ELSEVIER

COMPUTATIONAL
ANDSTRUCTURAL
BIOTECHNOLOGY
JOURNAL

journal homepage: www.elsevier.com/locate/csbj

Review

Evolution and host adaptability of plant RNA viruses: Research insights on compositional biases

Zhen He[a,b,*], Lang Qin[a], Xiaowei Xu[a], Shiwen Ding[a]

[a] *School of Horticulture and Plant Protection, Yangzhou University, Wenhui East Road No. 48, Yangzhou 225009, Jiangsu Province, PR China*
[b] *Joint International Research Laboratory of Agriculture and Agri-Product Safety of Ministry of Education of China, Yangzhou University, Wenhui East Road No. 48, Yangzhou 225009, Jiangsu Province, PR China*

ARTICLE INFO

Article history:
Received 12 March 2022
Received in revised form 10 May 2022
Accepted 12 May 2022
Available online 17 May 2022

Keywords:
Plant RNA viruses
Compositional biases
Codon usage
Host adaptation

ABSTRACT

During recent decades, many new emerging or re-emerging RNA viruses have been found in plants through the development of deep-sequencing technology and big data analysis. These findings largely changed our understanding of the origin, evolution and host range of plant RNA viruses. There is evidence that their genetic composition originates from viruses, and host populations play a key role in the evolution and host adaptability of plant RNA viruses. In this mini-review, we describe the state of our understanding of the evolution of plant RNA viruses in view of compositional biases and explore how they adapt to the host. It appears that adenine rich (A-rich) coding sequences, low CpG and UpA dinucleotide frequencies and lower codon usage patterns were found in the vast majority of plant RNA viruses. The codon usage pattern of plant RNA viruses was influenced by both natural selection and mutation pressure, and natural selection mostly from hosts was the dominant factor. The codon adaptation analyses support that plant RNA viruses probably evolved a dynamic balance between codon adaptation and deoptimization to maintain efficient replication cycles in multiple hosts with various codon usage patterns. In the future, additional combinations of computational and experimental analyses of the nucleotide composition and codon usage of plant RNA viruses should be addressed.

Contents

[*] Corresponding author.
E-mail address: hezhen@yzu.edu.cn (Z. He).

https://doi.org/10.1016/j.csbj.2022.05.021
2001-0370/

1. Introduction

In the last two decades, viromics (or viral metagenomics) have led to the discovery of many new RNA viruses in animals and plants through the development of deep-sequencing technology and big data analysis [1,2]. These findings largely changed our understanding of the origin, evolution and host range of plant RNA viruses. In general, several common forces drive the evolution of plant RNA viruses, including high mutational rates, strong purifying selection, genetic drift, and evolutionary arms races with infected hosts [1,3-12]. Consistent with animal RNA viruses, the evolutionary history of plant RNA viruses also comprises three possible hypotheses: horizontal gene transfers from the host genome, coevolution or codivergence with hosts, and parallel evolution with related genetic elements [1,13]. We can see that plant hosts had a significant influence on the evolutionary history and trends of RNA viruses. In fact, the recent frequent emergence or re-emergence of new viral diseases is driven by adaptive evolution corresponding to new ecological conditions, especially hosts [9,12,14-24].

In agriculture, several well-studied emerging plant RNA virus diseases have attracted much attention due to economic damage to crop hosts, such as from rice yellow mottle virus [23,25,26] and barley yellow dwarf virus[27]. During the process of emergence, the well-established original host species could be considered reservoir hosts. Elena et al. (2011, 2014) [11,12] described three temporal phases of emergence, such as host jumps to new species or the same species but in a new ecological condition, adaptation to the new host or environment, and epidemiology in the new host population, usually by adaptation to a new transmission mode or new vector species (Fig. 1). In summary, four groups of driving forces sharpen the emergence of viruses, including the genetic composition of the virus population, the genetic composition of the host population, the genetic composition of vectors for vectored viruses, and the ecology of viruses and/or host plants (Fig. 1) [12,22]. Thus, the genetic composition originating from virus, host and vector populations plays a key role in the evolution and host adaptability of plant RNA viruses.

In general, the four nucleotides (A, adenine, C cytosine, G guanine and U uracil) are not random in the genomes of viruses and the hosts they infect [28-33] (Fig. 2). This is often facilitated by synonymous codons (codons encoding the same amino acid), which allow for 61 triplet codons that encode 20 amino acids; for example, Asn, Asp, Cys, Glu, Gln, His, Phe, Tyr, and Lys are encoded by two codons; Ile is encoded by three codons; Ala, Gly, Thr, Pro, and Val are encoded by four codons; and Arg, Leu, Ser are encoded by six codons. These phenomena are termed codon degeneracy. Interestingly, the usage of codon degeneracy is also not randomly selected [34-40] (Fig. 2). In nature, the unequal preference for specific codons over other synonymous codons in various organisms creates a bias in codon usage [41-44]. Similar to codon usage, codon order is also not randomly selected because a ribosome decodes two codons simultaneously in the process of translation [45] (Fig. 2). In 1985, codon pair bias was first described in *Escherichia coli* [46] and then in bacteria, archaea, and eukaryotes [47]. Dinucleotide biases were considered the proposed explanation of nucleotide and codon preferences [48-51].

In the past three years, SARS-CoV-2 induced by COVID-19 has rapidly developed into a devastating global pandemic, causing nearly 5 million fatalities and more than 238 million cases, and now the daily number of people infected is also increasing rapidly [52,53]. Therefore, the evolution and host adaptation of animal RNA viruses have attracted great attention. Kustin and Stern (2020) described adenine rich (A-rich) coding sequences in the vast majority of animal RNA viruses and proposed possible reasons such as codon usage bias, weakened RNA secondary structures, and selection for a particular amino acid composition, concluding that similar biases in coding sequence composition across animal RNA viruses are possibly due to host immune pressures [29]. Gaunt and Digard (2021) reviewed the compositional biases mainly in RNA viruses in terms of the causes, consequences and applications [28]. For plant RNA viruses, several recent studies have reported nucleotide composition, codon usage bias, dinucleotide bias and host or vector adaptation [32,39,40,54-65]. In this review, we summarize evolution and host adaptation in plant RNA viral genomes by considering the compositional biases, and widely used software packages for compositional bias analyses. We also discuss future trends under the rapid development of big data and metagenomic analysis.

2. Nucleotide composition of plant RNA viruses

2.1. Nucleotide bias of plant RNA viruses

Ideally, the four bases A, T, C, and G occur at a frequency of 25% equally in an organism's genome. However, in nature,

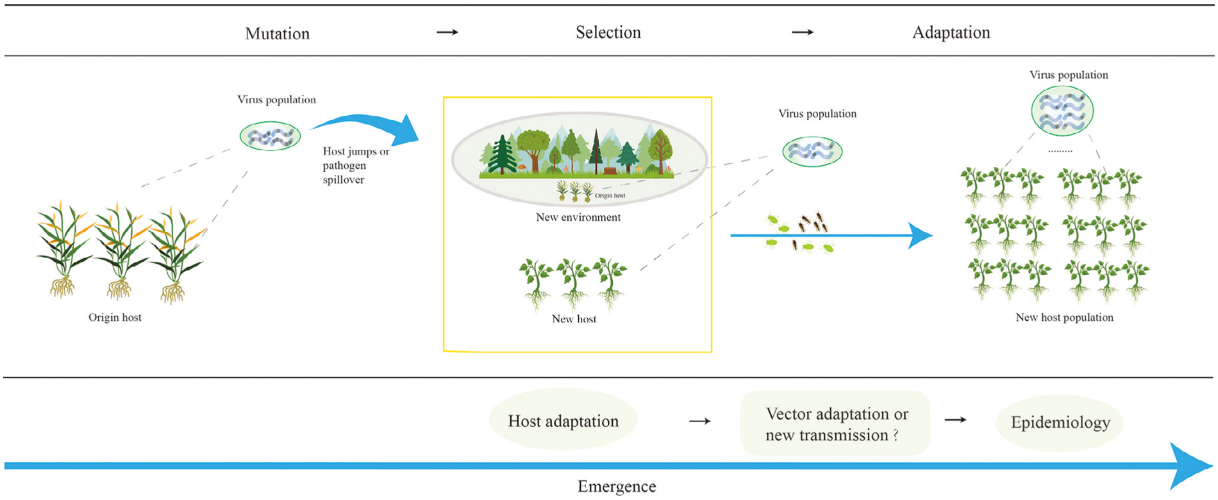

Fig. 1. Schematic overview on the emergence and host adaptation of plant RNA viruses.

