Molecular Phylogenetics and Evolution 141 (2019) 106618

Contents lists available at ScienceDirect

Molecular Phylogenetics and Evolution

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

Comprehensive codon usage analysis of porcine deltacoronavirus

Wei He[a], Ningning Wang[a], Jimin Tan[a], Ruyi Wang[a], Yichen Yang[a], Gairu Li[a], Haifei Guan[a], Yuna Zheng[a], Xinze Shi[a], Rui Ye[a], Shuo Su[a,*], Jiyong Zhou[b,c,d,*]

[a] *MOE Joint International Research Laboratory of Animal Health and Food Safety, Jiangsu Engineering Laboratory of Animal Immunology, Institute of Immunology and College of Veterinary Medicine, Nanjing Agricultural University, Nanjing 210095, China*

[b] *MOA Key Laboratory of Animal Virology, Department of Veterinary Medicine and Veterinary Medical Research Center, Zhejiang University, Hangzhou 310058, China*

[c] *Institute of Preventive Veterinary Sciences, Zhejiang University, Hangzhou 310058, China*

[d] *Collaborative Innovation Center and State Key Laboratory for Diagnosis and Treatment of Infectious Diseases, First Affiliated Hospital, Zhejiang University, Hangzhou 310003, China*

ARTICLE INFO

Keywords:
PDCoV
Codon usage
Mutation pressure
Nature selection
Evolution

ABSTRACT

Porcine deltacoronavirus (PDCoV) is a newly identified coronavirus of pigs that was first reported in Hong Kong in 2012. Since then, many PDCoV isolates have been identified worldwide. In this study, we analyzed the codon usage pattern of the S gene using complete coding sequences and complete PDCoV genomes to gain a deeper understanding of their genetic relationships and evolutionary history. We found that during evolution three groups evolved with a relatively low codon usage bias (effective number of codons (ENC) of 52). The factors driving bias were complex. However, the primary element influencing the codon bias of PDCoVs was natural selection. Our results revealed that different natural environments may have a significant impact on the genetic characteristics of the strains. In the future, more epidemiological surveys are required to examine the factors that resulted in the emergence and outbreak of this virus.

1. Introduction

Coronaviruses (CoVs) are the causative agents of major diseases in a variety of avian and mammalian species including humans. CoVs belong to the subfamily *Orthocoronavirinae* of the *Coronaviridae*, order *Nidovirales*. The *Orthocoronavirinae* subfamily is further divided into four genera including, Alphacoronavirus, Betacoronavirus, Gammacoronavirus, and the recently identified Deltacoronavirus (Chan et al., 2013; King et al., 2018). To date, six CoVs have been reported in pigs: transmissible gastroenteritis virus (TGEV), porcine respiratory coronavirus (PRCV), swine enteric alphacoronavirus (SeACoV), porcine epidemic diarrhea virus (PEDV), porcine hemagglutinating encephalomyelitis virus (PHEV), and porcine deltacoronavirus (PDCoV) (Pan et al., 2017; Homwong et al., 2016). PDCoV was first recorded as an emerging enteropathogenic coronavirus in pigs in Hong Kong in 2012 (Chan et al., 2013; Woo et al., 2012), and thereafter was isolated from a swine farm in Ohio, USA in 2014 (Wang et al., 2014a). Since then, PDCoV has been reported in many countries and regions, including USA, Canada, South Korea, mainland China, Mexico, Japan, Thailand, Viet Nam, and Lao PDR (Lee and Lee, 2014; Suzuki et al., 2018; Saeng-Chuto et al., 2017; Wang et al., 2014b; Ajayi et al., 2018; Perez-Rivera et al., 2019). A previous study showed that the global PDCoVs consist of the China lineage, the USA/Japan/South Korea lineage, and the Viet Nam/Laos/Thailand lineage (Zhang et al., 2019). PDCoV is an enveloped, positive-sense, and single-stranded RNA virus with a genome size of approximately 25.4 kb. The genome includes a 5′UTR, ORF1a/1b, the spike (S), the envelope (E), the membrane (M), nonstructural protein 6 (NS6), the nucleocapsid (N), the nonstructural protein 7 (NS7), and a 3′UTR (Lee and Lee, 2014).

The codon usage pattern is an important indicator of genome evolution. Except for methionine and tryptophan, more than one codon can encode an amino acid due to the redundancy of the genetic code. Codons encoding the same amino acid also are known as synonymous codons. Interestingly, the codon usage is not random and some codons are used more than others, a phenomenon referred to as codon usage bias (Marin et al., 1989). Codon usage bias has been reported for some RNA viruses. However, the degree of bias varies depending on the identity of the specific virus. For instance, Rubella virus and Rotavirus show strong codon usage biases, whereas Equine infectious anemia virus (EIAV), Ebola virus (EBOV), the N gene of Rabies virus (RABV),

[*] Corresponding authors at: MOE Joint International Research Laboratory of Animal Health and Food Safety, Jiangsu Engineering Laboratory of Animal Immunology, Institute of Immunology and College of Veterinary Medicine, Nanjing Agricultural University, Nanjing 210095, China (S. Su). MOA Key Laboratory of Animal Virology, Department of Veterinary Medicine and Veterinary Medical Research Center, Zhejiang University, Hangzhou 310058, China (J. Zhou).

E-mail addresses: shuosu@njau.edu.cn (S. Su), jyzhou@zju.edu.cn (J. Zhou).

https://doi.org/10.1016/j.ympev.2019.106618
Received 30 December 2017; Received in revised form 5 September 2019; Accepted 11 September 2019
Available online 16 September 2019

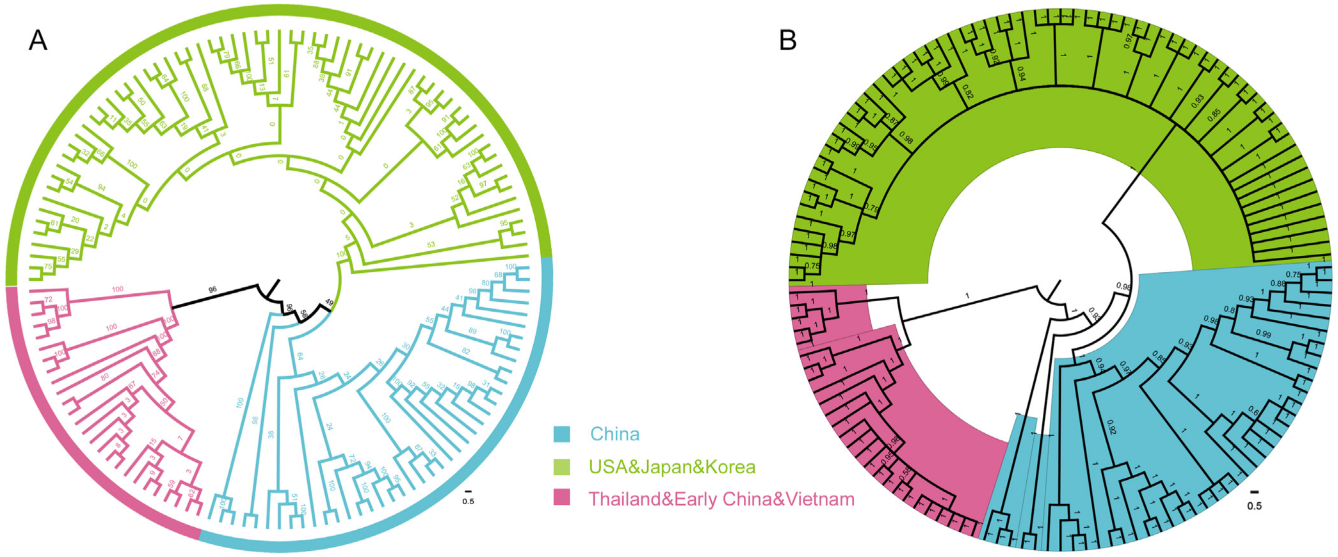

Fig. 1. (A) Maximum likelihood tree of the PDCoV S gene reconstructed by RAxML (v8.2.10). (B) Bayesian Inference tree of the PDCoV S gene reconstructed by MrBayes (v3.2.7). The China group, USA-Japan-Korea group, and Thailand-Early China-Vietnam group are represented in light blue, green, and pink, respectively.

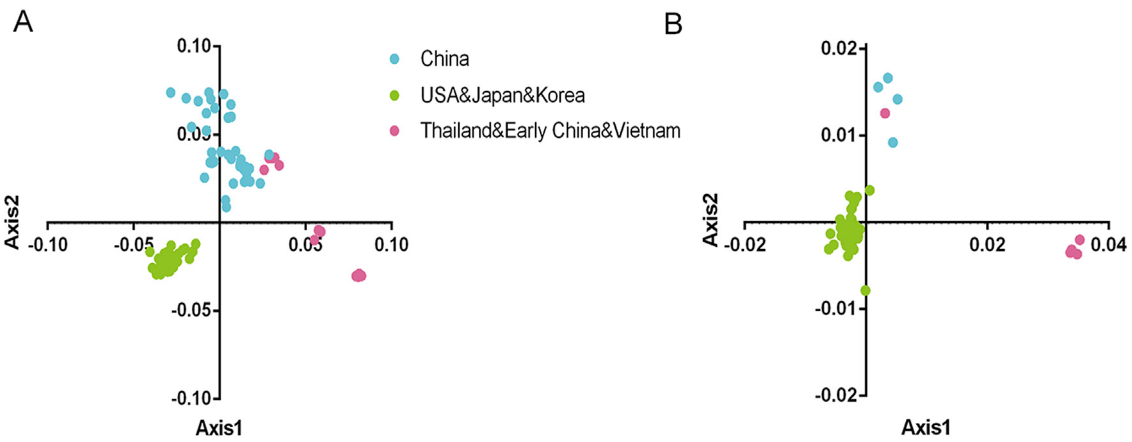

Fig. 2. Principal component analysis (PCA) of the PDCoV S gene (A) and complete coding genomes (B). The China, USA-Japan-Korea, and Thailand-Early China-Vietnam groups are represented in light blue, green, and pink, respectively.

and Porcine epidemic diarrhea virus (PEDV) show weak codon usage bias (Belalov and Lukashev, 2013; Yin et al., 2013; Cristina et al., 2015; Chen et al., 2014; He et al., 2017). Natural selection, mutation pressure, the abundance of tRNA, RNA structure, and gene length all contribute to the codon usage bias (Jenkins and Holmes, 2003; Parmley and Hurst, 2007; Hershberg and Petrov, 2008; Plotkin and Kudla, 2011). The virus and host can both influence codon usage, which likely affects the survival, evolution, fitness, and immune evasion of the virus from host defenses (Li et al., 2018b, 2019; He et al., 2019). Indeed, synonymous triplets are not used randomly, and factors such as natural selection and saltatorial bias can cause synonymous codon usage to diverge (Sharp and Li, 1986). Investigating the codon usage patterns of viruses could provide insights into their molecular evolution and viral gene expression regulation, assisting vaccine design, in which high levels of viral antigen expression are likely to be needed to produce immunity (Butt et al., 2014). Given the recent increase in PDCoV epidemics and the threat to pork production, in the present study, we reported an exhaustive genome-wide investigation of PDCoV codon usage and evaluated the possible influencing factors.

