Polymorphism of Genetic Ambigrams

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

***For correspondence:**
gytisdudas@gmail.com (GD);
greg.huber@czbiohub.org (GH);
michael.wilkinson@czbiohub.org
(MW); david.yllanes@czbiohub.org
(DY)

[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; [3]School of Mathematics and Statistics, The Open University, Walton Hall, Milton Keynes, MK7 6AA, UK

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 polymorphs of an ambigrammatic virus indicates that most of the open reading frame on the complementary strand does *not* code for a functional protein. This ambigrammatic gene was found to be associated with an apparently symbiotic companion RNA molecule, also ambigrammatic, termed 'Robin'. Our analysis of the polymorphism of Robin suggests that it 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 narnaviruses appear to be single genes, which code for an RNA-dependent RNA polymerase (abbreviated as RdRp) ***Hillman and Cai*** (***2013***). Some narnaviruses are found to have a genome for which there is an open reading frame (that is, a reading frame without stop codons) on the strand which is complementary to that which codes the gene for the RdRp. This reverse open reading frame has codon boundaries aligned with the forward reading frame. Because the genome can be read 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 observations on the polymorphism of the genetic code can distinguish between competing hypotheses on the function and nature of ambigrammatic viral genomes.

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 synonym substitutions which result from single-base mutations in order 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 on the forward gene. 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 poison the defence mechanisms of the host cell. 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 detection by cellular defence mechanisms. 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 the preliminary evidence is in favour of this second hypothesis, namely that most of the open reading frame 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, synonymous mutations occur more frequently than non-synonymous ones. 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 43 polymorphic variants of an ambigrammatic narnavirus, using data reported in ***Batson et al.*** (***2020***). We find that neither of our tests supports the hypothesis that the translated sequence of the complementary strand sequence is under selective pressure. We also apply these tests to a second ambigrammatic RNA sequence, termed the *Robin* sequence, which is closely associated with ambigrammatic narnavirus infection in mosquitos. We find that neither the of the two complementary open reading frames of Robin appears to be under selective pressure for its amino acid sequence. In the concluding section, we consider the interpretation of these observations, and discuss whether there may be implications for other viral families.

There are many examples of overlapping genes with staggered reading frames, and recent work by Nelson, Ardern and Wei ***Nelson et al.*** (***2020***) discusses how these can be identified. Our investigations indicate that the ambigrammatic genes 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.

First rule: 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 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 CU*, UUA, UUG (where * means any base). The synonyms of Ser are

UC*, 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.

Furthermore, it is found that complementary-strand stops cannot always be removed by synonym substitutions in the other two read frames for the complementary strand (this requires a longer argument, also given in ***DeRisi et al.*** (***2019***)). As a consequence of the two arguments, we need discuss only the complementary read frame with aligned codons.

Second rule: there exist double synonyms

Most synonymous mutations of the forward strand produce a non-synonymous change of 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 make a transversal mutation to CGG (Arg), 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.)

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 data on polymorphism of the 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 image of the complementary strand.

AA	Codon	Syns.: $S^{(n)} + S^{(v)}$	Dbl syns.: $D^{(n)} + D^{(v)}$	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

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 $S^{(n)} + S^{(v)}$ ($D^{(n)} + D^{(v)}$), where the superscript n denotes transitions, and superscript v transversions.

Tests of 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 polymorphs 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 $\mathcal{N}$ 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 *double-synonym locus*. The number of double-synonym loci is $\mathcal{N}_\mathrm{d}$.

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

Look for the existence of mutational 'hotspots'

We can look for evidence that the double-synonym loci are more active 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 double-synonym 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. In particular, if $\delta R = |1 - R|$, the deviation of R from unity is significant if $\delta R\sqrt{\mathcal{N}} \gg 1$.

Mutation frequencies test

We can also look at codon frequencies for different mutations at the double-synonym loci. If the complementary strand is coding, we expect to find that the frequency of mutations observed at double-synonym 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 $\mathcal{N}_\mathrm{a}$ *mutationally active double-synonym loci*, we can determine two numbers: $n_\mathrm{s}(k)$ is the numbers of singly-synonymous variants at locus k, and $n_\mathrm{d}(k)$ is the number of these variants which are also doubly-synonymous. (Clearly $n(k) \geq n_\mathrm{s}(k) \geq n_\mathrm{d}(k)$). If $n_\mathrm{d}(k) = n_\mathrm{s}(k)$, 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 mutationally active double-synonym loci, we then calculate

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

If the complementary strand is coding, we expect

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

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

We must estimate how large R can be before the complementary-strand coding hypothesis must be rejected. To this end, we shall estimate $N_{\rm exp}$, the expected value of $N_{\rm s}$, based upon the null hypothesis that the complementary strand is non-coding. If R_0 is the value of the ratio R that is derived from this null hypothesis, then:

$$N_{\rm exp} = R_0 N_{\rm d}\ . \tag{4}$$

We describe the calculation of R_0 below.

