Since January 2020 Elsevier has created a COVID-19 resource centre with free information in English and Mandarin on the novel coronavirus COVID-19. The COVID-19 resource centre is hosted on Elsevier Connect, the company's public news and information website.

Virus Research 178 (2013) 264–271

Contents lists available at ScienceDirect

Virus Research

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

The evolution of codon usage in structural and non-structural viral genes: The case of *Avian coronavirus* and its natural host *Gallus gallus*

Paulo Eduardo Brandão*

Department of Preventive Veterinary Medicine and Animal Health, School of Veterinary Medicine, University of São Paulo, Brazil

ARTICLE INFO

Article history:
Received 4 July 2013
Received in revised form
13 September 2013
Accepted 20 September 2013
Available online 30 September 2013

Keywords:
Codon usage
Avian coronavirus
Gallus gallus
Virus–host

ABSTRACT

To assess the codon evolution in virus–host systems, *Avian coronavirus* and its natural host *Gallus gallus* were used as a model. Codon usage (CU) was measured for the viral spike (S), nucleocapsid (N), non-structural protein 2 (NSP2) and papain-like protease (PL^{pro}) genes from a diverse set of *A. coronavirus* lineages and for *G. gallus* genes (lung surfactant protein A, intestinal cholecystokinin, oviduct ovomucin alpha subunit, kidney vitamin D receptor and the ubiquitary beta-actin) for different *A. coronavirus* replicating sites. Relative synonymous codon usage (RSCU) trees accommodating all virus and host genes in a single topology showed a higher proximity of *A. coronavirus* CU to the respiratory tract for all genes. The codon adaptation index (CAI) showed a lower adaptation of S to *G. gallus* compared to NSP2, PL^{pro} and N. The effective number of codons (*Nc*) and GC_{3s} revealed that natural selection and genetic drift are the evolutionary forces driving the codon usage evolution of both *A. coronavirus* and *G. gallus* regardless of the gene being considered. The spike gene showed only one 100% conserved amino acid position coded by an *A. coronavirus* preferred codon, a significantly low number when compared to the three other genes ($p < 0.0001$). Virus CU evolves independently for each gene in a manner predicted by the protein function, with a balance between natural selection and mutation pressure, giving further molecular basis for the viruses' ability to exploit the host's cellular environment in a concerted virus–host molecular evolution.

1. Introduction

Codon usage (CU) refers to the frequency of the occurrence of each codon for at least two-fold degenerate codons (Hershberg and Petrov, 2008), i.e., it is an indication of the 'preference' of a genome for one or more codons if more than one codon is possible for the same amino acid.

Natural selection for efficient protein synthesis speed and folding and genetic drift based on mutation pressure that leads to a homogeneous genome and the 3rd codon's GC%s are the most evident forces under codon usage evolution that could lead to detectable codon usage bias (CUB) (Yang and Nielsen, 2008), which has been increasingly used in studies on virus and host molecular evolution.

Avian coronavirus (*Nidovirales*: *Coronaviridae*: *Coronavirinae*: *Gammacoronavirus*), which originated approximately 4800 years ago (Woo et al., 2012) and has a large number of serotypes and genotypes, primarily infects the respiratory tract of laying hens, broilers and breeders but can also infect the kidneys, intestines and reproductive tracts of both females and males (Cook et al., 2012), depending on the pathotype. Though affinity to different classes of cell membrane glycans could be one of the explanations for the existence of the different viral pathotypes (Wickramasinghe et al., 2011), the exact mechanism for this level of diversity is still unknown.

The 27.6 kb single-stranded positive sense RNA of *A. coronavirus* encodes 23 proteins, and the first two-thirds of the genome contains ORF 1, which encodes 15 non-structural proteins involved in RNA transcription and replication (Masters, 2006; Ziebuhr and Snijder, 2007). Among these, the papain-like protease (PL^{pro}) is the proteolytic processor of the N-proximal domain of polyproteins pp1a and pp1ab (Ziebuhr et al., 2000). Non-structural protein 2 (NSP2), the first in ORF 1 because the *A. coronavirus* lacks NSP1, has a still undefined role, though a role on global RNA synthesis has been suggested (Graham et al., 2005).

Of the structural proteins, the spike glycoprotein (S) has a strong interaction with the host immune system and is so highly polymorphic that mutations in only 10 amino acids on the amino terminal ectodomain (S1) could result in the loss of cross-reactivity (Cavanagh, 2007). While S1 allows the virus to attach to α2,3Sia, which is widespread in chicken cells (Winter et al., 2008), the carboxy terminal S2 has the capacity to fuse virus-to-cell and cell-to-cell membranes (Masters, 2006).

The nucleocapsid (N) protein binds to the genomic RNA due to its positively charged amino acid domains, and though under a more

* Correspondence to: Av. Prof. Dr. Orlando M. Paiva, 87, Cidade Universitária, CEP 05508-270 São Paulo, SP, Brazil. Tel.: +55 11 3091 7655; fax: +55 11 3091 7928.
E-mail address: paulo7926@usp.br

0168-1702/$ – see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.virusres.2013.09.033

strict mutation constraint than S, positive selection plays a role in N evolution (Kuo et al., 2013; Masters, 2006).

The codon usage of *A. coronavirus* has been reported to be highly to moderately biased but closer to that found in the respiratory tract of *Gallus gallus* when compared to other tissues (Brandão, 2012). However, that report was limited because codon usage was measured based on only the spike gene.

The aim of this study was to assess the evolution of codon usage in viral structural and non-structural genes and their molecular relationship with host codon usage using *A. coronavirus* and its natural host *G. gallus* as a model.

2. Materials and methods

2.1. Sequences

2.1.1. A. coronavirus

For *A. coronavirus*, sequences were chosen to promote diversity of geographic origin and serotype/genotypes, including the archetypical strains, with an effort to keep the same datasets if possible. Because the number of complete genomes and genes for *A. coronavirus* available in GenBank did not allow for the representation of such diversity, only partial genes were used in this study instead of complete ones to have the most diverse dataset possible. As the accuracy of codon usage measurements is lower for short sequences, sequences <100 codons in length (Roth et al., 2012) were not included. Sequence redundancy was avoided by keeping only one sequence if any 100% nucleotide identity was found.

