Diversity of tRNA Clusters in the Chloroviruses

Garry A. Duncan[1], David D. Dunigan[1,2], James L. Van Etten[1,2]

[1]Nebraska Center for Virology and [2]Department of Plant Pathology, University of Nebraska-Lincoln, Lincoln, NE 68583-0900

Key index words: tRNAs, tRNA clusters, chloroviruses, algal viruses, codon usage bias, CUB

Author for Correspondence: James L Van Etten, email: jvanetten1@unl.edu

Running title: tRNA clusters in chloroviruses

ABSTRACT

Viruses rely on their host's translation machinery for the synthesis of their own proteins. Problems belie viral translation when the host has a codon usage bias (CUB) that is different from an infecting virus with differences in the GC content between the host/virus genome. Here, we evaluate the hypothesis that chloroviruses adapted to host CUB by acquisition and selection of tRNAs that at least partially favor their own CUB. The genomes of 41 chloroviruses comprising three clades of three different algal hosts have been sequenced, assembled and annotated. All 41 viruses not only encode tRNAs, but their tRNA genes are located in clusters. One tRNA gene was common to all three clades of chloroviruses, while differences were observed between clades and even within clades. By comparing the codon usage of one chlorovirus algal host, whose genome has been sequenced and annotated (67% GC content), to that of two of its viruses (40% GC content), we found that the viruses were able to at least partially overcome the host's CUB by encoding tRNAs that recognize AU-rich codons. In addition, 39/41 chloroviruses encode a putative lysidine synthase, which alters the anticodon of $tRNA^{met}$ that normally recognizes AUG to recognize the codon AUA, a codon for isoleucine. This is advantageous to the viruses because the AU-rich codon AUA is 12-13 times more common in the chloroviruses than their host. Evidence is presented that supports the concept that chlorovirus tRNA clusters were acquired prior to events that separated them into the three clades.

IMPORTANCE

Chloroviruses are members of a group of giant viruses that infect freshwater green algae around the world. More than 40 chloroviruses have been sequenced and annotated. In order to propagate efficiently, chloroviruses with low GC content must overcome the high GC content and codon usage bias (CUB) of their hosts. We provide support for one mechanism by which viruses can overcome host CUB. Specifically, the chloroviruses examined herein encode tRNAs whose cognate codons are common in the viruses but not in the host. Virus-encoded tRNAs that recognize AU-rich codons enable more efficient protein synthesis, thus enhancing viral propagation. The tRNA genes are located in clusters and the original tRNA gene cluster was acquired by the most recent common ancestor of the four chlorovirus clades. Furthermore, we show

some conservation among all clades, but also substantial variation between and within clades, demonstrating the dynamics of viral evolution.

INTRODUCTION

Viruses rely on most or all of their host's translation machinery to synthesize their proteins. A conflict for viruses occurs when they have a codon usage bias (CUB) different from their host. However, some viruses have genes that help adapt to the host's CUB in favor of their own CUB by encoding tRNAs. These include large dsDNA viruses infecting eukaryotic organisms (see Morgado and Vicente, 2019 for an extensive list) and bacteriophages (Abe *et al*., 2014). Recently, tRNA-encoding genes have also been reported in small ssDNA and ssRNA viruses (Morgado and Vicente, 2019).

Among the group of tRNA-encoding viruses are large viruses that infect algae (Michely *et al.,* 2013; Pagarete *et al*., 2013; Santini *et al*., 2013; Derelle *et al*., 2015), including the chloroviruses (family *Phycodnaviridae*) with a lytic life-style (Li *et al*., 1997; Nishida *et al*., 1999; Lee *et al.,* 2005. Fitzgerald *et al*., 2007, 2007a, 2007b; Jeanniard *et al*., 2013). The chloroviruses have genomes that are 290 to 370 kb in size and are predicted to encode up to 400 proteins (CDSs) and 16 tRNAs (Van Etten *et al*., 2020). They infect certain chlorella-like green algae that live in a symbiotic relationship with protists and metazoans (referred to as zoochlorellae), forming the holobiont. There are four known clades of chloroviruses based on the host they infect: viruses that infect *Chlorella variabilis* NC64A (referred to as NC64A viruses), viruses that infect *Chlorella variabilis* Syngen 2-3 (referred to as Osy viruses), viruses that infect *Chlorella heliozoae* SAG 3.83 (referred to as SAG viruses), and viruses that infect *Micractinium conductrix* Pbi (referred to as Pbi viruses).

The genomes of more than 40 chloroviruses, representing 3 of the 4 clades, have been sequenced, assembled and annotated (Jeanniard *et al*., 2013 and references cited therein). The genome GC content of the NC64A viruses ranges from 40%-41%; the GC content of the Pbi viruses ranges from 44-47%; while the GC content of the SAG viruses ranges from 48-52% (Jeanniard *et al.,* 2013). Their hosts, *C. variabilis* NC64A (Blanc *et al*., 2010) and *M. conductrix* Pbi (Arriola *et al*., 2018), both have nuclear genome GC contents of 67%, and it is assumed that the *C. heliozoae* has a similar GC content.

Two NC64A chloroviruses, PBCV-1 and CVK2, also have the interesting property of clustered tRNA genes (Lee *et al*., 2005; Nishida *et al*. 1999). Clusters of tRNA genes are found in some bacteriophage (Morgado and Vicente, 2019), as well as in a small percentage of organisms from all three domains of life (Bermudez-Santana *et al*., 2010; Morgado and Vicente, 2018, 2019, 2019a). Clusters of tRNA-genes have also been reported in mitochondria and plastids (Abe *et al*., 2014; Fan *et al*., 2017; Freidrich *et al*., 2012).

This current *in silica* investigation had several purposes. i) To evaluate chlorovirus tRNA genes with respect to clustering. ii) To compare and contrast the tRNA genes within and among the three clades of chloroviruses. iii) To determine if there was a CUB in the algal host *C. variabilis* NC64A, and likewise, if there was a CUB among the chloroviruses. iv) To evaluate whether the chlorovirus clades acquired their tRNA gene clusters before or after their divergences from their most recent common ancestor (MRCA).

MATERIALS AND METHODS

Genomic sequence data: In total, the genomes of 41 sequenced, assembled (in some cases to draft genomes) and annotated chloroviruses that represent three of the four known chlorovirus clades were used in this study: 14 NC64A viruses, 13 SAG viruses, and 14 Pbi viruses (Jenniard *et al*., 2013 and references cited therein). Table S1 provides the accession numbers, as well as the source of the viruses. The genome of *C. variabilis* NC64A, which is the host of the NC64A viruses, has been sequenced and annotated (project accession number ADIC00000000; Blanc *et al.,* 2010). *M. conductrix*, the host of the Pbi virus, has also recently been sequenced and annotated (Arriola *et al.,* 2018), but its sequence was not included in this study.

Identification and localization of the tRNA genes: A tRNA gene cluster was previously reported for chlorovirus PBCV-1 (Lee *et al*., 2005; Dunigan *et al.,* 2013), as well as CVK2 (Nishida *et al*., 1999). The other 40 chloroviruses were then examined to determine if they also had tRNA gene clusters, and their tRNA genes and gene order were documented. The tRNA genes for each virus were identified by entering the genome accession numbers into the Nucleotide database at NCBI (ncbi.nlm.nih.gov). The Graphics link was used to identify the tRNA genes and their locations, as well as the

surrounding non-tRNA genes. The putative tRNA sequences were verified using tRNAscan-SE (Lowe and Eddy, 1997; Lowe and Chan, 2016). Commonalities among the three clades of chloroviruses, as well as their differences were also recorded.

The non-tRNA genes immediately 5' and 3' to the tRNA clusters were sought in order to give insight as to whether the 41 chloroviruses acquired their tRNA gene clusters prior to or after their divergence into the three clades. The source of the data was the same NCBI website and it was processed as described above.

CUB in chloroviruses and their hosts: Because the chloroviruses have GC contents lower than their algal hosts, codon usage was examined for two NC64A viruses and one host, *C. variabilis* NC64A. Geneious 11.1.5 (https://www.geneious.com) was used to determine codon usage for the two viruses and their host.

Phylogenetic tree construction of three tRNA genes: Three tRNA genes encoding $tRNA^{tyr}$, $tRNA^{gly}$ and $tRNA^{arg}$ were common to all three clades of chloroviruses and were selected for phylogenetic analysis. The tRNA sequences were retrieved from NCBI following BLAST searches. Each set of sequences was aligned by MUSCLE and phylogenies were constructed using maximum parsimony (http://phylogeny.fr). The trees were saved in Newick format and MEGAX was used to produce the final trees (Stecher *et al.,* 2020).

RESULTS

Chlorovirus tRNA gene cluster locations: All 41 chloroviruses encode tRNA genes, and the tRNA genes in all 41 viruses were in clusters, usually with intergenic spacers of 1 to ~30 nucleotides. However, in a few cases non-tRNA genes were also present within the tRNA clusters. Collectively, the 14 NC64A viruses had from 7 (virus AR158) to 14 (MA-1E, CvsA1 and CviK1) tRNA genes in their clusters (Table 1A); the 13 SAG viruses had from 7 (NTS-1) to 13 (OR0704.3, Can0619SP, NE-JV-2) tRNA genes in their clusters (Table 1B); and, the 14 Pbi viruses had from 3 (NE-JV-1) to 11 (Fr5L) genes in their tRNA gene clusters (Table1C). If one assumes that the original clusters are the sum of all the tRNAs genes in a clade, the NC64A chlorovirus cluster would consist of 14 tRNA genes; indeed, 3 of the NC64A viruses had all 14 tRNA genes. However, the sum of the SAG viruses was 18 tRNA genes, but the largest number of

tRNA genes of any one virus was 13, and the sum of the Pbi viruses was 15 tRNA genes, but the largest number of tRNA genes of any one virus was 11 tRNA genes. Some of the tRNA genes likely represent gene duplications. For example, all 41 chloroviruses had 2 - 4 $tRNA^{asn}$ genes, whose encoded tRNAs recognize the same codon AAC.

The order of the chlorovirus tRNA genes within a cluster is also reported in the three heat map tables (Tables 1A-C). In general, there was synteny of tRNA genes among the viruses within a clade, but not between clades. (The exceptions to synteny within a clade are noted in the footnotes of Tables 1A-C.) Seven tRNA genes were common to one or more members of all three chlorovirus clades, although not present in every virus species within any of the three clades (Table 2). Two tRNA genes were unique to the NC64A viruses, 3 tRNA genes were unique to the Pbi viruses and 4 tRNA genes were unique to the SAG viruses. Three additional tRNA genes were common to one or more members of the NC64A and SAG clades. The $tRNA^{arg}$ gene was found in all 41 chloroviruses. Thirty-nine of the 41 chloroviruses had the $tRNA^{gly}$ gene. It is interesting to note that the position in order of encoded tRNAs in the cluster from the 5' end of the $tRNA^{gly}$ gene was invariant in all 39 viruses.

Differences and similarities in tRNA gene content were observed among individual chloroviruses within each of the three clades of chloroviruses. In some chloroviruses not all of the tRNA genes within a gene cluster were immediately contiguous to one another, leading to interrupted tRNA-gene clusters in which one or several non-tRNA genes were interspersed within the tRNA-gene cluster. This was especially true for the Pbi viruses in which at least four member viruses had interrupted tRNA-gene clusters (Table S2C).