We may assume that the codons at different loci may be modelled as are being statistical independent so that the numbers $N_{\rm s}$ and $N_{\rm d}$ are subject to Poissonian counting statistics. We can estimate the significance of the difference between $N_{\rm exp}$ and $N_{\rm s}$ by determining

$$\sigma = \frac{N_{\rm exp} - N_{\rm s}}{\sqrt{N_{\rm s}}}\ . \tag{5}$$

A large, positive, value of σ would indicate that the neutrality hypothesis can be rejected.

We assume that the M polymorphs 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 rates for transitions ($\rm A \leftrightarrow G$ or $\rm C \leftrightarrow U$) are different from those of transversions (other single-nucleotide mutations), and negligible for other types of mutation. The ratio of these rates is

$$\alpha = \frac{\mathcal{R}_{\rm transition}}{\mathcal{R}_{\rm transversion}}\ . \tag{6}$$

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

$$R_0 = \frac{\sum_{k\in\{k^*\}} \alpha S_k^{\rm (n)} + S_k^{\rm (v)}}{\sum_{k\in\{k^*\}} \alpha D_k^{\rm (n)} + D_k^{\rm (v)}} \tag{7}$$

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

Non-synonymous to synonymous ratio test

One standard test of whether a sequence codes for a protein is to look at the ratio of non-synonymous to synonymous mutations. We expect this ratio to be small when a readable base sequence is a functional gene coding for a well-conserved protein. Let us consider how to apply this test to the complementary strand of an ambigrammatic gene, which we assume is well-conserved in the forward direction. We can also consider another null-hypothesis for this test, in order to indicate whether the complementary strand codes for a functional protein.

Let us assume that the forward strand is perfectly conserved (so that only synonymous mutations are allowed), and the complementary strand is non-coding. We shall see that there are some non-synonymous mutations for the forward strand of the narnavirus RdRp gene, but nevertheless applying this test does give additional insight. Let $\tilde{N}_{\rm n}$ and $\tilde{N}_{\rm s}$ be the *total* numbers of non-synonymous and synonymous changes on the *complementary* strand, summed over all codon loci. We evaluate

$$\tilde{R} = \frac{\tilde{N}_{\rm n}}{\tilde{N}_{\rm s}} \tag{8}$$

and compare it with an estimate $\tilde{R}_0$ which is derived from the null hypothesis that the complementary strand is non-coding. The value of $\tilde{R}_0$ is obtained by noting that, because the mutations on the forward strand are assumed to be synonymous, the total rates for non-synonymous and synonymous transitions on the complementary strands are proportional to $\alpha S_k^{\rm (n)} + S_k^{\rm (v)}$ and $\alpha D_k^{\rm (n)} + D_k^{\rm (v)}$ respectively, so that

$$\tilde{R}_0 = \frac{\sum_k \alpha S_k^{\rm (n)} + S_k^{\rm (v)}}{\sum_k \alpha D_k^{\rm (n)} + D_k^{\rm (v)}} \tag{9}$$

where in this case we sum over *all* of the codon loci for the complementary strand. (In order to implement this test we need the values of the coefficients $S_k^{(\mathrm{n,v})}$ and $D_k^{(\mathrm{n,v})}$ for all of the codons (other than stop codons), not just those listed in table 1.)

Polymorphism of an ambigrammatic narnavirus

In a recent study ***Batson et al.*** (***2020***) of an ambigrammatic narnavirus, it was reported that this narnavirus system has the following properties:

1. There is a viral RNA segment which codes the for the RdRp, and which has the property of being ambigrammatic, with forward and reverse codons aligned, over very nearly the entire length (the forward strand has two stop codons close to the 3' end, and complementary strand has just one stop codon, which is also very close to its $3'$ end).
2. Infection with this sequence is strongly associated with the presence of another RNA sequence, which was referred to in ***Batson et al.*** (***2020***) as the 'Robin' sequence.
3. The Robin sequence is also ambigrammatic, over its entire length (about 850 nt), with the codons of the rORF aligned (in this case the strand that we designate as the forward has one stop close to the 3' end, and its complementary strand has two stop codons, which are very close to its $3'$ end). Neither forward nor reverse directions are homologous with known sequences.

In addition 43 polymorphic variants of both the narnavirus gene which codes for the RdRp and the Robin gene were sequenced. We determined the optimal optimal alignment of these 43 sequences, and identified the 'consensus' nucleotide (the one with the largest number of counts) at each locus. We also identified the set of variant nucleotides observed at each locus, and counted the numbers of transition and transversion mutations. The requirement that there be a minimal number of stops was used to determine the reading frames for both the forward and complementary sequences, and the consensus nucleotides were used to determine a consensus codon.