Following these criteria, this study included 64 S protein sequences, codons 1–169 (14.6% of the 1162 S codons); 25 N protein sequences, codons 301–409 (26.7% of the 409 N codons); 18 NSP2 sequences, codons 1–245 (36.4% of the 673 NSP2 codons); and 15 papain-like protease sequences, codons 3–437 (99.5% of the 437 PL^{pro} codons). The accession numbers are shown in Fig. 1. All indicated positions are relative to the complete genome of the *Avian infectious bronchitis virus* strain M41 (DQ834384.1).

2.1.2. G. gallus

Aiming to assess the codon usage of the different tissues in which *A. coronavirus* replicates in chicken, non-redundant complete codon sequences were retrieved from the GenBank database and from the *G. gallus* genome project for cholecystokinin, expressed in the duodenum (NM_001001741.1 and GFC_000002315.3); lung surfactant pulmonary-associated protein A1 (SFTPA1), expressed in the lungs (NM_204606.1 and GFC_000002315.3); vitamin D receptor, expressed in the kidneys (NM_205098.1 and GFC_000002315.3); and ovomucin alpha subunit, expressed in the oviduct (AB046524.1 and GFC_000002315.3). As a reference, the complete *G. gallus* beta-actin gene (L08165 and GFC_000002315.3) was included in the analyses as a ubiquitously expressed gene.

All sequences used in this study can be found in Supplementary material 1.

2.2. Relative synonymous codon usage (RSCU)

RSCU, the relationship between the observed and the expected frequency of a codon if the synonymous codon usage is random (Roth et al., 2012), was calculated for 59 codons, excluding the single codons of methionine and tryptophan and the three stop codons, using the equation $RSCU_i = X_i/(\Sigma_i X_i/m)$ (Nei and Kumar, 2000), where X_i is the total count for a given codon, $\Sigma_i X_i$ is the sum of the count for all synonymous codons regarding the amino acid under consideration and m is the number of possible isoacceptors for that amino acid, implemented in MEGA 5.0 (Tamura et al., 2011).

The continuous RSCU values from *A. coronavirus* and *G. gallus* genes were converted to binary data using the value *1* for RSCUs >1, when a given codon was preferred for a specific amino acid, or *0* for RSCUs ≤ 1, when the codon was not preferred (RSCU < 1) or was neutral (RSCU = 1). Finally, the combined dataset of the four viral and five host genes was used to build a binary 59 characters × 132 sequences matrix (Supplementary material 2) for the presence or absence of a preferred codon, which was used to build a neighbor-joining tree (1000 bootstrap replicates) using PAUP, version 4.1b (Swofford, 2000).

2.3. Codon adaptation index (CAI)

The CAI is a measure of codon usage derived from the geometric mean of the relative codon adaptiveness for each codon based on a set of translationally optimal codons used as a reference (Roth et al., 2012) and can be calculated according to the equation

$$\mathrm{CAI}_g = \prod_{k=1}^{61} W_k^{X_{k,g}}$$

Here, w_k is the relative adaptiveness of the kth codon (61 codons; the three stop codons were excluded), and $X_{k,g}$ is the fraction of the codon k relative to the total number of codons in the gene.

Values closer to 1 indicate a high fitness in terms of codon usage for a given codon sequence in relation to the reference system (Sharp and Li, 1987), i.e., a high adaptation of viral genes to the host.

The CAI was calculated for sequences from both *A. coronavirus* and *G. gallus* using a reference set of highly expressed *G. gallus* genes available in the ACUA 1.0 software (Vetrivel et al., 2007).

2.4. Effective number of codons (Nc)

Nc is a measure of the total number of different codons present in a sequence and shows the bias from equal use of all synonymous codons for a given amino acid, with each synonymous codon treated as an allele as in the calculation of the effective number of alleles in population genetics (Roth et al., 2012). *Nc* values range from 20 to 61, with values closer to 61 indicating a lower bias (Wright, 1990).

Nc was calculated according to the equation $Nc = 2 + (9/F2) + (1/F3) + (5/F4) + (3/F6)$, where F is the average homozygosity for equal use of each synonymous codon for each class of degeneracy ranging from 2 to 6, using ACUA 1.0 (Vetrivel et al., 2007).

2.5. Codon selection test

The expected effective number of codons (ENC), a measure of codon usage affected only by the $GC_{3\%}$ (the percentage of G or C at the third position of all codons in a sequence) as a result of mutation pressure and drift, was calculated using the equation $ENC_{expec} = 2 + s + 29[s^2 + (1-s)^2]^{-1}$ (Wright, 1990), where s is the $GC_{3\%}$ ranging from 0 to 100%.

The ENC and simulated $GC_{3\%}$ values were plotted as a curve together with the *Nc* and observed $GC_{3\%}$ values; an *Nc* × observed plot lying on the ENC × simulated curve indicates genetic drift/mutational bias, while plots outside the curve indicate natural selection (Wright, 1990).

2.6. Analysis of conserved amino acids coded by preferred codons

To assess the significance of each preferred codon on the molecular evolution of *A. coronavirus*, 100% conserved amino acid positions coded by the preferred codon(s), i.e., those with RSCUs >1,

were counted for each gene, and the significance of the differences was assessed with Fisher's exact test and the odds ratio (OR).

2.7. Protein selection test

To understand the relationship between codon and protein selection, the occurrence of purifying or positive selection on *A. coronavirus* S, N, NSP2 and PL^{pro} sequences was tested with Fisher's exact test of neutrality for sequence pairs using the Nei–Gojobori method (Nei and Gojobori, 1986) for the difference between the synonymous and non-synonymous substitution distances (dS–dN) using Mega 5 (Tamura et al., 2011).

3. Results

3.1. RSCU phylogeny

Fig. 1 shows that *G. gallus* RSCUs segregate in a tissue-specific manner in a topology supported by bootstrap values of 100 for each gene analyzed.

