous mutation, while the reverse complement of AGG, which is CCU (Pro) transforms to CCG (Pro) under the same mutation.

The full set of double synonyms in the standard genetic code are as follows:

- Two of the six synonyms of Ser are double synonyms, with reverse complements coding Arg. Conversely, two of the six synonyms of Arg are double synonyms, with reverse complement coding Ser.
- Two more of the six synonyms of Arg are double synonyms, with reverse complement Pro. Conversely, two of the four synonyms of Pro are double synonyms coding for Arg.
- Two of the six synonyms of Leu are double synonyms, with reverse complement Gln. Conversely, both synonyms of Gln are double synonyms, with reverse complement coding Leu.

Table 1 lists the sets of single and double synonyms for those amino acids that can have double synonyms. (We exclude the two synonyms of Ser and the one synonym of Leu for which the reverse complement is Stop, because these do not occur in ambigrammatic genes.)

Table 1. For each amino acid (AA) that can have double-synonym mutations, we list all of the possible codons which do not code for Stop on the complementary strand, indicating their reverse complement (Comp. AA). The codons that have a double synonym are marked with an asterisk. For each of these codons, we list the number of mutations which are synonymous, and the number of double synonym mutations. In each case the numbers of single (double) mutations are written $S^{(\mathrm{n})}+S^{(\mathrm{v})}$ ($D^{(\mathrm{n})}+D^{(\mathrm{v})}$), where the superscript n denotes transitions, and superscript v transversions. Also, double synonyms are counted in the list of single synonyms.

AA	Codon	$S^{(\mathrm{n})}+S^{(\mathrm{v})}$	$D^{(\mathrm{n})}+D^{(\mathrm{v})}$	Comp. AA
Leu	UUG*	$1+0$	$1+0$	Gln
	CUU	$1+1$	$0+0$	Lys
	CUC	$1+1$	$0+0$	Glu
	CUG*	$1+2$	$1+0$	Gln
Pro	CCU*	$1+2$	$0+1$	Arg
	CCC	$1+2$	$0+0$	Gly
	CCA	$1+2$	$0+0$	Trp
	CCG*	$1+2$	$0+1$	Arg
Gln	CAA*	$1+0$	$1+0$	Leu
	CAG*	$1+0$	$1+0$	Leu
Arg	CGU	$1+2$	$0+0$	Thr
	CGC	$1+2$	$0+0$	Ala
	CGA*	$1+3$	$0+1$	Ser
	CGG*	$1+3$	$0+1$	Pro
	AGA*	$1+1$	$0+1$	Ser
	AGG*	$1+1$	$0+1$	Pro
Ser	UCU*	$1+1$	$0+1$	Arg
	UCC	$1+1$	$0+0$	Gly
	UCG*	$1+2$	$0+1$	Arg
	AGU	$1+0$	$0+0$	Thr
	AGC	$1+0$	$0+0$	Ala

Implications

Our first rule shows that an ambigrammatic version of any gene can evolve, without making any changes to the amino acid sequence. This establishes how ambigrammatic sequences can arise, but it does not illuminate why they are favoured.

Combined with observed polymorphisms of narnaviruses, the second ambigram rule can give an indication of the utility of ambigrammatic sequences. In studies on the (usual) non-ambigrammatic genomes, the ratio of synonymous to non-synonymous mutations is used as an indicator of whether the nucleotide sequence codes for a protein: non-synonymous mutations are likely to be deleterious if the sequence codes for a functional protein. We shall adapt this approach to our study of ambigrammatic narnavirus genes. We assume that the forward direction is a coding sequence (usually for RdRp), and confine attention to those mutations which are synonymous in the forward direction. If the complementary strand codes for a functional protein, most of these synonymous mutations will inevitably result in changes of the complementary amino acid sequence. However, at many loci the evolutionarily favoured amino acid will be one that allows double synonyms. In these cases, there can be non-deleterious mutations between a pair of codons that preserve the amino acid sequence of both the forward and the complementary strands.

If the complementary strand codes for a functional protein, we expect studies of the polymorphism of the gene would show that these double-synonym loci will be mutational 'hotspots', where mutations occur more frequently. In addition, the double-synonym pairs would be represented far more frequently than other mutations at these loci. These observations lead to two distinct tests for whether there is evolutionary pressure on the translated sequence of the complementary strand.

Ambigrammatic narnavirus genes

We analysed data from samples of two ambigrammatic narnaviruses, *Culex narnavirus 1* (CNV, with 46 genomes) and *Zhejiang mosquito virus 3* (ZMV, with 10 genomes). Both narnaviruses have an ambigrammatic RdRp coding gene, denoted CNV-RdRp and ZMV-RdRp respectively. The reverse open reading frame has its codons aligned with the forward frame. In both forward and reverse reading frames any stop codons are close to the $3'$ end of the respective frame.