2. Materials and methods

2.1. Data analysis

We retrieved all PDCoVs sequences from the National Center for Biotechnology Information (NCBI) nucleotide database (http://www.ncbi.nlm.nih.gov) available up to April 2019. The detailed sequence information (serial number, strain name, accession number, location, and isolation year) for all 159 complete coding sequences of the S gene and 98 complete coding sequences (with the following concatenated order: ORF1ab-S-E-M-NS6-N-NS7) of PDCoV are displayed in supplementary materials (Table S1).

2.2. Recombination and phylogenetic group analysis

Potential recombination signals were detected using RDP4 (Recombination Detection Program version 4) (Martin et al., 2015) with default settings. Seven methods were chosen for the analysis, including RDP, GENECONV, Chimaera, MaxChi, BootScan, SiScan, and 3 Seq. In particular, four methods were firstly applied. Thereafter, the remaining sequences were run again with at least two methods until there was no recombination signal.

Phylogenetic trees were reconstructed in RAxML (v8.2.10)

Table 1
The nucleotide composition and properties of S gene of the PDCoV strains.

Strain	A%	U%	C%	G%	C% + G%	GC1s	GC2s	GC12s	GC3s	U3s	C3s	A3s	G3s	ENC
JQ065042	0.280	0.306	0.237	0.177	0.414	0.478	0.412	0.445	0.351	0.476	0.267	0.308	0.171	52.250
KP757891	0.279	0.305	0.239	0.177	0.415	0.480	0.413	0.446	0.354	0.471	0.270	0.309	0.172	52.410
KP757892	0.280	0.305	0.239	0.176	0.415	0.478	0.411	0.444	0.357	0.467	0.276	0.310	0.168	52.430
KR131621	0.280	0.307	0.237	0.176	0.413	0.476	0.412	0.444	0.352	0.471	0.270	0.311	0.169	52.300
KT336560	0.278	0.306	0.238	0.178	0.417	0.481	0.413	0.447	0.356	0.477	0.268	0.300	0.176	52.210
KU204694	0.279	0.306	0.237	0.178	0.415	0.480	0.413	0.447	0.351	0.472	0.267	0.311	0.170	52.060
KU204695	0.281	0.307	0.236	0.176	0.412	0.477	0.408	0.443	0.352	0.473	0.269	0.310	0.170	52.140
KU204696	0.279	0.308	0.235	0.178	0.413	0.480	0.408	0.444	0.351	0.475	0.268	0.310	0.169	52.050
KU204697	0.281	0.307	0.236	0.176	0.412	0.477	0.408	0.443	0.352	0.473	0.269	0.310	0.170	52.140
KU665558	0.280	0.305	0.238	0.177	0.415	0.478	0.413	0.445	0.353	0.472	0.269	0.309	0.172	52.420
KU981059	0.280	0.306	0.237	0.177	0.414	0.479	0.413	0.446	0.351	0.474	0.269	0.310	0.168	52.030
KX534090	0.278	0.307	0.238	0.177	0.415	0.480	0.414	0.447	0.352	0.472	0.268	0.310	0.169	52.100
KY065120	0.280	0.305	0.239	0.176	0.415	0.478	0.411	0.445	0.356	0.468	0.276	0.310	0.167	52.710
KY078905	0.280	0.307	0.237	0.176	0.413	0.476	0.413	0.444	0.351	0.474	0.267	0.310	0.171	52.100
KY078907	0.280	0.305	0.239	0.176	0.415	0.478	0.411	0.444	0.356	0.468	0.276	0.310	0.166	52.610
KY078909	0.280	0.305	0.239	0.176	0.415	0.479	0.410	0.445	0.356	0.468	0.278	0.308	0.163	52.360
KY078910	0.280	0.306	0.239	0.176	0.414	0.478	0.410	0.444	0.354	0.470	0.275	0.310	0.165	52.600
KY078911	0.280	0.306	0.239	0.176	0.414	0.478	0.411	0.444	0.354	0.470	0.275	0.310	0.165	52.580
KY078914	0.280	0.306	0.239	0.175	0.414	0.477	0.411	0.444	0.354	0.470	0.275	0.310	0.165	52.580
KY293677	0.279	0.305	0.238	0.178	0.416	0.481	0.413	0.447	0.353	0.471	0.270	0.310	0.169	52.400
KY293678	0.279	0.306	0.237	0.177	0.415	0.477	0.413	0.445	0.354	0.470	0.271	0.310	0.170	52.550
KY496312	0.280	0.306	0.237	0.177	0.414	0.477	0.411	0.444	0.353	0.470	0.270	0.311	0.171	52.380
KY513724	0.280	0.304	0.239	0.177	0.416	0.477	0.412	0.444	0.359	0.466	0.275	0.309	0.172	53.040
LC216914	0.279	0.306	0.237	0.178	0.415	0.480	0.412	0.446	0.353	0.473	0.269	0.308	0.171	52.130
MF037204	0.281	0.304	0.239	0.176	0.415	0.475	0.413	0.444	0.357	0.465	0.275	0.313	0.169	52.840
MF041982	0.279	0.305	0.239	0.177	0.416	0.480	0.413	0.446	0.357	0.470	0.274	0.306	0.170	52.350
MF280390	0.279	0.306	0.239	0.176	0.415	0.478	0.411	0.444	0.356	0.468	0.276	0.309	0.166	52.690
MF431742	0.278	0.307	0.237	0.178	0.415	0.479	0.413	0.446	0.354	0.477	0.267	0.301	0.174	52.150
MF431743	0.280	0.302	0.241	0.177	0.418	0.482	0.413	0.448	0.358	0.466	0.274	0.310	0.171	52.960
MF461406	0.281	0.307	0.236	0.176	0.413	0.476	0.408	0.442	0.353	0.470	0.269	0.312	0.172	51.910
MF461408	0.280	0.306	0.237	0.177	0.414	0.475	0.413	0.444	0.353	0.471	0.269	0.309	0.171	52.300
MF461409	0.280	0.307	0.237	0.177	0.414	0.475	0.413	0.444	0.353	0.471	0.269	0.309	0.171	52.350
MF948005	0.281	0.304	0.240	0.176	0.415	0.475	0.413	0.444	0.357	0.464	0.275	0.314	0.168	52.860
MG242062	0.279	0.304	0.239	0.177	0.416	0.479	0.413	0.446	0.356	0.468	0.272	0.310	0.172	52.880
MG832584	0.280	0.306	0.237	0.177	0.414	0.476	0.412	0.444	0.354	0.472	0.269	0.308	0.173	52.330
MH708123	0.279	0.306	0.237	0.177	0.415	0.479	0.411	0.445	0.354	0.472	0.269	0.308	0.173	52.560
MH708124	0.279	0.306	0.237	0.177	0.415	0.479	0.411	0.445	0.354	0.472	0.269	0.308	0.173	52.560
MH708125	0.279	0.306	0.237	0.177	0.415	0.479	0.411	0.445	0.354	0.472	0.269	0.308	0.173	52.560
MH715491	0.280	0.306	0.239	0.176	0.414	0.479	0.411	0.445	0.352	0.472	0.273	0.310	0.165	52.360
MK248485	0.280	0.306	0.237	0.177	0.414	0.478	0.411	0.444	0.354	0.472	0.269	0.308	0.173	52.510
NC_039208	0.277	0.304	0.240	0.178	0.418	0.481	0.413	0.447	0.361	0.469	0.275	0.301	0.175	52.360
LC260038	0.279	0.304	0.240	0.177	0.417	0.475	0.412	0.444	0.364	0.464	0.279	0.306	0.173	52.730
LC260039	0.279	0.303	0.241	0.177	0.418	0.477	0.412	0.444	0.365	0.462	0.281	0.305	0.173	52.810
LC260040	0.279	0.303	0.241	0.177	0.418	0.477	0.411	0.444	0.365	0.462	0.281	0.305	0.173	52.870
LC260041	0.279	0.303	0.240	0.177	0.417	0.476	0.412	0.444	0.364	0.464	0.279	0.305	0.173	52.710
LC260042	0.280	0.303	0.240	0.177	0.417	0.476	0.412	0.444	0.363	0.464	0.278	0.307	0.173	52.750
LC260043	0.280	0.304	0.240	0.176	0.416	0.475	0.412	0.444	0.362	0.465	0.279	0.307	0.171	52.520
LC260044	0.280	0.303	0.241	0.177	0.417	0.476	0.412	0.444	0.364	0.463	0.280	0.307	0.172	52.650
LC260045	0.280	0.301	0.242	0.177	0.419	0.477	0.413	0.445	0.367	0.457	0.285	0.308	0.171	52.700
KJ462462	0.280	0.303	0.241	0.176	0.417	0.477	0.412	0.444	0.363	0.464	0.280	0.307	0.171	52.660
KJ481931	0.280	0.303	0.240	0.177	0.417	0.477	0.412	0.444	0.362	0.465	0.277	0.307	0.173	52.680
KJ567050	0.279	0.304	0.240	0.177	0.417	0.477	0.411	0.444	0.363	0.465	0.279	0.305	0.173	52.520
KJ569769	0.279	0.303	0.241	0.177	0.418	0.477	0.414	0.445	0.364	0.462	0.280	0.307	0.171	52.750
KJ584355	0.279	0.304	0.240	0.177	0.417	0.476	0.412	0.444	0.362	0.466	0.277	0.305	0.173	52.690
KJ584356	0.280	0.303	0.240	0.177	0.417	0.475	0.411	0.443	0.364	0.464	0.280	0.306	0.172	52.620
KJ584357	0.279	0.303	0.240	0.177	0.417	0.476	0.413	0.444	0.363	0.464	0.278	0.306	0.173	52.760
KJ584358	0.279	0.303	0.240	0.177	0.417	0.476	0.412	0.444	0.364	0.464	0.279	0.305	0.173	52.690
KJ584359	0.279	0.304	0.240	0.177	0.417	0.476	0.412	0.444	0.363	0.465	0.278	0.305	0.173	52.720
KJ601777	0.279	0.304	0.240	0.177	0.417	0.476	0.412	0.444	0.362	0.465	0.278	0.306	0.172	52.720
KJ601778	0.279	0.304	0.240	0.177	0.417	0.476	0.412	0.444	0.362	0.465	0.278	0.306	0.172	52.780
KJ601779	0.280	0.303	0.240	0.177	0.417	0.477	0.412	0.444	0.362	0.465	0.277	0.307	0.173	52.680
KJ601780	0.279	0.303	0.241	0.177	0.418	0.477	0.412	0.445	0.365	0.463	0.280	0.305	0.173	52.880
KJ620016	0.280	0.304	0.240	0.177	0.417	0.476	0.412	0.444	0.362	0.465	0.278	0.307	0.172	52.600
KJ769231	0.280	0.304	0.240	0.176	0.416	0.476	0.411	0.444	0.361	0.466	0.278	0.307	0.171	52.550
KM012168	0.279	0.304	0.240	0.177	0.417	0.474	0.414	0.444	0.363	0.465	0.278	0.304	0.174	52.600
KP981395	0.279	0.304	0.240	0.177	0.417	0.474	0.414	0.444	0.363	0.465	0.278	0.304	0.174	52.600
KP995358	0.280	0.303	0.241	0.176	0.417	0.477	0.411	0.444	0.363	0.464	0.280	0.307	0.171	52.710
KR150443	0.280	0.304	0.239	0.177	0.416	0.475	0.412	0.444	0.361	0.466	0.277	0.305	0.173	52.640
KR265847	0.279	0.304	0.241	0.177	0.417	0.477	0.411	0.444	0.364	0.465	0.280	0.304	0.172	52.740
KR265848	0.279	0.304	0.240	0.177	0.416	0.476	0.410	0.443	0.363	0.465	0.280	0.305	0.172	52.740
KR265849	0.280	0.303	0.241	0.176	0.417	0.476	0.413	0.444	0.363	0.463	0.280	0.307	0.171	52.590
KR265850	0.280	0.303	0.241	0.176	0.417	0.476	0.413	0.444	0.363	0.463	0.280	0.307	0.171	52.590
KR265851	0.279	0.304	0.240	0.177	0.417	0.476	0.412	0.444	0.364	0.464	0.279	0.305	0.173	52.720
KR265852	0.279	0.304	0.240	0.177	0.417	0.476	0.411	0.444	0.363	0.465	0.278	0.305	0.173	52.710