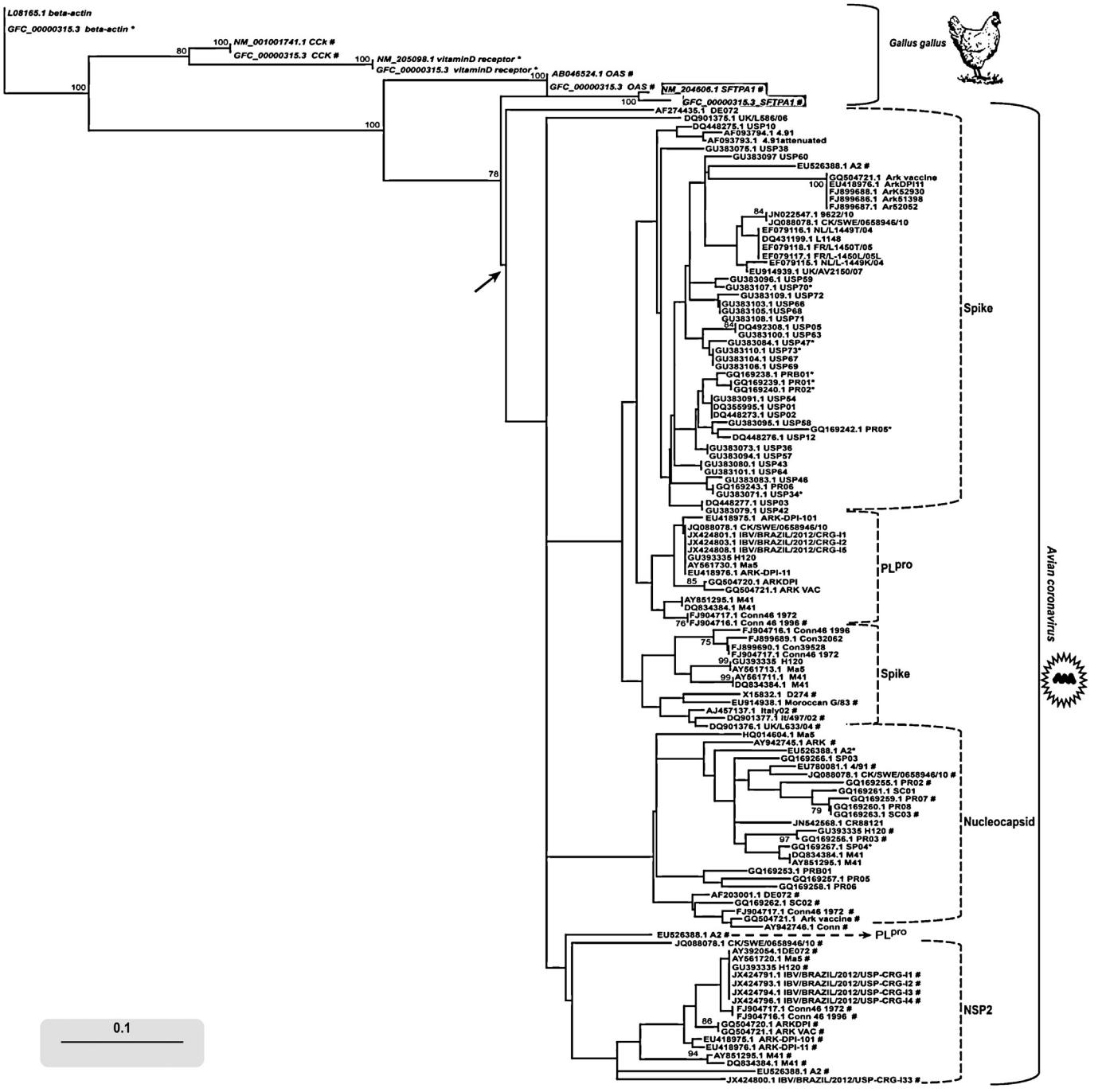

Fig. 1. Neighbor-joining distance tree for the relative synonymous codon usage (RSCU) for the *Avian coronavirus* spike (S), nucleocapsid (N), non-structural protein 2 (NSP2) and papain-like protease (PL^{pro}) genes and the *Gallus gallus* beta-actin, lung surfactant protein A (SFTPA1, gray background), intestinal cholecystokinin (CCK), oviduct ovomucin alpha subunit (OSA) and kidney vitamin D receptor genes. The tree was based on binary data using the value *1* for RSCUs > 1 (codon is preferred) or *0* for RSCUs ≤ 1 when the codon is not preferred (RSCU < 1) or is neutral (RSCU = 1). ENC (effective number of codons) values <40 and >45 are marked with an asterisk and a hash, respectively; sequences with ENC values between 40 and 45 have no marks. The arrow indicates the separation between *G. gallus* and *Avian coronavirus* clusters. Numbers at each node are bootstrap values (1000 replicates, only values >50 are shown). The bar represents the codon usage preferences distance.

For the *A. coronavirus* RSCUs, all genes segregated in gene-specific clusters, except for the sequence EU526388.1 A2 PL^{pro}, which segregated closer to the NSP2 cluster. All strains segregated in a cluster separated from *G. gallus*, with the internal nodes resulting in the genotype-specific sub-clusters for the S gene, including those for the archetypes Connecticut, Massachusetts and Arkansas, with two sub clusters and the PL^{pro} cluster between them. No pathotype-specific cluster was found.

Though the distinction between the *A. coronavirus* and *G. gallus* RSCUs clusters is also clear for the N, NSP2 and PL^{pro} genes, a less resolved topology emerges because the distinction among the different genotypes is not sustained.

For all four genes, *A. coronavirus* clusters show an increasing distance from the *G. gallus* clusters, with them being closer to SFTPA1 (from the respiratory tract) and more distant from cholecystokinin (from the intestine) and with the ubiquitous beta-actin cluster being the most distant from both *A. coronavirus* and the other *G. gallus* clusters.