The ambigrammatic feature is certainly a puzzle. There appear to be two classes of plausible explanations:

1. **The reverse open reading frame codes a protein**. This is logically possible, but if the RdRp gene is strongly conserved, there is very little flexibility in the rORF. However, in the absence of any additional evidence it is the explanation which requires the fewest additional hypotheses.
2. **The reverse open reading frame facilitates association of ribosomes with RNA**. This could conceivably convey advantages by providing a mechanism to protect viral RNA from degradation, but without further evidence this requires additional hypotheses.

Recently, additional evidence has emerged which may provide support for the second of these explanations. Specifically, the CNV infection has recently been shown to be associated with another ambigrammatic viral RNA segment, termed *Robin* (Batson et al., 2020; Retallack et al., 2020). It was reported that this segment, CNV-Robin, is ambigrammatic, with forward and reverse codons aligned, over very nearly the entire length (about 850 nt), where direction designation is determined by which amino acid sequence appears more conserved. Again, any stop codons occur close to the $3'$ end. Neither forward nor reverse directions of Robin are homologous with known sequences.

Because ambigrammatic genes are rare, finding two of them in the same system is a strong indication that their occurrence has a common explanation. This observation makes it appear unlikely that the reverse open reading frame is a device to 'pack in' an additional protein coding gene, and more likely that the ambigrammatic feature is associated with allowing ribosomes to associate with both strands of the viral RNA.

This reasoning suggests that the Robin gene may play a role in selecting for the ambigrammatic property (for example, it may facilitate protection by ribosomes of the viral RNA). If this surmise is correct, we should expect to see a version of the Robin gene associated with other ambigrammatic narnaviruses. It is possible that this might be detected by a search of archived sequence data. Only *Zhejiang mosquito virus 3* appeared to be observed multiple times to make detection of an additional Robin segment possible, so we concentrated on that system.

We were able to find evidence of an ambigrammatic RNA, of length approximately 900 nt, that co-occurs with ZMV RdRp segment across multiple samples recovered by at least two studies that, like CNV Robin, bears no recognisable homology to publicly available sequences or CNV Robin itself. Given the conjunction of these unusual features we strongly believe this ambigrammatic RNA to be the equivalent of a Robin segment in ZMV.

Methods

Tests for whether the complementary strand is coding

We have argued that doubly-synonymous mutations will give a signature of the reverse strand coding for a functional protein. If the reverse-direction code is functional, then the only assuredly non-deleterious mutations would be the double-synonym ones, where one codon is transformed by a single-nucleotide substitution to another codon which preserves the amino acid coded in both the forward and the reverse directions.

Assume that we have M sequences of an ambigrammatic gene, fully sequenced and maximally aligned with each other, and that one strand, referred to as the 'forward' strand, codes for a functional protein. We identify a 'consensus' codon at each of the N loci, and then enumerate the set of variant codons at each amino acid locus. If the consensus codon at a locus is one of the twelve double-synonym codons listed in table 1, we term this a *doubly-synonymous locus*. The number of doubly-synonymous loci is $N_{\rm ds}$.

There are two different approaches to testing whether double synonyms indicate that the complementary strand is coding:

Look for the existence of mutational 'hotspots'

We can look for evidence that the doubly-synonymous loci experience more substitutions than other loci.

For each codon locus k, we can determine the number of elements of the variant set, $n(k)$, and also the fraction of codons $f(k)$ which differ from the consensus codon. We then determine the averages of these quantities, $\langle n(k)\rangle$ and $\langle f(k)\rangle$, for the doubly-synonymous loci and for the other loci. If the ratios

$$R_n = \frac{\langle n(k)\rangle|_{\text{double syn. loci}}}{\langle n(k)\rangle|_{\text{other loci}}} \ , \quad R_f = \frac{\langle f(k)\rangle|_{\text{double syn. loci}}}{\langle f(k)\rangle|_{\text{other loci}}} \tag{1}$$

are large, this is evidence that the complementary strand is coding.

The null hypothesis, indicating that the reverse open reading frame is non-coding, is that the ratios R_n and R_f are sufficiently close to unity that the difference may be explained by statistical fluctuations.