The tRNA clusters in the NC64A and Pbi viruses were located near the center of their genomes except for the Pbi virus NE-JV-1, which was located in the last third of its genome; NE-JV-1 is unusual in many other aspects (Jenniard *et al*., 2013). The tRNA clusters in all of the SAG viruses were located in the first third of their genomes. It is interesting to note that all 13 SAG viruses had a $tRNA^{thr}$ gene located ~30kb beyond the 3' end of the tRNA gene cluster, placing the $tRNA^{thr}$ genes near the center of their respective genomes.

Viral tRNAs help to overcome host CUB: Because nuclear genomic codon usage data are available for *C. variabilis* NC64A and its viruses, we were able to compare

158 their respective codon usages (Fig 1 and Table 3; all codons are reported in Fig S1A,
159 S1B). Codon usage was available for 13/14 NC64A viruses, but only two NC64A viruses
160 were included herein as representatives. CUB favoring codons with high GC content were
161 noted in the host alga *C. variabilis* NC64A whose genome is 67% GC (green bars in Fig
162 1), while the two NC64A viruses, PBCV-1 and AN69C, have GC contents of 40%
163 (Jeanniard et al., 2013) and a CUB favoring AU (Table 3, ratio of codon frequencies
164 comparing virus usage to host usage). For example, in the standard universal code there
165 are four codons for the amino acid alanine, differing only in the third (3') base of the codon.
166 In *C. variabilis* NC64A the two codons whose third base was C or G (GCC and GCG)
167 were the most common, while the two most common in PBCV-1 and AN69C ended in A
168 and U (GCA and GCU) (Table 3). A parallel example occurs in the usage of the two
169 synonymous codons for glutamic acid; GAG was almost exclusively used by *C. variabilis*
170 NC64A, while GAA was the most common in the two NC64A viruses. The same was true
171 for all amino acids encoded by two synonymous codons (asparagine, aspartic acid,
172 cysteine, glutamic acid, glutamine, histidine, lysine, phenylalanine and tyrosine), as well
173 as other amino acids with more than two synonymous codons. Furthermore, the
174 isoleucine codon AUA (AU-rich) occured in 2.44% of all PBCV-1 codons, while it occured
175 in only 0.20% of codons in *C. variabilis* NC64A, a 12-fold difference (Table 3); hence, in
176 the virus, natural selection appears to have favored the retention of the gene that encodes
177 the cognate tRNA for this codon. Likewise, the lysine codon AAA (AU rich) occured in
178 4.7% of all PBCV-1 codons, but in only 0.25% of host codons, nearly a 20-fold difference.
179 The leucine codon UUA was the rarest codon used by *C. variabilis* NC64A (0.06% of all
180 codons) while it was a moderately common codon used by PBCV-1 (1.39% of all codons)
181 (Table 3). As a final example, there are six synonymous codons for the amino acid
182 arginine, but only one codon that had one guanine or cytosine (AGA), while the other five
183 codons had a minimum of two guanines and/or cytosines. 1.43% of all viral codons were
184 AGA for arginine, but only 0.25% of *C. variabilis* NC64A codons were AGA. Indeed, the
185 two arginine codons with three guanines and cytosines (CGC and CGG) were the two
186 most common in *C. variabilis* NC64A (3.22% and 2.11% respectively), while the same
187 two codons in the two viruses averaged 0.75% and 0.70%, respectively.

Acquisition of chlorovirus tRNA clusters: One question we wished to address was: did the three chlorovirus clades independently acquire their tRNA clusters by horizontal gene transfer (HGT) or was the tRNA gene cluster acquired by the MRCA, i.e., the last common ancestor prior to splitting into the three clades? We used several lines of inquiry to address this question. First, as described above, the three clades of chloroviruses had seven tRNA genes in common with one another (Table 2), including 2 - 4 $tRNA^{asn}$ genes, whose encoded tRNAs recognize the codon AAC. There were, however, considerable differences in composition and order between clades, unlike the similar composition and synteny within each of the three clades (Tables 1A-C).

The second line of inquiry focused on the protein-encoding genes that border the 5' and 3' sides of the tRNA clusters. Supplementary Tables S2A-C report the commonness of protein-encoding genes (orthologous genes) within each of the three clades, but the protein-encoding genes surrounding the tRNA clusters of each clade were not orthologous across the three clades.

A third line of inquiry was to examine phylogenetic tree constructs of three tRNA genes common to all three clades of chloroviruses: $tRNA^{gly}$, $tRNA^{arg}$ and $tRNA^{tyr}$ (Fig 2A, 2B and 2C, respectively). Most of the chloroviruses tended to cluster within their clade, but there were a number of exceptions, including the presence of subclades, as will be discussed below.

A fourth line of inquiry focused on the $tRNA^{tyr}$ gene found in 34/41 chloroviruses representing all three clades, all 34 of which had an intron (Table S3). We examined the locations and sequences of the introns and found that the introns were located in identical positions in all the $tRNA^{tyr}$ genes, one nucleotide from the anticodon (3' direction). While the intron lengths and sequences were nearly identical within a clade, there were slight differences in length between clades: NC64A 13-14 nt; Pbi 13 nt; SAG 10-11 nt in the three clades.

DISCUSSION

Structural features of the chlorovirus tRNA clusters: The results presented in this paper indicate that all 41 chloroviruses, representing three clades, had clusters of tRNAs with intergenic spacers usually of 1 to ~30 nucleotides. Collectively, the

chloroviruses encoded a total of 410 tRNAs of which there were 17 different tRNAs for 14 different amino acids (3 synonymous codons). The large numbers of viral tRNA genes in the chloroviruses should not be surprising because Morgado and Vicente (2019) report a positive correlation between viral genome length and the number of tRNA genes in viruses. In addition, they found that tRNA gene clusters are more common in viruses with larger genomes than in those with smaller genomes. All the individual chlorovirus-encoded tRNAs lacked the 3' terminal three nucleotides (CCA) necessary in order to be aminoacylated. Since fully functional tRNAs were reported for the NC64A chlorovirus CVK2 (Nishida *et al*., 1999), the chlorovirus tRNAs, like tRNAs from cellular organisms (Rak *et al*., 2018), must either use an unidentified virus enzyme(s) or the host tRNA nucleotidytransferase to add the CCA nucleotides prior to tRNA aminoacylation for functionality.

Due to the short intergenic spacers between clustered chlorovirus tRNA genes, we suspect that the tRNA gene clusters are transcribed as one transcript. Indeed, Nishida *et al*., (1999) reported that chlorovirus CVK2 transcribes its tRNA gene cluster of 14 tRNAs into one transcript; furthermore, they reported that the RNA transcript was precisely processed into individual tRNA species by either some unknown virus-encoded or host-encoded RNase. In this regards, PBCV-1 encodes a functional RNase III enzyme (Zhang *et al*., 2003) but its role in virus replication is unknown.

Other viruses in the *Phycodnaviridae* family also have tRNA gene clusters. *Micromonas pusilla* virus 12T encodes six tRNA genes, five of which are clustered (NC_020864.1). The sixth tRNA gene, $tRNA^{thr}$, is an orphan ~30kb beyond the $tRNA^{leu}$ gene, the 3' member of the tRNA cluster. The position of these two tRNA genes and the distance between them is the same as in the chlorovirus SAG clade; i.e., there was a 30kb gap between the genes that encode $tRNA^{leu}$ and $tRNA^{thr}$. The virus 12T also encodes a $tRNA^{tyr}$, as do the chloroviruses, but unlike the chloroviruses, this *Micromonas* virus does not have an intron in its $tRNA^{tyr}$ gene. On the other hand, *Ostreococcus lucimarinus* virus 7 (OlV7) (Derelle *et al*., 2015; KP874737), which has five tRNA-clustered genes, encodes a $tRNA^{tyr}$ that does have a 15 nt intron located in the same position as the chloroviruses. In fact, the tRNA genes in the *Micromonas* and *Ostreococcus* virus clusters are similar to some of the tRNA gene clusters in the NC64A

viruses, suggesting the two groups of viruses might have a common evolutionary ancestor. Thus, while not all viruses of the *Phycodnaviridae* family were examined, tRNA clusters appear to be common in this family.

Functional features of the chlorovirus tRNA clusters: The chlorovirus-encoded tRNAs were evaluated to determine if they help the viruses overcome the CUB ascribed to the host (Fig 1 and Table 3). That is, most of the chlorovirus encoded tRNAs had codons favoring AU, whereas, the host codons had a GC bias. Nishida *et al*. (1999) reached a similar conclusion for the chloroviruses. Therefore, we conclude that the chloroviruses partially solved the CUB problem by encoding some tRNAs that support a virus AU bias. At least some of the differences in tRNA gene content between clades is likely due to the evolutionary pressures of utilizing three different hosts. We also observed unexplained differences within clades. Only $tRNA^{arg}$ was common to all 41 viruses. In the SAG viruses the $tRNA^{thr}$ gene that recognized ACU was located ~30kb downstream of the clusters. The Pbi virus NE-JV-1 is very odd in many aspects (Jeanniard *et al*., 2013), including that it encoded only three tRNAs and the location of its tRNA cluster was far downstream of all the 40 other viruses. So, with the exclusion of NE-JV-1, 40/40 chloroviruses had genes that encoded $tRNA^{asn}$ and $tRNA^{gly}$, while 37/40 encoded $tRNA^{lys}$.

However, not all tRNA members of a cluster assist in overcoming the CUB of the host. For example, PBCV-1 encoded 11 tRNAs but only seven help it to overcome CUB; likewise, AN69C encoded ten tRNAs but only eight favor its own CUB (Table 1A). Thus, we speculate that the cluster as a unit is under natural selection because some of the tRNA genes in the clusters are neutral and do not bestow any positive selective advantage. A reasonable explanation for these observations is that the neutral tRNA genes are preserved by natural selection as hitchhikers due to their close linkage to tRNA genes that help the viruses overcome the CUB of the host. To illustrate this point, all three clades of chloroviruses had 2-4 $tRNA^{asn}$ genes whose tRNAs only recognize the codon AAC; none of the 41 chloroviruses encoded a $tRNA^{asn}$ that would recognize the alternate codon AAU, which was 10X more common in the two viruses than in *C. variabilis* (Table 3), and which would presumably enhance viral protein translation. Of no surprise, the most common of the two Asn codons in *C. variabilis* is AAC. The presence of the AAC codon cognate $tRNA^{asn}$ genes, which appear to be neutral in benefit to all 41

chloroviruses, suggests that they were acquired in the MRCA tRNA cluster due to an evolutionary accident. That is, their presence appears to be an artifact of evolutionary history by which some neutral tRNA genes were acquired in the tRNA cluster along with selectively advantageous genes by some mechanism, such as HGT. (This kind of event is similar to the frozen accident concept first proposed by Crick (1968).) In the case at hand, neutral $tRNA^{asn}$ genes may be maintained in the tRNA gene clusters seemingly due to their tight linkage to selectively advantageous viral tRNA genes in the cluster. The same argument might explain the presence in PBCV-1 of the $tRNA^{lys}$ gene for the cognate codon AAG. PBCV-1 encoded both of the tRNAs genes that recognize the two lysine codons, AAG and AAA. However, the AAA codon was >20X more common in PBCV-1 than in *C. variabilis*, while AAG was the most common lysine codon in *C. variabilis*; hence, the former appears to be maintained by positive selection while the latter appears to provide a neutral benefit to PBCV-1.

