수의학 박사학위논문

Genetic analysis of human coronavirus (SARS-CoV-2) and porcine coronavirus (PEDV) and their genetic mutations having potential to affect viral antigenicity and diagnosis

사람 코로나바이러스 (SARS-CoV-2)와 돼지 코로나바이러스 (PEDV)의 유전학적 분석과 유전적 변이가 바이러스 항원성과 진단에 영향을 미칠 가능성

2021 년 2 월

서울대학교 대학원

수의학과 수의미생물학 전공

김 성 재

Genetic analysis of human coronavirus (SARS-CoV-2) and porcine coronavirus (PEDV) and their genetic mutations having potential to affect viral antigenicity and diagnosis

지도교수 박 용 호

이 논문을 수의학 박사학위논문으로 제출함

2020 년 10 월

서울대학교 대학원

수의학과 수의미생물학 전공

김 성 재

김성재의 박사학위논문을 인준함

2020 년 12 월

위 원 장	최 강 석	(인)
부위원장	박 용 호	(인)
위 원	박 봉 균	(인)
위 원	한 정 희	(인)
위 원	박 건 택	(인)

Genetic analysis of human coronavirus (SARS-CoV-2) and porcine coronavirus (PEDV) and their genetic mutations having potential to affect viral antigenicity and diagnosis

By

Kim, Sung Jae

February, 2021

Department of Veterinary Medicine

The Graduate School of

Seoul National University

Genetic analysis of human coronavirus (SARS-CoV-2) and porcine coronavirus (PEDV) and their genetic mutations having potential to affect viral antigenicity and diagnosis

By

Kim, Sung Jae

Supervisor: Prof. Park, Yong Ho, D.V.M., M.Sc., Ph.D.

A dissertation submitted to the faculty of the Graduate School of Seoul National University in partial fulfillment of the requirements for the degree of Doctor of Philosophy in Veterinary Microbiology

February, 2021

Department of Veterinary Medicine
The Graduate School of
Seoul National University

Genetic analysis of human coronavirus (SARS-CoV-2) and porcine coronavirus (PEDV) and their genetic mutations having potential to affect viral antigenicity and diagnosis

By

Kim, Sung Jae

(Supervised by Prof. Park, Yong Ho)

Veterinary Microbiology, Department of Veterinary Medicine, the Graduate School of Seoul National University

Abstract

Aspect of virus evolution, viruses have continued to fight with host immune through genetic mutation facilitating immune evasion and this strategy of viruses for their survival will continue in the future. Genetic mutation, which may change structural form of viral proteins by non-synonymous changes, may alter antigenicity of a viral epitope and in turn can lead to decreased efficacy of previously developed vaccine for the virus. Furthermore, such genetic mutation can hamper diagnostic accuracy of polymerase chain reaction (PCR) and enzyme-linked immunosorbent assay (ELISA),

which are routinely used diagnostic techniques. Thus, it is very important work to investigate and track critical genetic events along with virus evolution. These efforts can give worthy information and insight to establish appropriate prevent and diagnostic strategy for viruses. Herein, human and animal coronaviruses causing sever disease were investigated by genetic and phylogenetic analysis.

As stated in chapter I, The S glycoprotein of coronaviruses is important for viral entry and pathogenesis with most variable sequences. Therefore, we analyzed the S gene sequences of SARS-CoV-2 to better understand the antigenicity and immunogenicity of this virus in this study. In phylogenetic analysis, two subtypes (SARS-CoV-2a and -b) were confirmed within SARS-CoV-2 strains. These two subtypes were divided by a novel non-synonymous mutation of D614G. This may play a crucial role in the evolution of SARS-CoV-2 to evade the host immune system. The region containing this mutation point was confirmed as a B-cell epitope located in the S1 domain, and SARS-CoV-2b strains exhibited severe reduced antigenic indexes compared to SARS-CoV-2a in this area. This may allow these two subtypes to have different antigenicity. If the two subtypes have different serological characteristics, a vaccine for both subtypes will be more effective to prevent COVID-19. Thus, further study is urgently required to confirm the antigenicity of these two subtypes.