Mutation frequencies test

We can also look at codon frequencies for different mutations at doubly-synonymous loci. If the complementary strand is coding, we expect to find that the frequency of mutations observed at doubly-synonymous loci will heavily favour double-synonym codons over single synonyms. We consider the subset of double-synonym loci where mutations are observed (that is, where $n(k) > 1$). For each of these $N_{\rm a}$ *variable doubly-synonymous loci*, we can determine two numbers: $n_{\rm s}(k)$ is the numbers of singly-synonymous variants at locus k, and $n_{\rm d}(k)$ is the number of these variants which are also doubly-synonymous. (Clearly $n(k) \geq n_{\rm s}(k) \geq n_{\rm d}(k)$). If $n_{\rm d}(k) = n_{\rm s}(k)$, that means that the mutations preserve the complementary-strand amino acid, which is an indication that the reverse strand is coding. If $\{k^*\}$ is the set of variable doubly-synonymous loci, we then calculate

$$N_{\rm s} = \sum_{k\in\{k^*\}} n_{\rm s}(k) \ , \quad N_{\rm d} = \sum_{k\in\{k^*\}} n_{\rm d}(k) \ . \tag{2}$$

If the complementary strand is coding, we expect

$$R \equiv \frac{N_{\rm s}}{N_{\rm d}} \tag{3}$$

to be close to unity.

However, there will also be beneficial or neutral mutations which do change the amino acids, so that not all mutations will be between sets of doubly-synonymous codons. We need to be able to quantify the extent to which finding other than double-synonym mutations is an indication that the reverse strand is non-coding. We must do this by comparison with a null hypothesis, in which the reverse strand is non-coding.

Null hypothesis for mutation frequencies

Let R_0 be the value of the ratio R that is derived from this null hypothesis that the complementary strand is non-coding. In order to compute the expected $N_\mathrm{s}/N_\mathrm{d}$ ratio, R_0, we adopt the following approach. We assume that the M sequences are sufficiently similar that only a small fraction of loci have undergone mutations. We adopt the Kimura model (Kimura, 1980), which assumes that the mutation rate r_n for transitions (A $\leftrightarrow$ G or C $\leftrightarrow$ U) is different from the rate r_v for transversions (other single-nucleotide mutations), and negligible for other types of mutation. The ratio of these rates is

$$\alpha = \frac{r_\mathrm{n}}{r_\mathrm{v}} \ . \tag{4}$$

If the numbers of single (double) synonyms of the consensus nucleotide at locus k leading to transitions or transversions are respectively $S_k^{(\mathrm{n})}$ and $S_k^{(\mathrm{v})}$ ($D_k^{(\mathrm{n})}$, $D_k^{(\mathrm{v})}$), then we estimate

$$R_0 = \frac{\sum_{k\in\{k^*\}} \alpha S_k^{(\mathrm{n})} + S_k^{(\mathrm{v})}}{\sum_{k\in\{k^*\}} \alpha D_k^{(\mathrm{n})} + D_k^{(\mathrm{v})}} \tag{5}$$

The numbers $S_k^{(\mathrm{n})}$, $S_k^{(\mathrm{v})}$, $D_k^{(\mathrm{n})}$, $D_k^{(\mathrm{v})}$ are given in table 1 for all of the double-synonym codons.

Finding the Robin segment of *Zhejiang mosquito virus 3*

We looked through assembled contig datasets from two metagenomic mosquito studies (three from China and six from Australia) (Shi et al., 2016, 2017), kindly provided to us by Mang Shi and Edward C Holmes. We clustered contigs from the nine datasets by similarity using CD-HIT (Fu et al., 2012) with a threshold of 90% and looked for clusters that contained contigs from at least 6 samples, that did not have standard deviation in contig length greater than 1200, and had fewer than 200 contigs. Of the hundreds of clusters filtered this way only a handful also possessed sequences ambigrammatic across at least 90% of their length and only two clusters were mostly comprised of ambigrammatic sequences, while the rest were clearly recognisable as mosquito contigs. Of the two clusters one was identifiable as the RdRp of *Zhejiang mosquito virus 3*, while we presume the other to be an unrecognisably distant orthologue of *Culex narnavirus 1* Robin, on account of its co-occurrence with ZMV RdRp, ambigrammaticity, and length.

Results

Next we report the results of our studies of polymorphism of the four ambigrammatic narnavirus genes. We discuss what can be learned from applying standard techniques, before discussing the results of our tests for whether the reverse open reading frame codes for a protein.

Forward reading frame

Each sequence was trimmed to a length of $3N$ nucleotides. We identified a consensus nucleotide at each locus, and determined the set of variant nucleotides at each locus. We determined the total number of transition and transversion mutations which are observed, N_n and N_v respectively. We also determined the total number of mutations at each position in the codon, (n_1, n_2, n_3). We estimated the average number of variable sites r as the total number of nucleotide variants, divided by the product of the number of sequences and alignment length. We also estimated the ratio α of the rate of selected transition mutations to the rate of transversions:

$$r \equiv \frac{n_1+n_2+n_3}{3NM} \ , \quad \alpha \equiv \frac{r_\mathrm{n}}{r_\mathrm{v}} = \frac{2N_\mathrm{n}}{N_\mathrm{v}} \tag{6}$$

(recall that there are twice as many transversions as transitions). We also determined a 'normalised' triplet of variable sites for each position within the codon: $(z_1 : z_2 : z_3) = 3(n_1 : n_2 : n_3)/(n_1+n_2+n_3)$. Our results on the nucleotide-level investigation of polymorphism are summarised in table 2.