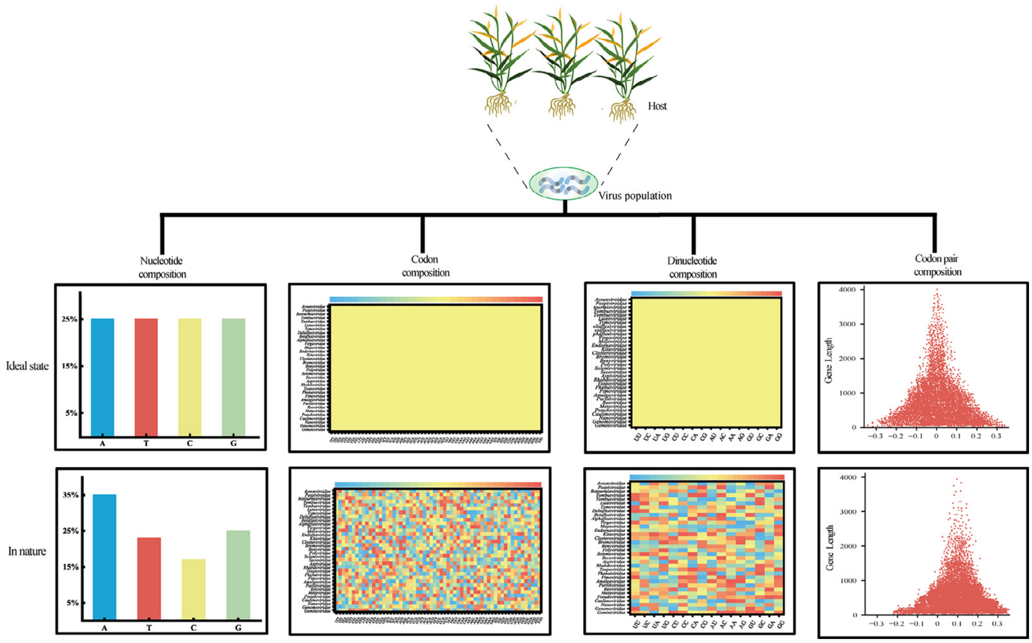

Fig. 2. Schematic overview on the compositional biases of plant RNA viruses in ideal and nature conditions.

nucleotide bias is frequently seen across almost all genomes. Adenine rich (A-rich) coding sequences have been found in the vast majority of animal RNA viruses, accompanied by a strong diminution of C [29]. The highest A (49%) was found in VPg sequences of Rhinovirus [29]. Consistently, in the family *Potyviridae*, A-rich composition has been found in all genera, with the highest value (35%) in *Arepavirus* [66] (Fig. 3A). For single plant virus species, A-rich composition has also been found in all or partial coding region sequences of potato virus Y (PVY) [40], citrus tristeza virus (CTV) [64], sugarcane mosaic virus (SCMV) [62], rice black streak dwarf virus (RBSDV) [61], narcissus degeneration virus (NDV), narcissus late season yellows virus (NLSYV) and narcissus yellow stripe virus (NYSV) [128] (Fig. 3). Uracil rich (U-rich) coding sequences were found in the two open reading frames (ORFs) of broad bean wilt virus 2 (BBWV-2) [60], the cysteine-rich nucleic acid binding protein (NABP) gene of potato virus M (PVM) [63], P8 protein coding sequences of RBSDV [61], coat protein (CP) of CTV [54], cowpea mild mottle virus (CpMMV) [127] and banana bract mosaic virus (BBrMV) [129] (Fig. 3C). Similarly, U3-rich (uracil at the third codon position) has been found in most coding sequences of these plant RNA viruses (Fig. 3D). More U3S-rich sequences (the third position's nucleotide composition of synonymous codons) were also found in these plant RNA virus coding sequences (Fig. 3E). Overall, AU-rich coding sequences were found in these plant RNA viruses, and the highest AU (65.50%) was found in the P8 coding sequences of RBSDV [61] (Fig. 3F). For viruses, the AU- or GC-rich composition tends to correlate with their RSCU patterns [60–63,67,68,132]. For example, an AU-rich composition of SCMV genomes contains codons that frequently end with A and U [62]. Codon usage bias, weakened RNA secondary structures, and selection for a particular amino acid composition possibly explain adenine rich (A-rich) coding sequences in these plant RNA viruses [29]. However, extensive G, G3, G3s and GC were observed in the CP gene of PVM, reflecting the influence of mutation pressure [63] (Fig. 3). Codon W, MEGA, BioEditor, DnaSP and CAIcal SERVER can calculate the base composition of plant RNA viruses (Table 1).

2.2. Codon bias of plant RNA viruses

The codon usage bias of viruses is not randomly selected [28,29,34,69-72], including that of plant RNA viruses [39,40,60,63-65]. Most reported animal RNA viruses show a low codon usage bias [37,64,67,68,73], which allows for efficient replication in the host cell by lowering the level of competition with the host genes. For plant RNA viruses, Adams and Antoniw (2004) found low codon usage bias in CP gene sequences of several genera, such as *Potyvirus*, *Cucumovirus*, *Sobemovirus*, and *Polerovirus* [56]. More recently, a lower codon usage pattern was also found in complete or partial gene coding sequences of several plant RNA viruses, such as BBWV2, CTV, PRSV, PVM, PVX, RSV and SCMV [39,60-64,74]. These lower codon usage patterns indicate a low degree of preference in plant RNA viruses.

Similar to eukaryotic life, the codon usage patterns of viruses are shaped by mutation, natural selection, drift, compositional constraints, gene length and function, secondary protein structure, selective transcription, replication and hydrophobicity [41,43-45,75-79]. Several codon usage pattern analyses, including ENC-plot, neutrality plot, PR2, and regression analyses between ENC, GC, GC3S and ARO, GRAVY values indicated that plant RNA viruses were influenced by both natural selection and mutation pressure, and natural selection was the dominant factor shaping the codon usage pattern of plant RNA viruses. Chen et al. (2020) found that virus codon usage bias (CUB) tended to be more similar to that of symptomatic hosts than that of asymptomatic natural hosts, indicating a general dissimilation of CUB in virus–host coevolution due to translational selection (Fig. 4) [80]. Codon W, DnaSP, MEGA, Chips, cusp, EncPrime, CodonO, SMS and CAIcal SERVER can calculate the codon usage of plant RNA viruses (Table 1).

2.3. Codon pair bias of plant RNA viruses

Consistent with the codon usage bias, some codon pairs are used more frequently than others in prokaryotic and eukaryotic genomes, and the phenomenon was described as codon pair bias (CPB) [45,81,82]. CPB has been summarized for bacteria, archaea,

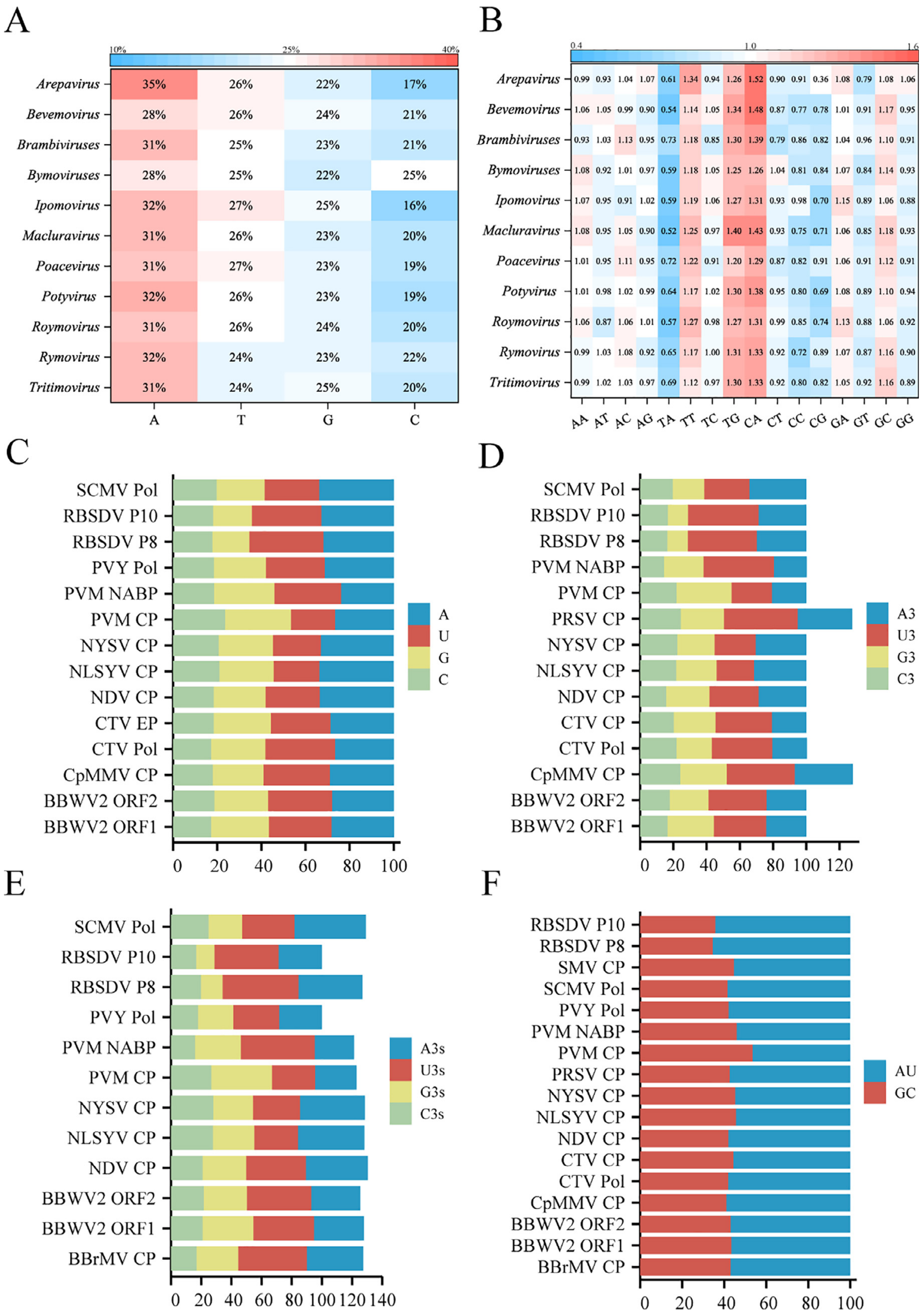

Fig. 3. Nucleotide composition of recently reported plant RNA viruses. Source from Biswas et al. (2019), Chakraborty et al. (2015), Cheng et al. (2012), Gómez et al. (2020), He et al. (2019, 2020, 2021, 2022), Prádena et al (2020), Patil et al. (2017), Yang et al. (2022), Huang et al. (2015).