Results on the RdRp gene

In order to bring all of the sequences into alignment we had to insert a few 'dummy' codons. We inserted either one or three dummy codons at nucleotide locus 395 into most of the sequences, and another dummy codon at nucleotide locus 2924 into 19 sequences. For each nucleotide locus, we determined a consensus nucleotide, and determined the set of variants that were seen at each site. We found a total of 619 transitions and 395 transversions across the 3166 nucleotide loci, indicating a mutation rate ratio $\alpha = 3.13$, and a rate of selected mutations equal to 0.0074 mutations per nucleotide per polymorph.

For each codon locus k, we determined 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 identified the set of double synonym loci, for which the consensus codon is one of the starred codons listed in table 1.

We first tested for whether there are mutational hotspots. We determined average values of $n(k)$ and $f(k)$ for the double synonym sites and for the other sites. The results are listed in table 2. From these data we find do not suggest that the double-synonym sites are mutational 'hotspots'.

Sample	$\langle n(k) \rangle$	$\langle f(k) \rangle$
Double syns., RdRp	0.943	0.193
Other codons, RdRp	0.898	0.187

Gene	$\mathcal{N}$	$\mathcal{N}_\mathrm{d}$	R_n	R_f
RdRp	1054	227	1.049	1.034

Table 2. Summary of results of '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: $\mathcal{N}$ is the number of loci in the alignment, $\mathcal{N}_\mathrm{d}$ is the number of double-synonym loci, and R_n, R_f are the ratios defined in equation (1). The differences of these ratios from unity do not appear significant.

Sample	$\mathcal{N}$	$\mathcal{N}_\mathrm{a}$	N_s	N_d	R_0	σ
RdRp	939	136	315	99	3.08	−0.52

Table 3. Results for the mutational codon frequency test: $\mathcal{N}$ is the number of loci in the alignment, $\mathcal{N}_\mathrm{a}$ is the number of mutationally active double-synonym loci, and N_s, N_d are, respectively, the numbers of single and double synonym mutations. The values of R_0 are evaluated at $\alpha = 3.13$.

Strand	N_syn	N_nsyn	N_mult	$R = N_\mathrm{nsyn}/N_\mathrm{syn}$	R_0
Robin-fwd	112	225	92	2.01	2.36
Robin-comp	68	261	89	3.84	2.42
RdRp-fwd	621	318	140	0.512	–
RdRp-comp	138	801	140	5.80	11.2

Table 4. Numbers of synonymous and non-synonymous mutations, N_s and N_n, for two different reading directions of both genes.

We then tried the mutational frequency test. Our results are presented in table 3. We found that $N_\mathrm{s} \gg N_\mathrm{d}$, indicating that there is not strong selection pressure on the complementary sequence. Table 3 also includes results on the application of the null hypothesis. The value of R_0 depends upon α: in table 3 we used $\alpha = 3.13$, as derived from our observed nucleotide mutations. The results are consistent with the null hypothesis, that the complementary sequence is non-coding.

We conclude that there is no significant evidence that there are mutational hotspots at potentially doubly-synonymous sites, and that the frequencies of double synonym codons are compatible with what would be expected by chance, based upon the null hypothesis that the complementary strand is non-coding.

Results on the Robin gene

We also investigated polymorphism of the Robin sequences, obtained from the same 43 samples as the narnavirus RdRp genes. In order to bring all of the sequences into alignment we had to insert a few 'dummy' codons. We inserted two dummy codons at nucleotide locus 408 into 25 of the sequences, and another dummy codon at nucleotide locus 560 into 37 sequences. For each nucleotide locus, we determined a consensus nucleotide, and determined the set of variants that were seen at each site. We found a total of 218 transitions and 156 transversions across the 860 nucleotide loci, indicating a mutation rate ratio $\alpha = 2.79$, and a rate of selected mutations equal to 0.0101 mutations per nucleotide per polymorph.

The next step was to identify which is the coding direction of the Robin gene. One expects that this can be identified by determining which read direction has a high proportion of synonymous mutations. We identified the consensus sequence for Robin, and evaluated the number of variants $n(k)$ at each locus, and the number of synonymous and non-synonymous variants, $n_\mathrm{syn}(k)$ and $n_\mathrm{nsyn}(k)$. In table 4 we list the total of the number of non-synonymous and synonymous variants

$$N_\mathrm{nsyn} = \sum_k n_\mathrm{nsyn}(k)\ , \quad N_\mathrm{syn} = \sum_k n_\mathrm{syn}(k) \tag{10}$$

for the two different reading directions, denoted by Robin A and Robin B, together with comparable data for the forward and reverse reads of the RdRp gene.