As stated in chapter I, Porcine epidemic diarrhea virus *(*PEDV*)* causes continuous, significant damage to the swine industry worldwide. By RT-PCR-based methods, this study demonstrated the ongoing presence of PEDV in pigs of all ages in Korea at the average detection rate of 9.92%. By the application of Bayesian phylogenetic analysis,

it was found that the nucleocapsid (N) gene of PEDV could evolve at similar rates to the spike (S) gene at the order of 10^{-4} substitutions/site/year. Based on branching patterns of PEDV strains, three main N gene-based genogroups (N1, N2, and N3) and two sub-genogroups (N3a, N3b) were proposed in this study. By analyzing the antigenic index, possible antigenic differences also emerged in both the spike and nucleocapsid proteins between the three genogroups. The antigenic indexes of genogroup N3 strains were significantly lower compared with those of genogroups N1 and N2 strains in the B-cell epitope of the nucleocapsid protein. Indeed, there is different antigenicity between the genogroups based on the N gene, it may affect diagnostic results using commercial ELISA kits based on N1 protein. Similarly, significantly lower antigenic indexes in some parts of the B-cell epitope sequences of the spike protein (COE, S1D, and 2C10) were also identified. PEDV mutants derived from genetic mutations of the S and N genes may cause severe damage to swine farms by evading established host immunities.

In conclusion, the crucial genetic variations, which may induce immune evasion or diagnostic error, were revealed in these coronavirus. It is expected that these results provide better understanding for preventing viral infection and more precise diagnosis. Also, constant surveillances through genetic analysis should maintain to appropriately respond to coronavirus evolution in the future.

Keywords: coronavirus, genetic mutation, phylogenetic analysis, antigenicity, diagnosis

Student number: 2016-30968

Contents

List of Figures

List of Tables

General introduction

Virus are submicroscopic infectious particles, of which range are mostly in size from 5 to 300 nanometers. These parasitic agents can't reproduce by themselves, but only replicates inside the living cells. Viruses can infect all types of living organisms from mammal and plants to microorganisms including bacteria and archaea [1]. Viruses are ubiquitously found in almost ecosystem on Earth [2]. When they infect to susceptible cells, the infected cells are forced to rapidly produce numerous numbers of identical copies of the original virus. When they are not inside an infected cell or in the process of infecting a cell, viruses exist in the form of independent particles, or virions, consisting of genetic molecules and several structural proteins. Viruses have either RNA or DNA as a viral genome, which are present in the form of single- or double-stranded. Essential components, which are whole viral genome, replicase and structural protein, are produced from the viral genome for replication of virus inside a cell, and then these components are assembled for forming virus progeny. Finally, virus progeny are released outside cells [1].

Viruses evolve by a series of change in sequence of bases of DNA or RNA in a genome. These changes somewhat quite rapidly happen [3, 4]. As a result of this process, certain virus undergoes positive selection where best adapted variants quickly overwhelm their less fit counterparts in their population [5]. The way of viruses reproduced in their host cells makes them are prone to the genetic mutations that help to drive their evolution [4, 6]. Most mutations remains in silence and do not result in any obvious changes to the progeny viruses, but some mutations give advantages increasing the fitness of the viruses

in the environment. These could make the viruses escape from their host immune system or resistant to antiviral drugs [7].

Recently, a new novel beta-coronavirus named SARS-CoV-2, which is speculated to be originated from bats, first broke out in Wuhan city, China. The first case of this virus was reported in December 2019 and this virus has subsequently spread explosively worldwide and severely threatened human health [8, 9]. As well as human, coronaviruses are causing severe diseases in animals. Porcine epidemic diarrhea virus (PEDV), which is an alpha-coronavirus, is one of the major pathogens causing acute enteritis disease, which is characterized by vomiting and watery diarrhea and commonly leads to high rates of mortality and morbidity in suckling piglets [10, 11].

The phylogenetic analysis is popularly performed in various virus research parts including epidemiology, diagnostics, forensic studies, phylogeography, evolutionary studies, and virus taxonomy. This work can give an evolutionary perspective on variation of any trait that can be measured for a group of viruses [12-14].

Aspect of virus evolution, viruses have continued to fight with host immune through genetic mutation facilitating immune evasion and this strategy of viruses for their survival will continue in the future. Also, genetic variations, which may change structural form of viral proteins by non-synonymous changes, can hamper diagnostic accuracy of polymerase chain reaction (PCR) and enzyme-linked immunosorbent assay (ELISA), which are routinely used diagnostic techniques. Thus, it is very important work to investigate and track critical genetic events along with virus evolution. These efforts can give worthy information and insight to establish appropriate prevent and diagnostic

strategy for these viruses. Herein, coronaviruses causing sever disease in human and porcine (economic animal) by genetic and phylogenetic analysis.