Table 1
Software for nucleotide composition and codon adaptation analyses.

Name	Description and advantages	Uses	Availability	URL	Reference
Software for nucleotide composition and codon analyses					
BioEditor	BioEditor is an application that enables scientists and educators to prepare and present structure annotations containing formatted text, graphics, sequence data, and interactive molecular views.	BioEditor can be used to analyse codon and base composition.	Local installation	https://bioeditor.sdsc.edu	[110]
chips	Nc provides an intuitive and meaningful measure of the degree of codon bias in genes. Low values indicate strong codon bias and high values indicate low bias (probably noncoding regions).	Chips computes Frank Wright's Nc statistic for nucleotide sequences.	Local installation	https://emboss.sourceforge.net/apps/release/6.6/emboss/apps/chips.html	[111]
CodonW	Codon W is a software package for codon usage analysis. It is designed to simplify multivariate analysis (MVA) of codon usage. The MVA method employed in codon W is COA, the most popular MVA method for codon usage analysis. Codon W can generate COAs for codon usage, relative synonymous codon usage, or amino acid usage. Other analyses of codon usage include studies of optimal codons, codon and dinucleotide bias and/or base composition.	Codon W applies correspondence analysis (COA), the most popular MVA method for codon usage analysis. Codon W can generate COA for codon usage, relative synonymous codon usage or amino acid usage analyses.	Local installation	https://sourceforge.net/projects/codonw/	[112]
cusp	Cusp computes a codon usage table for one or more nucleotide coding sequences and writes the table to a file. The codon usage table gives each codon: i. sequence of codons. ii. The encoded amino acid. iii. The proportion of codon usage in its redundant set, i.e., the set of codons encoding the amino acid of that codon. iv. Given the input sequence, the expected number of codons per 1000 bases. v. The number of codons observed in the sequence.	Creates a codon usage table from a nucleotide sequence.	Local installation	https://emboss.sourceforge.net/apps/release/6.6/emboss/apps/cusp.html	[111]
DnaSP	DnaSP is a software package for the analysis of DNA polymorphism data.	The present version allows for analysis of the evolutionary pattern of preferred and unpreferred codons.	Local installation	https://www.ub.es/dnasp	[113]
EncPrime	A program to calculate the summary statistic Nc' of codon usage bias.	Calculates the ENC metric.	Local installation	https://github.com/jnovembre/ENCprime	
SMS (Sequence Manipulation Suite)	The program can compares the frequency of codons encoding the same amino acid (synonymous codons)	SMS can be used to assess whether sequences show a preference for particular synonymous codons.	Web	https://www.bioinformatics.org/sms2/codon_usage.html	[114]
MEGA 11	Molecular Evolutionary Genetics Analysis (MEGA) software has matured to contain a large collection of methods and tools of computational molecular evolution.	MEGA now contains methods for analyses of codons, RSCU and base composition.	Local installation	https://www.megasoftware.net/citations	[115]
Software for codon pair analysis					
ANACONDA	The Anaconda software package provides a set of statistical, bioinformatics and data visualization tools for gene primary structure analysis.	It can be used for analysis of genomic codon preference and codon pair preference	Local installation	https://bioinformatics.ua.pt/software/anaconda/	[116]
CoCoPUTs	CoCoPUT is a table of codon and codon pair usage derived from all available GenBank and RefSeq data. When searching for species, the search takes precedence over RefSeq, so that if the RefSeq assembly is available, it will automatically extract data from that source. If searching for a species without RefSeq assemblies, use the taxonomic ID of the organism for best results.	The codon usage table is a measure of codon usage bias, such as the relative frequency with which different codons are used in genes of a given species. Likewise, the codon pair usage table shows counts for each codon pair in the CDS of a given species and is a measure of codon pair usage bias.	Web	https://hive.biochemistry.gwu.edu/review/codon2	[117]
CPS (codon pair score)	Measures codon pair bias, defined analogously to the RSCU.	It can be used to determine the level of similarity in codon pair preferences between viruses and their host.	R package	https://rdrr.io/github/alex-sbu/CPBias/man/CPScalc.html	[48]

Table 1 (*continued*)

Name	Description and advantages	Uses	Availability	URL	Reference
CPO (codon pair optimization)	A software tool to provide codon pair optimization for synthetic gene design.	CPO provides a simple and efficient means for customizing codon optimization based on the codon pair bias of Pichia pastoris.	R package	https://microbialcellfactories.biomedcentral.com/articles/10.1186/s12934-021-01696-y#Sec15	[118]
Software for codon adaptation analysis					
CAIcal	It includes a complete set of CAI related utilities. The server provides useful important functions such as computational and graphical representation of CAI, representation along single sequences or protein multiple sequence alignments translated into DNA. The CAIcal tool also includes automatic calculation of the CAI and its expected value.	The CAIcal server provides a complete set of tools to assess codon usage adaptation and aid in genome annotation.	Web	https://genomes.urv.es/CAIcal	[107]
CBI (codon bias index)	Optimal codon usage is measured using the ratio between the number of optimal codons in the gene and the total number of codons in the gene. It uses the expected usage as a scaling factor.	It can calculate the presence of components with high CUB in a particular gene.	Local installation	https://codonw.sourceforge.net/index.html	[119]
COOL	COOL was designed as an adaptable web-based interface that provides a wide range of functions. Users can completely customize the synthetic gene design process through a step-by-step job submission process, which allows for them to specify their optimal parameter settings.	COOL supports a simple and flexible interface for customizing various codon optimization parameters such as the codon adaptation index, single codon usage, and codon pairing.	Web	https://bioinfo.bti.a-star.edu.sg/COOL/	[120]
coRdon	Codon usage bias can be used to predict the relative expression levels of genes by comparing the CU bias of a gene to the CU bias of a set of genes known to be highly expressed. This method can be effectively used to predict highly expressed genes in a single genome, and it is particularly useful at a higher level of the whole metagenome. By analysing the CU deviation of the macrogenome, we can identify the genes with high predictive expression in the whole microbial community, and determine the enrichment functions in the community, that is, their “functional fingerprint”.	It can calculation of different CU bias statistics and CU-based gene expression predictions, gene set enrichment analysis of annotated sequences, and several methods for displaying CU and enrichment analysis results.	R package	https://www.bioconductor.org/packages/devel/bioc/vignettes/coRdon/inst/doc/coRdon.html	
COUSIN	Calculates codon usage for user-supplied Sequences.	COUSIN allows for easy and complete analysis of cuprefs, including seven other indices, and provides functions such as statistical analysis, clustering and cuprefs optimization of gene expression.	Web or install	https://cousin.ird.fr/index.php	[121]
HEG-DB	Database of the CAI index of HEGs for 200 genomes	Calculates the CAI.	Web	https://genomes.urv.cat/HEG-DB/	[122]
Jcat (Java Codon Adaptation Tool)	Further choices for Jcat codon adaptation include the avoidance of unwanted cleavage sites for restriction enzymes and Rho-independent transcription terminators. Compared with existing tools, Jcat does not need to manually define high-expression genes, so it is a very fast and simple method.	A novel method for the adaptation of target gene codon usage to most sequenced prokaryotes and selected eukaryotic gene expression hosts to improve heterologous protein production.	Web	http://www.jcat.de/Start.jsp	[123]
OPTIMIZER	OPTIMIZER allows for three optimization methods and uses several valuable new reference sets. It can be used to optimize the expression levels of genes, assess the fitness of foreign genes inserted into the genome, or design new genes from protein sequences.	Optimizes the codon usage of a DNA sequence to increase its expression level.	Web	https://genomes.urv.es/OPTIMIZER/	[124]

(*continued on next page*)

Table 1 (*continued*)

Name	Description and advantages	Uses	Availability	URL	Reference
stAI (species-specific tRNA adaptation index)	The tRNA adaptation index (tAI) is a widely used measure of the efficiency with which the intracellular tRNA pool recognizes coding sequences. The index includes weights representing the wobble interactions between codons and tRNA molecules. The software presents a new method to adjust tAI weights to any target model organism without the need for gene expression measurements. The method is based on optimizing the correlation between tAI and codon usage bias measures.	The calculator includes optimized tAI weights for 100 species from three life domains, as well as a stand-alone software package to optimize weights for new organisms.	Web	https://www.cs.tau.ac.il/~tamirtul/stAIcalc/stAIcalc.html	[125]
Synthetic Gene Designer	Synthetic Gene Designer includes three main stages of genetic design. Given it a gene of interest and the target genome in which it is expressed.	Synthetic Gene Designer offers enhanced functionality compared to existing software options; for example, it enables users to use nonstandard genetic codes, user-defined codon usage patterns, and an expanded set of codon optimization methods.	Web	https://www.evolvingcode.net/codon/sgd/index.php	[126]