In table 4 we also indicate the number of mutations which involve more than changes to more than one nucleotide, N_mult.

In the case of the Robin gene, we see that the number of synonymous mutations 112 is slightly higher than the null hypothesis for the forward strand, which predicts 100 mutations. The difference is not statistical significant. For the complementary strand there were 68 synonymous mutations, which is significantly less than the predicted number of 96. These results indicate that, if the Robin gene codes for a protein, it has been under very little selective pressure.

However, there is evidence that the Robin gene has been under similar selective pressure to the the RdRp gene. The RdRp and Robin sequences are obtained from the same set of samples. The RdRp gene is regarded as a stable object which evolves very slowly. If Robin were under weak selective pressure, it would be expected to be more tolerant of mutations, so that its rate of selected mutations would be much higher. However, the rate of mutation per nucleotide per polymorph are quite close, differing by approximately 30%.

Another line of evidence that Robin is under strong selective pressure comes from the number of codons which have undergone mutations at more than one nucleotide. Despite the fact that the rate of selected mutations per nucleotide per polymorph is small, the fraction of codons with multiple mutations is significant, for both the RdRp and Robin genes: we find that the fraction of multi-nucleotide mutations is ? for Robin-fwd, ? for Robin-comp and ? for RdRp-fwd. These data suggest that the Robin sequence has been under even more selective pressure than the RdRp gene.

While these lines of evidence are not conclusive, they do indicate that the Robin gene has been under selective pressure, without conserving the aminoacid sequence. The inference is that the Robin gene is not translated into a functional protein. Because these results indicate that the Robin gene is not translated in either direction, so there is no point in testing for whether double synonyms are mutational hotspots.

Discussion

We have argued that the recent observations of polymorphism of an ambigrammatic narnavirus ***Batson et al.*** (***2020***) favour the hypothesis that the reverse read of an ambigrammatic sequence does not code for a functional proteins. Furthermore, the Robin gene, which is symbiotic with the RdRp narnavirus gene, does not appear to code for a functional protein in either direction.

Other, circumstantial, evidence favours the interpretation that the complementary strand is non-coding. Ambigrammatic sequences have been observed in a variety of simple RNA virus genomes , but they are undoubtedly a rare phenomenon. If the rORF (reverse open reading frame) of both the RdRp and the partner fragment had evolved to code for a functional protein, each RNA sequence 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 complementary strand open reading frames are translated into functional proteins.

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 is 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 strands able to be covered by ribosomes may provide some protection for the viral RNA against degradation by the defence mechanisms of the host cell.

The 'Robin' sequence identified in ***Batson et al.*** (***2020***) appears to play an important role in the infection, because of its strong association with the RdRp coding sequence. A plausible hypothesis is that it somehow prevents ribosomes from detaching from the $3'$ end of viral RNA, so that the viral RNA can become shrouded in a coating of ribosomes. The evidence from polymorphism studies is consistent with the Robin RNA being able to bind directly with the RdRp gene, rather than being translated into a protein. Further studies of 'ribosome profiling' of this system have that potential to reveal evidence of whether ribosomes accumulate on the viral RNA in this system.

References

Batson, J., Dudas, G., Haas-Stapleton, E., Kistler, A. L., Li, L. M., Logan, P., Ratnasiri, K., and Retallack, H. (2020). 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.

Cobián Güemes, A. G., Youle, M., Cantú, V. A., Felts, B., Nulton, J., and Rohwer, F. (2016). Viruses as winners in the game of life. *Annual Review of Virology*, 3(1):197–214. PMID: 27741409.

DeRisi, J., Huber, G., Kistler, A., Retallack, H., Wilkinson, M., and Yllanes, D. (2019). An exploration of ambigrammatic sequences in narnaviruses. *Sci. Rep.*, 9:17982.

Hillman, B. I. and Cai, G. (2013). The family *narnaviridae*: simplest of RNA viruses. In Ghabrial, S. A., editor, *Mycoviruses*, volume 86 of *Advances in Virus Research*, pages 149–176.

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

Nelson, C. W., Ardern, Z., and Wei, X. (2020). Olgenie: Estimating natural selection to predict functional overlapping genes. *Molecular Biology and Evolution*, 37:2440–2449.

Acknowledgements

We thank Hanna Retallack and Joe DeRisi for discussions of their experimental studies of narnaviruses and Amy Kistler and Gitys Dudas for assistance with narnaviral genomes. MW thanks the Chan Zuckerberg Biohub for its hospitality.

Author contributions statement

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.

Additional information

There are no competing interests.