3.2. Codon adaptation index (CAI)

Mean CAI values for the *A. coronavirus* S, N, NSP2 and PL^{pro} genes were 0.66 (sd 0.01), 0.77 (sd 0.01), 0.69 (sd 0.01) and 0.7 (sd 0.01), respectively, while, for the *G. gallus* genes, the mean CAI was 0.81 (sd 0.06), ranging from 0.71 for the pulmonary gene SFTPA1 to 0.88 for the renal vitamin D receptor (mean values for two sequences).

A boxplot representation of *G. gallus* and *A. coronavirus* CAIs (Fig. 2) shows that, in relation to *G. gallus*, S has the lowest values (0.64–0.7) and N has the highest values (0.75–0.79), while NSP2 and PL^{pro} have intermediate values (0.69–0.71), with non-overlapping medians.

3.3. Effective number of codons (Nc)

The mean *Nc* values for *A. coronavirus* S, N, NSP2 and PL^{pro} were 43 (sd 2.31), 44.9 (sd 3.64), 51.33 (sd 1.56) and 43.79 (sd 0.86), respectively, and for *G. gallus*, the mean *Nc* values were 33.59 for vitamin D receptor, 40.03 for beta-actin, 46.48 for cholecystokinin, 50.21 for SFTPA1 and 53.01 for ovomucin.

3.4. Codon selection test

The *Nc* x $GC_{3\%}$ graphs (Fig. 3) show that, regardless of the *A. coronavirus* gene under consideration (S, N, NSP2 or PL^{pro}), all plots fall either just below or in the vicinity of the ENC × $GC_{3\%}$ expected curve. This same pattern was also found for the *G. gallus* genes, though with plots dislocated to the right side of the graph due to a higher $GC_{3\%}$ content.

3.5. Analysis of conserved amino acids coded by preferred codons

The number of 100% conserved amino acid positions coded by the preferred codons for genes S, N, NSP2 and PL^{pro} was one, 20, 28 and 71, respectively. Fisher's exact test showed that only the S gene presented a statistically significant lower number of occurrences (Table 1) when compared to the other 3 genes ($p < 0.0001$), with ORs

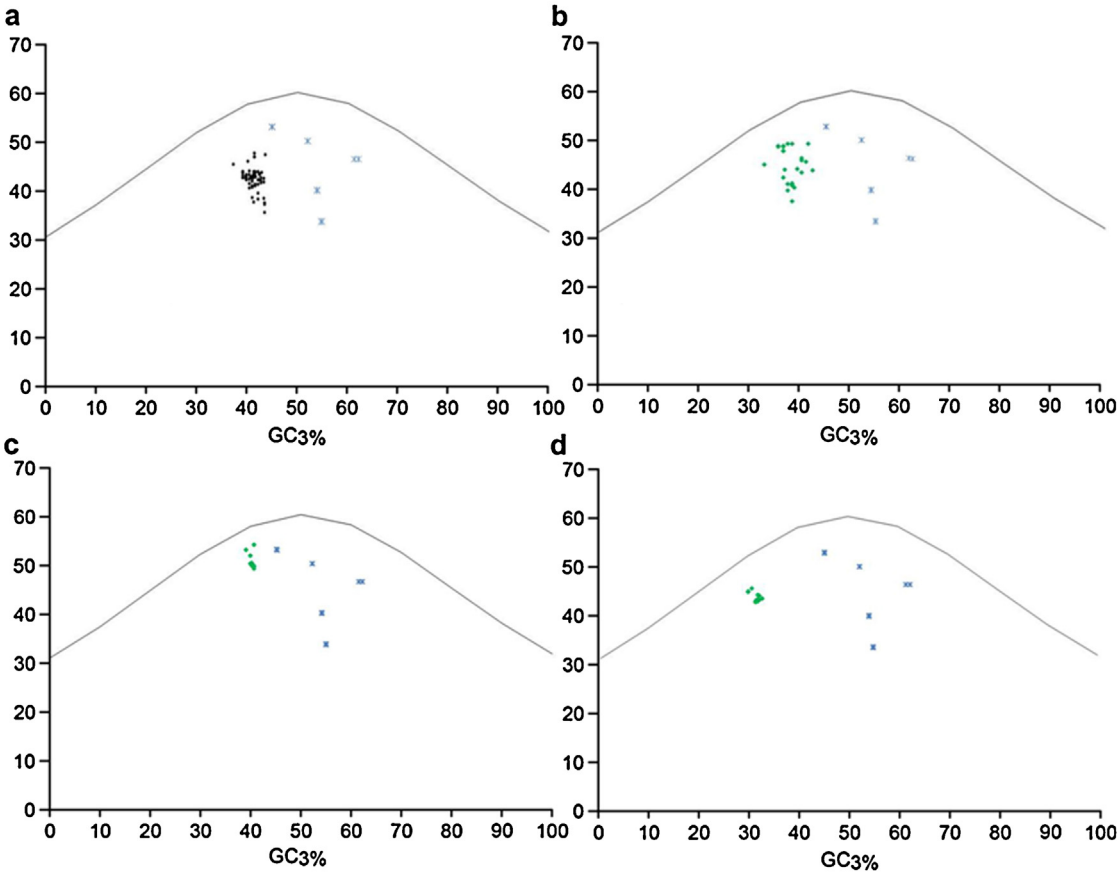

Fig. 2. Four graphs showing the expected (seen in the curves of each graph) and observed (seen in the points of each graph) effective number of codons (ENC and *Nc*, respectively) (*Y* axis) and the expected and observed $GC_{3\%}$ (*X* axis) for (a) *Avian coronavirus* spike (S); (b) nucleocapsid (N); (c) non-structural protein 2 (NSP2) and (d) papain-like protease (PL^{pro}) (*dots*) and *Gallus gallus* beta-actin, lung surfactant protein A, intestinal cholecystokinin, oviduct ovomucin alpha subunit and kidney vitamin D receptor (*asterisks*).