Literature review

1. Coronavirus

1.1. General overview

Coronaviruses are largely classified to four genera (alpha, beta, gamma and delta) infect a wide variety of animals and are common throughout the world (Figure 1). They can cause respiratory, enteric, hepatic, and neurological diseases with variable severity, from asymptomatic to severe. Coronaviruses that infect mammals except pigs belong mainly to two genetic groups which are Alpha- and Betacoronavirus genera [15].

It has been well known that alphacoronavirus can be infected with a wide variety of species. Each species-specific alphacoronaviruses within the same sub-genera are speculated to be diverged from a certain common ancestor in the relatively near past. For instance, two types of alphacoronavirus 1, feline coronavirus (FCoV) and canine coronavirus (CCoV), are known to exist in two serotypes. Serotype II binds Aminopeptidase N, while the receptor of serotype I is unknown. The difference of their recepters is due to a different spike protein [16]. There is a common ancestor for FCoV and CCoV. This ancestor gradually evolved into FCoV I and CCoV I. An S protein from an unknown virus was recombined into the ancestor and gave rise to CCoV II. CCoV II once again recombined with FCoV to create FCoV II. CCoV II gradually evolved into TGEV. A spike deletion in TGEV creates PRCV. All these viruses are sorted into the subgenus Tegacovirus [16].

The betacoronaviruses of the greatest clinical importance concerning humans are OC43 and HKU1 (which can cause the common cold) of lineage A, SARS-CoV and SARS-CoV-2 casuing severe respiratory disease, COVID-19 of lineage B [17], and MERS-

CoV of lineage C. MERS-CoV is the first betacoronavirus belonging to lineage C that is known to infect humans. It is known that these human betacoronaviruses was originated from bat betacoronaviruses [18].

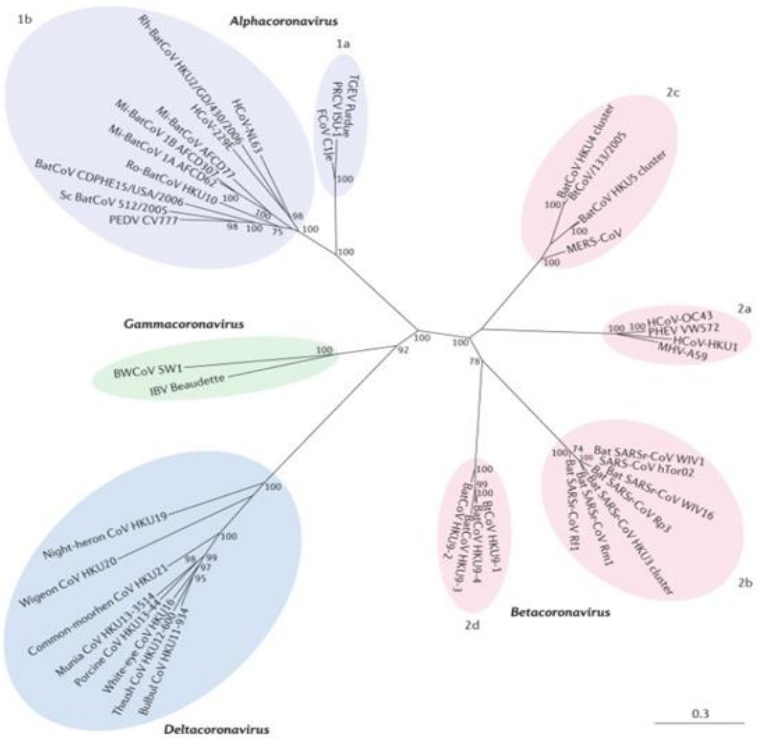


Figure 1. Phylogenetic relationships in the *Coronavirinae* subfamily. The highly human-pathogenic coronaviruses belong to the subfamily *Coronavirinae* from the family *Coronaviridae*. The viruses in this subfamily group into four genera: *Alphacoronavirus* (purple), *Betacoronavirus* (pink), *Gammacoronavirus* (green) and *Deltacoronavirus* (blue). Classic subgroup clusters are labelled 1a and 1b for the alphacoronaviruses and 2a–2d for the betacoronaviruses. The tree is based on published trees of *Coronavirinae*, and reconstructed with sequences of the complete RNA-dependent RNA polymerase-