(*continued on next page*)

Table 1 (*continued*)

Strain	A%	U%	C%	G%	C% + G%	GC1s	GC2s	GC12s	GC3s	U3s	C3s	A3s	G3s	ENC
KR265853	0.279	0.303	0.241	0.177	0.417	0.476	0.411	0.444	0.365	0.462	0.281	0.306	0.173	52.850
KR265854	0.279	0.304	0.240	0.177	0.417	0.475	0.412	0.444	0.364	0.463	0.280	0.306	0.172	52.790
KR265855	0.279	0.304	0.240	0.177	0.417	0.475	0.411	0.443	0.364	0.463	0.280	0.306	0.172	52.740
KR265856	0.279	0.303	0.240	0.177	0.417	0.476	0.411	0.444	0.365	0.463	0.280	0.305	0.173	52.820
KR265857	0.279	0.303	0.240	0.177	0.417	0.476	0.411	0.444	0.365	0.463	0.280	0.305	0.173	52.820
KR265858	0.280	0.304	0.240	0.177	0.417	0.476	0.411	0.444	0.363	0.465	0.279	0.305	0.172	52.580
KR265859	0.279	0.303	0.240	0.177	0.417	0.476	0.412	0.444	0.364	0.464	0.279	0.305	0.173	52.710
KR265860	0.279	0.304	0.240	0.177	0.417	0.476	0.412	0.444	0.363	0.465	0.278	0.305	0.173	52.720
KR265861	0.279	0.304	0.240	0.177	0.417	0.476	0.412	0.444	0.363	0.465	0.278	0.305	0.173	52.720
KR265862	0.279	0.303	0.241	0.177	0.417	0.477	0.413	0.445	0.363	0.464	0.279	0.307	0.171	52.760
KR265863	0.280	0.303	0.241	0.176	0.417	0.477	0.412	0.444	0.362	0.464	0.280	0.307	0.170	52.590
KR265864	0.280	0.303	0.240	0.176	0.416	0.475	0.410	0.443	0.364	0.464	0.280	0.305	0.172	52.680
KR265865	0.280	0.304	0.240	0.176	0.416	0.474	0.412	0.443	0.363	0.464	0.278	0.306	0.174	52.660
KT381613	0.280	0.304	0.240	0.176	0.417	0.476	0.412	0.444	0.362	0.465	0.279	0.307	0.171	52.530
KX022602	0.280	0.303	0.240	0.177	0.417	0.476	0.412	0.444	0.363	0.464	0.278	0.307	0.173	52.550
KX022603	0.280	0.303	0.240	0.177	0.417	0.477	0.412	0.444	0.363	0.464	0.278	0.307	0.173	52.550
KX022604	0.280	0.303	0.240	0.177	0.417	0.477	0.412	0.444	0.363	0.464	0.278	0.307	0.173	52.550
KX022605	0.280	0.304	0.240	0.177	0.417	0.477	0.412	0.444	0.361	0.466	0.276	0.307	0.173	52.470
MK478380	0.280	0.303	0.241	0.177	0.417	0.476	0.412	0.444	0.364	0.463	0.280	0.307	0.172	52.670
MK478381	0.279	0.304	0.241	0.176	0.417	0.477	0.411	0.444	0.362	0.464	0.280	0.307	0.169	52.640
MK478382	0.280	0.304	0.241	0.176	0.417	0.476	0.412	0.444	0.362	0.467	0.279	0.304	0.171	52.570
MK478383	0.279	0.303	0.241	0.176	0.417	0.476	0.413	0.444	0.364	0.464	0.281	0.305	0.170	52.690
KM820765	0.279	0.303	0.241	0.177	0.417	0.476	0.412	0.444	0.365	0.463	0.280	0.305	0.173	52.730
KR060082	0.279	0.303	0.241	0.177	0.418	0.477	0.412	0.445	0.365	0.462	0.281	0.305	0.173	52.900
KR060083	0.279	0.303	0.241	0.177	0.418	0.477	0.413	0.445	0.365	0.463	0.280	0.305	0.173	52.840
KX710201	0.280	0.305	0.239	0.176	0.415	0.474	0.411	0.443	0.360	0.465	0.277	0.308	0.171	52.670
KX710202	0.280	0.305	0.239	0.176	0.415	0.474	0.411	0.443	0.361	0.465	0.277	0.307	0.172	52.680
KY354363	0.280	0.305	0.239	0.176	0.415	0.474	0.411	0.443	0.360	0.465	0.277	0.308	0.171	52.670
KY354364	0.280	0.305	0.239	0.176	0.415	0.474	0.411	0.443	0.360	0.465	0.277	0.308	0.171	52.670
KY364365	0.279	0.304	0.240	0.177	0.417	0.474	0.413	0.444	0.363	0.465	0.278	0.304	0.174	52.670
KY926511	0.279	0.305	0.240	0.176	0.415	0.473	0.412	0.442	0.361	0.465	0.278	0.306	0.172	52.650
KY926512	0.279	0.306	0.239	0.176	0.415	0.473	0.413	0.443	0.360	0.466	0.277	0.306	0.171	52.540
KU051641	0.280	0.304	0.240	0.176	0.416	0.483	0.412	0.448	0.353	0.473	0.271	0.309	0.169	52.940
KU051642	0.280	0.304	0.240	0.176	0.416	0.483	0.412	0.448	0.353	0.473	0.271	0.309	0.169	52.940
KU051643	0.280	0.304	0.240	0.176	0.416	0.483	0.412	0.448	0.353	0.474	0.270	0.308	0.169	52.940
KU051644	0.281	0.303	0.240	0.176	0.416	0.483	0.412	0.448	0.353	0.473	0.271	0.309	0.169	52.920
KU051645	0.281	0.303	0.240	0.176	0.416	0.483	0.412	0.448	0.353	0.473	0.271	0.309	0.169	52.920
KU051646	0.281	0.303	0.240	0.176	0.416	0.483	0.412	0.448	0.353	0.473	0.271	0.309	0.169	52.920
KU051647	0.281	0.303	0.240	0.176	0.416	0.483	0.412	0.448	0.353	0.473	0.271	0.309	0.169	52.920
KU051648	0.281	0.303	0.240	0.176	0.416	0.483	0.412	0.448	0.353	0.473	0.271	0.309	0.169	52.920
KU051649	0.281	0.303	0.240	0.176	0.416	0.483	0.412	0.448	0.353	0.473	0.271	0.309	0.169	52.920
KU051650	0.280	0.304	0.240	0.176	0.416	0.483	0.412	0.448	0.353	0.473	0.271	0.309	0.169	52.940
KU051651	0.280	0.304	0.240	0.176	0.416	0.483	0.412	0.448	0.354	0.473	0.271	0.308	0.169	52.900
KU051652	0.280	0.304	0.240	0.176	0.416	0.483	0.412	0.448	0.353	0.473	0.271	0.309	0.169	52.940
KU051653	0.280	0.303	0.240	0.176	0.416	0.483	0.412	0.448	0.354	0.472	0.272	0.309	0.169	53.000
KU051654	0.280	0.303	0.240	0.176	0.416	0.483	0.412	0.448	0.354	0.472	0.272	0.309	0.169	53.000
KU051655	0.280	0.305	0.239	0.176	0.415	0.483	0.411	0.447	0.352	0.475	0.270	0.307	0.169	52.940
KU051656	0.280	0.304	0.239	0.176	0.416	0.483	0.412	0.448	0.352	0.474	0.270	0.309	0.169	52.920
KU984334	0.280	0.302	0.241	0.177	0.417	0.484	0.413	0.448	0.356	0.469	0.274	0.309	0.168	53.140
KX118627	0.281	0.303	0.240	0.177	0.416	0.482	0.412	0.447	0.354	0.469	0.273	0.311	0.168	53.210
KX834351	0.279	0.308	0.237	0.177	0.413	0.483	0.411	0.447	0.346	0.485	0.264	0.303	0.167	52.270
KX834352	0.279	0.308	0.237	0.176	0.413	0.482	0.411	0.447	0.346	0.484	0.265	0.304	0.166	52.320
KY078906	0.279	0.308	0.236	0.177	0.413	0.484	0.411	0.447	0.343	0.486	0.261	0.306	0.167	52.080
MF642324	0.280	0.306	0.238	0.177	0.414	0.476	0.412	0.444	0.355	0.472	0.273	0.306	0.169	52.780
MF642323	0.279	0.306	0.238	0.176	0.415	0.475	0.413	0.444	0.356	0.472	0.274	0.305	0.169	52.770
MF642322	0.279	0.306	0.238	0.177	0.415	0.477	0.414	0.445	0.355	0.472	0.273	0.306	0.169	52.830
MF642325	0.278	0.307	0.237	0.178	0.415	0.478	0.412	0.445	0.355	0.472	0.272	0.305	0.171	52.830
KP757890	0.279	0.305	0.239	0.177	0.416	0.479	0.414	0.447	0.354	0.475	0.270	0.306	0.170	52.460
Average	0.280	0.304	0.239	0.177	0.416	0.478	0.412	0.445	0.358	0.468	0.275	0.307	0.171	52.630
SD	0.001	0.001	0.001	0.001	0.001	0.003	0.001	0.002	0.005	0.005	0.005	0.002	0.002	0.253