Table 1
Conserved amino acid (aa) positions in the *Avian coronavirus* spike (S), nucleocapsid (N), non-structural protein 2 (NSP2) and papain-like protease (PL^{pro}) genes coded by a preferred codon and the preferred codon for each aa in the *Gallus gallus* beta-actin (B-act), lung surfactant protein A (SFTPA1), intestinal cholecystokinin (CCK), oviduct ovomucin alpha subunit (Ovo) and kidney vitamin D receptor (ViTD rec) genes. Tryptophan and methionine, coded by a single codon, were excluded. Codon preference was indicated by relative synonymous codon usage (RSCU) >1. Positions are provided only for *Avian coronavirus* genes as *G. gallus* genes were used as the reference for comparison.

	G gallus					*Avian coronavirus*							
aa	B-act	CCK	Ovo	SFTPA1	VitD rec	S	Position	N	Position	Nsp2	Position	PL^{Pro}	Position
F	UUC	UUC	UUU/UUC	UUU	UUC	UUU	155	UUU	313, 390	UUU	52, 64, 101, 138, 200, 217	UUU	28, 54, 57, 102, 144, 211, 220, 270, 326, 369, 393
L	CUC/CUG	CUC/CUG	CUG	UUG/CUU/CUA/CUG	CUC/CUG	NC	NC	NC	NC	NC	NC	CUU	16, 47, 58, 94, 111, 129, 273, 288, 315, 413
I	AUU/AUC	AUC	AUU/AUC	AUU	AUC	NC	NC	AUU	319, 397	AUU	23, 199, 210	AUU	52, 133, 194, 226, 383, 429
V	GUG	GUG	GUU/GUG	GUU/GUG	GUC/GUG	NC	NC	NC	NC	GUU	154, 163, 226, 228, 235	GUU	191, 192, 317, 323, 324, 371, 394, 430, 435
S	UCU/UCC/AGC	UCC/AGC	UCU/UCC/UCA/AGU/AGC	UCU/AGU/AGC	UCC/AGC	NC	NC	UCA	340, 344	NC	NC	NC	NC
P	CCU/CCC	CCC	CCU/CCC/CCA	CCU	CCC	NC	NC	CCA	338	NC	NC	CCU	178, 294, 338
T	ACC/ACA	ACA	ACU/ACC/ACA	ACU/ACA	ACC/ACG	NC	NC	NC	NC	ACU	123, 167, 241	NC	NC
A	GCC	GCU/GCG	GCU/GCC/GCA	GCU/GCA	GCC	NC	NC	GCA	376	NC	NC	NC	NC
Y	UAC	UAC	UAU/UAC	UAU/UAC	UAC	NC	NC	UAU	316	NC	NC	NC	NC
H	CAC	CAC	CAU/CAC	CAU	CAC	NC	NC	NF	NF	NC	NC	CAU	143, 201
Q	CAG	CAG	CAA/CAG	CAA	CAG	NC	NC	CAG	312, 369, 387	NC	NC	NC	NC
N	AAC	All RSCUs = 1	AAC	AAU	AAC	NC	NC	AAU	315, 385, 407	NC	NC	AAU	27, 82, 97, 140, 186, 296, 343
K	AAG	AAG	AAA	AAA	AAG	NC	NC	NC	NC	AAA	6, 21, 86	NC	NC
D	GAU	GAU	GAU/GAC	GAC	GAC	NC	NC	GAU	314, 374	NC	NC	GAU	5, 105, 160, 176, 182, 184, 217, 258, 281
E	GAG	All RSCUs = 1	GAA	GAG	GAG	NC	NC	NC	NC	GAA	98, 136, 142, 165	GAA	130, 164, 185, 342
C	UGC	UGC	UGU	UGU	UGC	NC	NC	UGU	320, 323	UGU	68, 242	UGU	132, 154, 183, 202, 439
R	CGU/AGA	CGC/CGG/AGG	AGA/AGG	CGA/AGA	CGC/CGG/AGG	NC	NC	AGA	349	CGU	54, 111	NC	NC
G	GGU/GGC	GGC	GGA	GGA	GGC	NC	NC	NC	NC	NC	NC	GGU	56, 86, 177, 319, 402

NC: no 100% conserved amino acids positions coded by the preferred codon; NF: amino acid not found in the sequence.

Table 2
The mean number of amino acid residues in the sequences used for this study from the *Avian coronavirus* spike (S), nucleocapsid (N), non-structural protein 2 (NSP2) and papain-like protease (PL^{pro}) genes coded by a preferred codon and the preferred codon for each aa in the *Gallus gallus* beta-actin (B-act), lung surfactant protein A (SFTPA1), intestinal cholecystokinin (CCK), oviduct ovomucin alpha subunit (Ovo) and kidney vitamin D receptor (ViTD rec) genes.

Amino acid	*G. gallus*				*Avian coronavirus*				
	S	N	Nsp2	PL^{pro}	B-act	CCK	Ovo	SFTPA1	VitD rec
F	9.2	3	15.9	21.8	13	3	87	5	24
L	14.6	4.28	26.7	37.7	27	12	111	27	43
I	5.8	2.04	13.1	19.3	28	5	112	8	20
V	14.3	7.08	23.1	38.3	22	5	137	10	22
S	19.8	7.24	15.2	33.3	25	16	163.5	14	45
P	7.0	9.96	8.1	14.7	19	8	116	9	24
T	13.1	5.52	13.6	26.6	26	4	147	11	19
A	14.2	5.6	28.0	36.3	29	13	85.5	15	24.5
Y	11.0	1.2	3.0	19.1	15	5	76	12	7
H	5.3	0	1.0	5.3	9	4	46.5	1	13
Q	7.1	6.12	13.9	11.9	12	8	77.5	12	21
N	12.6	5.8	3.9	28.1	9	2	103.5	14	13
K	6.0	11.48	21.2	33.7	19	3	133.5	14	28
D	2.6	13.6	12.2	28.5	23	6	118	7	33
E	1.8	9.96	12.9	22.3	26	6	131	19	33.5
C	7.7	2.16	5.0	11.9	6	2	201	8	13
R	2.9	8.24	11.8	13.1	18	10.5	60	7	26
G	11.4	4.56	9.1	22.3	28	13	142	20	18
M	4.3	0	5.4	2.3	17	3	37	5	22
W	2.6	1	2.0	10.0	4	1.5	23	4	2

of 21.7, 32.8 and 37.8 when compared to N, NSP2 and PL^{pro} genes, respectively, while differences among N, NSP2 and PL^{pro} were not significantly different ($p > 0.05$). The mean number of amino acids for each sequence is shown in Table 2.