coding region of the representative coronaviruses (maximum likelihood method under the GTR + I + Γ model of nucleotide substitution as implemented in PhyML, version 3.1. Only nodes with bootstrap support above 70% are shown. IBV, infectious bronchitis virus; MERS-CoV, Middle East respiratory syndrome coronavirus; MHV, mouse hepatitis virus; PEDV, porcine enteric diarrhoea virus; SARS-CoV, severe acute respiratory syndrome coronavirus; SARSr-CoV, SARS-related coronavirus [18].

1.2. SARS-CoV-2: Human betacoronavirus

Over the past 20 years, a few novel betacoronaviruses originated from bats have been transmitted to humans and caused severe respiratory syndrome. SARS-CoV and MERS-CoV were first introduced to humans in 2002 and 2012, respectively [8, 9, 17]. Recently, a new novel beta-coronavirus named SARS-CoV-2 first broke out in Wuhan city, China. The first case of this virus was reported in December 2019 and this virus has subsequently spread explosively worldwide and severely threatened human health [9].

CoV particles are roughly spherical, have a diameter of 120–160 nm and consist of a core, which is also called nucleocapsid, surrounded by a protective coat or envelope. The nucleocapsid contains the viral genome complexed with the N protein. CoV genomes consist of a single RNA molecule of positive polarity with a length of ~26.4 to ~31.7 kb. CoV is generally composed of four major structural proteins: nucleocapsid protein (N), membrane (M), envelope (E), and spike glycoprotein (S) (Figure 2) [9, 19].

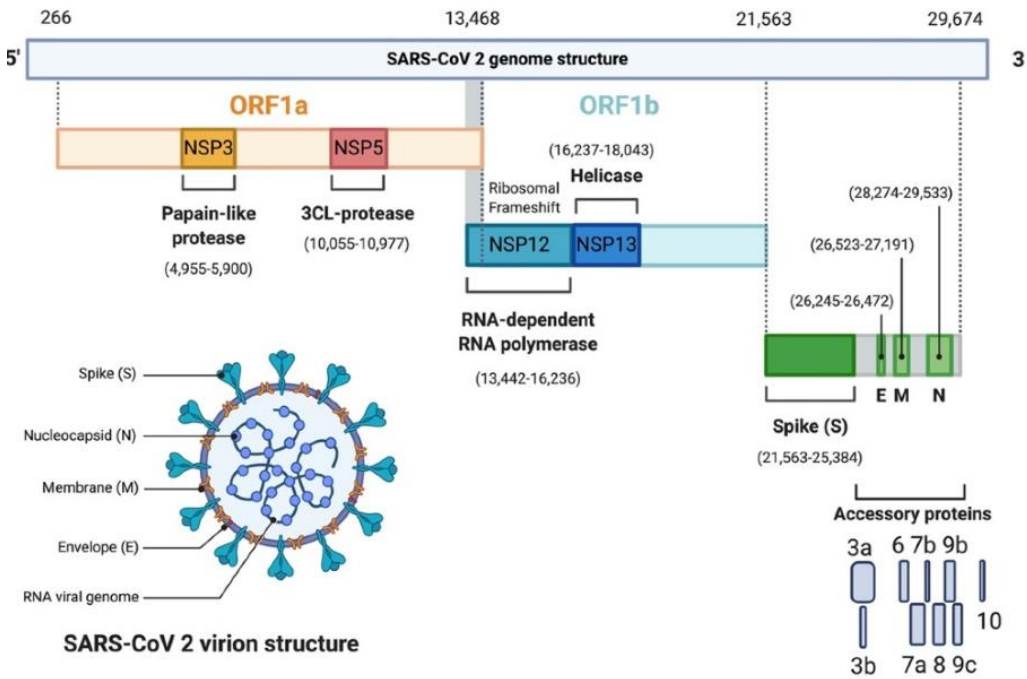


Figure 2. SARS-CoV 2 genome and virion structure [9]**.**

Among these proteins, the S glycoprotein plays major roles in viral entry and pathogenesis as its widely exposed structure forms large petal-shaped spikes on the surface of the virion. The S protein is involved in the binding of the virus to target cells and has a crucial role in the penetration of these cells by mediating membrane fusion [9, 20]. Virus particles first bind to its receptor angiotensin I converting enzyme 2 (ACE2), and then the S protein is cleaved by transmembrane serine protease 2 (TMPRSS2). Subsequently, the viral envelope fuses with the plasma membrane of the target cell, resulting in the delivery of the viral genome inside the cell (Figure 3). Also, this spike protein is an antigenic determinant as a main target for neutralizing antibodies against the virus [9, 20].