(Stamatakis, 2014) and MrBayes (v3.2.7) (Ronquist et al., 2012) using non recombinant sequences. The GTR + Gamma substitution model was used to reconstruct the ML tree with a total of 1000 bootstraps. For the Bayesian inference (BI) tree, 1000,000 generations were run, with the first 25% of burn in. The final trees were displayed in Figtree (v1.4.4) (http://tree.bio.ed.ac.uk/software/figtree/).

2.3. Principal component analysis (PCA)

To study the relationship between the multivariate and sample, a multidimensional statistical method, PCA, was applied. PCA is mainly a mathematical transformation process that converts the relevant variables (dependent on the relative synonymous codon usage (RSCU) values) into a smaller number of irrelevant variables (called the principal components). Every coding sequence was split into a 59-dimensional vector, and each dimension represented the matching dedication of the RSCU values of 59 different synonymous codons, which included only a specific amino group, without AUG, UGG and the three stop codons. The parameters used for the PCA were calculated in program Codon W (http://codonw.sourceforge.net/).

Table 2
The relative synonymous codon usage (RSCU) of the S gene and complete genomes of PDCoV strains. The numbers in bold denote the eighteen abundant codons of three genotypic groups and all sequences.

	China (S)	Thailand &Early China &Vietnam (S)	USA &Japan & Korea (S)	All(S)	China (complete genome)	Thailand &Early China & Vietnam (complete genome)	USA &Japan & Korea (complete genome)	All(complete genome)
UUU(F)	**1.29**	**1.21**	**1.24**	**1.25**	**1.22**	**1.21**	**1.23**	**1.22**
UUC(F)	0.71	0.79	0.76	0.75	0.78	0.79	0.77	0.78
UUA(L)	1.04	0.84	1.04	1	0.77	0.76	0.79	0.79
UUG(L)	0.73	0.71	0.74	0.73	0.86	0.84	0.87	0.87
CUU(L)	**1.83**	**1.89**	**1.79**	**1.82**	**1.99**	**1.96**	**1.96**	**1.96**
CUC(L)	1.04	1.03	1.04	1.04	1.03	1.02	1.05	1.04
CUA(L)	0.79	0.83	0.8	0.8	0.65	0.7	0.64	0.65
CUG(L)	0.57	0.69	0.6	0.61	0.69	0.72	0.69	0.7
AUU(I)	**1.53**	**1.51**	**1.51**	**1.52**	**1.54**	**1.53**	**1.53**	**1.53**
AUC(I)	0.64	0.63	0.66	0.65	0.73	0.74	0.75	0.74
AUA(I)	0.83	0.87	0.82	0.83	0.73	0.73	0.73	0.73
GUU(V)	**2.13**	**2.13**	**2.15**	**2.14**	**1.79**	**1.79**	**1.77**	**1.77**
GUC(V)	0.67	0.72	0.65	0.67	0.72	0.73	0.73	0.73
GUA(V)	0.77	0.79	0.71	0.74	0.77	0.76	0.76	0.76
GUG(V)	0.43	0.35	0.5	0.45	0.72	0.72	0.74	0.74
UCU(S)	**2.11**	**1.94**	**1.95**	**1.99**	**1.95**	**1.91**	**1.91**	**1.91**
UCC(S)	0.87	0.97	1.09	1	0.71	0.73	0.75	0.75
UCA(S)	1.25	1.34	1.29	1.29	1.49	1.57	1.48	1.49
UCG(S)	0.29	0.3	0.26	0.28	0.36	0.32	0.38	0.37
AGU(S)	0.78	0.77	0.71	0.74	0.98	0.95	0.97	0.97
AGC(S)	0.69	0.68	0.71	0.7	0.51	0.52	0.52	0.52
CCU(P)	**1.89**	**1.9**	**1.95**	**1.92**	**1.59**	**1.61**	**1.62**	**1.61**
CCC(P)	0.67	0.84	0.69	0.71	0.65	0.64	0.63	0.63
CCA(P)	1.04	0.88	0.98	0.98	1.43	1.42	1.45	1.44
CCG(P)	0.4	0.38	0.39	0.39	0.33	0.33	0.31	0.31
ACU(T)	**1.89**	**2.03**	**1.95**	**1.95**	**1.68**	**1.75**	**1.69**	**1.7**
ACC(T)	0.96	0.84	0.9	0.91	1	0.93	0.98	0.97
ACA(T)	0.98	0.98	0.96	0.97	1.05	1.05	1.05	1.05
ACG(T)	0.17	0.14	0.2	0.18	0.27	0.28	0.28	0.28
GCU(A)	1.38	1.47	1.39	1.4	**1.72**	**1.75**	**1.73**	**1.73**
GCC(A)	0.76	0.74	0.77	0.76	0.67	0.66	0.66	0.66
GCA(A)	**1.76**	**1.67**	**1.79**	**1.76**	1.3	1.28	1.3	1.3
GCG(A)	0.09	0.12	0.05	0.08	0.31	0.31	0.31	0.31
UAU(Y)	**1.12**	**1.18**	0.99	**1.07**	**1.07**	**1.08**	**1.06**	**1.06**
UAC(Y)	0.88	0.82	**1.01**	0.93	0.93	0.92	0.94	0.94
CAU(H)	0.95	0.97	0.96	0.96	**1.15**	**1.19**	**1.16**	**1.16**
CAC(H)	**1.05**	**1.03**	**1.04**	**1.04**	0.85	0.81	0.84	0.84
CAA(Q)	0.97	**1.02**	0.94	0.96	0.98	**1.01**	0.96	0.96
CAG(Q)	**1.03**	0.98	**1.06**	**1.04**	**1.02**	0.99	**1.04**	**1.04**
AAU(N)	**1.15**	**1.16**	**1.19**	**1.17**	**1.08**	**1.06**	**1.08**	**1.08**
AAC(N)	0.85	0.84	0.81	0.83	0.92	0.94	0.92	0.92
AAA(K)	**1.12**	**1.16**	**1.1**	**1.12**	0.93	0.96	0.94	0.94
AAG(K)	0.88	0.84	0.9	0.88	**1.07**	**1.04**	**1.06**	**1.06**
GAU(D)	**1.19**	**1.23**	**1.16**	**1.18**	**1.1**	**1.14**	**1.1**	**1.1**
GAC(D)	0.81	0.77	0.84	0.82	0.9	0.86	0.9	0.9
GAA(E)	**1.03**	**1.06**	**1.03**	**1.04**	0.96	0.93	0.96	0.96
GAG(E)	0.97	0.94	0.97	0.96	**1.04**	**1.07**	**1.04**	**1.04**
UGU(C)	**1.35**	**1.3**	**1.28**	**1.31**	**1.14**	**1.15**	**1.12**	**1.12**
UGC(C)	0.65	0.7	0.72	0.69	0.86	0.85	0.88	0.88
CGU(R)	1.36	1.14	1.26	1.26	**1.72**	**1.69**	**1.75**	**1.75**
CGC(R)	0.52	0.73	0.56	0.58	1.14	1.18	1.13	1.13
CGA(R)	0.43	0.41	0.43	0.43	0.5	0.49	0.48	0.48
CGG(R)	0.87	0.81	0.83	0.84	0.53	0.51	0.53	0.53
AGA(R)	**1.96**	**1.83**	**1.96**	**1.93**	1.36	1.34	1.36	1.36
AGG(R)	0.86	1.08	0.97	0.96	0.76	0.79	0.75	0.76
GGU(G)	1.64	**1.71**	1.64	1.65	**1.87**	**1.88**	**1.88**	**1.88**
GGC(G)	**1.65**	1.57	**1.71**	**1.66**	1.03	1.03	1.03	1.03
GGA(G)	0.55	0.59	0.59	0.58	0.87	0.85	0.86	0.86
GGG(G)	0.16	0.13	0.07	0.11	0.23	0.24	0.23	0.23

2.4. Compositional and principal parameters analysis

The compositional characteristics of the PDCoV coding sequences of the S gene and complete genomes, were calculated. The frequency of all nucleotides (GC%, AU%, A%, U%, G% and C%) was estimated using BioEdit (http://www.softpedia.com/get/Science-CAD/BioEdit.shtml). The A, C, G, and U frequencies in synonymous codons at different sites (GC1%, GC2%, GC3%, GC12%, A3%, U3%, G3%, C3%, AU3%) of each sequence were computed using CUSP (http://emboss.toulouse.inra.fr/cgi-bin/emboss/cusp) and Codon W (http://codonw.sourceforge.net/).