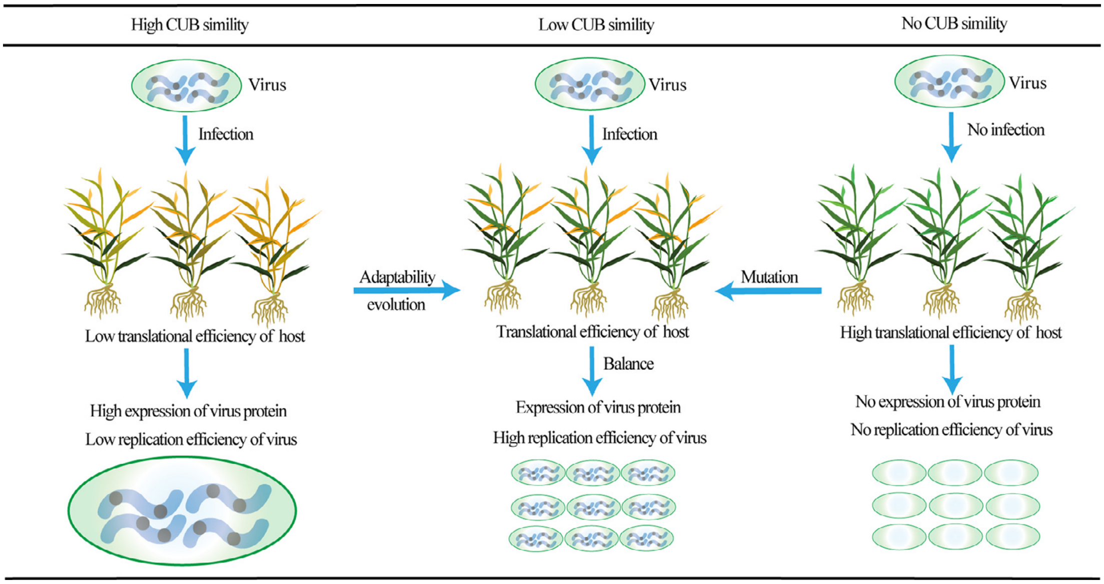
High CUB simility
Low CUB simility
No CUB simility
Virus
Infection
Virus
Infection
Virus
No infection
Adaptability evolution
Mutation
Low translational efficiency of host
Translational efficiency of host
High translational efficiency of host
Balance
High expression of virus protein
Low replication efficiency of virus
Expression of virus protein
High replication efficiency of virus
No expression of virus protein
No replication efficiency of virus

Fig. 4. Schematic overview on the regulatory role of plant RNA viruses' CUB and its evolutionary implication.

and eukaryotes [47,81]. For viruses, CPB was first described in poliovirus [45], followed by classical swine fever virus [83], human immunodeficiency virus type 1 [84,85], porcine reproductive and respiratory syndrome virus [86], dengue virus type 2 [87], influenza A/Puerto Rico/8/34 (H1N1) virus [88], Marek's disease herpesvirus [89,90], Zika virus [91], influenza A virus (IAV) [92], influenza B virus [92], and influenza C virus [92]. However, there have been no reports on CPB in plant viruses until now. In general, CpG/UpA dinucleotide and translational selection shape codon pair usage in protein coding sequences [47,48,82,87]. In prokaryotic and eukaryotic genomes, the most frequently preferred codon pairs are nnGCnn, nnCAnn and nnUnCn [47]. The most frequently avoided codon pairs are nnGGnn, nnUAnn, nnCGCn, nnGnnC, GUCCnn, CUCCnn, UUCGnn and nnCnnA [47]. ANACONDA, CoCoPUT, and CPS software can calculate the codon pair bias of plant RNA viruses (Table 1).

2.4. Dinucleotide bias of plant RNA viruses

Normally, the dinucleotide (two consecutive nucleotides) frequencies in different or even the same organisms usually do not match that of the nucleotide composition [93-98]. In other words, dinucleotides are also not randomly present in organisms. Recent studies have revealed that low CpG and UpA dinucleotide frequencies in animal RNA viruses could avoid specific host defences [99-101]. Similarly, UpA and CpG were largely underrepresented in the genomes of rice stripe virus [65], potato virus X [102], SCMV [59] and other potyvirids [66] (Fig. 3B). Prádena et al. (2020) showed that an increase in UpA frequency strongly diminishes virus accumulation and fitness [130] using plum pox virus (PPV) as a model. They also demonstrated that host RNA polymerase II plays a key role in the anticorrelation between UpA frequency and RNA accumulation in the genome of PPV. Codon W can calculate the dinucleotide bias of plant RNA viruses (Table 1).

3. Codon adaptation to the host

3.1. Relative synonymous codon usage (RSCU) analysis

The adaptation, evolution, fitness and survival of viruses are affected by codon usage bias [37,54,64,67,68]. The RSCU value of a codon for viruses and their hosts is the ratio between the observed usage frequency and the expected usage frequency [103]. Generally, hosts have a significant effect on the selection of optimal codons in viruses. Both coincident and antagonistic codon usage between viruses and their hosts have been reported [54,61,64,68]. It is accepted that coincident codon usage allows for the corresponding amino acids to be translated efficiently, whereas antagonistic codon usage suggests viral proteins are folded properly, regardless of whether the translation efficiency of the corresponding amino acids might be reduced [68,104]. He et al. (2020) compared the RSCU patterns of RBSDV with those of its hosts and vector and showed that RBSDV had evolved complete antagonistic codon usage patterns relative to its host and a mixture of coincident and antagonistic codon usage patterns relative to its vector [61]. Similar results were also found in CTV and its host citrus [54,64]. These results indicate that the selection pressure exerted by hosts has greatly influenced the codon usage patterns of plant RNA viruses. Codon W, MEGA, and CAIcal SERVER can calculate the RSCU of plant RNA viruses (Table 1).

3.2. Codon adaptation index (CAI) and relative codon deoptimization index (RCDI) analyses

The codon adaptation index (CAI) is used to measure the synonymous codon usage bias for a DNA or RNA sequence, including viral DNA or RNA sequences. Several reports show that the CAI is frequently used to assess the adaptation of viral genes to their hosts [38,60,63,64,67,105]. Generally, if the CAI value is high, then the codon usage bias is extremely high [106,107]. For example, CAI analysis showed that CTV might have evolved millions of years ago in *Citrus reticulata* and then vertically or horizontally transmitted to later citrus species [64], SCMV genes were strongly adapted to maize compared to sugarcane and canna [62], and RBSDV was strongly adapted to rice, followed by maize, wheat and its vector *Laodelphax striatellus* [61]. Similar to CAI analysis, RCDI was performed to calculate codon deoptimization by comparing the codon usage similarity of a gene and a reference genome sequence [108], and a low RCDI value indicates strong adaptation to a host [108]. Based on the CAI and RCDI analyses, it is proposed that plant RNA viruses have probably evolved a dynamic balance between codon adaptation and codon deoptimization to maintain efficient replication cycles in multiple hosts with various codon usage patterns. Codon W, DnaSP, COOL, CUSIN, HEG-DB, Jact, CAIcal and RCDI SERVER can calculate the CAI and RCDI of plant RNA viruses (Table 1).

3.3. Similarity index (SiD) analysis

SiD analysis can reflect the influence of the codon usage bias of hosts on viral genes. The function SiD, ranging from 0 to 1.0, indicates the potential effect of the entire codon usage of the host on the different clades of the viral genes. Normally, a higher SiD value shows that the host plays a key role in the usage of virus codons. For example, during SCMV evolution, maize had a greater impact on the virus than canna or sugarcane because the highest SiD values were observed in maize based on the complete polyprotein and eleven protein coding sequences of SCMV [62]. Several recent studies also report similar SiD analyses on plant RNA viruses and their hosts [39,61-65].

3.4. CpG and UpA dinucleotide bias

Recently, CpG and UpA dinucleotide motifs have been found to be markedly underrepresented in RNA viruses, including plant RNA viruses [39,40,59,65,66]. The avoidance of CpG dinucleotide was observed in several potyviruses and other plant RNA viruses [39,40,59,65,66], possibly due to the outcome of selection on nucleotide composition. Moreover, using PVY as a model, Ibrahim et al. (2019) indicated that increased CpG dinucleotide frequencies in the PVY genome showed a reduction in systemic spread and pathogenicity and attenuated replication kinetics in tobacco plants [109]. Similarly, UpA is also underrepresented in plant RNA viruses. In the *Potyviridae* family, one of the most important plant RNA virus groups, the UpA odds ratio was observed with a mean frequency of 0.632 (±0.066) [66]. An increase in UpA frequency in the genome of plum pox virus (PPV) strongly diminishes virus accumulation and viral fitness. Furthermore, Prádena et al (2020) showed that the anticorrelation between UpA frequency and RNA accumulation applies to mRNA-like fragments produced by the host RNA polymerase II, and indicated that the host controls diverse RNAs in a dinucleotide-based system in plant cells, including plant RNA viruses [66].