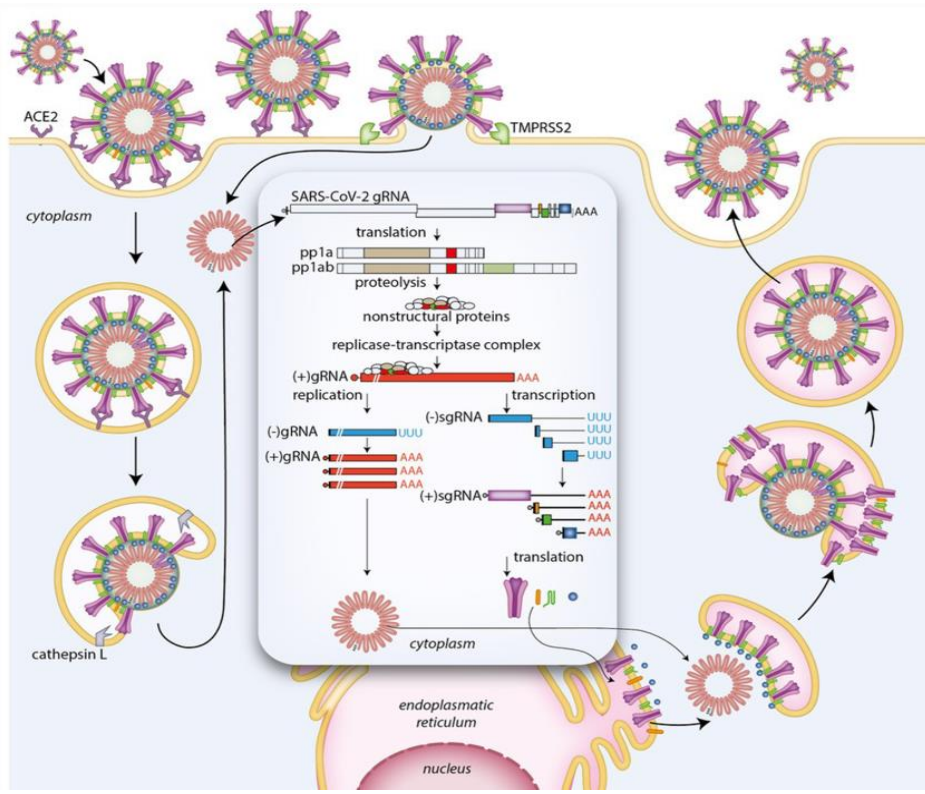


Figure 3. Schematic drawing of SARS-CoV-2 life and infectious cycle [20].

1.3. Porcine epidemic diarrhea virus: Animal alphacoronavirus

Porcine epidemic diarrhea virus (PEDV) is an enveloped, single-stranded RNA virus belonging to the family *Coronaviridae*, subfamily *Coronavirinae*, genus *Alphacoronavirus* and *subgenus Pedacovirus* [10]. PEDV is one of the major pathogens causing acute enteritis disease, which is characterized by vomiting and watery diarrhea and commonly leads to high rates of mortality and morbidity in suckling piglets. It is known that PEDV cannot be transmitted to humans, nor contaminate the human food supply [10].

PEDV is enveloped, pleomorphic and 95–190 nm in diameter including the projections on viral surface, which are approximately 18 nm in length [21]. PEDV has a single-stranded positive-sense RNA genome of approximately 28 kb in size that encodes four structural proteins, namely, S, envelope (E), membrane (M) and nucleocapsid (N) proteins, sixteen nonstructural proteins (nsp1-nsp16) and an accessory protein ORF3 (Figure 4) [22, 23]. The S protein is critical for interactions with the specific host cell receptor to mediate viral binding and entry and the formation of syncytia, and for inducing neutralizing antibodies [24].