Thirty-nine of the 41 chloroviruses encode another putative enzyme involved in codon usage, tRNA isoleucine lysidine synthase (TilS). The methionine codon AUG is normally recognized by its cognate $tRNA^{met}$ with the 3' UAC 5' anticodon. The TilS enzyme ligates lysine to the cytidine in the 5' position of the tRNA anticodon; this modified cytidine becomes lysidine, which is complementary to adenine in the 3' position of the codon, rather than guanine (Fig 3). As such, this modified tRNA then behaves as a $tRNA^{ile}$ and recognizes the isoleucine AUA codon (Nakanishi et al., 2005; Suzuki and Miyauchi, 2010). The AUA codon was 12-13 times more common in the two NC64A viruses than in the host, *C. variabilis* (Table 3). Thus, we suspect this enzyme provides one additional mechanism that helps the viruses overcome CUB of the host by enabling more efficient viral protein synthesis, diminishing the chances of a ribosome stall during elongation when AUA codons are encountered. In addition, all chloroviruses encode a homolog of translational elongation factor 3 (EF-3) (Jeanniard et al., 2013). EF-3 plays a role in optimizing the accuracy of mRNA decoding at the ribosomal acceptor site during protein synthesis in fungi (Belfield and Tuite, 1993; Belfield *et al*., 1995); EF-3 has been reported recently in algae (Mateyak *et al*., 2018). The role this putative enzyme plays in chlorovirus translation is unknown.

Evolutionary history of the chlorovirus tRNA clusters: The origin(s) of the chlorovirus tRNA clusters is intriguing, i.e., was there a single HGT event in the MRCA prior to the evolution of the chlorovirus clades, or did individual HGT events occur independently in each clade post-MRCA? A previous study suggests that the chloroviruses had a common host prior to their current three hosts (Jenniard *et al*., 2013). Support for the argument that the viruses may have acquired the tRNA clusters shortly after they separated into the three clades, includes: i) while the three clades of chloroviruses had some tRNA genes in common with one another, there were differences in the composition and gene order among the clades, unlike the somewhat uniform composition and order within each independent clade. ii) The protein-encoding genes that border the 5' and 3' sides of the tRNA clusters were orthologous for the viruses within each of the three clades but differed across the three clades.

In contrast, four lines of evidence favor the notion that the viruses acquired the tRNA cluster by a single HGT event in the MRCA prior to the evolution of all three clades of chloroviruses. i) 34/41 chloroviruses had a $tRNA^{tyr}$ gene and in all 34 cases the gene had a tRNA intron (Schmidt and Matera, 2019), which was located in an identical position among the 34 viruses representing all three clades, suggesting that the intron was inserted as a single event. While the lengths and sequences of the introns were nearly identical within a clade there were small differences in length between clades: NC64A 13-14 nt; Pbi 13 nt; SAG 10-11 nt (Table S3). Point mutations and indels over evolutionary time could explain the differences within and between the clades. ii) All three clades had several tRNA genes in common, as described above and seen in Tables 1A-C. Perhaps the most interesting is the $tRNA^{asn}$ gene, which occured in 2-4 copies in 40/40 chloroviruses (excluding NE-JV-1). Every single $tRNA^{asn}$ recognizes the codon AAC, but none of the 40 viruses encoded a $tRNA^{asn}$ cognate for the codon AAU, which was the most common codon in the viruses and least common in the host (Table 3). Independent acquisition of the same neutral tRNA gene by each of the three clades seems unlikely. iii) The third line of support can be seen in Figures 2A, 2B and 2C, which display the phylogenetic tree constructs of $tRNA^{gly}$, $tRNA^{tyr}$ and $tRNA^{arg}$ genes, genes that were common to all three clades. While most of the viruses assorted within their respective clades, it was the exceptions that support a single HGT event hypothesis. For example,

the tRNA^{tyr} genes of four of the NC64A viruses were more similar to the SAG viruses than to other NC64A viruses (Fig 2A). Likewise, for the tRNA^{gly} gene, half of the NC64A viruses were more similar to SAG viruses, while the other half of the NC64A viruses were more similar to Pbi viruses (Fig 2B); furthermore, there were two Pbi and two SAG viruses that were more similar to a subclade of NC64A viruses than to other members of their own clade. iv) The fourth line of support involves the tRNA^{thr} gene that was found in all of the SAG and Pbi viruses. In all 14 of the Pbi viruses, this gene was the most 3' member of the cluster, but for all 13 SAG viruses, the tRNA^{thr} gene was located ~30kb beyond the 3' end of the tRNA cluster. One possible explanation is that the tRNA cluster in the SAG viruses was translocated in the 5' direction, but without the tRNA^{thr} gene, prior to speciation within the SAG clade. Indeed, unlike the NC64A and Pbi viruses whose tRNA clusters were located near the center of their genomes, the tRNA clusters in the SAG viruses were located more towards the 5' direction of their genomes – i.e., in the first third of their genomes. If a translocation did take place, then the original location of the cluster in the ancestral SAG virus would have been more towards the center of its genome. It is perhaps of note that the average genome size of the NC64A, Pbi and SAG clades is as follows: 326, 321 and 307 kb, respectively (Jenniard *et al*., 2013). A translocation event could not only explain the relocation of the SAG tRNA gene cluster, but also the reduction in genome size of the ancestral SAG virus.

Therefore, we feel that the evidence most strongly supports a single origin for the chloroviruses tRNA gene clusters. Regardless of the origin, it is clear that many evolutionary changes have occurred. The ability of tRNA genes to proliferate is thought to be similar to the mechanism by which mobile elements can lead to intragenomic gene duplications (Velandia-Huerto *et al*., 2016). Duplications and losses are clearly evident among and within all three chlorovirus clades; for example, the Pbi virus CZ-2 had four tRNA^{asn} genes while almost all of the other Pbi viruses had just two copies. As well, the NC64A virus MA-1E had three tRNA^{lys} genes, while some others in the same clade had just one, and AR158 had none. There are other similar examples among the three chlorovirus clades.

Pope *et al*. (2014) implicates a homing endonuclease (HNH) in the generation of tRNA genes in mycobacteriophages. While not immediately adjacent to the tRNA gene

clusters, all of the chloroviruses encode at least two putative HNHs (e.g., orthologs of A087R and A422R in PBCV-1). A second source of HNH or other endonucleases might have been from co-infecting viruses that had such genes in their repertoire that generated tRNA gene duplications, losses and translocations within the chloroviruses. Previously, we proposed three potential sources of HGT for the chlorovirus protein-encoding genes that might also explain the tRNA clusters in the chloroviruses: (i) viral host(s), although there are only a few NC64A chlorovirus genes that have likely been acquired from *C. variabilis* NC64A but the viruses probably had at least one other host through evolutionary time; (ii) bacteria, because some of the chlorovirus genes appear to be of bacterial origin; and, (iii) from other host-competing viral species (Jeanniard *et al.*, 2013). Plastids and mitochondria might also have contributed to the viruses via HGT, at least for the tRNA gene clusters, because those organelles have tRNA gene clusters, as well. These results are consistent with the analyses and conclusions of Fan *et al*. (2017) who sequenced the mitochondria and plastids of the three chlorovirus hosts, *C. variabilis*, *C. heliozoae* and *M. conductrix.* Indeed, Margado and Vicente (2019) propose that viruses with tRNA clusters might be the source of dissemination of such clustered structures in the three domains of life. Acquisition of clustered genes appears to be a common occurrence among the chloroviruses; six of the chloroviruses examined by Filée et al. (2008) appear to have acquired genes from both bacteria and eukarotes, in many cases they were in clusters (see Supplementary Material; Additional File 1 in the Filée reference). The chlorovirus genomes appear to have been randomly inserted with those acquired genes. Two of the six chloroviruses that were examined had numerous insertion elements, which could explain the movement of genes between genomes and within genomes. Additionally, gene gangs are conserved clusters of colinear monocistronic chlorovirus genes, some of which have an apparent common origin (Seitzer et al., 2018).

In the larger picture about viruses encoding constituents of the protein synthesis machinery, it is clear that many of the recently discovered giant viruses encode components of the protein synthetic machinery (e.g., Brandes and Linial, 2019). The first giant virus to be discovered, *Acanthamoeba polyphaga mimivirus* (AMPV), encodes four putative aminoacyl tRNA synthetases (aaRS) and six tRNAs (Raoult *et al*., 2004), which, unlike the chloroviruses, are not clustered. One of the four mimivirus aaRSs, the tyrosine

tRNA synthetase, has been crystalized and shown to function (Abergel *et al*., 2005). More amazing, two recently described giant Tupanviruses, a close relative of the mimiviruses, encode up to 70 tRNA genes, many in clusters of ~15 tRNAs, and 20 aaRS genes, plus many more protein synthesis genes (Abrahao *et al*., 2018). Thus, these newly discovered giant viruses are apparently solving the CUB issues by encoding several elements of the protein synthetic machinery. It will be exciting to find out if all of these putative proteins have their predicted activities.

In summary, the two hosts for the chloroviruses that have been genomically sequenced have a high GC content (67%) and natural selection tends to favor GC-rich codons, whereas in the chloroviruses natural selection favored AU-rich codons. The viruses appear to overcome the CUB of their hosts by maintaining a cluster of tRNA-encoding genes that favor cognate codons richer in AU. However, the evolutionary events have differed among the viruses, even within the same clade, because very few of the tRNAs were conserved among all of the chloroviruses. We are aware that tRNAs are turning out to play other roles in cells (e.g., Lyons et al., 2018) and so it is always possible that the viral encoded tRNAs have some other functions.

ACKNOWLEDGEMENTS

This research was supported by the National Science Foundation under Grant No. 1736030 and the University of Nebraska-Lincoln Agricultural Research Division and the Office of Research and Economic Development.

LITERATURE CITED

Abe, T., Inokuchi, H., Yamada, Y., Muto, A., Iwasaki, Y., Ikemura, T. tRNADB-CE: tRNA gene database well-timed in the era of big sequence data. *Front. Genet.* **2014**, 5, 114.

Abergel, C., Chenivesse, S., Byrne, D., Suhre, K., Arondel, V., Claverie, J.M. Mimivirus TryRS: preliminary structural and functional characterization of the first amino-acyl tRNA synthetase found in a virus. *Acta Crystallogr Sect F Struct. Biol. Cryst. Commun*. **2005**, 61, 212-215.

Abrahoa, J., Silva, L., Silva, L.S., Khalil, J.Y.B., Rodrigues, R., Arantes, T., Assis, F., Bratto, P., Andrade, M., Kroon, E.G., Ribeiro, B., Bergier, I., Seligmann, H., Ghigo,

E., Colson, P., Levasseur, A., Kroemer, G., Raoult, D., Sa Scola, B. Tailed giant Tupanvirus possesses the most complete translational apparatus of the known virosphere. *Nat. Commun*. **2018**, 9, 749. Doi: 110.1038/s41467-018-03168-1.

Arriola, M.B., Velmurugan, N., Zhang, Y., Plunkett, M.H., Hondzo, H., Barney, B.M. Genome sequences of *Chlorella sorokiniana* UTEX 1602 and *Micractinium conductrix* SAG 241.80: implications to maltose excretion by a green alga. *Plant J.* **2018**, 93, 566-586.