2.5. Relative dinucleotide abundance analysis

The relative dinucleotides abundances were computed according to a previously reported method (Karlin and Burge, 1995). The odds ratio of the ability of the observed frequencies of the 16 dinucleotides was computed using the equation below:

$$P_{xy} = \frac{f_{xy}}{f_y f_x}$$

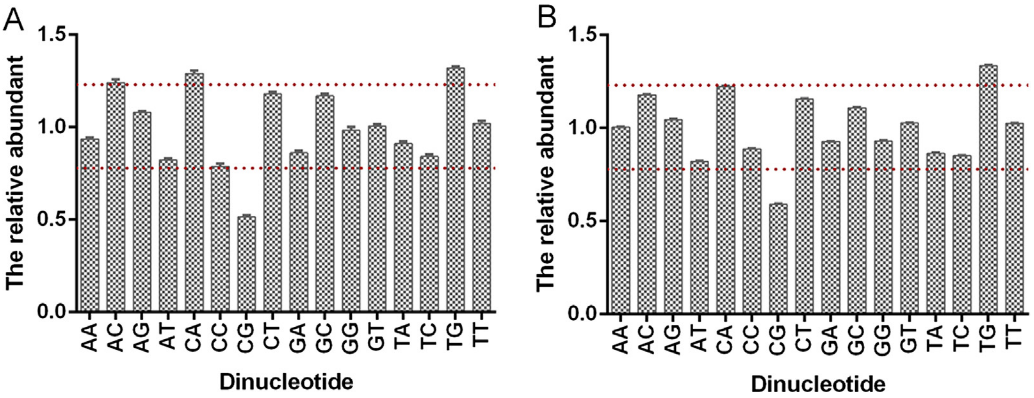

Fig. 3. Dinucleotide abundancy of the PDCoV S gene (A) and the complete coding genomes (B).

Table 3
Relative dinucleotides frequencies among different groups of S gene and complete genomes of PDCoV strains.

	China (S)	Thailand &Early China &Vietnam (S)	USA &Japan & Korea (S)	All (S)	China (complete genome)	Thailand &Early China &Vietnam (complete genome)	USA &Japan & Korea (complete genome)	All (complete genome)
AA	0.938 ± 0.009	0.947 ± 0.01	0.931 ± 0.004	0.936 ± 0.009	1.004 ± 0.002	1.01 ± 0.003	1.004 ± 0.001	1.005 ± 0.002
AC	1.246 ± 0.012	1.212 ± 0.011	1.244 ± 0.011	1.241 ± 0.018	1.18 ± 0.004	1.175 ± 0.004	1.178 ± 0.003	1.178 ± 0.003
AG	1.0815 ± 0.007	1.073 ± 0.007	1.077 ± 0.007	1.079 ± 0.006	1.048 ± 0.002	1.043 ± 0.003	1.046 ± 0.001	1.046 ± 0.002
AU	0.815 ± 0.007	0.838 ± 0.006	0.822 ± 0.009	0.821 ± 0.011	0.819 ± 0.003	0.821 ± 0.005	0.821 ± 0.002	0.821 ± 0.003
CA	1.292 ± 0.011	1.274 ± 0.02	1.293 ± 0.011	1.291 ± 0.016	1.219 ± 0.006	1.213 ± 0.008	1.223 ± 0.003	1.222 ± 0.005
CC	0.784 ± 0.015	0.806 ± 0.011	0.792 ± 0.006	0.788 ± 0.015	0.891 ± 0.005	0.888 ± 0.008	0.887 ± 0.003	0.888 ± 0.004
CG	0.513 ± 0.011	0.506 ± 0.013	0.514 ± 0.004	0.514 ± 0.011	0.591 ± 0.005	0.59 ± 0.004	0.59 ± 0.002	0.59 ± 0.003
CU	1.184 ± 0.016	1.188 ± 0.009	1.176 ± 0.01	1.18 ± 0.011	1.155 ± 0.002	1.162 ± 0.005	1.154 ± 0.002	1.155 ± 0.003
GA	0.853 ± 0.01	0.855 ± 0.01	0.867 ± 0.008	0.861 ± 0.011	0.928 ± 0.001	0.925 ± 0.002	0.928 ± 0.001	0.928 ± 0.002
GC	1.171 ± 0.013	1.177 ± 0.019	1.166 ± 0.005	1.17 ± 0.012	1.106 ± 0.004	1.115 ± 0.001	1.108 ± 0.002	1.108 ± 0.002
GG	0.983 ± 0.018	0.996 ± 0.033	0.976 ± 0.014	0.981 ± 0.019	0.93 ± 0.003	0.926 ± 0.012	0.929 ± 0.002	0.929 ± 0.005
GU	1.012 ± 0.009	0.997 ± 0.01	1.003 ± 0.006	1.005 ± 0.01	1.028 ± 0.003	1.026 ± 0.002	1.028 ± 0.001	1.028 ± 0.002
UA	0.915 ± 0.009	0.919 ± 0.011	0.908 ± 0.006	0.912 ± 0.011	0.867 ± 0.005	0.868 ± 0.007	0.864 ± 0.002	0.864 ± 0.003
UC	0.841 ± 0.015	0.852 ± 0.012	0.838 ± 0.007	0.842 ± 0.012	0.849 ± 0.001	0.85 ± 0.003	0.852 ± 0.002	0.852 ± 0.002
UG	1.316 ± 0.011	1.322 ± 0.01	1.326 ± 0.006	1.321 ± 0.008	1.333 ± 0.004	1.34 ± 0.003	1.335 ± 0.002	1.336 ± 0.002
UU	1.02 ± 0.015	1.004 ± 0.021	1.028 ± 0.006	1.02 ± 0.014	1.027 ± 0.003	1.02 ± 0.004	1.025 ± 0.002	1.025 ± 0.002

where the frequency of nucleotide X is represented by f_x, the frequency of nucleotide Y is represented by f_y, the expected frequency of the dinucleotide XY is represented by $f_y f_x$, and the frequency of the dinucleotide XY is represented by f_{xy}. As an universal standard, for < 0.78 or xy > 1.23, we considered that the XY pair was under-represented or over-represented respectively, compared with the random association of single nucleotides and according to its relative abundance (Butt et al., 2016).

2.6. Relative synonymous codon usage (RSCU)

RSCU refers to the relative probability of a specific synonymous codon, which indicates whether the codon usage is influenced by the amino acid composition. In the case where all synonymous codons of a particular amino acid are assumed to be used equally, the RSCU value of a sequence is the ratio of the frequency at which the codon is actually observed at its expected frequency (Chen and Chen, 2014). The RSCU is calculated as:

$$\mathrm{RSCU} = \frac{g_{ij}}{\sum_{j}^{ni} g_{ij}} \mathrm{ni}$$

where g_{ij} is the derived value of the ith codon for the jth amino acid with n_i kinds of synonymous codons. RSCU values = 1.0, > 1.0, and < 1.0, represent no bias, positive codon usage bias, and negative codon usage bias, respectively. The RSCU was calculated using MEGA7 (https://www.megasoftware.net/).

2.7. Effective number of codons (ENC) analysis

The degree of codon usage bias, measured by the ENC, was estimated taking into account the number of amino acids and the gene length. The ENC values vary between 20 and 61, with values closer to 20 indicating a high codon usage bias and values closer to 61 indicating a low codon usage bias. The ENC value can reflect the preference of a synonymous codon in a family of codons. Highly expressed genes often show a high codon usage bias, whereas poorly expressed genes contain more rare codons and thus a lower codon usage bias. Generally, the codon usage is considered to show strong bias when the ENC value is less than or equal to 45 (Comeron and Aguade, 1998). We used the following equation to calculate the ENC (Fuglsang, 2006):

$$\mathrm{ENC} = 2 + \frac{9}{\bar{F}_2} + \frac{1}{\bar{F}_3} + \frac{5}{\bar{F}_4} + \frac{3}{\bar{F}_6}$$

where the average value of F_i (i = 2, 3, 4, 6) for the i-fold degenerate amino acids is represented by F. The following equation was used to calculate F_i values:

$$\bar{F}_i = \frac{n\sum_{j=1}^{i}\left(\frac{n_j}{n}\right)^2 - 1}{n-1}$$

where the total number of appearances of the codons for that amino acid is represented by n and the total number of appearances of the j_{th} codon for that amino acid is represented by n_j.