4. Summary and discussion

Presently, numerous studies have shown the diverse nucleotide composition, codon usage and adaptation in animal- and human-infecting viruses. We are now in a position to better explore the evolution of plant RNA viral genomes by considering the compositional biases and how they adapt to the host. It appears that adenine rich (A-rich) coding sequences were found in the vast majority of plant RNA viruses. A lower codon usage pattern was also found in the gene coding sequences of plant RNA viruses, indicating a low degree of preference in plant RNA viruses. The codon usage pattern of plant RNA viruses was influenced by both natural selection and mutation pressure, and natural selection was the dominant factor. Low CpG and UpA dinucleotide frequencies were also found in plant RNA viruses, possibly to avoid specific host defences. The codon adaptation analyses support that plant RNA viruses have probably evolved a dynamic balance between codon adaptation and deoptimization to maintain efficient replication cycles in multiple hosts with various codon usage patterns.

Generally, the nucleotide composition of plant RNA viruses is determined by mutation and drift, resulting in a diversity of codons, dinucleotides, and codon pairs [31]. Meanwhile, tRNA selection preference of host affects the diversity of codons, dinucleotides and codon pairs of plant RNA viruses [80]. Tian et al (2018) showed that viruses can invade a narrow spectrum only (NSTVs) had a higher degree of matching to their hosts' tRNA pools than others can invade a broad spectrum of hosts (BSTVs) [131]. Andmore, Chen et al. (2020) found that virus CUB tended to be more similar to that of symptomatic hosts than that of asymptomatic natural hosts. Thus, the hypothesis we considered that for viruses with narrow host range or high pathogenicity, host tRNA selection bias has a great influence on the virus, making the virus codon bias highly similar to the host, while for viruses with wide host range or weak pathogenicity, host tRNA selection bias has a balance to virus CUB, which makes the virus CUB similar to the host but different to some extent (Fig. 4). However, Cardinale et al. (2013) showed that the codon bias of plant RNA viruses is not only affected by mutation and drift of its own genome and the selection of host tRNA, but also influenced by the genomic architecture and secondary structure of the virus. In future, more factors should be considered in evaluation the CUB of plant RNA viruses.

5. Outstanding questions

With the increase in studies on evolution and host adaptation of plant RNA viruses, our understanding of the evolutionary changes of plant RNA viruses has been greatly improved. However, many outstanding questions remain: (i) Synonymous mutations do not change the amino acid encoded by the sequence, so synonymous mutations are generally considered to be neutral mutations. While, several studies have found that synonymous mutations can promote the adaptive evolution of animal viruses (for example influenza A virus, vesicular stomatitisvirus, and Qβ bacteriophage) [131,133,134] and one plant virus, tobacco etch virus [136]. However, it remains unclear how synonymous mutations affect the adaptive evolution of viruses, especially plant RNA viruses. (ii) Genetic drift is a key factor on the evolution of viruses, while the effect of drift on the CUB of plant RNA viruses is unclear. (iii) For single strand plant RNA viruses, the genes or coding protein regions of viruses appeared more stronger effect than host tRNA selection on their nucleotide, and dinucleotide composition [40,59,60,63,64]. How about the segment plant RNA viruses and satellites? (iv) Only one study showed that increased UpA frequency greatly reduces plant RNA virus replication in the host [66], more and systemic experimental analyses on the effect of UpA frequency in plant RNA virus replication is eagerly needed. (v). How about other dinucleotide frequency affect host adaptation and replication of plant RNA viruses. (vi) Codon pair bias has significant affect on the evolution of animal and human viruses [45,83-92], however, there have been no reports on codon pair bias in plant RNA viruses. (vii) The CUB and dinucleotide bias is also related to amino acid conservation, gene length, protein structure and hydrophobicity level [32,135], new methods of analyzing CUB and dinucleotide bias are required. Therefore, future work must be performed to combine computational and experimental analyses on the evolution and host adaptability of plant RNA viruses.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This work was supported by grants from the Natural Science Foundation of Jiangsu Province of China (BK20211323), National Natural Science Foundation of China (No. 31601604), Yangzhou Municipal Key S&T Program (YZ2019051), and the Qing Lan Project of Yangzhou University.

Data accessibility statement

All data presented in this study are available on request from the corresponding authors.

Author Contributions

Zhen HE contributed to conceptual design, acquiring funding, acquiring research permits, collecting data, analyzing data, creating figures, and writing the manuscript. Lang QIN contributed to conceptual design, collecting data, and collecting data, analyzing data, creating figures. Xiaowei XU and Shiwen DING contributed to acquiring research permits, creating figures and writing the manuscript.

References

[1] Lefeuvre P, Martin DP, Elena SF, Shepherd DN, Roumagnac P, Varsani A. Evolution and ecology of plant viruses. Nat Rev Microbiol 2019:12–5. https://doi.org/10.1038/s41579-019-0232-3.

[2] Ladau J, Eloe-Fadrosh EA. Spatial, temporal, and phylogenetic scales of microbial ecology. Trends Microbiol 2019;27:662–9. https://doi.org/10.1016/j.tim.2019.03.003.

[3] Roossinck MJ. Mechanisms of plant virus evolution. Annu Rev Phytopathol 1997;35:191–209. https://doi.org/10.1146/annurev.phyto.35.1.191.

[4] Holmes EC. What does virus evolution tell us about virus origins? J Virol 2011;85:5247–51. https://doi.org/10.1128/JVI.02203-10.

[5] Holmes EC. What can we predict about viral evolution and emergence? Curr Opin Virol 2013;3:180–4. https://doi.org/10.1016/j.coviro.2012.12.003.

[6] Moya A, Elena SF, Bracho A, Miralles R, Barrio E. The evolution of RNA viruses: A population genetics view. Proc Natl Acad Sci U S A 2000;97:6967–73.

[7] Holmes EC. The evolutionary genetics of emerging viruses. Annu Rev Ecol Evol Syst 2009;40:353–72. https://doi.org/10.1146/annurev.ecolsys.110308.120248.

[8] Elena SF, Agudelo-Romero P, Lalic J. The evolution of viruses in multi-host fitness landscapes. Open Virol J 2009;3:1–6. https://doi.org/10.2174/1874357900903010001.

[9] Holmes EC. The evolution of viral emergence. Proc Natl Acad Sci 2006;103:4803–4. https://doi.org/10.1073/pnas.0601166103.

[10] Nelson MI, Holmes EC. The evolution of epidemic influenza. Nat Rev Genet 2007;8:196–205. https://doi.org/nrg2053 [pii]10.1038/nrg2053.

[11] Elena SF, Bedhomme S, Carrasco P, Cuevas JM, de la Iglesia F, Lafforgue G, et al. The Evolutionary genetics of emerging plant RNA viruses. Mol Plant-Microbe Interact 2011;24:287–93. https://doi.org/10.1094/MPMI-09-10-0214.

[12] Elena SF, Fraile A, García-Arenal F. Evolution and emergence of plant viruses. Adv Virus Res 2014;88:161–91. https://doi.org/10.1016/B978-0-12-800098-4.00003-9.

[13] Nasir A, Caetano-Anollés G. A phylogenomic data-driven exploration of viral origins and evolution. Sci Adv 2015;1. https://doi.org/10.1126/sciadv.1500527.

[14] Davino S, Willemsen A, Panno S, Davino M, Catara A, Elena SF, et al. Emergence and phylodynamics of Citrus tristeza virus in Sicily, Italy. PLoS One 2013;8:e66700. https://doi.org/10.1371/journal.pone.0066700.

[15] Lefeuvre P, Moriones E. Recombination as a motor of host switches and virus emergence: Geminiviruses as case studies. Curr Opin Virol 2015;10:14–9. https://doi.org/10.1016/j.coviro.2014.12.005.

[16] Gibbs AJ, Nguyen HD, Ohshima K. The 'emergence' of turnip mosaic virus was probably a 'gene-for-quasi-gene' event. Curr Opin Virol 2015;10:20–6. https://doi.org/10.1016/j.coviro.2014.12.004.

[17] Alicai T, Omongo CA, Maruthi MN, Hillocks RJ, Baguma Y, Kawuki R, et al. Re-emergence of Cassava Brown Streak disease in Uganda. Plant Dis 2007;91:24–9. https://doi.org/10.1094/PD-91-0024.

[18] Monjane AL, Dellicour S, Hartnady P, Oyeniran KA, Owor BE, Bezeidenhout M, et al. Symptom evolution following the emergence of maize streak virus. Elife 2020;9:e51984. https://doi.org/10.7554/eLife.51984.

[19] Fereres A. Insect vectors as drivers of plant virus emergence. Curr Opin Virol 2015;10:42–6. https://doi.org/10.1016/j.coviro.2014.12.008.

[20] Holmes EC. Evolution in health and medicine Sackler colloquium: The comparative genomics of viral emergence. Proc Natl Acad Sci U S A 2010;107 (Suppl):1742–6. https://doi.org/10.1073/pnas.0906193106.

[21] Roossinck MJ, García-Arenal F. Ecosystem simplification, biodiversity loss and plant virus emergence. Curr Opin Virol 2015;10:56–62. https://doi.org/10.1016/j.coviro.2015.01.005.