Belfield, G.P., Ross-Smith, N.J., Tuite, M.F. Translation elongation factor-3 (EF-3): an evolving eukaryotic ribosomal protein? *J. Mol. Evol*. **1995**, 41, 376-387.

Belfield, G.P., Tuite, M.F. Translation elongation factor-3: a fungus-specific translation factor? *Mol Microbiol*. **1993**, 9, 411-418.

Bermudez-Santana, C., Attolini, C.S.-O., Kirsten, T., Engelhardt, J., Prohaska, S.J., Steigele, S., Stadler, P.F. Genomic organization of eukaryotic tRNAs. *BMC Genom*. **2010**, 11, 1-14.

Blanc, G., Duncan, G., Agarkova, I., Borodovsky, M., Gurnon, J., Kuo, A., Lindquist, E., Lucas, S., Pangilinan, J., Salamov, A., Terry, A., Yamada, T., Dunigan, D.D., Grigoriev, I.V., Claverie, J.M., Van Etten, J.L. The *Chlorella variabilis* NC64A genome reveals adaptation to photosymbiosis, coevolution with viruses and cryptic sex. *Plant Cell* **2010**, 22, 2943-2955.

Brandes, N., Linial, M. Giant viruses-big surprises. *Viruses*. **2019**, 11, 404. doi: 10.3390/v11050404.

Crick, F.H.C. The origin of the genetic code. *J. Mol. Biol.* **1968**, 38, 367–379

Derelle, E., Monier, A., Cooke, R., Worden, A.Z., Grimsley, N.H., Moreau, H. Diversity of viruses infecting the green microalga *Ostreococcus lucimarinus*. *J Virol.* **2015**, 89, 5812-5821.

Dunigan, D.D., Cerny, R.L., Bauman, A.T., Roach, J.C., Lane, L.C., Agarkova, I.V., Wulser, K., Yanai-Balser, G.M., Gurnon, J.R., Vitek, J.C., Kronschnabel, B.J., Jeanniard, A., Blanc, G., Upton, C., Duncan, G.A., McClung, O.W., Ma, F., Van Etten, J.L. *Paramecium bursaria* chlorella virus 1 proteome reveals novel architectural and regulatory features of a giant virus. J *Virol.* **2012**, 86, 8821-8834.

Fan, W., Guo, W., Van Etten, J.L., Mower, J.P. Multiple origins of endosymbionts in *Chlorellaceae* with no reductive effects on the plastid or mitochondrial genomes. *Sci Rep* **2017**, 7, 1-10.

Filée, J., Pouget, N., Chandler, M. 2008. Phylogenetic evidence of extensive lateral acquisition of genes by Nucleocytoplasmic large DNA viruses. BMC Evol Biol. 8:320

Fitzgerald, L.A., Graves, M.V., Li, X., Feldblyum, T., Nierman, W.C., Van Etten, J.L. Sequence and annotation of the 369-kb NY-2A and the 345-kb AR158 viruses that infect *Chlorella* NC64A. *Virology* **2007**, 358, 472–484.

Fitzgerald, L.A., Graves, M.V., Li, X., Hartigan, J., Pfitzner, A.J.P., Hoffart, E., Van Etten, J.L. Sequence and annotation of the 288-kb ATCV-1 virus that infects an endosymbiotic chlorella strain of the heliozoon *Acanthocystis turfacea*. *Virology* **2007a**, 362, 350–361.

Fitzgerald, L.A., Graves, M.V., Li, X., Feldblyum, T., Hartigan, J., Van Etten, J.L. Sequence and annotation of the 314-kb MT325 and the 321-kb FR483 viruses that infect Chlorella Pbi. *Virology* **2007b**, 358, 459–471.

Friedrich, A., Jung, P.P., Hou, J., Neuvéglise, C., Schacherer, J. Comparative mitochondrial genomics within and among yeast species of the *Lachancea* genus. *PLoS ONE*, **2012**, 7, e47834.

Jeanniard, A., Dunigan, D.D., Gurnon, J.R., Agarkova, I.V., Kang, M., Vitek, J., Duncan G., McClung, O.M., Larsen, M., Claverie, J.M., Van Etten, J.L., Blanc, G. Towards defining the chloroviruses: a genomic journey through a genus of large DNA viruses. *BMC Genomics*. **2013**, 14, 158.

Lee, D.Y., Graves, M. V., Van Etten, J. L., Choi, T-J. Functional implication of tRNA genes encoded in the chlorella virus PBCV-1 genome. *Plant Pathol. J.* **2005**, 21, 334-342.

Li, Y., Lu, Z., Sun, L., Ropp, S., Kutish, G.F., Rock, D.L.,Van Etten J.L. Analysis of 74 kb of DNA located at the right end of the 330-kb chlorella virus PBCV-1 genome. *Virology* **1997**, 237, 360-377.

Lowe, T.M., Chan, P.P. tRNAscan-SE on-line: Search and contextual analysis of transfer RNA genes. *Nucleic Acids Res.* **2016**, 44, W54-57.

494 Lowe, T.M., Eddy, S.R. tRNAscan-SE: a program for improved detection of transfer RNA
495 genes in genomic sequence. *Nucleic Acids Res.* **1997**, 25, 955–964.
496 Mateyak, M.K., Pupek, J.K., Garino, A.E., Knapp, M.C., Colmer, S.F., Kinzy, T.G.,
497 Kunaway, S. Demonstration of translation elongation factor 3 activity from a non-
498 fungal species, *Phytophthora infestans*. *PLoS One*. **2018**, 13, e0190524.
499 Michely, S., Toulza, E., Subirana, L., John, U., Cognat, V., Marechal-Drouard, L.,
500 Grimsley N., Moreau, H., and Piganeau, G. Evolution of codon usage in the
501 smallest photosynthetic eukaryotes and their giant viruses. *Genome Biol. Evol*.
502 **2013**, 5, 848–859.
503 Morgado, S.M., Vicente, A.C.P. Beyond the limits: tRNA array units in *Mycobacterium*
504 genomes. *Front. Microbiol*. **2018**, 9, 1042.
505 Morgado, S.M., Vicente, A.C.P. Global *in-silico* scenario of tRNA genes and their
506 organization in virus genomes. *Viruses* **2019,** 11, 180.
507 Morgado, S.M. and Vicente, A.C.P. Exploring tRNA gene cluster in archaea. *Mem Inst*
508 *Oswaldo Cruz*, Rio de Janeiro, **2019a**, 114, e180348.
509 Nakanishi, K, Fukai, S., Ikeuchi, Y., Soma, A., Sekine, Y., Suzuki T., Nureki O. Structural
510 basis for lysidine formation by ATP pyrophosphatase accompanied by a lysine-
511 specific loop and a tRNA recognition domain. *Proc. Natl. Acad. Sci. USA*. **2005**,
512 102, 7487-7492.
513 Nishida, K, Kawasaki, T., Fujie, M., Usami, S., Yamada, T. Aminoacylation of tRNAs
514 encoded by *Chlorella* virus CVK2. *Virology* **1999**, 263, 220-229.
515 Pagarete, A., Lanzen, A., Puntervoll, P., Sandaa, R.A., Larsen, A., Larsen, J.B., Allen,
516 M.J., Bratbak, G. Genomic sequence and analysis of EhV-99B1, a new
517 Coccolithovirus from the Norwegian Fjords. *Intervirology.* **2013**, 56, 60-66.
518 Pope, W.H., Anders, K.R., Baird, M., Bowman, C.A., Boyle, M.M., Broussard, G.W.,
519 Chow, T., Clase, K.L., Cooper, S., Cornely, K.A., *et al*. Cluster M
520 mycobacteriophages Bongo, PegLeg, and Rey with unusually large repertoires of
521 tRNA isotypes. *J. Virol*. **2014**, 88, 2461–2480.
522 Raoult, D., Audic, S., Robert, C., Abergel, C., Renesto, P., Ogata, H., La Scola, B., Suzan,
523 M., Claverie, J.M. The 1.2-megabase genome sequence of Mimivirus. *Science*
524 **2004**, 306, 1344-1350.

525 Rak, R., Dahan, O., Pilpel, Y. Repertoires of tRNAs: the couplers of genomics and
526 proteomics. *Ann. Rev. Cell Devel. Biol*. **2018**, 34, 239-264.
527 Santini, S., Jeudy, S., Bartoli, J., Poirot, O., Lescot, M., Abergel, C., Barbe, V., Wommack,
528 K.E., Noordeloos, A.A., Brussaard, C.P., Claverie, J.M. Genome of *Phaeocystis*
529 *globosa* virus PgV-16T highlights the common ancestry of the largest known DNA
530 viruses infecting eukaryotes. *Proc. Natl. Acad. Sci. USA*. **2013**, 110, 10800-10805.
531 Schmidt, C.A.,, Matera, A.G. tRNA introns: Presence, processing, and purpose. *Wiley*
532 *Interdiscip Rev RNA*. **2020**;11(3):e1583. doi:10.1002/wrna.1583
533 Seitzer, P., Jeanniard, A., Ma, F., Van Etten, J.L., Facciotti, M.T., Dunigan, D.D. Gene
534 gangs of the chloroviruses: Conserved clusters of collinear monocistronic
535 genes. *Viruses*. **2018**;10(10):576. doi:10.3390/v10100576
536 Stecher, G., Tamura, T., Kumar, S. 2020. Molecular evolutionary genetics analysis
537 (MEGA) for macOS. *Mol. Biol. Evol.* **2020***,* 37, 1237-1239.
538 Suzuki, T., Miyauchi, K. Discovery and characterization of tRNAIle lysidine synthetase
539 (TilS). *FEBS Lett*. **2010**, 584, 272-277.
540 Tran, T.T., Belahbib, H., Bonnefoy, V., Talla, E. A comprehensive tRNA genomic survey
541 unravels the evolutionary history of tRNA arrays in prokaryotes. *Genome Biol.*
542 *Evol*. **2015**, 8, 282–295.
543 Van Etten, J.L., Agarkova, I.V., Dunigan, D.D. Chloroviruses. *Viruses* 2020, 12, 20.
544 Velandia-Huerto, C.A., Berkemer, S.J., Hoffman, A., Retzlaff, N., Romero Marroquin,
545 L.C., Hernandez-Rosales, M., Stadler, P.F., Bermudez-Santans, C.I. Orthologs,
546 turn-over, and remolding of tRNAs in primates and fruit flies. *BMC Genom.* **2016**,
547 17, 617.
548 Yanai-Balser, G.M., Duncan, G.A., Eudy, J.D., Wang, D., Li, X., Agarkova, I.V., Dunigan,
549 D.D., Van Etten, J.L. Microarray analysis of Paramecium bursaria chlorella virus 1
550 transcription. *J. Virol.* **2010**, 84, 532–542.
551 Zhang, Y., Calin-Jageman, I., Gurnon, J.R., Choi, T.J., Adams, B., Nicholson, A.W., Van
552 Etten, J.L. Characterization of a chlorella virus PBCV-1 encoded ribonuclease III.
553 *Virology* **2003**, 317, 73-83.

Table 1. A) The order of clustered tRNA genes from NC64A viruses. **B)** The order of clustered tRNA genes from SAG viruses. **C)** The order of clustered tRNA genes from Pbi viruses.

Table 1A.