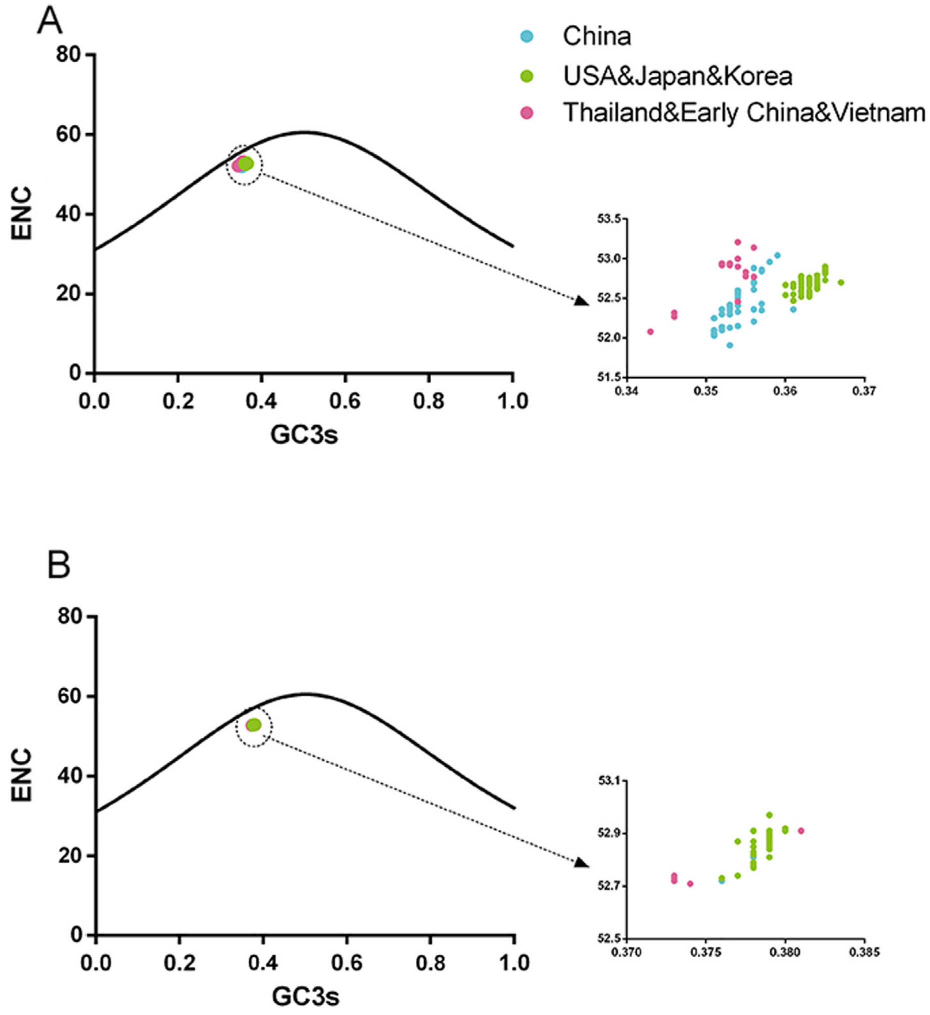

Fig. 4. ENC-plot analysis (GC3s plotted against ENC) of the PDCoV S gene (A) and complete coding genomes (B). The China, USA-Japan-Korea, and Thailand-Early China-Vietnam groups are represented in light blue, green, and pink, respectively.

2.8. ENC-plot analysis

ENC-plot analysis is commonly used to determine the factors influencing the codon usage bias (i.e. mutation pressure). The ENC values relative to the GC3 values (the frequency of guanine or cytosine at the third codon position of synonymous codons excluding Met, Trp and stop codons) were plotted (Karlin and Burge, 1995). When the codon usage is limited only to the GC3 mutation, the expected ENC value falls on a theoretical curve (the functional relationship between the ENC expectation curve and the GC3 value). When the actual ENC-plot values of these sequences are lower than the standard curve, it is suggestive of natural selection playing a role in driving codon usage bias (Fuglsang, 2008). The theoretical ENC values in ENC-plot analysis were calculated as follows.

$$\mathrm{ENC}_{\mathrm{expected}} = 2 + \mathrm{s} + \frac{29}{\mathrm{s}^2 + (1-\mathrm{s})^2}$$

where s denotes the frequency of C or G at the synonymous codons third position (i.e. GC3).

2.9. Neutrality plot analysis

Neutrality analysis or neutrality evolution analysis was carried out to compare and define the effect of natural selection and mutation pressure on the PDCoV codon usage patterns by comparing the value of GC12s of synonymous codons with the GC3s value using diagonal analysis. In the graph, the plot regression coefficient is considered as the mutation selection balance coefficient, and the evolutionary rates caused by natural selection pressure and mutation pressure are represented by the slope of the regression line. If all points are distributed along the diagonal and there is no significant difference in the three codon positions, this indicates that there is only weak or no external selection pressure. However, if the regression curve is parallel or tilted to the horizontal axis, this would indicate that the correlation between the changes of GC12 and GC3 is very low. Thus, the regression curve shows that the effect of natural selection evolution effectively balances the degree of neutrality (Kumar et al., 2016).

2.10. Parity rule 2 (PR2) analysis

PR2 analysis was used to investigate the effect of selection and mutation pressure on gene codon usage. PR2 is a gene map with AU deviation [A3/ (A3 + U3)] as the ordinate and GC deviation [G3/ (G3 + C3)] as the abscissa. At the center of the graph, the values of the two coordinates are 0.5, which means that G = C and A = U (PR2), and there is no deviation between the mutation effect and selectivity (substitution rate) (Sueoka, 1996).

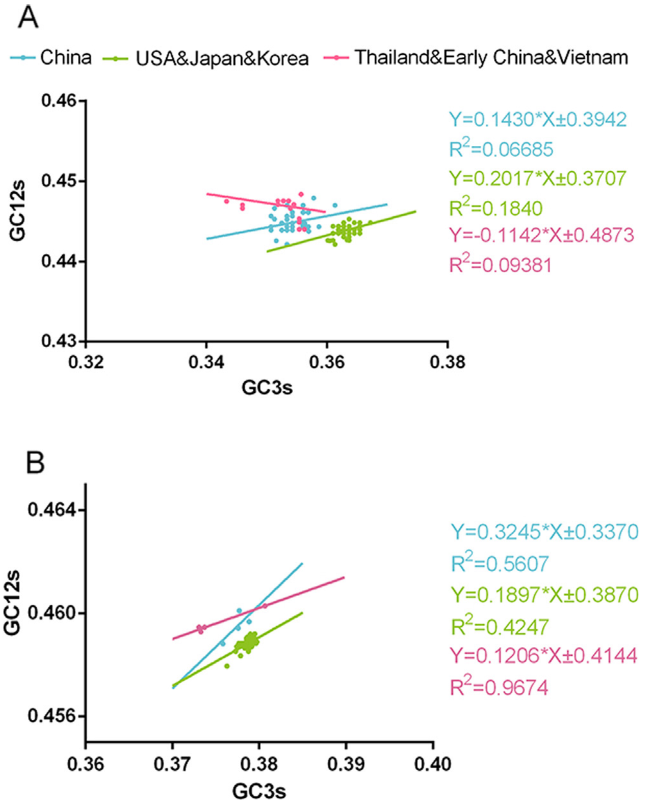

Fig. 5. Neutrality analysis (GC12 against GC3) of the PDCoV S gene (A) and complete coding genomes (B). The China, USA-Japan-Korea and Thailand-Early China-Vietnam groups are represented in light blue, green, and pink, respectively.

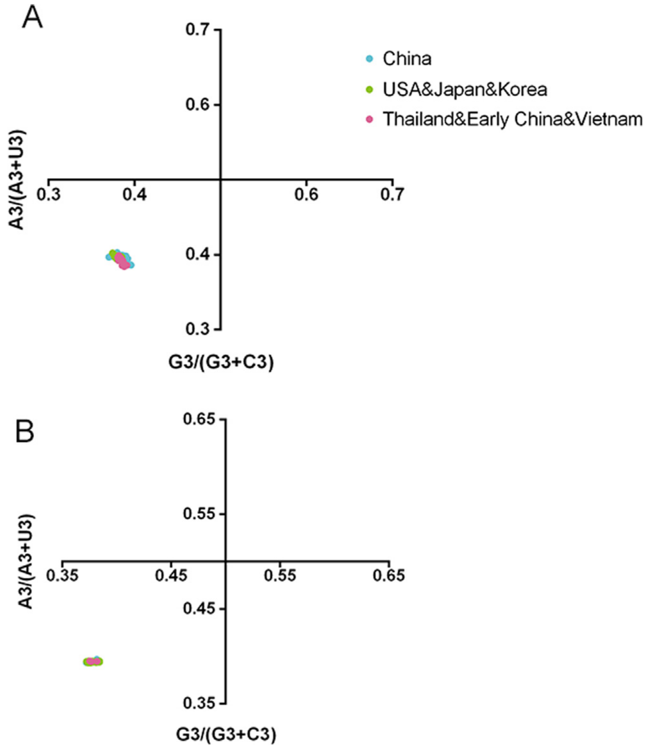

Fig. 6. Parity Rule 2 (PR2)-bias plot [A3/(A3 + U3) against G3/(G3 + C3)]. The PR2 bias plot was calculated for the S gene (A) and complete coding genomes (B). The China, USA-Japan-Korea, and Thailand-Early China-Vietnam groups are represented in light blue, green, and pink, respectively.

3. Results

3.1. Recombination and phylogenetic analysis

After removal of recombinant sequences, 132 S gene and 64 complete genomes were left for further analysis. Phylogenetic analysis of S gene based on ML (Fig. 1A) and BI (Fig. 1B) trees revealed three individual PDCoV groups including, China, USA-Japan-Korea, and Thailand-Early-China-Vietnam groups. We then used these three groups to investigate into codon usage and associations.

3.2. Principle component analysis (PCA)

PCA showed that the three groups clustered separately, especially the USA-Japan-Korea group, although several overlaps existed between the USA-Japan-Korea and Thailand-Early China-Vietnam groups (Fig. 2). For whole genomes, the three groups clustered separately too, except for several overlaps between the USA-Japan-Korea and the Thailand-Early China-Vietnam groups.

3.3. Nucleotide composition of PDCoV S gene and complete genomes

The nucleotide U was the most abundant in the S gene, followed by A, C and G, regardless of the individual phylogenetic group (Table 1). The detailed information of the nucleotide composition is shown in Table S2. The nucleotide composition of synonymous codons at the third position of (A3, C3, G3, U3) showed that the frequencies of U3 and A3 were higher than C3 and G3. The percentage content of AU and GC were indicative of AU-rich component in the coding sequences of PDCoV. Analysis of the synonymous codons at the first, second and third position showed that the values of GC1 were the highest, followed by GC2 and GC3 (Table S2). The same pattern was identified for whole genomes. Overall, these results illustrated that a relatively large part of the PDCoV coding sequence comprises A and U nucleotides.

3.4. PDCoV relative synonymous codon usage

All of the PDCoV 18 optimal synonymous codons for the corresponding amino acids of the S gene ended with U (Perez-Rivera et al., 2019) (Table 2). A total of 7 of the 18 priority codons had RSCU values greater than 1.6 (CUU (L), GUU (V), UCU (S), CCU (P), ACU (T), AGA (R), and GGC (G)). However, the remaining codons had RSCU values less than 1.6, with no underrepresented codons observed within the preferred codons. For whole genomes, U-ended codons were also the preferred codons among the 18 most abundant synonymous codons (Table 2). The RSCU analyses and the nucleotide composition revealed that the compositional constraints (the nucleotides U in this case) had the most influence on the selection of the preferred codons.