[22] Jones RAC. Plant virus emergence and evolution: origins, new encounter scenarios, factors driving emergence, effects of changing world conditions, and prospects for control. Virus Res 2009;141:113–30. https://doi.org/10.1016/j.virusres.2008.07.028.

[23] Pinel-Galzi A, Traoré O, Séré Y, Hébrard E, Fargette D. The biogeography of viral emergence: rice yellow mottle virus as a case study. Curr Opin Virol 2014;10C:7–13. https://doi.org/10.1016/j.coviro.2014.12.002.

[24] Webster CG, Frantz G, Reitz SR, Funderburk JE, Mellinger HC, McAvoy E, et al. Emergence of Groundnut ringspot virus and Tomato chlorotic spot virus in Vegetables in Florida and the Southeastern United States. Phytopathology 2015;105:388–98. https://doi.org/10.1094/PHYTO-06-14-0172-R.

[25] Traore O, Sorho F, Pinel A, Abubakar Z, Banwo O, Maley J, et al. Processes of diversification and dispersion of Rice yellow mottle virus inferred from large-scale and high-resolution phylogeographical studies. Mol Ecol 2005;14:2097–110. https://doi.org/10.1111/j.1365-294X.2005.02578.x.

[26] Fargette D, Pinel-Galzi A, Sérémé D, Lacombe S, Hébrard E, Traoré O, et al. Diversification of rice yellow mottle virus and related viruses spans the history of agriculture from the neolithic to the present. PLoS Pathog 2008;4:e1000125. https://doi.org/10.1371/journal.ppat.1000125.

[27] Walls J, Rajotte E, Rosa C. The past, present, and future of barley yellow dwarf management. Agriculture 2019;9:23. https://doi.org/10.3390/agriculture9010023.

[28] Gaunt ER, Digard P. Compositional biases in <scp>RNA</scp> viruses: Causes, consequences and applications. WIREs RNA 2021:1–24. https://doi.org/10.1002/wrna.1679.

[29] Kustin T, Stern A. Biased mutation and selection in RNA viruses. Mol Biol Evol 2021;38:575–88. https://doi.org/10.1093/molbev/msaa247.

[30] Lyons DM, Lauring AS. Mutation and epistasis in influenza virus evolution. Viruses 2018;10:1–13. https://doi.org/10.3390/v10080407.

[31] Belalov IS, Lukashev AN. Causes and implications of codon usage bias in RNA viruses. PLoS ONE 2013;8:e56642.

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

[33] Lauring AS, Acevedo A, Cooper SB, Andino R. Codon usage determines the mutational robustness, evolutionary capacity, and virulence of an RNA virus. Cell Host Microbe 2012;12:623–32. https://doi.org/10.1016/j.chom.2012.10.008.

[34] Deb B, Uddin A, Chakraborty S. Analysis of codon usage of Horseshoe Bat Hepatitis B virus and its host. Virology 2021;561:69–79. https://doi.org/10.1016/j.virol.2021.05.008.

[35] Hussain S, Shinu P, Islam MM, Chohan MS, Rasool ST. Analysis of codon usage and nucleotide bias in Middle East respiratory syndrome coronavirus genes. Evol Bioinforma 2020;16:117693432091886. https://doi.org/10.1177/1176934320918861.

[36] He W, Wang N, Tan J, Wang R, Yang Y, Li G, et al. Comprehensive codon usage analysis of porcine deltacoronavirus. Mol Phylogenet Evol 2019;141:. https://doi.org/10.1016/j.ympev.2019.106618106618.

[37] Yan Z, Wang R, Zhang L, Shen B, Wang N, Xu Q, et al. Evolutionary changes of the novel Influenza D virus hemagglutinin-esterase fusion gene revealed by the codon usage pattern. Virulence 2019;10:1–9. https://doi.org/10.1080/21505594.2018.1551708.

[38] He W, Zhao J, Xing G, Li G, Wang R, Wang Z, et al. Genetic analysis and evolutionary changes of Porcine circovirus 2. Mol Phylogenet Evol 2019;139:. https://doi.org/10.1016/j.ympev.2019.106520106520.

[39] He M, He C-Q, Ding N-Z. Evolution of Potato virus X. Mol Phylogenet Evol 2022;167:. https://doi.org/10.1016/j.ympev.2021.107336107336.

[40] Gómez MM, de Mello VE, Assandri IR, Peyrou M, Cristina J. Analysis of codon usage bias in potato virus Y non-recombinant strains. Virus Res 2020;286:. https://doi.org/10.1016/j.virusres.2020.198077198077.

[41] Sueoka N. Directional mutation pressure and neutral molecular evolution. Proc Natl Acad Sci 1988;85:2653–7. https://doi.org/10.1073/pnas.85.8.2653.

[42] Sharp PM, Cowe E. Synonymous codon usage in Saccharomyces cerevisiae. Yeast 1991;7:657–78. https://doi.org/10.1002/yea.320070702.

[43] Comeron JM, Aguadé M. An evaluation of measures of synonymous codon usage bias. J Mol Evol 1998;47:268–74. https://doi.org/10.1007/PL00006384.

[44] Hershberg R, Petrov DA. Selection on codon bias. Annu Rev Genet 2008;42:287–99. https://doi.org/10.1146/annurev.genet.42.110807.091442.

[45] Coleman JR, Papamichail D, Skiena S, Futcher B, Wimmer E, Mueller S. Virus attenuation by genome-scale changes in codon pair bias. Science (80-) 2008;320:1784–7. https://doi.org/10.1126/science.1155761.

[46] Yarus M, Folley LS. Sense codons are found in specific contexts. J Mol Biol 1985;182:529–40. https://doi.org/10.1016/0022-2836(85)90239-6.

[47] Tats A, Tenson T, Remm M. Preferred and avoided codon pairs in three domains of life. BMC Genomics 2008;9:463. https://doi.org/10.1186/1471-2164-9-463.

[48] Kunec D, Osterrieder N. Codon pair bias is a direct consequence of dinucleotide bias. Cell Rep 2016;14:55–67. https://doi.org/10.1016/j.celrep.2015.12.011.

[49] Nussinov R. Eukaryotic dinucleotide preference rules and their implications for degenerate codon usage. J Mol Biol 1981;149:125–31. https://doi.org/10.1016/0022-2836(81)90264-3.

[50] Kariin S, Burge C. Dinucleotide relative abundance extremes: a genomic signature. Trends Genet 1995;11:283–90. https://doi.org/10.1016/S0168-9525(00)89076-9.

[51] Rima BK, McFerran NV. Dinucleotide and stop codon frequencies in single-stranded RNA viruses. J Gen Virol 1997;78:2859–70. https://doi.org/10.1099/0022-1317-78-11-2859.

[52] Zhang YZ, Holmes EC. A genomic perspective on the origin and emergence of SARS-CoV-2. Cell 2020:1–5. https://doi.org/10.1016/j.cell.2020.03.035.

[53] Li J, Lai S, Gao GF, Shi W. The emergence, genomic diversity and global spread of SARS-CoV-2. Nature 2021:1–11. https://doi.org/10.1038/s41586-021-04188-6.

[54] Cheng X, Wu X, Wang H, Sun Y, Qian Y, Luo L. High codon adaptation in citrus tristeza virus to its citrus host. Virol J 2012;9:113. https://doi.org/10.1186/1743-422X-9-113.

[55] Xu X, Liu Q, Fan L, Cui X, Zhou X. Analysis of synonymous codon usage and evolution of begomoviruses. J Zhejiang Univ Sci B 2008;9:667–74. https://doi.org/10.1631/jzus.B0820005.

[56] Adams MJ, Antoniw JF. Codon usage bias amongst plant viruses. Arch Virol 2003;149:113–35. https://doi.org/10.1007/s00705-003-0186-6.

[57] Zu H, Zhang H, Yao M, Zhang J, Di H, Zhang L, et al. Molecular characteristics of segment 5, a unique fragment encoding two partially overlapping ORFs in the genome of rice black-streaked dwarf virus. PLoS ONE 2019;14:e0224569.

[58] Zhou Y, Zhang L, Zhang X, Zu H, Di H, Dong L, et al. Rice black-streaked dwarf virus Genome in China: diversification, phylogeny, and selection. Plant Dis 2017;101:1588–96. https://doi.org/10.1094/PDIS-12-16-1814-RE.

[59] He Z, Qin L, Wang W, Ding S, Xu X, Zhang S. The dinucleotide composition of sugarcane mosaic virus is shaped more by protein coding regions than by host species. Infect Genet Evol 2022;97:. https://doi.org/10.1016/j.meegid.2021.105165105165.

[60] He Z, Dong Z, Qin L, Gan H. Phylodynamics and codon usage pattern analysis of broad bean wilt virus 2. Viruses 2021;13:1–21. https://doi.org/10.3390/v13020198.

[61] He Z, Dong Z, Gan H. Comprehensive codon usage analysis of rice black-streaked dwarf virus based on P8 and P10 protein coding sequences. Infect Genet Evol 2020;86:. https://doi.org/10.1016/j.meegid.2020.104601104601.

[62] He Z, Dong Z, Gan H. Genetic changes and host adaptability in sugarcane mosaic virus based on complete genome sequences. Mol Phylogenet Evol 2020;149:. https://doi.org/10.1016/j.ympev.2020.106848106848.