NC64A viruses	Codons recognized by NC64A virus tRNAs														Total tRNAs
	Leu-1 UUG	Ile AUA	Leu-2 UUA	Asn-1 AAC	Gly GGA	Asn-2 AAC	Lys AAG	Gln CAG	Tyr UAC[a]	Lys-1 AAA	Lys-2 AAG	Arg AGA	Asp GAC	Val GUU	
MA-1E															14
CvsA1															14
CviKI															14
KS1B															12
PBCV-1					b						c				11
IL-3A															11
MA-1D															12
NE-JV-4								d							11
AN69C															10
NY-2B					e							f			9
IL-5-2s1					e							f			9
NY-2A					e							f			8
NYs-1					e							g			8
AR158			h		e										7

The heading of each column identifies the cognate amino acid and codon for each encoded tRNA. Red color indicates same tRNA gene; orange indicates a pseudogene; green indicates a tRNA gene substitution; white indicates absense of tRNA gene.
[a]tRNA^{tyr} contains an intron; [b]tRNA^{lys} substitution for tRNA^{gly}; [c]Silent mutation codon change from AAG to AAA; [d]tRNA^{asn} substitution for tRNA^{gln}; [e]Appears to have been transposed, follows tRNA^{arg}; [f]The gene that encodes this tRNA^{arg} follows Leu-2; [g]Follows Asn-1; [h]Follows Asn-1.

Table 1B. The order of clustered tRNA genes from SAG viruses.

SAG viruses	Codons recognized by SAG virus tRNAS																		
	Ile-1 AUA	Ser AGU	Arg AGA	Asn-1 AAC	Gly GGA	Ile-2 AUU	Asn-2 AAC	Met AUG	Asp GAC	Val-1 GUU	Val-2 GUU	Asn-3 AAC	Tyr UAC[a]	Lys AAG	Leu-1 UUG	Asn-4 AAC	Leu-2 UUA	Thr ACU[b]	**Total tRNAs**
OR0704.3																			13
Can0610SP																			13
NE-JV-2																			13
NE-JV-3																			12
ATCV-1																	c		11
WI0606																			11
MO0605SPH																			11
GM0701.1																			10
Br0604L																			9
TN603.4.2																			9
Canal-1									d										10
MN0810.1																			9
NTS-1																			7

The heading of each column identifies the cognate amino acid and codon for each encoded tRNA. Red indicates the presence of a tRNA gene; orange indicates a pseudogene; white indicates the absence of the tRNA gene.
[a]$tRNA^{tyr}$ contains an intron; [b]This tRNA gene is 30,000-31,502 nt downstream of the tRNA cluster; [c]This tRNA gene is adjacent to the other tRNA genes but is within the protein encoding gene Z256R; [d]Translocated towards the 3' end of the cluster, between Asn-4 and Leu-2 tRNA genes.

Table 1C. The order of clustered tRNA genes from Pbi viruses.

Pbi Viruses	Codons recognized by Pbi virus tRNAS															Total tRNAs
	Ile AUA	Leu UUA	Phe UUC	Arg AGA	Gly GGA	Asn-1 AAC	Tyr-1 UAC[a]	Lys-1 AAG	Asn-2 AAC	Asn-3 AAC	Asn-4 AAC	Tyr-2 UAC[a]	Lys-2 AAG	Thr-1 ACG	Thr-2 ACG	
Fr5L			■	■	■	■	■	■	■	■		■		■	■	11
CZ-2			■	■	■	■			■	■	■	■	■	■		10
MT325	■	■	■	■	■	■			■			■	■	■		10
Can18-4	■	■	■	■	■	■			■			■	■	■		10
CVB-1	■	■	■	■	■	■			■			■	■	■		10
FR483	■	■		■	■	■			■			■	■	■		9
CVG-1	■		■	■	■	■			■			■	■	■		9
CVR-1	■		■	■	■	■			■			■	■	■		9
CVA-1	■		■	■	■	■			■			■	■	■		9
AP110A	■		■	■	■	■			■			■	■	■		9
CVM-1	■		■	■	■	■			■			■	■	■		9
NW665.2	■			■	■	■			■			■	■	■		8
OR0704.2.2				■	■	■			■			■	■	■		7
NE-JV-1	■			■										■		3

The heading of each column identifies the cognate amino acid and codon for each encoded tRNA. Red indicates the presence of a tRNA gene; white indicates the absence of the tRNA gene.
[a]tRNAtyr contains an intron.

Table 2. tRNA genes common to one another and unique to each chlorovirus clade[1].

tRNA	Codon	NC64A	SAG	Pbi
Ile-1	AUA	c	c	c
Leu-1	UUA	c	c	c
Asn-1	AAC	c	c	c
Gly-1	GGA	c	c	c
Lys-1	AAG	c	c	c
Tyr-1	UAC	c	c	c
Arg-1	AGA	c	c	c
Asp-1	GAC	d	d	
Val-1	GUU	d	d	
Leu-2	UUG	d	d	
Gln-1	CAG	u		
Lys-2	AAA	u		
Ser-1	AGU		u	
Ile-2	AUU		u	
Met-1	AUG		u	
Thr-1	ACU		u	
Phe-1	UUC			u
Thr-1	ACG			u
Thr-2	ACG			u

[1](c) means that some members in all three clades of the chloroviruses have the gene. (d) means that some members in the NC64A and SAG chlorovirus clades have the gene. (u) means that the gene is unique to some viruses in one of the three clades of chloroviruses.

Table 3. Codon frequency use comparison: virus to host

Codon	AA	PBCV-1 % usage	AN69C % usage	C. variabilis NC64A % usage	Ratio of codon frequency #
GCA	A	1.92	1.15	2.42	2.24
GCC	A	0.86	0.64	5.54	0.47
GCG	A	1.29	0.84	5.46	0.68
GCT	A	1.30	0.82	1.72	2.17
TGC	C	0.67	1.08	1.77	1.65
TGT	C	1.23	1.88	0.32	16.24
GAC	D	1.97	1.31	3.04	1.89
GAT	D	3.02	1.91	1.06	8.19
GAA	E	3.69	2.47	0.54	20.12
GAG	E	1.26	1.08	4.92	0.82
TTC	F	2.53	2.40	1.55	5.47
TTT	F	2.94	3.49	0.95	11.50
GGA	G	1.71	1.28	0.76	6.89
GGC	G	0.61	0.61	5.88	0.36
GGG	G	1.12	0.92	2.07	1.70
GGT	G	2.10	1.42	0.67	9.27
CAC	H	0.89	1.21	1.89	1.87
CAT	H	1.26	1.87	0.52	10.12
ATA	I	2.44	2.64	0.20	44.36
ATC	I	2.00	1.80	1.68	3.90
ATT	I	2.87	2.88	0.44	22.58
AAA	K	4.70	3.76	0.25	58.94
AAG	K	2.48	1.91	2.53	3.01
CTA	L	0.85	0.93	0.27	11.36
CTC	L	1.26	1.09	1.47	2.76
CTG	L	0.85	1.08	6.85	0.48
CTT	L	1.65	1.67	0.50	11.46
TTA	L	1.39	1.81	0.06	84.90
TTG	L	1.76	2.06	0.56	11.69
ATG	M	2.76	1.97	1.84	4.49
AAC	N	2.62	2.28	1.48	5.73
AAT	N	3.16	2.78	0.30	34.36
CCA	P	1.42	1.24	1.08	4.27
CCC	P	1.07	0.91	2.55	1.34
CCG	P	0.96	0.97	2.30	1.44
CCT	P	1.33	0.89	0.96	4.06
CAA	Q	1.84	2.09	0.59	11.36
CAG	Q	0.87	1.06	4.93	0.67
AGA	R	1.43	1.71	0.25	21.15
AGG	R	0.68	0.84	0.92	2.81
CGA	R	0.66	1.48	0.44	7.99
CGC	R	0.66	0.84	3.22	0.79
CGG	R	0.45	0.94	2.11	1.08
CGT	R	1.05	1.38	0.44	9.35
AGC	S	0.71	0.82	3.03	0.86
AGT	S	1.26	1.24	0.27	15.81
TCA	S	1.48	1.78	0.43	12.81
TCC	S	0.98	1.34	1.33	2.94
TCG	S	1.10	1.38	1.05	3.98
TCT	S	1.88	1.69	0.51	11.98
ACA	T	2.01	1.87	0.61	10.87
ACC	T	1.32	1.38	1.98	2.34
ACG	T	1.62	1.57	1.29	4.24
ACT	T	1.57	1.31	0.39	12.67
GTA	V	1.80	1.54	0.25	23.06
GTC	V	1.35	1.25	1.16	3.88
GTG	V	1.57	1.37	4.23	1.20
GTT	V	2.34	2.28	0.40	19.96
TGG	W	1.09	1.24	1.61	2.48
TAC	Y	1.52	1.46	1.42	3.62
TAT	Y	2.24	2.66	0.38	22.02

#, Mean virus codon frequency per Kb genome / Codon frequency per Kb genome NC64A. Heat map indicates the ratio of codon frequencies of virus compared to host, red = low ratio, green = high ratio.

Red-flagged codon and amino acid indicates that Chlorovirus AN69C encodes the cognate tRNA.

Green-flagged codon and amino acid indicates that both Chlorovirus AN69C and PBCV-1 encodes the cognate tRNA.