We then assigned a consensus codon at each codon locus, selecting the frame by the criterion of minimising the number of stop codons. For each of the N codons, we determined the variant set of codons which were observed in each of the M sequences. The total number of synonymous and non-synonymous single-nucleotide changes in the variant sets was N_sy and N_ns respectively. The total number of mutations encountered in the variant sets where two or three nucleotides were changed was N_mult. For each codon there are numbers of possible non-synonymous mutations which are transistions and transversions, $n_k^{(\mathrm{n})}$ and $n_k^{(\mathrm{v})}$, and numbers of synonymous mutations which are transitions and transversions, $s_k^{(\mathrm{n})}$ and $s_k^{(\mathrm{v})}$ (with $s_k^{(\mathrm{n})} + n_k^{(\mathrm{n})} + s_k^{(\mathrm{v})} + n_k^{(\mathrm{v})} = 9$). Under the null hypothesis that the sequence is non-coding, the expected value of the ratio

$$R = \frac{N_\mathrm{ns}}{N_\mathrm{sy}} \tag{7}$$

is

$$R_\mathrm{exp} = \frac{\sum_{k=1}^{N} \alpha n_k^{(\mathrm{n})} + n_k^{(\mathrm{v})}}{\sum_{k=1}^{N} \alpha s_k^{(\mathrm{n})} + s_k^{(\mathrm{v})}} \ . \tag{8}$$

We also determined the fraction of codons where multi-nucleotide mutations are observed, $f_\mathrm{mult} = N_\mathrm{mult}/N$. We present our results for the codon-level mutations in table 3, which includes information for both the forward and the complementary read directions (with codon boundaries aligned for the complementary direction).

The alignments are *ambigrammatic*, in the sense that there are no stop codons in the interior of the sequence. None of the individual sequences had stop codons in the body of the sequence in either direction.

We also computed ORF-wide $\mathrm{d}N/\mathrm{d}S$ values (plotted in figure 1(**b**)), by assuming that every mutation in the alignment has occurred only once to be conservative. This was motivated by the presence of pairs of sites with four haplotypes between them (4G sites), an indication that recombination may be a potential issue with narnavirus sequences. Normalising the number of observed non-synonymous and synonymous mutations was done by assuming a transition/transversion ratio of 2, consistent with equation (6). These values $\mathrm{d}N/\mathrm{d}S$ values are slightly different from the R/R_exp ratios in table 3 because the latter excludes mutations where more than one base differs from the consensus codon. In all but one of the cases $\mathrm{d}N/\mathrm{d}S$ is higher than R/R_exp, because the multiple nucleotide mutations which are included in $\mathrm{d}N/\mathrm{d}S$ are predominantly non-synoymous.

Based upon these tables, we can make the following observations and deductions:

1. **Diversity**. We observe that both RdRp and Robin segments are comparable in their diversity, for both CNV and ZMV. As expected, RdRp sequences are highly conserved at the amino acid level. Robin, on the other hand, appears far more relaxed at the amino acid level and, consistent with this, diverged beyond recognition between CNV and ZMV.
2. **Relative mutation rate by codon position**. For RdRp sequences, more mutations are observed at the third nucleotide in each codon, as expected for a sequence that preserves the amino acid sequence (because most synonymous mutations involve the third nucleotide of a codon). In the case of Robin sequences, the frequencies of mutation are much closer to being equal, to the extent that for CNV-Robin the null hypothesis that the rates are equal is not definitively rejected. However, mutations at different codon sites are sufficiently weighted towards the third position that we shall assume that Robin does code for a functional protein.

   While the values of $(z_1 : z_2 : z_3)$ are very different for RdRp and Robin, their values are comparable for CNV and ZMV, which is an indication that the selective pressures on both viruses are the same.
3. **Rate of multiple-nucleotide mutations**. The fraction of multiple-nucleotide mutations is higher for Robin sequences than it is for RdRp sequences. This may be an indication that the Robin sequence is under strong selective pressure, because some aminoacid substitutions can only be achieved through multiple nucleotide mutations.
4. **Transition to transversion ratio**. Three of the values of α were similar to each other, while the value for ZMV-RdRp was higher than the others. Because transitions occur at a higher intrinsic rate, a lower value of α indicates that observed mutations are biased in favour of the rarer transversions, which is an indication of unusual selective pressures. The fact that the values of α for the Robin segments are comparable to, or lower than, the values for RdRp are a further indication that Robin is under similar selective pressure too.
5. **Ratio of non-synonyms to synonyms**. For the RdRp segments the values of $R = N_{\mathrm{ns}}/N_{\mathrm{s}}$ are much smaller than the values R_0 predicted (equation (8)) by the null hypothesis that mutations are random. This indicates that the selective pressure on RdRp acts to preserve the amino acid sequence. For Robin segments, the values of R are much larger, but still smaller than the prediction from the null hypothesis. This indicates that while points 1-4 above indicate that Robin is under some selective pressure, the amino acid sequence is not strongly conserved. This is consistent with the hypothesis that the selection acting on Robin is relaxed.