[63] He Z, Gan H, Liang X. Analysis of synonymous codon usage bias in potato virus M and its adaption to hosts. Viruses 2019;11:752. https://doi.org/10.3390/v11080752.

[64] Biswas K, Palchoudhury S, Chakraborty P, Bhattacharyya U, Ghosh D, Debnath P, et al. Codon usage bias analysis of Citrus tristeza virus: higher codon adaptation to Citrus reticulata host. Viruses 2019;11:331. https://doi.org/10.3390/v11040331.

[65] He M, Guan SY, He CQ. Evolution of rice stripe virus. Mol Phylogenet Evol 2017;109:343–50. https://doi.org/10.1016/j.ympev.2017.02.002.

[66] González de Prádena A, Sánchez Jimenez A, San León D, Simmonds P, García JA, Valli AA. Plant virus genome is shaped by specific dinucleotide restrictions that influence viral infection. MBio 2020;11:1–16. https://doi.org/10.1128/mBio.02818-19.

[67] Zhang W, Zhang L, He W, Zhang X, Wen B, Wang C, et al. Genetic evolution and molecular selection of the HE gene of influenza C virus. Viruses 2019;11:167. https://doi.org/10.3390/v11020167.

[68] Butt AM, Nasrullah I, Qamar R, Tong Y. Evolution of codon usage in Zika virus genomes is host and vector specific. Emerg Microbes Infect 2016;5:1–14. https://doi.org/10.1038/emi.2016.106.

[69] Das JK, Roy S. Comparative analysis of human coronaviruses focusing on nucleotide variability and synonymous codon usage patterns. Genomics 2021;113:2177–88. https://doi.org/10.1016/j.ygeno.2021.05.008.

[70] Fros JJ, Visser I, Tang B, Yan K, Nakayama E, Visser TM, et al. The dinucleotide composition of the Zika virus genome is shaped by conflicting evolutionary pressures in mammalian hosts and mosquito vectors. PLOS Biol 2021;19:e3001201.

[71] Si F, Jiang L, Yu R, Wei W, Li Z. Study on the characteristic codon usage pattern in porcine epidemic diarrhea virus genomes and its host adaptation phenotype. Front Microbiol 2021;12:1–18. https://doi.org/10.3389/fmicb.2021.738082.

[72] Mordstein C, Cano L, Morales AC, Young B, Ho AT, Rice AM, et al. Transcription, mRNA export and immune evasion shape the codon usage of viruses. Genome Biol Evol 2021:1–30. https://doi.org/10.1093/gbe/evab106.

[73] Zhang X, Cai Y, Zhai X, Liu J, Zhao W, Ji S, et al. Comprehensive analysis of codon usage on rabies virus and other lyssaviruses. Int J Mol Sci 2018;19:2397. https://doi.org/10.3390/ijms19082397.

[74] Chakraborty P, Das S, Saha B, Sarkar P, Karmakar A, Saha A, et al. Phylogeny and synonymous codon usage pattern of Papaya ringspot virus coat protein gene in the sub-Himalayan region of north-east India. Can J Microbiol 2015;61:555–64. https://doi.org/10.1139/cjm-2015-0172.

[75] Hasegawa M, Yasunaga T, Miyata T. Secondary structure of MS2 phage RNA and bias in code word usage. Nucleic Acids Res 1979;7:2073–9. https://doi.org/10.1093/nar/7.7.2073.

[76] Kyte J, Doolittle RF. A simple method for displaying the hydropathic character of a protein. J Mol Biol 1982;157:105–32. https://doi.org/10.1016/0022-2836(82)90515-0.

[77] Duret L, Mouchiroud D. Expression pattern and surprisingly, gene length shape codon usage in Caenorhabditis, Drosophila, and Arabidopsis. Proc Natl Acad Sci 1999;96:4482–7. https://doi.org/10.1073/PNAS.96.8.4482.

[78] Sueoka N. Translation-coupled violation of Parity Rule 2 in human genes is not the cause of heterogeneity of the DNA G+C content of third codon position. Gene 1999;238:53–8. https://doi.org/10.1016/S0378-1119(99)00320-0.

[79] Fuglsang A. Accounting for background nucleotide composition when measuring codon usage bias: brilliant idea, difficult in practice. Mol Biol Evol 2006;23:1345–7. https://doi.org/10.1093/molbev/msl009.

[80] Chen F, Wu P, Deng S, Zhang H, Hou Y, Hu Z, et al. Dissimilation of synonymous codon usage bias in virus–host coevolution due to translational selection. Nat Ecol Evol 2020;4:589–600. https://doi.org/10.1038/s41559-020-1124-7.

[81] Buchan R, Stansfield I. Codon pair bias in prokaryotic and eukaryotic genomes. BMC Bioinf 2005;6:P4. https://doi.org/10.1186/1471-2105-6-S3-P4.

[82] Wang FP, Li H. Codon-pair usage and genome evolution. Gene 2009;433:8–15. https://doi.org/10.1016/j.gene.2008.12.016.

[83] Leifer I, Hoeper D, Blome S, Beer M, Ruggli N. Clustering of classical swine fever virus isolates by codon pair bias. BMC Res Notes 2011;4:521. https://doi.org/10.1186/1756-0500-4-521.

[84] Martrus G, Nevot M, Andres C, Clotet B, Martinez MA. Changes in codon-pair bias of human immunodeficiency virus type 1 have profound effects on virus replication in cell culture. Retrovirology 2013;10:78. https://doi.org/10.1186/1742-4690-10-78.

[85] Jordan-Paiz A, Franco S, Martinez MA. Synonymous codon pair recoding of the HIV-1 env gene affects virus replication capacity. Cells 2021;10:1636. https://doi.org/10.3390/cells10071636.

[86] Gao L, Wang L, Huang C, Yang L, Guo X-K, Yu Z, et al. HP-PRRSV is attenuated by de-optimization of codon pair bias in its RNA-dependent RNA polymerase nsp9 gene. Virology 2015;485:135–44. https://doi.org/10.1016/j.virol.2015.07.012.

[87] Simmonds P, Tulloch F, Evans DJ, Ryan MD. Attenuation of dengue (and other RNA viruses) with codon pair recoding can be explained by increased CpG/UpA dinucleotide frequencies. Proc Natl Acad Sci 2015;112:E3633–4. https://doi.org/10.1073/pnas.1507339112.

[88] Broadbent AJ, Santos CP, Anafu A, Wimmer E, Mueller S, Subbarao K. Evaluation of the attenuation, immunogenicity, and efficacy of a live virus vaccine generated by codon-pair bias de-optimization of the 2009 pandemic H1N1 influenza virus, in ferrets. Vaccine 2016;34:563–70. https://doi.org/10.1016/j.vaccine.2015.11.054.

[89] Khedkar PH, Osterrieder N, Kunec D. Codon pair bias deoptimization of the major oncogene meq of a very virulent Marek's disease virus. J Gen Virol 2018;99:1705–16. https://doi.org/10.1099/jgv.0.001136.

[90] Eschke K, Trimpert J, Osterrieder N, Kunec D. Attenuation of a very virulent Marek's disease herpesvirus (MDV) by codon pair bias deoptimization. PLOS Pathog 2018;14:e1006857.

[91] Li P, Ke X, Wang T, Tan Z, Luo D, Miao Y, et al. Zika virus attenuation by codon pair deoptimization induces sterilizing immunity in mouse models. J Virol 2018;92(17):e00701–e718. https://doi.org/10.1128/JVI.00701-18.

[92] Plant EP, Ye Z. A codon-pair bias associated with network interactions in influenza A, B, and C genomes. Front Genet 2021;12:1–6. https://doi.org/10.3389/fgene.2021.699141.

[93] Wang Y, Hill K, Singh S, Kari L. The spectrum of genomic signatures: From dinucleotides to chaos game representation. Gene 2005;346:173–85. https://doi.org/10.1016/j.gene.2004.10.021.

[94] Karlin S, Doerfler W, Cardon LR. Why is CpG suppressed in the genomes of virtually all small eukaryotic viruses but not in those of large eukaryotic viruses? J Virol 1994;68:2889–97. https://doi.org/10.1128/jvi.68.5.2889-2897.1994.

[95] Karlin S, Ladunga I, Blaisdell BE. Heterogeneity of genomes: measures and values. Proc Natl Acad Sci 1994;91:12837–41. https://doi.org/10.1073/pnas.91.26.12837.

[96] Karlin S, Mrazek J. Compositional differences within and between eukaryotic genomes. Proc Natl Acad Sci 1997;94:10227–32. https://doi.org/10.1073/pnas.94.19.10227.

[97] Di Giallonardo F, Schlub TE, Shi M, Holmes EC. Dinucleotide composition in animal RNA viruses is shaped more by virus family than by host species. J Virol 2017;91:1–15. https://doi.org/10.1128/JVI.02381-16.

[98] Gu H, Fan RLY, Wang D, Poon LLM. Dinucleotide evolutionary dynamics in influenza A virus. Virus Evol 2019;5:vez038. https://doi.org/10.1093/ve/vez038.