In the late 2010s, new and highly pathogenic strains, which is named as G2 strains having distinct genetic features to G1 strains (old strains), were reported in China. These new strains were pathologically more critical than the classic strains, resulting in morbidities of 80%–100% and mortality rates of 50%–100% in infected suckling piglets [25]. In May 2013, these G2 strains moved from China to the USA and rapidly spread across the country, massively impacting the swine industry; they affected more than 4,000 farms, accounting for more than 7 million piglets within the year [26, 27]. Subsequently, these strains have become pandemic (Figure 5) [23, 28].

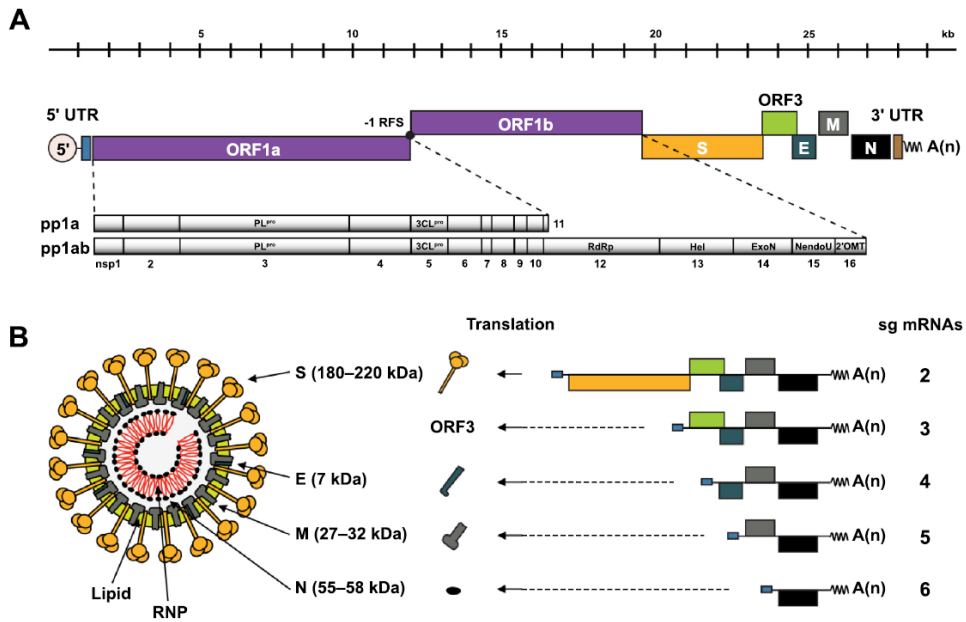


Figure 4. Schematic representations of PEDV genome organization and virion structure. **A**: The structure of PEDV genomic RNA. The 5'-capped and 3'-polyadenylated genome of approximately 28 kb is shown at the top. The viral genome is flanked by UTRs and is polycistronic, harboring replicase ORFs 1a and 1b followed by the genes encoding the envelope proteins, the N protein, and the accessory ORF3 protein. S, spike; E, envelope; M, membrane; N, nucleocapsid. Expression of the ORF1a and 1b yields two known polyproteins (pp1a and pp1ab) by −1 programmed RFS, which are co-translationally or post-translationally processed into at least 16 distinct nsps designated nsp1–16 (bottom). PLpro, papain-like cysteine protease; 3CLpro, the main 3C-like cysteine protease; RdRp; RNA-dependent RNA polymerase; Hel, helicase; ExoN, 3'→5' exonuclease; NendoU, nidovirus uridylate-specific endoribonuclease; 2'OMT, ribose-2'-O-methyltransferase. **B**: Model of PEDV structure. The structure of the PEDV virion is illustrated on the left. Inside the virion is

the RNA genome associated with the N protein to form a long, helical ribonucleoprotein (RNP) complex. The virus core is enclosed by a lipoprotein envelope, which contains S, E, and M proteins. The predicted molecular sizes of each structural protein are indicated in parentheses. A set of corresponding sg mRNAs (sg mRNA; 2–6), through which canonical structural proteins or nonstructural ORF3 protein are exclusively expressed via a co-terminal discontinuous transcription strategy, are also depicted on the right [23].