Figure 2 illustrates the distribution of mutations across the forward and reverse reading frames of all four ORFs for both CNV and ZMV. As expected, there is evidence that some regions accumulate mutations more readily than others. The pattern is consistent with what would be expected from the statistical reductions in the tables.

Complementary reading frame

We determined the set of N_{ds} doubly-synonymous codons in the consensus sequence, and the subset of N_{a} of these which have variant codons.

1. **Mutational hotspots test**. We applied the mutational hotspots test to all four sequences, as described by equations (1) above. The results (tables 4) show no evidence that the doubly synonymous sites are undergoing more frequent mutations, or that their mutations are more widely spread across the dataset.
2. **Mutation rate test**. We examined the number of mutations in the set of N_{a} doubly-synonymous sites which were variable. We found (table 5) that many more of the observed mutations at these sites are only singly synonymous, when a doubly-synonymous mutation is possible, which is further evidence that the complementary strand is non-coding. The numbers of doubly-synonymous mutations were quite low, and so it was not possible to make a reliable comparison of the ratio $N_{\mathrm{s}}/N_{\mathrm{d}}$ with the null hypothesis.
3. **Ratio of non-synonyms to synonyms**.

   The ratios of non-synonymous to synonymous mutations, presented in table 3 and figure 1(**b**), were lower than the null hypothesis for the forward direction. This is readily explained as an indication that the forward ORF codes for a functional protein. However the $N_{\mathrm{ns}}/N_{\mathrm{s}}$ ratios for the reverse direction were all higher than the null hypothesis. This observation is explained, qualitatively, as follows. If the forward direction strictly conserves the amino acid sequence, then all of the mutations which are synomymous on the reverse strand are doubly-synonymous. Because only 12 of the 64 codons allow for doubly-synonymous mutations, the $N_{\mathrm{ns}}/N_{\mathrm{s}}$ ratio would be very high for the complementary strand if the forward sequence were to be exactly conserved. We computed this ratio, and found 11.2 for CNV-RdRp, and similar values for the other sequences. This theoretical ratio is considerably higher than the measured value of 4.97, because the forward sequence is not exactly conserved. For Robin segments, the value of R for the reverse ORF is only slightly higher than the null hypothesis, because the amino acid sequence is only weakly conserved.

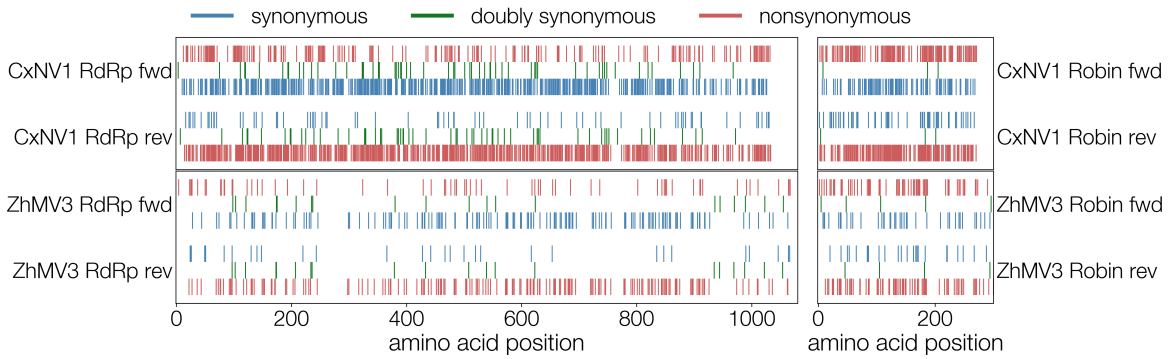

Fig. 2. Distribution of synonymous (blue), non-synonymous (red) substitutions, and doubly synonymous sites (green) in CNV (upper plots) and ZMV (lower plots) RdRp (left) and Robin (right) segments in both directions (forward towards top, reverse towards bottom). Translated reverse ORFs are shown backwards (segment coordinate space). Double synonyms don't overlap perfectly because forward and reverse ORFs differ in length and begin and end at different positions along the segment.