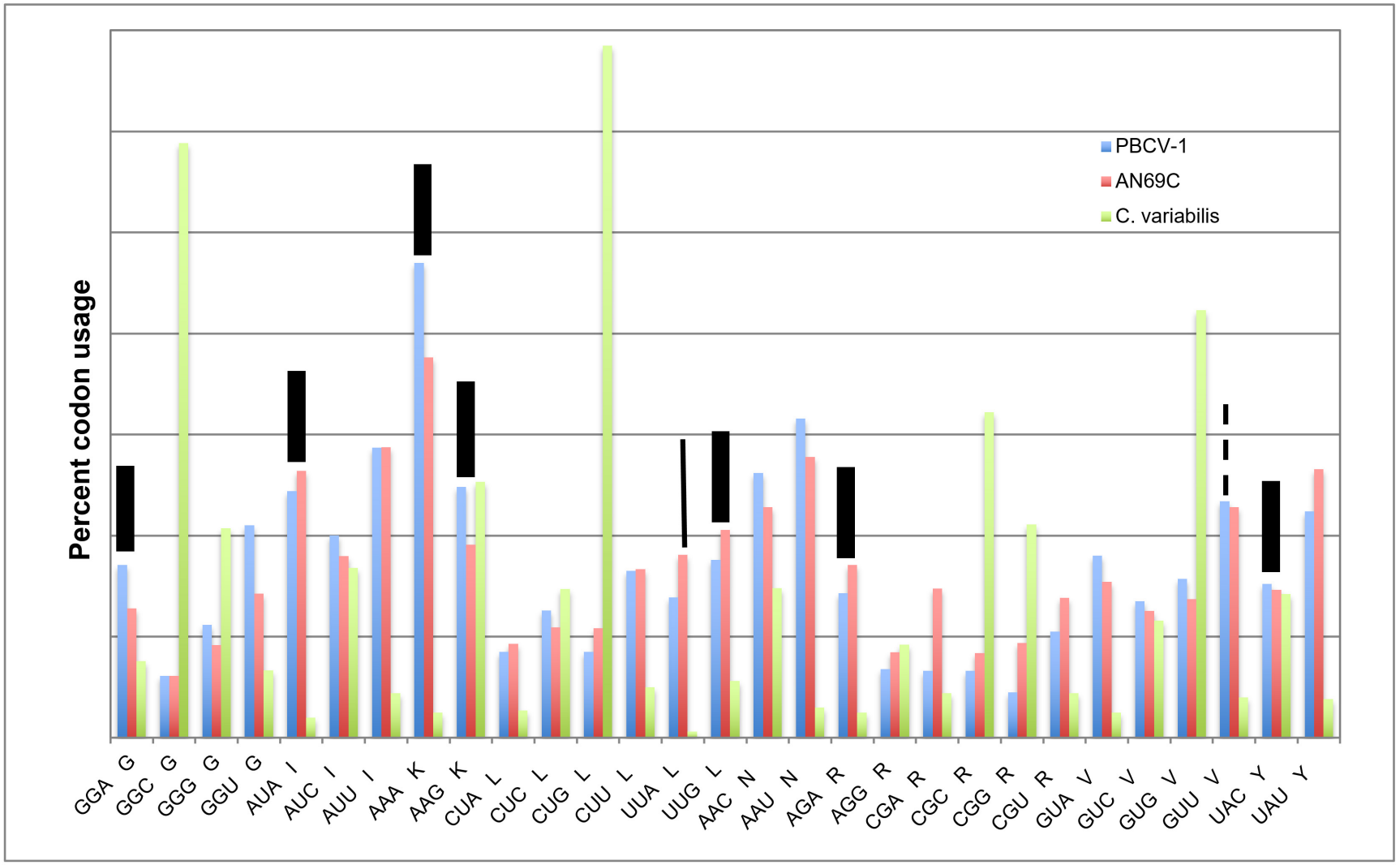

Figure 1. Comparison of the frequency of codon usage recognized by tRNAs encoded by PBCV-1 and AN69C, compared with their host *C. variabilis* NC64A. Wide bars denote the codons recognized by tRNAs encoded by PBCV-1 and AN69A. Narrow solid bar denotes codon recognized by tRNA encoded by AN69C but not PBCV-1. Narrow dash bar denotes codon recognized by tRNA encoded by PBCV-1 but not AN69C.

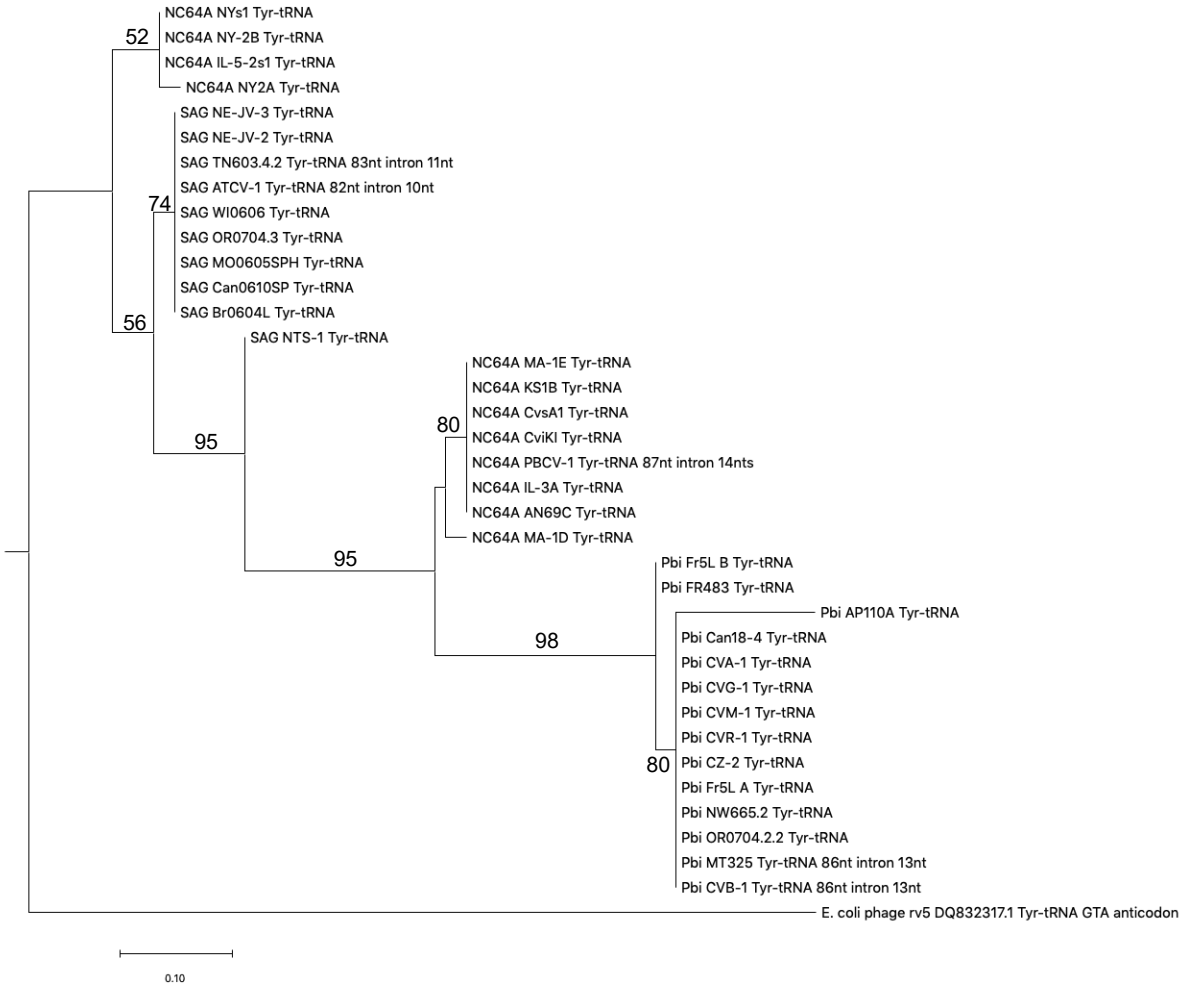

Figure 2A. **A**) Phylogenetic tree of tRNA^{tyr} genes from 35 chloroviruses representing all three clades, NC64A, Pbi and SAG. Six chloroviruses lacked the tRNA^{tyr} gene. The tRNA^{tyr} gene from *E. coli* phage rv5 was used as the outgroup. One subclade of NC64A viruses are more similar to SAG viruses, while the other subclade of NC64A viruses are more similar to the Pbi viruses. **B**) Phylogenetic tree of tRNA^{gly} genes from 39 chloroviruses representing all three clades, NC64A, Pbi and SAG. Two chloroviruses lacked the tRNA^{gly} gene. The tRNA^{gly} gene from cyanophage NATL1A was used as the outgroup. One NC64A subclade is more similar to SAG viruses than the other subclade of NC64A viruses. Two SAG viruses and two Pbi viruses are more similar to a second subclade of NC64A than they are to viruses in their own clades. **C**) Phylogenetic tree of tRNA^{arg} genes from all 41 chloroviruses representing all three clades, NC64A, Pbi and SAG. The tRNA^{arg} gene from enterobacteria phage EU330206.1 was used as the outgroup. One NC64A subclade and two SAG viruses are more similar to Pbi viruses than they are to viruses in their own clade. Bootstrap values greater than 50 are reported. The sequences were aligned with MUSCLE and the trees were constructed using the maximum likelihood algorithm.

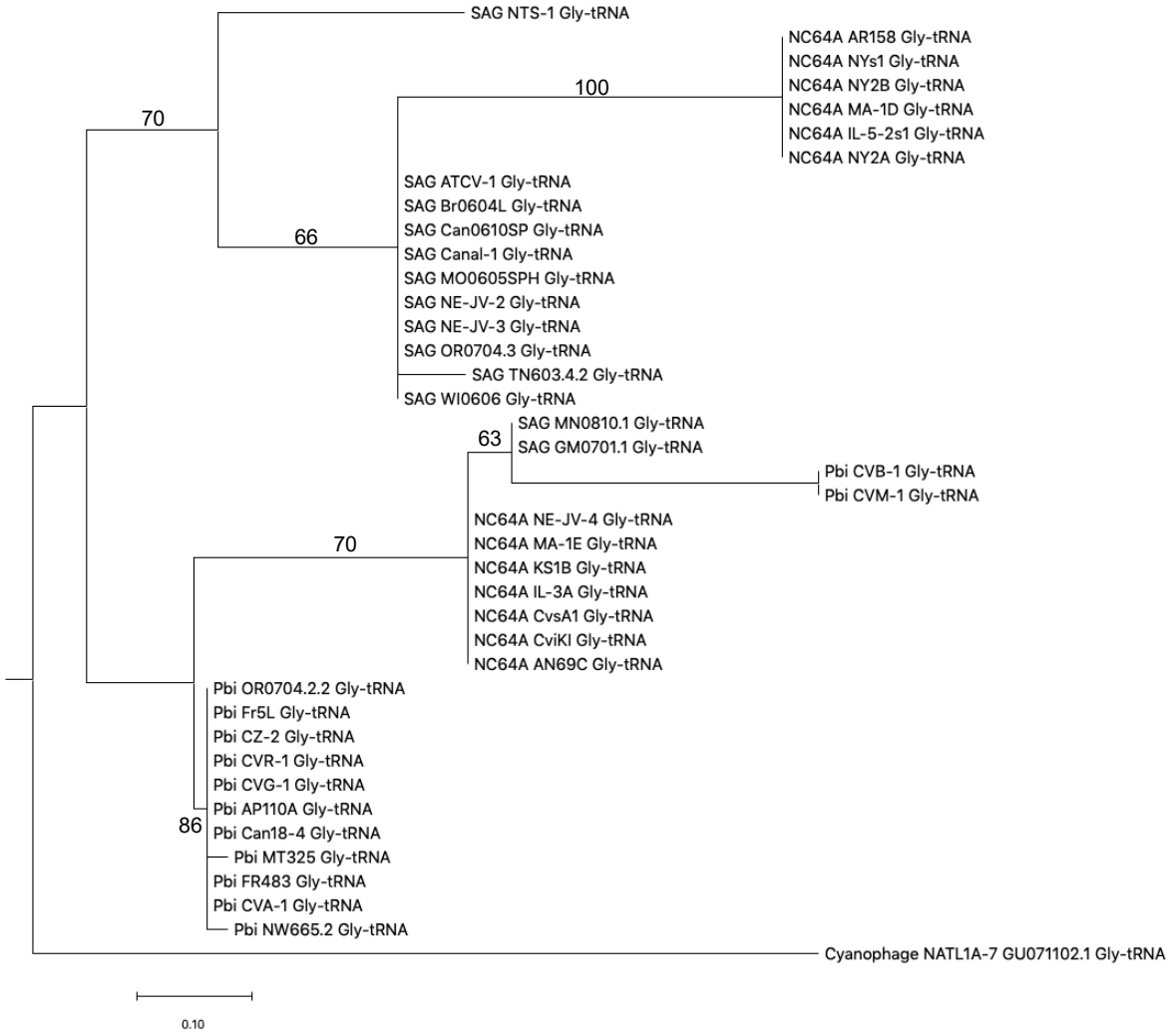

Figure 2B. Phylogenetic tree of $tRNA^{gly}$ genes from 39 chloroviruses representing all three clades, NC64A, Pbi and SAG. Two chloroviruses lacked the $tRNA^{gly}$ gene. The $tRNA^{gly}$ gene from cyanophage NATL1A was used as the outgroup. One NC64A subclade is more similar to SAG viruses than the other subclade of NC64A viruses. Two SAG viruses and two Pbi viruses are more similar to a second subclade of NC64A than they are to viruses in their own clades. Bootstrap values greater than 50 are reported. The sequences were aligned with MUSCLE and the trees were constructed using the maximum likelihood algorithm.