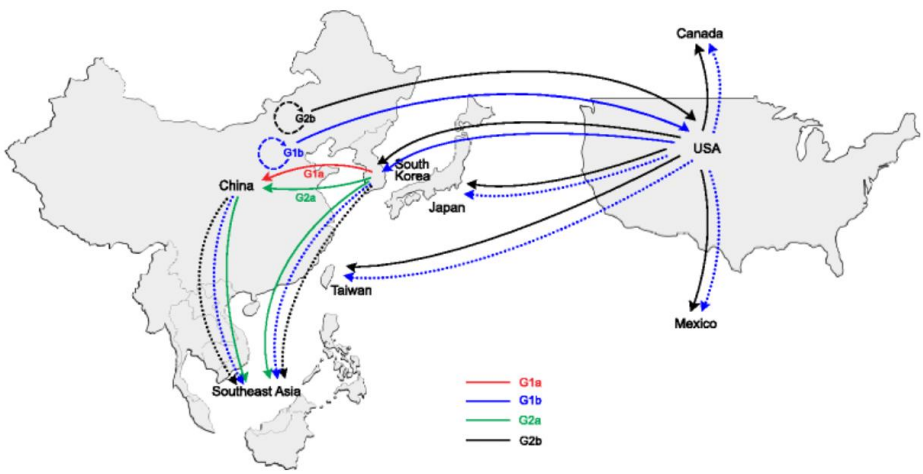


Figure 5. Potential international PEDV transmission routes. Solid lines indicate PEDV spreads that have already occurred between countries; dotted lines indicate PEDV spreads that are expected to happen eventually; dashed circular arrows denote genetic mutations or recombination events that lead to the emergence of the novel subtypes [23].

2. Virus evolution and genetic analysis on viruses

2.1. Antigenic drift and genetic shift

Viruses can shuffle their genes with other viruses when two similar strains simultaneously infect the same cell. This process is so called ‘genetic shift’, and can emerging of new and more virulent strains. It is specific cases of reassortments or viral shifts that confer phenotypic changes [29]. As another way for viral evolution, viral genome can be changed more slowly as gradually accumulation of genetic mutations over time, a process so called ‘antigenic drift’ [6]. Especially, antigenic drift on viral surface antigens results in a new strain of virus particles that is not effectively inhibited by antibodies that prevented infection by previous strains. It makes viruses easier for the changed virus to spread throughout a partially immune population [5].

2.2. Genetic mutation and recombination

Viruses evolve by a series of change in sequence of bases of DNA or RNA in a genome. These changes somewhat quite rapidly happen [3, 4]. As a result of this process, certain virus undergoes positive selection where best adapted variants quickly overwhelm their less fit counterparts in their population [5]. The way of viruses reproduced in their host cells makes them are prone to the genetic mutations that help to drive their evolution [4, 6]. In host cell, there are mechanisms such as ‘proof reading’ for correcting errors when DNA replicates and these mechanisms kick in whenever cells subdivide. These important mechanisms play a crucial role to prevent potentially lethal mutations from being inherited to progeny [30]. The mutations occur during replication of the viral genome due

to errors induced by the polymerase enzymes that replicate DNA or RNA. Unlike DNA polymerase, RNA polymerase is prone to errors because it is not capable of “proofreading” its work. Mutations of RNA viruses are faster than those of DNA viruses [31]. The changes in their genes are occasionally introduced in error, some of which are lethal. One virus can produce nearly millions of progeny viruses in just one cycle of replication. Thus, the production of a few "dud" viruses is not a severe problem for their survival [7, 32]. Most mutations remains in silence and do not result in any obvious changes to the progeny viruses, but some mutations give advantages increasing the fitness of the viruses in the environment. These could make the viruses escape from their host immune system or resistant to antiviral drugs [7].

Recombination can be defined as physical exchange of fragments originated form non-parental genetic material [29, 33]. It was considered as a crucial factor producing genetic diversity where natural selection can operate [33]. Recombination events can occur in both RNA and DNA viruses, but molecular events behind DNA and RNA recombination differ in many aspects [34]. Recombination in DNA viruses is known to be accomplished by cellular enzymatic activities [35, 36]. There are two general types of genetic DNA recombination in the cell as follows. One is homologous recombination regarded as general recombination and another is non-homologous recombination. Non-homologous or site-specific recombination relatively rarely happens and requires special proteins recognizing specific DNA sequences to promote recombination [37]. Homologous recombination happens between two DNA sequences being same or very similar in the region of crossovers [38]. The recombination process in positive-strand RNA virus can occur at the RNA level, as these viruses highly likely do not pass DNA steps in their

replication cycles. There are three classes of RNA recombination which are homologous, aberrant homologous and non-homologous [39-41]. The homologous recombination happens between two related RNA molecules at corresponding sites, although homologous RNA recombination can also happen within a common region shared by otherwise unrelated RNA sequences. The aberrant homologous recombination involves crossovers between related RNAs, but does not happen at corresponding sites, provoking sequence deletion or insertions [40]. The non-homologous recombination happens between unrelated RNA molecules [41].