Table 2. Nucleotide-level statistics of mutations. The consensus sequence has N codons. Among the mutations observed in M polymorphs, there are $N_{\rm n}$ transitions, $N_{\rm v}$ transversions, with overall rate r and transition/transversion rate ratio α. The numbers total mutations at each base position is $(n_1 : n_2 : n_3)$, and normalising these to ratios via equation (6) yields $(z_1 : z_2 : z_3)$.

Strand	N	M	$N_{\rm n}$	$N_{\rm v}$	r	α	(n_1, n_2, n_3)	$(z_1 : z_2 : z_3)$
CNV-RdRp	1033	46	606	362	0.0068	3.35	(181, 140, 645)	(0.56 : 0.44 : 2.00)
ZMV-RdRp	1075	12	210	39	0.0064	10.80	(47, 29, 173)	(0.57 : 0.35 : 2.08)
CNV-Robin	272	46	213	146	0.0096	2.92	(107, 100, 152)	(0.89 : 0.84 : 1.27)
ZMV-Robin	304	10	84	48	0.0145	3.50	(35, 31, 66)	(0.80 : 0.70 : 1.50)

Discussion

We have argued that doubly synonymous codons provide a key to understanding whether ambigrammatic viral RNA segments code for two functional proteins. If there were two coding genes, doubly synonymous mutations would be mutational hotspots, because they are unambiguously non-deleterious. We applied our analysis to recent observations of polymorphisms in two ambigrammatic narnaviruses: *Culex narnavirus 1* and *Zhejiang mosquito virus 3*. There was no evidence that doubly synonymous sites are mutational hotspots, or that there is a prevalence of mutations to other doubly-synonymous codons at these sites. Other, circumstantial, evidence favours the interpretation that the complementary strand is non-coding. Ambigrammatic sequences have been observed in other narnaviruses, but they are undoubtedly a rare phenomenon. If the rORF (reverse open reading frame) of both RdRp and Robin segments had evolved to code for a functional protein, each RNA segment would code for two genes. Given that ambigrammatic sequences are rare (DeRisi et al., 2019), finding a system where two had evolved independently would be highly improbable. Moreover, because the ambigrams are full length, each of the ambigrammatically coded sequences would code for two genes which have the same length as each other.

An observation of the simultaneous detection of two or more ambigrammatic genes would strongly favour models where there is an advantage in evolving an ambigrammatic sequence which is independent of whether the reverse open reading frames are translated into functional proteins. This argument led us to discover the Robin segment of ZMV, and suggests that more ambigrammatic narnaviruses with at least two segments will be discovered by metagenomic surveys, when suitable data sets become available. Similarly, the elusive Robin segment should already be hiding in datasets of narnaviruses descended from the common ancestor of CNV and ZMV.

Table 3. Summary of results for codon-level mutations. The numbers of single-nucleotide synonymous and non-synonymous mutations are $N_{\rm sy}$ and $N_{\rm ns}$ respectively, $N_{\rm mult}$ is the number of mutations with more than one base changed, $R_{\rm exp}$ is the null value of $R = N_{\rm ns}/R_{\rm sy}$, and $f_{\rm mult}$ if the fraction of mutations which have multiple-nucleotide changes.

Strand	$N_{\rm sy}$	$N_{\rm ns}$	$N_{\rm mult}$	$R = N_{\rm ns}/N_{\rm sy}$	$R_{\rm exp}$	$R/R_{\rm exp}$	$f_{\rm mult}$
CNV-RdRp-fwd	623	189	123	0.303	2.37	0.128	0.12
ZMV-RdRp-fwd	170	59	13	0.347	2.14	0.162	0.012
CNV-Robin-fwd	112	141	89	1.26	2.34	0.538	0.45
ZMV-Robin-fwd	49	61	14	1.24	2.35	0.528	0.046
CNV-RdRp-comp	136	676	123	4.97	2.43	2.04	0.12
ZMV-RdRp-comp	50	179	13	3.58	2.14	1.67	0.012
CNV-Robin-comp	66	187	89	2.83	2.39	1.23	0.45
ZMV-Robin-comp	32	78	14	2.43	2.28	1.07	0.046

Table 4. Summary of results of the mutational hotspots test. Left panel: values of the average number of elements of the variant set, $\langle n(k)\rangle$ and of the average fraction of non-consensus codons, $\langle f(k)\rangle$, for double-synonym sites, and for the other sites. Right panel: N is the number of loci in the alignment, N_{ds} is the number of double-synonym loci, and R_n, R_f are the ratios of $\langle n(k)\rangle$ and $\langle f(k)\rangle$ at double-synonym sites to their values at other sites. The differences of these ratios from unity do not appear significant.