3.5. Factors driving dinucleotide frequency abundance

The relative abundances of the 16 dinucleotides of PDCoV coding sequences were calculated. We found that dinucleotides were not present randomly. None of the dinucleotide relative abundance values corresponded to the theoretical frequency (i.e., 1.0) (Fig. 3, Table 3). Furthermore, in the S gene, CpA (1.29 ± 0.0016) and UpG (1.32 ± 0.008) showed different degrees (marginal or peripheral) of overrepresentation. Only CpG (0.514 ± 0.011) was underrepresented. For whole genomes, the overrepresented and underrepresented dinucleotides were UpG (1.34 ± 0.002) and CpG (0.59 ± 0.003), respectively.

3.6. ENC analysis

ENC values were estimated to evaluate the extent of codon usage deviation within coding sequences of different PDCoV isolates. This analysis showed that PDCoV coding sequences were relatively conserved and stable in terms of the S coding sequences or whole genomes with a low codon usage bias. The ENC values of the S coding sequences ranged from 52.71 to 52.97, with an average of 52.853 (ENC > 40) (Table 1). The ENC values of complete genome coding sequences were also within the range of the S gene, with no obvious difference in relation to phylogenetic groups.

3.7. Influence of mutation pressure on the PDCoV codon usage pattern

ENC-plot analysis was carried out to reveal the constraint of mutation pressure on the PDCoV codon usage pattern. The values of GC3 were plotted against the ENC values according to individual phylogenetic group. We found that all points regardless of group concentrated on the left side and near to the expected curve for the S gene (Fig. 4A). For whole genome coding sequences, all the points were also under but close to the standard curve (Fig. 4B).

3.8. Influence of natural selection on the PDCoV codon usage pattern

Here, neutrality analysis or diagonal analysis was used, between the GC3s and GC12s values, to judge the effects of natural selection and mutation pressure (Fig. 5). In the S gene, the relationships between GC3s and GC12s were calculated based on the three phylogenetic groups. The correlation coefficient in the USA-Japan-Korea group, China group, and Thailand-Early China-Vietnam group were the 0.2017 ± 0.3707, 0.143 ± 0.3942, and 0.1142 ± 0.4873, respectively. Thus, the percentages of constrain of natural selection were 79.83%, 85.7%, and 88.58% for the S gene (Fig. 5A). For whole genomes, GC12s and GC3s significantly correlated, with a correlation coefficient of 0.1897 ± 0.387 according to the USA-Japan-Korea group, indicating an 81.03% limit for natural selection or 18.97% of GC3 relative binding (100% neutral or 0% constraint) (Fig. 5B). Overall, the above results indicate that the effect of mutation pressure is in all codon positions, but natural selection plays a major role driving the codon usage bias of PDCoV. Considering the limited number of sequences in the China and Thailand-Early China-Vietnam groups, they were excluded from the results.

In addition, PR2 analysis was carried out (Fig. 6). We found that the A ≠ U, C ≠ G, for both the S gene and whole genomes, which indicates the inequivalent role of mutation pressure and natural selection in shaping the codon usage of PDCoV.

4. Discussion

PDCoV is an emerging coronavirus that infects the whole of the small intestine, especially the jejunum and ileum, causing severe enteritis, diarrhea, and vomiting in piglets. PDCoV was first discovered in Hong Kong, China in 2012 (Woo et al., 2012). At the beginning of 2014, PDCoV was first reported in the USA, after which at least 17 USA states confirmed its presence as of December 2014. In recent years, China, South Korea, Thailand, and other Asian countries have suffered from recurrent outbreaks (Lorsirigool et al., 2016; Janetanakit et al., 2016; Dong et al., 2015; Lee et al., 2016). Phylogenetic analysis is well studied to demonstrate the evolution of virus (He et al., 2018; Li et al., 2018a; Su et al., 2017, 2016) Here, we first analyzed the codon usage patterns of the S coding sequences, as well as whole genome coding sequences of PDCoVs isolated from around the world to determine the factors driving codon usage, and provided a comprehensive understanding of the characteristics and evolution of PDCoV whole coding genes.

Phylogenetic analysis of the S gene revealed that sequences clustered into three different groups similarly to a previous study (Zhang et al., 2019), but with more accuracy since more methods were applied and recombinant sequences were excluded. Additionally, PCA analysis also indicated three potential evolutionary groups.

Based on the S coding gene and complete coding genomes, we found a significant preference for A and U nucleotides, rather than G and C. The contents of AU and GC were not equal and were more inclined towards the usage of AU nucleotides. If the use of a synonymous codon was affected only by mutation pressure, the frequency of U and A nucleotides in the third codon position should be equal to the frequency of G and C (van Hemert et al., 2016). Thus, we can conclude that there was a low bias in the usage of nucleotides in all PDCoV strains. RSCU analysis revealed that PDCoV genomes have a tendency towards U-ending codons. In addition, the relative probability distribution of 16 dinucleotides showed that codons and dinucleotides were used unequal and followed certain rules. Dinucleotide abundance influences the codon usage bias in certain organisms, including RNA and DNA viruses (Rothberg and Wimmer, 1981). Dinucleotide sequences may be derived from odd partial of amino acid changes or codon usage bias; therefore, we analyzed dinucleotide composition distribution (Plotkin et al., 2004; Cristina et al., 2015). The translational selection pressure on a dinucleotide is the entropy cost of a given set of constraints that alter the number of dinucleotide occurrences, in this case the amino acid sequence of the given protein sequence and the cost of the codon usage bias (Cristina et al., 2015). Analyses of the frequencies of codons and dinucleotides revealed that translation selection also played a part in the codon usage of PDCoVs. These initial observations prompted further investigation to assess the extent of codon usage bias using ENC analyses. For PDCoV, the ENC value based on the S gene or complete coding genomes was 52, indicative of slight bias and that different PDCoVs are relatively conserved and stable. Previous studies indicated that ENC values correlate negatively with gene expression (van Hemert and Berkhout, 2016). Thus, a higher ENC value indicates lower gene expression and lower codon preference. A low codon bias could be explained by the need to better adapt towards efficient replication and survival in the host, and to reduce the energy required for virus biosynthesis while avoiding competition with host protein synthesis (van Hemert et al., 2016). When the ENC and GC3 values of PDCoVs were plotted, mutation pressure was revealed as a moderate factor influencing the PDCoV codon usage pattern. According to previous reports, both natural selection and mutation pressure can affect the ENC value, which indicates that the relative contribution of selection and mutation on the codon usage pattern are not robust (Chen et al., 2014; Gu et al., 2004). It is worth mentioning that the codon usage bias of species with A/U biased genomes is different from that of genomes with a G/C bias. Therefore, simple ENC-GC3 map analysis might be misleading. Generally, mutation pressure will always have a role in driving the codon usage of viruses. Here, using neutrality plots we found that natural selection was a more dominant factor compared with mutation pressure (Shi et al., 2013). Natural selection can lead to weak codon usage bias while the virus is trying to adapt to the host cells (Matsumoto et al., 2016). PR2 bias plot analysis showed that both natural selection and mutation pressure contributed to the observed codon bias consistent with the neutrality analysis.

In summary, we found that the codon usage of the S gene was similar to the complete coding genome. To open new perspectives, a further exploration of the function and features of functional genes is worth studying.

5. Conclusion

Here, we found that, to a large extent, the codon usage pattern and the sequences characteristics of PDCoVs were restricted by evolutionary processes. Briefly, PDCoV has a low codon usage bias, which was affected by natural selection, mutation pressure, and dinucleotide abundancy. The primary element affecting the PDCoV codon usage pattern was natural selection. Additionally, the results of PCA and phylogenetic analysis were highly consistent suggesting that the codon usage pattern study can reveal the evolutionary clustering relationship between strains based on their genetic composition. This study suggests that monitoring the updated sequences of this novel, emerging virus would provide clues to better understand viral evolution and the disease.

Acknowledgments

This paper was supported in part by the National Key Research and Development Program of China [Grant No. 2017YFD0500101]; National Natural Science Foundation of China (Grant No. 31802195), the Natural Science Foundation of Jiangsu Province (Grant No. BK20170721), the China Association for Science and Technology Youth Talent Lift Project (2017-2019), the Bioinformatics Center of Nanjing Agricultural University and the Priority Academic Program Development of Jiangsu Higher Education Institutions.

Declaration of Competing Interest

The authors declare no competing financial interest.

Appendix A. Supplementary material

Supplementary data to this article can be found online at https://doi.org/10.1016/j.ympev.2019.106618.

References

Ajayi, T., Dara, R., Misener, M., Pasma, T., Moser, L., Poljak, Z., 2018. Herd-level prevalence and incidence of porcine epidemic diarrhoea virus (PEDV) and porcine deltacoronavirus (PDCoV) in swine herds in Ontario, Canada. Transbound. Emerg. Dis. 65 (5), 1197–1207.

Belalov, I.S., Lukashev, A.N., 2013. Causes and implications of codon usage bias in RNA viruses. Plos One 8 (2), e56642.

Butt, A.M., Nasrullah, I., Tong, Y., 2014. Genome-wide analysis of codon usage and influencing factors in chikungunya viruses. Plos One 9 (3), e90905.

Butt, A.M., Nasrullah, I., Qamar, R., Tong, Y.G., 2016. Evolution of codon usage in Zika virus genomes is host and vector specific. Emerg. Microbes Infec. 5 (Oct).

Chan, J.F., To, K.K., Tse, H., Jin, D.Y., Yuen, K.Y., 2013. Interspecies transmission and emergence of novel viruses: lessons from bats and birds. Trends Microbiol. 21 (10), 544–555.

Chen, Y.H., Chen, Y.F., 2014. Extensive homologous recombination in classical swine fever virus: A re-evaluation of homologous recombination events in the strain AF407339. Saudi J. Biol. Sci. 21 (4), 311–316.

Chen, Y., Shi, Y.Z., Deng, H.J., Gu, T., Xu, J., Ou, J.X., et al., 2014. Characterization of the porcine epidemic diarrhea virus codon usage bias. Infect. Genet. Evol. 28 (Dec), 95–100.