The number of amino acids in the *G. gallus* proteins that presented the same codons used by at least one of the *A. coronavirus* genes in 100% conserved aa positions ranged from 1 (for vitamin D receptor) to 15 (for ovomucin alpha), and the most conserved preferred codon, found for all *A. coronavirus* genes, was UUU for F (Table 1). The positions of each of the conserved amino acids coded by preferred codons for *A. coronavirus* are also shown in Table 1.

3.6. Protein selection test

The sequences of N, NSP2 and PL^{pro} from all the strains in this study were found to be under purifying selection as the *p* values from Fisher's exact test were all above 0.05, with mean values of 0.99 for each gene and sd values of 0.06, 0.05 and 0.08, respectively. For S sequences, the mean *p* value was 0.97 (sd 0.13), but *p* values <0.05 were found between the groups of sequences FJ899690.1 Conn39528/FJ899689.1 Conn32062/FJ904716.1 Conn461996/FJ904717.1 Conn46197 and AY561711.1 M41/DQ834384.1 M41, indicating positive selection for these strains.

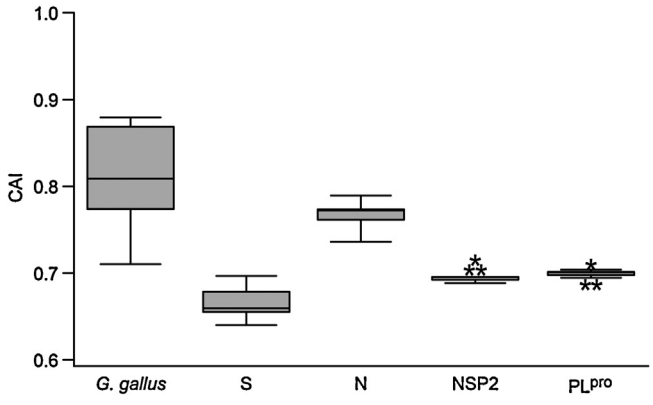

Fig. 3. Boxplot distribution for the codon adaptation index (CAI) for *Avian coronavirus* spike (S), nucleocapsid (N), non-structural protein 2 (NSP2) and papain-like protease (PL^{pro}) and *Gallus gallus* beta-actin, lung surfactant protein A, intestinal cholecystokinin, oviduct ovomucin alpha subunit and kidney vitamin D receptor (represented together in a single boxplot).

4. Discussion

Regardless of the gene being considered, all *A. coronavirus* sequences segregated in an exclusive cluster in the RSCU tree, which, despite being consistently separate from the *G. gallus* cluster, was closer to the SFTPA1 (a gene expressed in the respiratory tract of chicken) cluster. Taking the codon usage for these genes as a reflection of the codon usage in the respiratory tract, both structural and non-structural genes show a codon usage closer to the chicken respiratory tissue translational environment than to the reproductive, renal and enteric ones.

This similar codon usage could allow for an improved viral replication in the respiratory tract as a first site of viral replication, a feature common to all *A. coronavirus* strains in chickens, before the virus reaches other replication sites for each pathotype, as a result of the natural selection for codons and a more efficient translation of virus proteins, as already suggested for the S gene alone (Brandão, 2012).

Evidence of natural selection for codon usage as an evolutionary force acting upon *A. coronavirus* was found in the $Nc \times GC_{3\%}$ graphs (Fig. 2) because for all four viral genes, observed $GC_{3\%}$ points fell outside the curve, indicating that codon usage for all the strains under analysis was not the sole result of the random accumulation of mutations.

Nonetheless, the $Nc \times GC_{3\%}$ plots show that *A. coronavirus* codon usage could also be a consequence of mutation pressure, as the points were in the vicinity of the curve, meaning that the GC% at the synonymous 3rd codon position follows the viral genomic GC% to some degree.

It must be considered that both genetic drift derived from the mutation pressure and natural selection detected for *A. coronavirus* could also harbor some relationship with genomic RNA secondary structure constraints and not only codon usage, as synonymous 3rd base mutations, though synonymous in terms of amino acid codification, could result in altered RNA secondary structure

(Cardinale et al., 2013) and, consequently, impaired viral transcription, replication and assembly. As signals for RNA replication and genome packaging in coronaviruses are RNA secondary structure-dependent (Narayanan and Makino, 2007; Williams et al., 1999), such structures must be under intense evolutionary constraints that balance with codon usage evolution.

From the host side, mutation bias has also been shown to be the major driving force of *G. gallus* codon usage evolution, with minor participation of natural selection (Rao et al., 2011), in agreement with the results presented herein, suggesting a common evolutionary path for both virus and host.

A marked difference was noticed regarding the degree of codon usage bias for each *A. coronavirus* gene studied: for S, N and PL^{pro}, all mean values were just above 40, indicating a moderate bias (Gu et al., 2004), but for NSP2, the mean *Nc* (53.01) indicated a lower codon usage bias. These results provide evidence that *A. coronavirus* genes have taken different codon evolution pathways depending on the function that each protein possesses.

The function of NSP2 is still not clearly defined, but a role has been suggested as a co-factor for RNA synthesis (Graham et al., 2005), possibly in the early stages of virus replication. Despite the limited number of studies on NSP2 evolution, it can be speculated that a less biased codon usage for a protein involved in early stages of viral replication would allow for a less restricted tRNA preference and thus a more efficient start to the viral cycle.

The finding that the most biased gene was S (mean $Nc = 43$) might be linked to its relationship with the *G. gallus* immune system. The spike protein is the main target for neutralizing antibodies, and thus, theoretically, the more S protein that is expressed, the higher the generation of a humoral immune response against S and the lower cell infection by *A. coronavirus*.