Sample	$\langle n(k)\rangle$	$\langle f(k)\rangle$
Double syns., CNV-RdRp	0.954	0.161
Other codons, CNV-RdRp	0.968	0.155
Double syns., ZMV-RdRp	1.20	0.042
Other codons, ZMV-RdRp	1.23	0.050
Double syns, CNV-Robin	1.76	0.195
Other codons, CNVRobin	1.48	0.169
Double syns, ZMV-Robin	0.889	0.096
Other codons, ZMV-Robin	0.960	0.097

Gene	N	N_{ds}	R_n	R_f
CNV-RdRp	1033	220	0.986	1.044
ZMV-RdRp	1075	219	0.975	0.840
CNV-Robin	272	54	1.19	1.16
ZMV-Robin	304	81	0.926	0.978

Table 5. Results for the mutational codon frequency test: N is the number of loci in the alignment, N_{a} is the number of mutationally active double-synonym loci, and N_{s}, N_{d} are, respectively, the numbers of single and double synonym mutations.

Sample	N	N_{a}	N_{s}	N_{d}	R	R_0	R/R_0
CNV-RdRp	1033	136	151	60	2.51	3.02	0.83
ZMV-RdRp	1075	219	33	20	1.65	3.21	0.51
CNV-Robin	272	40	24	3	8.00	3.21	2.49
ZMV-Robin	304	59	20	4	4.00	4.04	0.99

Our studies of polymorphisms in the forward direction indicate that both RdRp and Robin are under purifying selection. In the case of RdRp the amino acid sequence is strongly conserved, but the Robin sequence is not.

The role of the RdRp coding fragment is already understood. This makes it plausible that the other fragment plays a role which facilitates the evolution of ambigrams. If the lack of stop codons on the complementary strand is not required to allow protein synthesis, we can surmise that its role is to allow ribosomes to associate with the complementary strand. Having RNA segments able to be covered by ribosomes may provide some protection for the viral RNA against degradation.

Recent experiments indicate that ambigrammatic narnavirus genes display unusual ribosome profiles, with a 'plateau' structure (Retallack et al., 2020). It has been argued (Wilkinson et al., 2021) that the plateaus indicate that the ribosomes attached to the viral RNA become stalled, creating a cover (see also Cepelewicz (2020)). The ambigram property allows binding of ribosomes to both strands, hiding the viral RNA from host defence and degradation mechanisms. We can surmise that there exists a molecule which binds to the 3′ end of the viral RNA, preventing release of ribosomes (Wilkinson et al., 2021). It is possible that Robin plays a role in this process, by creating a protein which blocks ribosome detachment at 3′ end. Alternatively, it might be proposed that the ribosome 'traffic jam' is a consequence of the structure of the RdRp itself, due to formation of RNA hairpins. However, these would have to trade off against RdRp function. The proposed mechanism involving Robin making a blocking protein has the advantage that the RdRp works efficiently when the viral RNA concentration is small. Later, after it has duplicated many copies of itself and of Robin, the Robin protein attaches to the viral RNA and creates stalled polysomes, protecting the viral RNA from degradation.

There may, however, be additional viral genes involved in ambigrammatic narnavirus infections, and there are many possible roles for the Robin gene. It could code a protein which inhibits the mechanism of 'no-go-decay', which releases stalled ribosomes, play a role in the viral suppression of RNAi (Mierlo et al., 2014) or in formation of syncytia or viral particles. Without a better understanding of the narnavirus lifecycle in arthropods it is not certain whether Robin does code for a protein which blocks detachment of ribosomes.

We did search the CNV dataset for further fragments of ambigrammatic viral RNA, which might be candidates for coding additional genes. A search for additional ambigrammatic sequences greater than 200nt in length did not produce any candidates.

A recent preprint (Retallack et al., 2020) presents evidence that inserting mutations in the RdRp sequence which are synonymous in the forward reading frame but introduce stop codons in the reverse frame reduces the fitness of the virus. The mutations were clustered close to the 3′ end of the RdRp gene. These observations could be interpreted as indicating that the reverse reading frame codes for a functional protein or that all ORFs in the cell may be translated in a 'leaky' way. However, changing the RNA sequence may also interfere with the action of molecules which bind to the RdRp strand.

Competing interests

There is NO Competing Interest.

Author contributions

GD devised and directed the search for an analog of Robin in the ZMV sequence archive. MW produced a draft of the manuscript following discussions with the other authors about the recent discovery of a narnavirus system which has two ambigrammatic genes. All authors contributed to writing the manuscript, and reviewed the manuscript before submission.

Acknowledgments

We thank Hanna Retallack and Joe DeRisi for discussions of their experimental studies of narnaviruses and Amy Kistler for assistance with narnaviral genomes and for comments on a draft. We would like to thank Mang Shi and Edward C Holmes for sharing assembled contigs from Australian and Chinese mosquito metagenomic datasets. G.H. and D.Y. were supported by the Chan Zuckerberg Biohub; MW thanks the Chan Zuckerberg Biohub for its hospitality.