[99] Fros JJ, Dietrich I, Alshaikhahmed K, Passchier TC, Evans DJ, Simmonds P. CpG and upA dinucleotides in both coding and non-coding regions of echovirus 7 inhibit replication initiation post-entry. Elife 2017;6:1–29. https://doi.org/10.7554/eLife.29112.

[100] Takata MA, Gonçalves-Carneiro D, Zang TM, Soll SJ, York A, Blanco-Melo D, et al. CG dinucleotide suppression enables antiviral defence targeting non-self RNA. Nature 2017;550:124–7. https://doi.org/10.1038/nature24039.

[101] Trus I, Udenze D, Berube N, Wheler C, Martel M-J, Gerdts V, et al. CpG-recoding in Zika virus genome causes host-age-dependent attenuation of infection with protection against lethal heterologous challenge in mice. Front Immunol 2020;10:3077. https://doi.org/10.3389/fimmu.2019.03077.

[102] He M, He CQ, Ding NZ. Evolution of potato virus X. Mol Phylogenet Evol 2021:107336. https://doi.org/10.1016/j.ympev.2021.107336.

[103] Sharp PM, Li WH. An evolutionary perspective on synonymous codon usage in unicellular organisms. J Mol Evol 1986;24:28–38. https://doi.org/10.1007/BF02099948.

[104] Hu J, Wang Q, Zhang J, Chen H, Xu Z, Zhu L, et al. The characteristic of codon usage pattern and its evolution of hepatitis C virus. Infect Genet Evol 2011;11:2098–102. https://doi.org/10.1016/j.meegid.2011.08.025.

[105] He W, Auclert LZ, Zhai X, Wong G, Zhang C, Zhu H, et al. Interspecies transmission, genetic diversity, and evolutionary dynamics of pseudorabies virus. J Infect Dis 2019;219:1705–15. https://doi.org/10.1093/infdis/jiy731.

[106] Xia X. An improved implementation of codon adaptation index. Evol Bioinforma 2007;3:117693430700300. https://doi.org/10.1177/117693430700300028.

[107] Puigbò P, Bravo IG, Garcia-Vallve S. CAIcal: A combined set of tools to assess codon usage adaptation. Biol Direct 2008;3:38. https://doi.org/10.1186/1745-6150-3-38.

[108] Puigbò P, Aragonès L, Garcia-Vallvé S. RCDI/eRCDI: a web-server to estimate codon usage deoptimization. BMC Res Notes 2010;3:87. https://doi.org/10.1186/1756-0500-3-87.

[109] Ibrahim A, Fros J, Bertran A, Sechan F, Odon V, Torrance L, et al. A functional investigation of the suppression of CpG and UpA dinucleotide frequencies in plant RNA virus genomes. Sci Rep 2019;9:18359. https://doi.org/10.1038/s41598-019-54853-0.

[110] Hall TA. BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. Nucleic Acids Symp Ser 1999;41:95–8.

[111] Rice P, Longden I, Bleasby A. EMBOSS: the European Molecular Biology Open Software Suite. Trends Genet 2000;16:276–7. https://doi.org/10.1016/s0168-9525(00)02024-2.

[112] Peden JF. Analysis of codon usage. University of Nottingham; 2000.

[113] Rozas J, Sanchez-DelBarrio JC, Messeguer X, Rozas R. DnaSP, DNA polymorphism analyses by the coalescent and other methods. Bioinformatics 2003;19:2496–7. https://doi.org/10.1093/bioinformatics/btg359.

[114] Stothard P. The sequence manipulation suite: JavaScript programs for analyzing and formatting protein and DNA sequences. Biotechniques 2000;28:1102–4. https://doi.org/10.2144/00286ir01.

[115] Tamura K, Stecher G, Kumar S. MEGA11: Molecular Evolutionary Genetics Analysis Version 11. Mol Biol Evol 2021;38:3022–7. https://doi.org/10.1093/molbev/msab120.

[116] Pinheiro M, Afreixo V, Moura G, Freitas A, Santos MAS, Oliveira JL. Statistical, computational and visualization methodologies to unveil gene primary structure features. Methods Inf Med 2006;45:163–8. https://doi.org/10.1055/s-0038-1634061.

[117] Alexaki A, Kames J, Holcomb DD, Athey J, Santana-Quintero LV, Lam PVN, et al. Codon and Codon-Pair Usage Tables (CoCoPUTs): facilitating genetic variation analyses and recombinant gene design. J Mol Biol 2019;431:2434–41. https://doi.org/10.1016/j.jmb.2019.04.021.

[118] Huang Y, Lin T, Lu L, Cai F, Lin J, Jiang Y, et al. Codon pair optimization (CPO): a software tool for synthetic gene design based on codon pair bias to improve the expression of recombinant proteins in Pichia pastoris. Microb Cell Fact 2021;20:209. https://doi.org/10.1186/s12934-021-01696-y.

[119] Bennetzen JL, Hall BD. Codon selection in yeast. J Biol Chem 1982;257:3026–31.

[120] Chin JX, Chung BKS, Lee DY. Codon optimization onLine (COOL): a web-based multi-objective optimization platform for synthetic gene design. Bioinformatics 2014;30:2210–2. https://doi.org/10.1093/bioinformatics/btu192.

[121] Bourret J, Alizon S, Bravo IG. COUSIN (COdon Usage Similarity INdex): A normalized measure of codon usage preferences. Genome Biol Evol 2019;11:3523–8. https://doi.org/10.1093/gbe/evz262.

[122] Puigbo P, Romeu A, Garcia-Vallve S. HEG-DB: a database of predicted highly expressed genes in prokaryotic complete genomes under translational selection. Nucleic Acids Res 2007;36:D524–7. https://doi.org/10.1093/nar/gkm831.

[123] Grote A, Hiller K, Scheer M, Munch R, Nortemann B, Hempel DC, et al. JCat: a novel tool to adapt codon usage of a target gene to its potential expression host. Nucleic Acids Res 2005;33:W526–31. https://doi.org/10.1093/nar/gki376.

[124] Puigbo P, Guzman E, Romeu A, Garcia-Vallve S. OPTIMIZER: a web server for optimizing the codon usage of DNA sequences. Nucleic Acids Res 2007;35:W126–31. https://doi.org/10.1093/nar/gkm219.

[125] Sabi R, Volvovitch Daniel R, Tuller T. stAIcalc: tRNA adaptation index calculator based on species-specific weights. Bioinformatics 2016:. https://doi.org/10.1093/bioinformatics/btw647btw647.

[126] Wu G, Bashir-Bello N, Freeland SJ. The Synthetic Gene Designer: A flexible web platform to explore sequence manipulation for heterologous expression. Protein Expr Purif 2006;47:441–5. https://doi.org/10.1016/j.pep.2005.10.020.

[127] Yang SQ, Liu Y, Wu XY, Cheng XF, Wu XX. Synonymous codon pattern of cowpea mild mottle virus sheds light on its host adaptation and genome evolution. Pathogens 2022;11:419. https://doi.org/10.3390/pathogens11040419.

[128] He Z, Ding SW, Guo JY, Qin L, Xu XW. Synonymous codon usage analysis of three narcissus potyviruses. Viruses 2022;14:846. https://doi.org/10.3390/v14050846.

[129] Patil AB, Dalvi VS, Mishra AA. Analysis of synonymous codon usage bias and phylogeny of coat protein gene in banana bract mosaic virus. Virus Dis 2017;28(2):156–63. https://doi.org/10.1007/s13337-017-0380-x.

[130] Prádena AGD, Jimenez AS, León DS, Simmonds P, Valli AA. Plant virus genome is shaped by specific dinucleotide restrictions that influence viral infection. mBio 2020;11(1):e02818–e2819. https://doi.org/10.1128/mBio.02818-19.

[131] Tian L, Shen XJ, Murphy RW, Shen YY. The adaptation of codon usage of +ssRNA viruses to their hosts. Infect Genet Evol 2018;63:175–9. https://doi.org/10.1016/j.meegid.2018.05.034.

[132] Huang SH, Yuan Y, Wang CK, Ren R, He ZW, Zhi HJ. Analysis on codon usage of CP gene in soybean mosaic virus. Chin J Oil Crop Sci 2015;37(2):148–53. https://doi.org/10.7505/j.issn.1007-9084.2015.02.004.

[133] Wu NC, Young AP, Al-Mawsawi LQ, Olson CA, Feng J, Qi H, et al. High-throughput profiling of influenza a virus hemagglutinin gene at single-nucleotide resolution. Sci Rep 2014;4942. https://doi.org/10.1038/srep04942.

[134] Domingo-Calap P, Cuevas JM, Sanjuán R. The fitness effects of random mutations in single-stranded DNA and RNA bacteriophages. PLoS Genet 2009;5(11):e1000742. https://doi.org/10.1371/journal.pgen.1000742.

[135] Sanjuán R, Moya A, Elena SF. The distribution of fitness effects caused by single-nucleotide substitutions in an RNA virus. Proc Natl Acad Sci U S A 2004;101(22):8396–401. https://doi.org/10.1073/pnas.0400146101.

[136] Carrasco P, De lIF, Elena SF. Distribution of fitness and virulence effects caused by single-nucleotide substitutions in tobacco etch virus. J Virol 2007; 81(23), 12979-84. https://doi.org/10.1128/JVI.00524-0.