Considering this stronger codon bias of S, the fact that S showed the lowest CAI value when compared to the other three genes and the fact that genes with lower CAIs are expressed less efficiently (Roth et al., 2012), a deoptimization of S expression could have been selected for with the advantage of lower S expression, providing further evidence that viral proteins that participate in host recognition might have a codon usage less similar to that presented by the host (Bahir et al., 2009).

Regarding CAI values for N, NSP2 and PL^{pro}, Fig. 3 suggests that the distributions were mostly above those for S, with the highest values for N (0.75–0.79). N protein plays a chief role in nucleocapsid assembly that is dependent on the association of positively charged amino acids with the genomic RNA of coronaviruses (Masters, 2006) and is thus under strong purifying selection, as shown herein by the Fisher's exact test on dS–dN values. Optimization of the codon usage in a manner closer to that of the host would endow *A. coronavirus* with a more efficient and accurate synthesis of the nucleocapsid protein.

The distribution of CAIs for NSP2 and PL^{pro} stayed between those for N and S (Fig. 3). Considering that PL^{pro} is a protease acting on the N-terminus domains of replicase polyproteins pp1a and pp1ab (Ziebuhr et al., 2000), an intermediate adaptation to the host's translational environment could have evolved as a balance between the conservation of structure of the enzymatic domain and the plasticity to follow amino acid mutations occurring on the PL^{pro} cleavage sites of diverse *A. coronavirus* types as compensatory mutations, showing that epistasis could also be detected at the codon usage evolution level.

It is noteworthy that none of the *A. coronavirus* strains showed no possible combinations of simultaneous occurrence maximum/minimum CAI or *Nc* (data not shown) for any of the four genes, meaning that CAI and *Nc* might be driven to different evolutionary pathways and that strains with a high CAI, i.e., highly adapted to the host's transcription environment, are not necessarily the ones with the lower bias, i.e., with higher *Nc*.

The distribution of 100% conserved amino acid positions coded by the preferred codon is noteworthy when one compares the S gene with N, NSP2 or PL^{pro}, as a single position was found in a region outside antigenic and hypervariable regions (Cavanagh et al., 1988; Kant et al., 1992) in the S gene, while for the other three genes, these positions ($n = 20$, 28 and 71, respectively) were scattered throughout the regions considered, with statistically significant differences when compared to S ($p < 0.0001$, OR = 21.7–37.8).

This low number of conserved amino acid positions coded by the preferred codon in S could be an additional molecular evolutionary mechanism for S antigenic diversity, as fine-tuning translation kinetics could result in high deoptimization of codon usage and a consequent increased fitness (Aragonés et al., 2010).

On the other hand, possibly due to strong structural and functional constraints, N, NSP2 and PL^{pro} have a higher number of amino acid positions coded by the preferred codon, which is the same codon preferred by the host (Table 1), which would allow higher fitness to the host transcription environment (Zhou et al., 2012) in a concerted virus–host molecular evolution.

Thus, taking conserved amino acid positions coded by the preferred codons as a selection unit, it follows from the above mentioned differences that natural selection could either be positive for these positions, leading a protein under purifying selection (e.g., N, NSP2 and PL^{pro}) to show the same codons as the host for that amino acid, or negative if a protein is under positive selection (as shown for S).

The most probable reason for the fact that non-100% conserved amino acid positions coded by a preferred codon for that amino acid (noted as NC in Table 1) were only found in *A. coronavirus* genes and not in the *G. gallus* genes is that host genes are less susceptible to both the occurrence of putative amino acids and codon usage polymorphisms, contrary to what is observed and expected for virus genes.

Nc might be considered to be an accurate indicator of codon usage bias because the frequency of amino acids is normalized during the analysis and does not add bias; however, similarly to the CAI, the outcome of the *Nc* analysis is a single number, leading to a loss of deep evolutionary information similar to the loss of evolutionary information in nucleotide or amino acid distance-based phylogenetic analyses.

Taking into account informative sites during codon evolution studies, for instance, 100% conserved amino acid positions coded by the preferred codon for that amino acid, could unveil data that would otherwise be lost in the analysis and that could be used to gain a more comprehensive understanding of molecular evolution in association with the codon usage bias indicators and selection analysis.

It would be interesting to use the analyses presented herein not only for a better understanding of virus evolution but also as supporting predictors of spill-over events, such as influenza (Wahlgren, 2011) and the new human coronavirus (Kindler et al., 2013) now named MERS-CoV, for which the role of codon usage evolution in virus adaptation to new hosts has been widely ignored.

In conclusion, *A. coronavirus* codon usage evolves independently for each gene in a manner predictable by the protein function. Proteins with high functional and structural constraints are more adapted to *G. gallus*, its natural host, with a balance between natural selection and mutation pressure, giving further molecular basis for the virus' ability to exploit the host's environment.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.virusres.2013.09.033.

References

Aragonés, L., Guix, S., Ribes, E., Bosch, A., Pintó, R.M., 2010. Fine-tuning translation kinetics selection as the driving force of codon usage bias in hepatitis A virus capsid. PLOS Pathogens 6, e1000797.

Bahir, I., Fromer, M., Yosef, P., Linial, M., 2009. Viral adaptation to host: a proteome-based analysis of codon usage and amino acid preferences. Molecular Systems Biology 5, 311.

Brandão, P.E., 2012. *Avian coronavirus* spike glycoprotein ectodomain shows a low codon adaptation to *Gallus gallus* with virus-exclusive codons in strategic amino acids positions. Journal of Molecular Evolution 75 (1–2), 19–24.

Cardinale, D.J., DeRosa, K., Duffy, S., 2013. Base composition and translational selection are insufficient to explain codon usage bias in plant viruses. Viruses 5 (1), 162–181.

Cavanagh, D., 2007. Coronavirus avian infectious bronchitis virus. Veterinary Research 38 (2), 281–297.

Cavanagh, D., Davis, P.J., Mockett, A.P., 1988. Amino acids within hypervariable region1 of *Avian coronavirus* IBV (Massachusetts serotype) spike glycoprotein are associated with neutralization epitopes. Virus Research 11 (2), 141–150.