Data availability

References

B. G. Barrell, G. M. Air, and C. A. Hutchison. Overlapping genes in bacteriophage phiX174. *Nature*, 264:34–41, 1976. doi: 10.1038/264034a0.

J. Batson, G. Dudas, E. Haas-Stapleton, A. L. Kistler, L. M. Li, P. Logan, K. Ratnasiri, and H. Retallack. Single mosquito metatranscriptomics recovers mosquito species, blood meal sources, and microbial cargo, including viral dark matter. bioRxiv: https://doi.org/10.1101/2020.02.10.942854, 2020.

J. Cepelewicz. New clues about 'ambigram' viruses with strange reversible genes. *Quanta Magazine*, 2020. URL `https://www.quantamagazine.org/new-clues-about-ambigram-viruses-with-strange-reversible-genes-20200212/`.

N. Chirico, A. Vianelli, and R. Belshaw. Why genes overlap in viruses. *Proc Biol Sci.*, 277:1701, 2010. doi: 10.1098/rspb.2010.1052.

A. G. Cobián Güemes, M. Youle, V. A. Cantú, B. Felts, J. Nulton, and F. Rohwer. Viruses as winners in the game of life. *Annual Review of Virology*, 3(1):197–214, 2016. doi: 10.1146/annurev-virology-100114-054952. URL `https://doi.org/10.1146/annurev-virology-100114-054952`.

J. DeRisi, G. Huber, A. Kistler, H. Retallack, M. Wilkinson, and D. Yllanes. An exploration of ambigrammatic sequences in narnaviruses. *Sci. Rep.*, 9:17982, 2019. doi: 10.1038/s41598-019-54181-3. URL `https://doi.org/10.1038/s41598-019-54181-3`.

A. M. Dinan, N. I. Lukhovitskaya, I. Olendraite, and A. E. Firth. A case for a negative-strand coding sequence in a group of positive-sense rna viruses. *Virus Evolution*, 6:veaa007, 2020. doi: https://doi.org/10.1093/ve/veaa007.

L. Fu, B. Niu, Z. Zhu, S. Wu, and W. Li. CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics*, 28(23):3150–3152, Dec. 2012. ISSN 1367-4811. doi: 10.1093/bioinformatics/bts565.

B. I. Hillman and G. Cai. The family *narnaviridae*: simplest of RNA viruses. In S. A. Ghabrial, editor, *Mycoviruses*, volume 86 of *Advances in Virus Research*, pages 149–176. 2013. doi: 10.1016/B978-0-12-394315-6.00006-4.

M. Kimura. A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *J. Molecular Evolution*, 16:111–20, 1980.

J. T. v. Mierlo, G. J. Overheul, B. Obadia, K. W. R. v. Cleef, C. L. Webster, M.-C. Saleh, D. J. Obbard, and R. P. v. Rij. Novel Drosophila Viruses Encode Host-Specific Suppressors of RNAi. *PLOS Pathogens*, 10(7):e1004256, July 2014. ISSN 1553-7374. doi: 10.1371/journal.ppat.1004256. URL `https://journals.plos.org/plospathogens/article?id=10.1371/journal.ppat.1004256`.

C. W. Nelson, Z. Ardern, and X. Wei. Olgenie: Estimating natural selection to predict functional overlapping genes. *Molecular Biology and Evolution*, 37:2440–2449, 2020. doi: https://doi.org/10.1093/molbev/msaa087.

H. Retallack, K. D. Popova, M. T. Laurie, S. Sunshine, and J. L. DeRisi. Persistence of ambigrammatic narnaviruses requires translation of the reverse open reading frame. bioRxiv preprint, doi: http://10.1101/2020.12.18.423567 2020.

M. Shi, X.-D. Lin, J.-H. Tian, L.-J. Chen, X. Chen, C.-X. Li, X.-C. Qin, J. Li, J.-P. Cao, J.-S. Eden, J. Buchmann, W. Wang, J. Xu, E. C. Holmes, and Y.-Z. Zhang. Redefining the invertebrate RNA virosphere. *Nature*, Nov. 2016. ISSN 1476-4687. doi: 10.1038/nature20167.

M. Shi, P. Neville, J. Nicholson, J.-S. Eden, A. Imrie, and E. C. Holmes. High-Resolution Metatranscriptomics Reveals the Ecological Dynamics of Mosquito-Associated RNA Viruses in Western Australia. *Journal of Virology*, 91(17), Sept. 2017. ISSN 1098-5514. doi: 10.1128/JVI.00680-17.

M. Wilkinson, D. Yllanes, and G. Huber. Polysomally protected viruses. 2021. URL `https://arxiv.org/abs/2102.00316`. arXiv:2102.00316.