Comeron, J.M., Aguade, M., 1998. An evaluation of measures of synonymous codon usage bias. J. Mol. Evol. 47 (3), 268–274.

Cristina, J., Moreno, P., Moratorio, G., Musto, H., 2015. Genome-wide analysis of codon usage bias in Ebolavirus. Virus Res. 196 (Jan), 87–93.

Dong, N., Fang, L.R., Zeng, S.L., Sun, Q.Q., Chen, H.C., Xiao, S.B., 2015. Porcine Deltacoronavirus in Mainland China. Emerg. Infect. Dis. 21 (12), 2254–2255.

Fuglsang, A., 2006. Estimating the "Effective number of codons": The Wright way of determining codon homozygosity leads to superior estimates. Genetics 172 (2), 1301–1307.

Fuglsang, A., 2008. Impact of bias discrepancy and amino acid usage on estimates of the effective number of codons used in a gene, and a test for selection on codon usage. Gene 410 (1), 82–88.

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

He, W., Zhang, H., Zhang, Y., Wang, R., Lu, S., Ji, Y., et al., 2017. Codon usage bias in the N gene of rabies virus. Infect. Genet. Evol. 54 (Oct), 458–465.

He, W., Auclert, L.Z., Zhai, X., Wong, G., Zhang, C., Zhu, H., et al., 2018. Interspecies transmission, genetic diversity, and evolutionary dynamics of pseudorabies virus. J. Infect. Dis. 219 (11), 1705–1715.

He, W., Zhao, J., Xing, G., Li, G., Wang, R., Wang, Z., et al., 2019. Genetic analysis and evolutionary changes of Porcine circovirus 2. Mol. Phylogenet. Evol. 139, 106520 2019/10/01/.

Hershberg, R., Petrov, D.A., 2008. Selection on Codon Bias. Annu. Rev. Genet. 42, 287–299.

Homwong, N., Jarvis, M.C., Lam, H.C., Diaz, A., Rovira, A., Nelson, M., et al., 2016. Characterization and evolution of porcine deltacoronavirus in the United States. Prev Vet Med. 123 (Jan), 168–174.

Janetanakit, T., Lumyai, M., Bunpapong, N., Boonyapisitsopa, S., Chaiyawong, S., Nonthabenjawan, N., et al., 2016. Porcine Deltacoronavirus, Thailand, 2015. Emerg. Infect. Dis. 22 (4), 757–759.

Jenkins, G.M., Holmes, E.C., 2003. The extent of codon usage bias in human RNA viruses and its evolutionary origin. Virus Res. 92 (1), 1–7.

Karlin, S., Burge, C., 1995. Dinucleotide relative abundance extremes - a genomic signature. Trends Genet. 11 (7), 283–290.

King, A.M.Q., Lefkowitz, E.J., Mushegian, A.R., Adams, M.J., Dutilh, B.E., Gorbalenya, A.E., et al., 2018. Changes to taxonomy and the International Code of Virus Classification and Nomenclature ratified by the International Committee on Taxonomy of Viruses (2018). Arch. Virol. 163 (9), 2601–2631.

Kumar, N., Bera, B.C., Greenbaum, B.D., Bhatia, S., Sood, R., Selvaraj, P., et al., 2016. Revelation of influencing factors in overall codon usage bias of equine influenza viruses. PloS one 11 (4).

Lee, J.H., Chung, H.C., Nguyen, V.G., Moon, H.J., Kim, H.K., Park, S.J., et al., 2016. Detection and Phylogenetic Analysis of Porcine Deltacoronavirus in Korean Swine Farms, 2015. Transbound. Emerg. Dis. 63 (3), 248–252 Jun.

Lee, S., Lee, C., 2014. Complete genome characterization of Korean porcine deltacoronavirus strain KOR/KNU14-04/2014. Genome Announc. 2 (6).

Li, G., He, W., Zhu, H., Bi, Y., Wang, R., Xing, G., et al., 2018a. Origin, genetic diversity, and evolutionary dynamics of novel porcine circovirus 3. Adv. Sci. 5 (9), 1800275.

Li, G., Wang, H., Wang, S., Xing, G., Zhang, C., Zhang, W., et al., 2018b. Insights into the genetic and host adaptability of emerging porcine circovirus 3. Virulence 9 (1), 1301–1313 2018/12/31.

Li, G., Zhang, W., Wang, R., Xing, G., Wang, S., Ji, X., et al., 2019. Genetic analysis and evolutionary changes of the torque teno sus virus. Int. J. Mol. Sci. 20 (12).

Lorsirigool, A., Saeng-chuto, K., Temeeyasen, G., Madapong, A., Tripipat, T., Wegner, M., et al., 2016. The first detection and full-length genome sequence of porcine deltacoronavirus isolated in Lao PDR. Arch. Virol. 161 (10), 2909–2911.

Marin, A., Bertranpetit, J., Oliver, J.L., Medina, J.R., 1989. Variation in G + C-content and codon choice: differences among synonymous codon groups in vertebrate genes. Nucl. Acids Res. 17 (15), 6181–6189.

Martin, D.P., Murrell, B., Golden, M., Khoosal, A., Muhire, B., 2015. RDP4: detection and analysis of recombination patterns in virus genomes. Virus Evol. 1 (1) vev003-vev.

Matsumoto, T., John, A., Baeza-Centurion, P., Li, B.Y., Akashi, H., 2016. Codon usage selection can bias estimation of the fraction of adaptive amino acid fixations. Mol. Biol. Evol. 33 (6), 1580–1589.

Pan, Y., Tian, X., Qin, P., Wang, B., Zhao, P., Yang, Y.L., et al., 2017. Discovery of a novel swine enteric alphacoronavirus (SeACoV) in southern China. Vet. Microbiol. 211 (Nov), 15–21.

Parmley, J.L., Hurst, L.D., 2007. How do synonymous mutations affect fitness? BioEssays: News Rev. Mol., Cell. Develop. Biol. 29 (6), 515–519.

Perez-Rivera, C., Ramirez-Mendoza, H., Mendoza-Elvira, S., Segura-Velazquez, R., Sanchez-Betancourt, J.I., 2019. First report and phylogenetic analysis of porcine deltacoronavirus in Mexico. Transbound. Emerg. Dis. 66 (4), 1436–1441.

Plotkin, J.B., Kudla, G., 2011. Synonymous but not the same: the causes and consequences of codon bias. Nat. Rev. Genet. 12 (1), 32–42.

Plotkin, J.B., Robins, H., Levine, A.J., 2004. Tissue-specific codon usage and the expression of human genes. P. Natl. Acad. Sci. USA 101 (34), 12588–12591.

Ronquist, F., Teslenko, M., van der Mark, P., Ayres, D.L., Darling, A., Höhna, S., et al., 2012. MrBayes 3.2: efficient bayesian phylogenetic inference and model choice across a large model space. System. Biol. 61 (3), 539–542.

Rothberg, P.G., Wimmer, E., 1981. Mononucleotide and dinucleotide frequencies, and codon usage in poliovirion RNA. Nucl. Acids Res. 9 (23), 6221–6229.

Saeng-Chuto, K., Lorsirigool, A., Temeeyasen, G., Vui, D.T., Stott, C.J., Madapong, A., et al., 2017. Different lineage of porcine deltacoronavirus in Thailand, Vietnam and Lao PDR in 2015. Transbound Emerg Dis. 64 (1), 3–10.

Sharp, P.M., Li, W.H., 1986. An evolutionary perspective on synonymous codon usage in unicellular organisms. J. Mol. Evol. 24 (1–2), 28–38.

Shi, S.L., Jiang, Y.R., Liu, Y.Q., Xia, R.X., Qin, L., 2013. Selective pressure dominates the synonymous codon usage in parvoviridae. Virus Genes 46 (1), 10–19.

Stamatakis, A., 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics (Oxford, England) 30 (9), 1312–1313.

Su, S., Wong, G., Shi, W., Liu, J., Lai, A.C.K., Zhou, J., et al., 2016. Epidemiology, genetic recombination, and pathogenesis of coronaviruses. Trends Microbiol. 24 (6), 490–502 2016/06/01/.

Su, S., Gu, M., Liu, D., Cui, J., Gao, G.F., Zhou, J., et al., 2017. Epidemiology, evolution,and pathogenesis of H7N9 influenza viruses in five epidemic waves since 2013 in China. Trends Microbiol. 25 (9), 713–728 2017/09/01/.

Sueoka, N., 1996. Intrastrand parity rules of DNA base composition and usage biases of synonymous codons (vol 40, pg 318, 1995). J. Mol. Evol. 42 (2), 323.

Suzuki, T., Shibahara, T., Imai, N., Yamamoto, T., Ohashi, S., 2018. Genetic characterization and pathogenicity of Japanese porcine deltacoronavirus. Infect. Genet. Evol. 61 (Jul), 176–182.

van Hemert, F., Berkhout, B., 2016. Nucleotide composition of the Zika virus RNA genome and its codon usage. Virol. J. (Jun), 13.

van Hemert, F., van der Kuyl, A.C., Berkhout, B., 2016. Impact of the biased nucleotide composition of viral RNA genomes on RNA structure and codon usage. J. Gen. Virol. 97 (Oct), 2608–2619.

Wang, L., Byrum, B., Zhang, Y., 2014a. Detection and genetic characterization of deltacoronavirus in pigs, Ohio, USA, 2014. Emerg. Infect. Dis. 20 (7), 1227–1230 Jul.

Wang, L., Byrum, B., Zhang, Y., 2014b. Porcine coronavirus HKU15 detected in 9 US States, 2014. Emerg. Infect. Dis. 20 (9), 1594–1595.

Woo, P.C., Lau, S.K., Lam, C.S., Lau, C.C., Tsang, A.K., Lau, J.H., et al., 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. J. Virol. 86 (7), 3995–4008.

Yin, X., Lin, Y., Cai, W., Wei, P., Wang, X., 2013. Comprehensive analysis of the overall codon usage patterns in equine infectious anemia virus. Virol. J. 10, 356.

Zhang, Y., Cheng, Y., Xing, G., Yu, J., Liao, A., Du, L., et al., 2019. Detection and spike gene characterization in porcine deltacoronavirus in China during 2016–2018. Infect. Genet. Evol. 73 (Apr), 151–158.