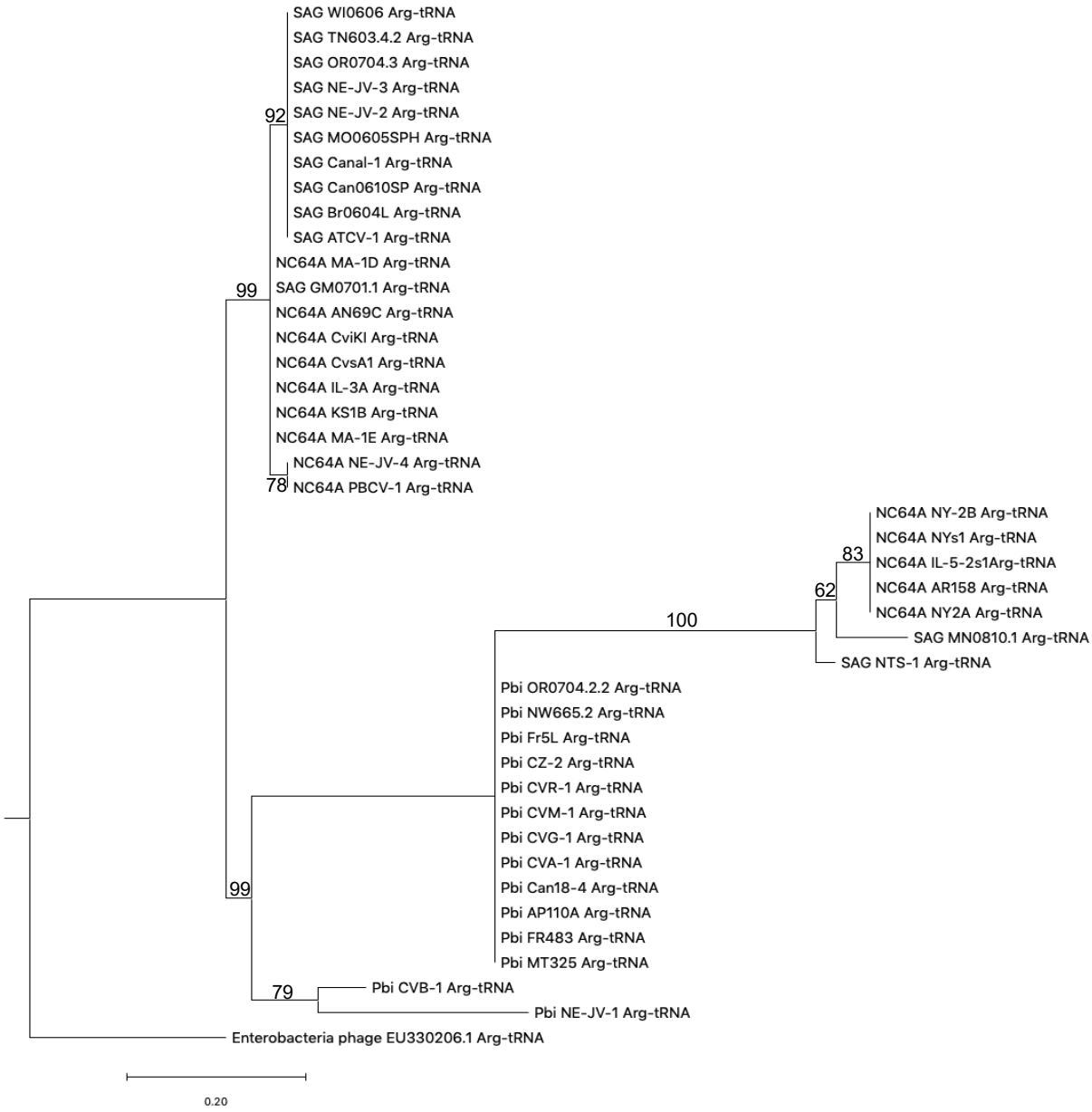

Figure 2C. Phylogenetic tree of $tRNA^{arg}$ genes from all 41 chloroviruses representing all three clades, NC64A, Pbi and SAG. The $tRNA^{arg}$ gene from enterobacteria phage EU330206.1 was used as the outgroup. One NC64A subclade and two SAG viruses are more similar to Pbi viruses than they are to viruses in their own clade. Bootstrap values greater than 50 are reported. The sequences were aligned with MUSCLE and the trees were constructed using the maximum likelihood algorithm.

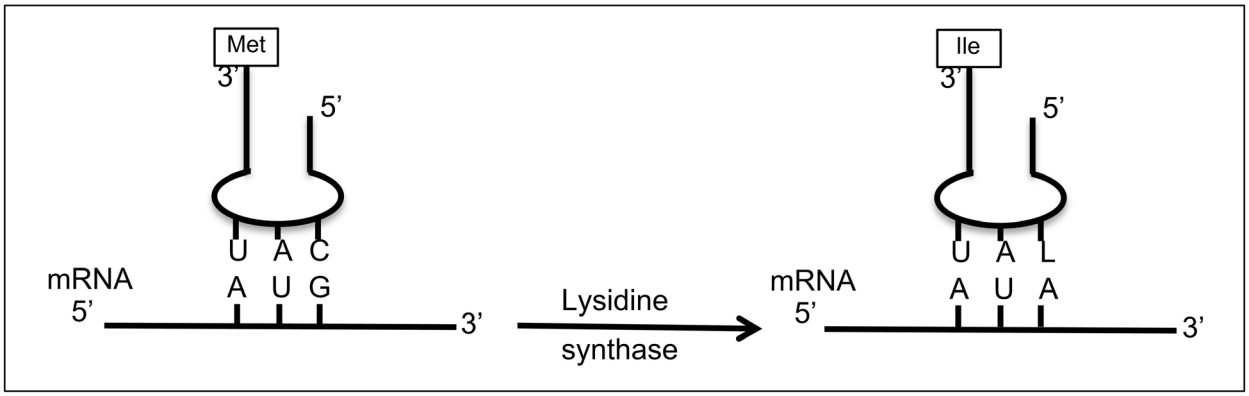

Figure 3. Simplified diagram of tRNA anticodon base pairing with the mRNA codon. The $tRNA^{met}$ anticodon 3' UAC 5' recognizes the codon 5' AUG 3' (left figure). The enzyme tRNA ile lysidine synthase (TilS) attaches lysine to the 5' cytosine of the tRNA, which becomes lysidine. Lysidine base pairs with adenine. As such, the modified tRNA now recognizes the isoleucine codon AUA.

Table S1. Accession numbers for 41 chloroviruses grouped into three clades that have three different hosts.

NC64A viruses	Accession number	Sampling Location[1]	Chlorovirus group
MA-1E	JX997173	Massachussetts, USA	NC64A
CvsA1	JX997165	Sawara, Japan	NC64A
CviKI	JX997162	Kyoto, Japan	NC64A
KS1B	JX997171	Kansas, USA	NC64A
PBCV-1	JF411744.1	North Carolina, USA	NC64A
IL-3A	JX997169	Illinois, USA	NC64A
MA-1D	JX997172	Massachussetts, USA	NC64A
NE-JV-4	JX997179	Nebraska, USA	NC64A
AN69C	JX997153	Canberra, Australia	NC64A
NY-2B	JX997182	New York state, USA	NC64A
IL-5-2s1	JX997170	Illinois, USA	NC64A
NY-2A	DQ491002.1	New York state, USA	NC64A
NYs-1	JX997183	New York state, USA	NC64A
AR158	DQ491003.2	Buenos Aires, Argentina	NC64A
Fr5L	JX997167	France	Pbi
CZ-2	JX997166	Czech Republica	Pbi
MT325	DQ491001.1	Montana, USA	Pbi
Can18-4	JX997157	Canada	Pbi
CVB-1	JX997160	Berlin, Germany	Pbi
FR483	DQ890022.1	France	Pbi
CVG-1	JX997161	Göttingen, Germany	Pbi
CVR-1	JX997164	Rauschenberg, Germany	Pbi
CVA-1	JX997159	Amönau, Germany	Pbi
AP110	JX997154	Unknown	Pbi
CVM-1	JX997163	Marburg, Germany	Pbi
NW665.2	JX997181	Norway	Pbi
OR0704.2.2	JX997184	Oregon, USA	Pbi
NE-JV-1	JX997176	Nebraska, USA	Pbi
OR0704.3	JX997185	Oregon, USA	SAG
Can0610SP	JX997156	British Columbia, Canada	SAG
NE-JV-2	JX997177	Nebraska, USA	SAG
NE-JV-3	JX997178	Nebraska, USA	SAG
ATCV-1	EF101928.1	Stuttgart, Germany	SAG
WI0606	JX997187	Wisconsin, USA	SAG
MO0605SPH	JX997175	Missouri, USA	SAG
GM0701.1	JX997168	Guatemala	SAG
Br0604L	JX997155	SaoPaolo, Brazil	SAG
TN603.4.2	JX997186	Tennessee, USA	SAG
Canal-1	JX997158	Nebraska, USA	SAG
MN0810.1	JX997174	Minnesota, USA	SAG
NTS-1	JX997180	Nebraska, USA	SAG

[1]Sampling locations obtained from Jeanniard et al., 2013.

Table S2 A) NC64A viruses: 5' and 3' genes closest to the tRNA gene cluster. **B**) SAG viruses: 5' and 3' genes closest to the tRNA gene cluster. **C**) Pbi viruses: 5' and 3' genes closest to the tRNA gene cluster.

S2A

NC64A virus	5' upstream (~1068 nt)	5' upstream (~291 nt)	tRNA cluster	3' downstream (~1299 nt)	3' downstream (~1140)	3' downstream[1] (~1191 nt)
MA-1E	390L	**393R**			**396R**	407L
CvsA1	361L	**364R**			**368R**	380L
CviKI	353L	**356R**			**359R**	370L
KS1B	310L	**311R**				**314L**
PBCV-1	A328L	**A329R**		**A330R**		A333L
IL-3A	368L	**369R**		**371R**	375R	386L
MA-1D	**347L**				**355R**	367L
NE-JV-4	384L	**385R**		**388R**		390L
AN69C	377L	**378R**		**380R**	384R	395L
NY-2B	**465L**				473R	484L
IL-5-2s1	**484L**				**492R**	503/506L
NY-2A	B458L	**B460R**			B465R	B480L
NYs-1	**474L**				**483R**	495L
AR158	**C406L**				C413R	C423L

The genes in each column are orthologs to one another; approximate nt length of each ortholog is noted in parentheses. The genes in bold font are the closest 5' and 3' genes for each of the 14 NC64A viruses. There is a common tRNA cluster location among the NC64A viruses.
[1]This column is included because the KS1B gene 314L is the closest 3' gene. Orthologs of 314L are present in the other NC64A, but they are further downstream.