Cook, J.K., Jackwood, M., Jones, R.C., 2012. The long view: 40 years of infectious bronchitis research. Avian Pathology 41 (3), 239–250.

Graham, R.L., Sims, A.C., Brockway, S.M., Baric, R.S., Denison, M.R., 2005. The nsp2 replicase proteins of Murine hepatitis virus and Severe acute respiratory syndrome coronavirus are dispensable for viral replication. Journal of Virology 79 (21), 13399–13411.

Gu, W., Zhou, T., Ma, J., Sun, X., Lu, Z., 2004. Analysis of synonymous codon usage in SARS Coronavirus and other viruses in the Nidovirales. Virus Research 101 (2), 155–161.

Hershberg, R., Petrov, D.A., 2008. Selection on codon bias. Annual Review of Genetics 42, 287–299.

Kant, A., Koch, G., van Roozelaar, D.J., Kusters, J.G., Poelwijk, F.A., van der Zeijst, B.A., 1992. Location of antigenic sites defined by neutralizing monoclonal antibodies on the S1 avian infectious bronchitis virus glycopolypeptide. Journal of General Virology 73 (Pt. 3), 591–596.

Kindler, E., Jónsdóttir, H.R., Muth, D., Hamming, O.J., Hartmann, R., Rodriguez, R., Geffers, R., Fouchie, R.A., Drosten, C., Müller, M.A., Dijkman, R., Thiel, V., 2013. Efficient replication of the novel Human Betacoronavirus EMC on primary human epithelium highlights its zoonotic potential. MBio 4 (1), pii: e00611-12.

Kuo, S.M., Kao, H.W., Hou, M.H., Wang, C.H., Lin, S.H., Su, H.L., 2013. Evolution of infectious bronchitis virus in Taiwan: positively selected sites in the nucleocapsid protein and their effects on RNA-binding activity. Veterinary Microbiology 162 (2–4), 408–418.

Masters, P., 2006. The molecular biology of coronavirus. Advances in Virus Research 66, 193–292.

Narayanan, K., Makino, S., 2007. Coronavirus genome packaging. In: Thiel, V. (Ed.), Coronaviruses: Molecular and Cellular Biology. Caister Academic Press, Norfolk, UK, pp. 131–142.

Nei, M., Gojobori, T., 1986. Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. Molecular Biology and Evolution 3 (5), 418–426.

Nei, M., Kumar, S., 2000. Molecular Evolution and Phylogenetics. Oxford University Press, New York.

Rao, Y., Wu, G., Wang, Z., Chai, X., Nie, Q., Zhang, X., 2011. Mutation bias is the driving force of codon usage in the *Gallus gallus* genome. DNA Research 18 (6), 499–512.

Roth, A., Anisimova, M., Cannarozzi, G.M., 2012. Measuring codon bias. In: Cannarozzi, G.M., Schneider, A. (Eds.), Codon Evolution. Oxford University Press, New York, pp. 189–217.

Sharp, P.M., Li, W.H., 1987. The codon Adaptation Index—a measure of directional synonymous codon usage bias, and its potential applications. Nucleic Acids Research 15 (3), 1281–1295.

Swofford, D.L., 2000. PAUP*, Phylogenetic analysis using parsimony (*and Other Methods). Version 4. Sinauer Associates, Sunderland.

Tamura, K., Peterson, D., Peterson, N., Stecher, G., Nei, M., Kumar, S., 2011. MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. Molecular Biology and Evolution 28 (10), 2731–2739.

Vetrivel, U., Arunkumar, V., Dorairaj, S., 2007. ACUA: a software tool for automated codon usage analysis. Bioinformation 2 (2), 62–63.

Wahlgren, J., 2011. Influenza A viruses: an ecology review. Infection Ecology and Epidemiology 1, http://dx.doi.org/10.3402/iee.v1i0.6004.

Wickramasinghe, I.N., de Vries, R.P., Gröne, A., de Haan, C.A., Verheije, M.H., 2011. Binding of *Avian coronavirus* spike proteins to host factors reflects virus tropism and pathogenicity. Journal of Virology 85 (17), 8903–8912.

Williams, G.D., Chang, R.Y., Brian, D.A., 1999. A phylogenetically conserved hairpin-type 3' untranslated region pseudoknot functions in coronavirus RNA replication. Journal of Virology 73 (10), 8349–8355.

Winter, C., Herrler, G., Neumann, U., 2008. Infection of the tracheal epithelium by infectious bronchitis virus is sialic acid dependent. Microbes and infection 10 (4), 367–373.

Woo, P.C., Lau, S.K., Lam, C.S., Lau, C.C., Tsang, A.K., Lau, J.H., Bai, R., Teng, J.L., Tsang, C.C., Wang, M., Zheng, B.J., Chan, K.H., Yuen, K.Y., 2012. Discovery of seven novel mammalian and avian coronaviruses in the genus deltacoronavirus supports bat coronaviruses as the gene source of alphacoronavirus and betacoronavirus and avian coronaviruses as the gene source of gammacoronavirus and deltacoronavirus. Journal of Virology 86 (7), 3995–4008.

Wright, F., 1990. The 'effective number of codons' used in a gene. Gene 87 (1), 23–29.

Yang, Z., Nielsen, R., 2008. Mutation-selection models of codon substitution and their use to estimate selective strengths on codon usage. Molecular Biology and Evolution 25 (3), 568–579.

Ziebuhr, J., Snijder, E.J., 2007. The coronavirus replicase gene: special enzymes for special viruses. In: Thiel, V. (Ed.), Coronaviruses: Molecular and Cellular Biology. Caister Academic Press, Norfolk, UK, pp. 33–63.

Ziebuhr, J., Snijder, E.J., Gorbalenya, A.E., 2000. Virus-encoded proteinases and proteolytic processing in the Nidovirales. Journal of General Virology 81 (Pt. 4), 853–879.

Zhou, Y., Chen, X., Ushijima, H., Frey, T.K., 2012. Analysis of base and codon usage by rubella virus. Archives of Virology 157 (5), 889–899.