Table S2B. SAG viruses: 5' and 3' genes closest to the tRNA gene cluster.

SAG viruses	5' upstream (~285 nt)	5' upstream (~837 nt)	tRNA cluster	3' downstream (~1041 nt)	3' downstream (~3,774 nt)
OR0704.3	**307R**				**301L**
Can0610SP	**308R**			**309R**	313L
NE-JV-2	338R	**339R**			**341L**
NE-JV-3	**301R**				**303L**
ATCV-1	**Z254R**				**Z257L**
WI0606	**329R**				**332L**
MO0605SPH	**313R**				**316L**
GM0701.1	**305R**			**309R**	312L
Br0604L	**306R**				**308L**
TN603.4.2	**303R**				**307L**
Canal-1	**302R**				**304L**
MN0810.1	**337R**				**340L**
NTS-1	**345R**				**351L**

The genes in each column are orthologs to one another; approximate nt length of each ortholog is noted in parentheses. The genes in bold font are the closest 5' and 3' genes for each of the 13 SAG viruses. There is a common tRNA cluster location among the SAG viruses.

Table S2C Pbi viruses: 5' and 3' genes closest to the tRNA gene cluster.

Pbi viruses	5' upstream (~726 nt)	5' upstream (~270 nt)	5' upstream (870 nt)	5' upstream (267 nt)	tRNA cluster	3' downstream (789 nt)	3' downstream (~297 nt)[1]	3' downstream (~420 nt)	3' downstream (~552 nt)
Fr5L	393R	**397R**						**401L**	
CZ-2	352R	**355R**						**358L**	
MT325	**M3422R**						**M344L**		
Can18-4	**414R**				418R		**419L**	422L	
CVB-1	**406R**				408R		**411L**	413L	
FR483	**N351R**						**N345L**	n356L	
CVG-1	**385R**						**390L**	392L	
CVR-1	**400R**						**404L**	406L	
CVA-1	**392R**						**396L**	398L	
AP110A	**403R**					**407R**	411L	413L	
CVM-1	**421R**						**425L**	428L	
NW665.2	**375R**				378R		**381L**	383L	
OR0704.2.2	349R	352R	**353R**					**356L**	
NE-JV-1				**683R**	688R				**690L**

The genes in each column are orthologs to one another; approximate nt length of each ortholog is noted in parentheses. The genes in bold font are the closest 5' and 3' genes for each of the 13 SAG viruses. There is a common tRNA cluster location among the Pbi viruses, with the exception of NE-JV-1, which continues to be unique. Several viruses have protein-encoding genes within the tRNA cluster, which are not orthologous to one another. The HGT events that led to these gene insertions (blue) resulted in the loss of one or more tRNA genes in each case.
[1]MT325 and FR483 have longer genes (525 nt) but the core of 297 is present with high identity.

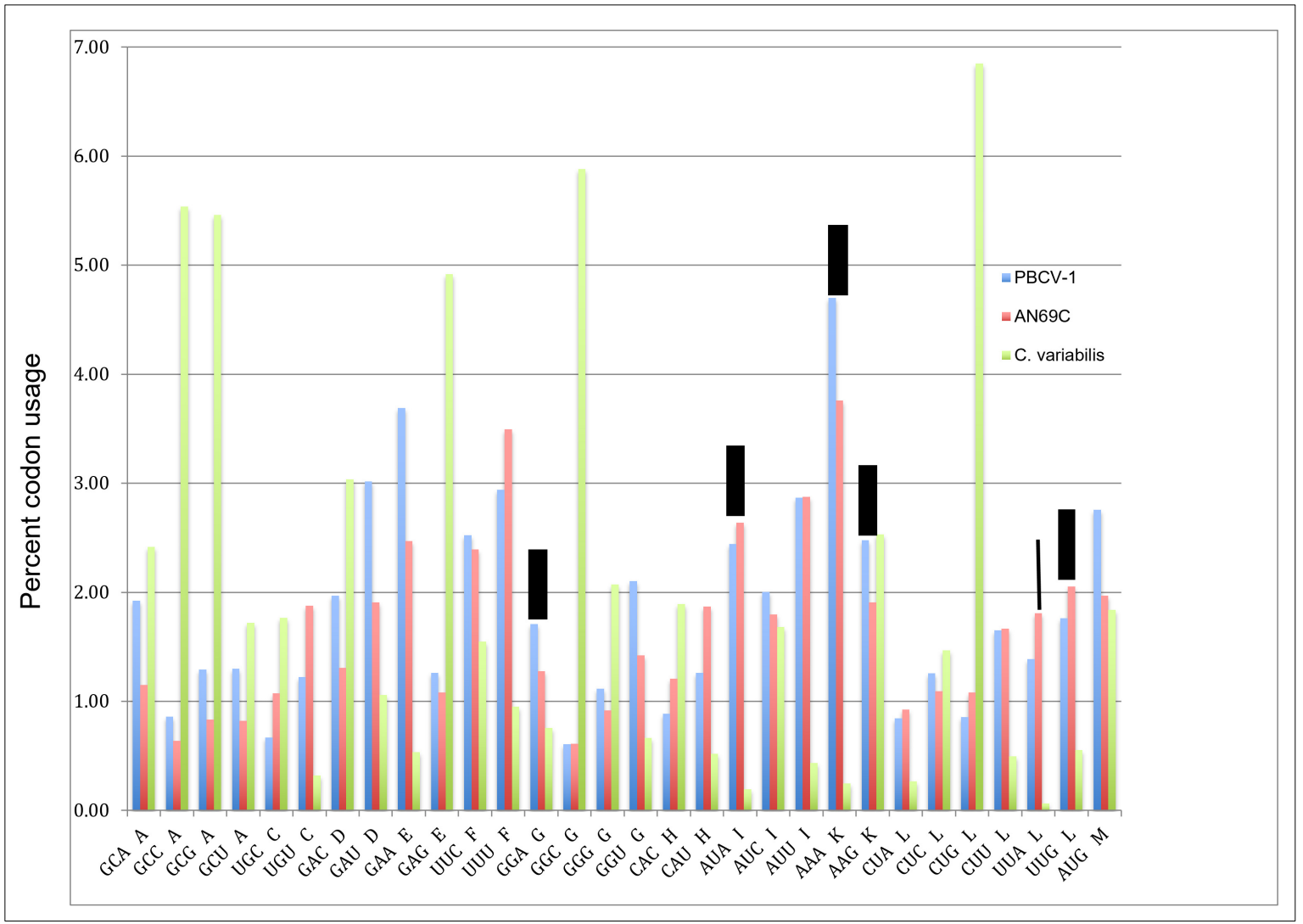

Figure S1A. **A**) First half of genetic code book, comparing frequency of codon usage in the NC64A viruses PBCV-1 and AN69C to their host, *C. variabilis* NC64A. **B**) Second half of genetic code book, comparing frequency of codon usage in the NC64A viruses PBCV-1 and AN69C to their host, *C. variabilis* NC64A. Wide bars denote the codons recognized by tRNAs encoded by PBCV-1 and AN69C. Narrow dashed bar denotes codon recognized by tRNA encoded by PBCV-1 but not AN69C.

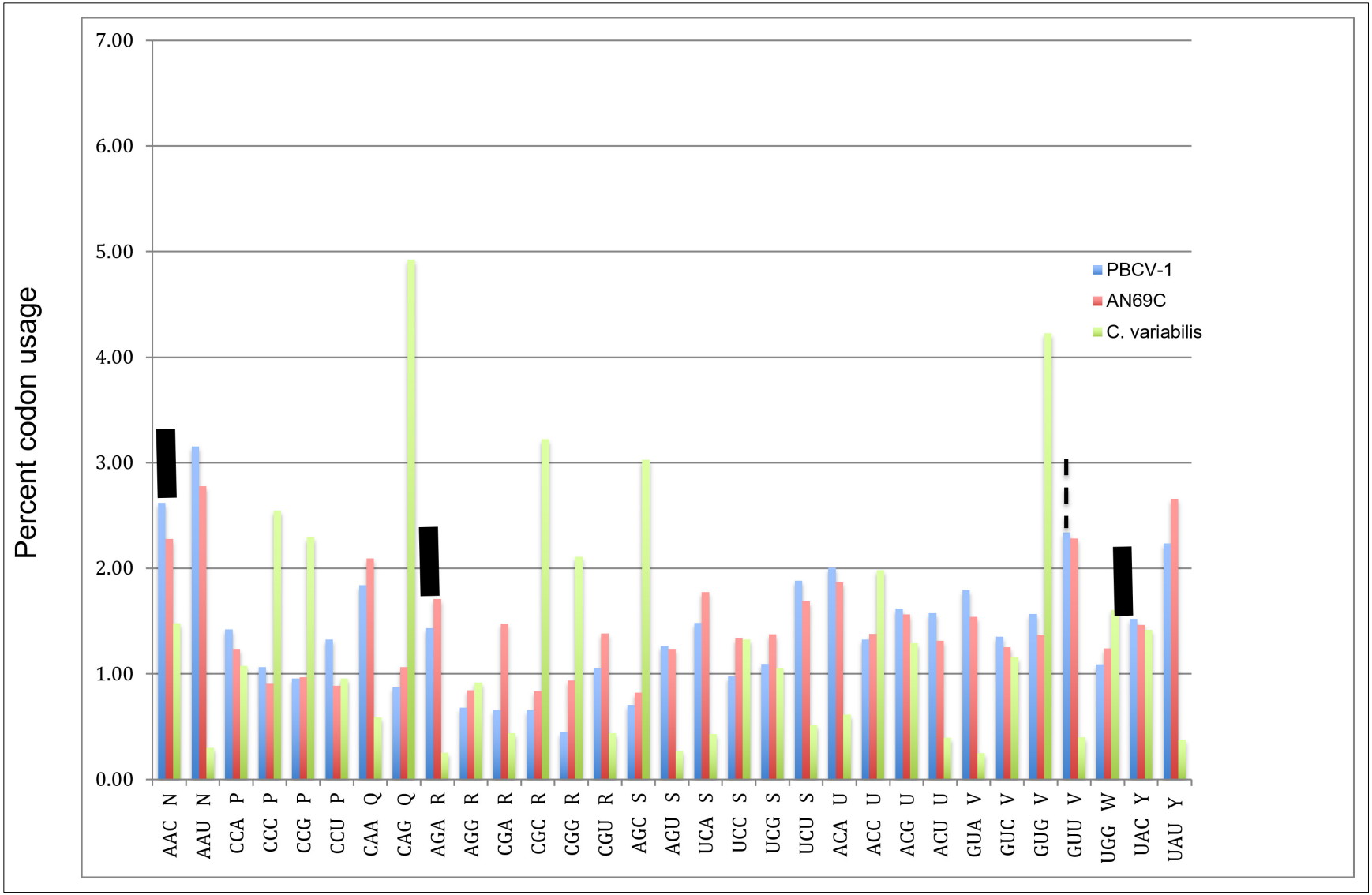

Figure. S1B. Second half of genetic code book, comparing frequency of codon usage in the NC64A viruses PBCV-1 and AN69C to their host, *C. variabilis* NC64A. Wide bars denote the codons recognized by tRNAs encoded by PBCV-1 and AN69C. Narrow dashed bar denotes codon recognized by tRNA encoded by PBCV-1 but not AN69C.