2.3. Mutation rate of DNA and RNA viruses

DNA viruses have relatively low mutation rates similar to those of eukaryotic cells, because their replication enzymes have proofreading functions as like eukaryotic DNA polymerases [30]. The mutation rate in DNA viruses has been calculated as from 10^{-8} to 10^{-11} errors per incorporated nucleotide. With this low mutation rate, replication of even the most complex DNA viruses having 2×10^5 to 3×10^5 nucleotide pairs per genome rarely generate mutants [31, 32, 42, 43]. Whereas, RNA lacking a proofreading function of their replication enzymes, and some have mutation rates that are much higher 10^{-3} to 10^{-4} errors per incorporated nucleotide. Even the simplest RNA viruses having approximately 7,400 nucleotides per genome can generate mutants, perhaps as often as once per genome copy [31, 32, 42, 43]. Not all mutations generated persist in the virus population. Mutations, which interfere with the essential functions of attachment, penetration, uncoating, replication, assembly, and release, are not allowed to be

replicated and are rapidly appeared from the population. Only mutations that do not hamper essential viral functions can persist in a virus population [42].

2.4. Phenotypic Variation by Mutations

Mutation can generate novel antigenic determinants. For instance, mutations within hemagglutinin (HA) gene of influenza virus can make an altered antigenic site considered as epitope of HA [44]. If provided attachment function of certain new hemagglutinin is intact, the mutant virus may be able to infect in a host immune against viruses expressing the previous hemagglutinin. This relatively modest process of antigenic change by genetic mutation, so called antigenic drift, may allow a virus to break down host defenses and cause disease in previously immune individuals [7].

2.5. Phylogenetic analysis on viruses

Viruses are continuously changing through generations and over time in the process known as evolution. Viruses can evolve at high, uneven, and fluctuating rates among genome sites [3, 4]. The accumulated changes by either mutation or recombination events are fixed in the genome of successful individuals subsequently giving rise to genetic lineages [5]. The relationship among biological lineages relating to common descent which is so called ‘phylogeny’. For inference of phylogeny, the differences between aligned sequences of genomes and proteins are quantified and depicted in forms of phylogenetic trees where contemporary species and their intermediate and common ancestors are indicated in the terminal nodes, internal nodes, and the root, respectively. The trees are characterized by topologies, length of branches, shapes, and the root

positions [13, 14]. A complex mathematical apparatus has been devised for phylogeny inference that can evaluate and validate inter-species differences. Also, it can facilitate tree building and comparison of trees, and assessing a best fit model among data and inferred trees, through typically computationally intensive calculations [12]. A reconstructed tree is an approximation of the true phylogeny that practically remains unknown. The phylogenetic analysis is popularly performed in various virus research parts including epidemiology, diagnostics, forensic studies, phylogeography, evolutionary studies, and virus taxonomy. This work can give an evolutionary perspective on variation of any trait that can be measured for a group of viruses [12-14].

3. Impact of genetic mutation on viral antigenicity and diagnosis

Most genetic mutations remains in silence and do not result in any obvious changes to the progeny viruses, but some mutations give advantages increasing the fitness of the viruses in the environment. These could make the viruses escape from their host immune system or resistant to antiviral drugs [7]. There are two types (synonymous and non-synonymous) of genetic mutation in DNA and RNA viruses. Synonymous substitution is nucleotide changes with no changes in amino acids in the encoded protein, so it is often called a silent substitution. But, despite being silent in protein sequence, this mutation can be targeted by natural selection directly at the DNA or RNA level [45]. For instance, different synonymous codons can have different translation efficiencies because transfer RNAs vary in concentration within the cell, typically leading to codon usage biases [46].

Non-synonymous substitution gives more powerful impact on phenotypic features in viruses compared to synonymous substitution. It leads to changes in amino acids in the encoded protein and these changes may result in viral antigenicity [47]. For instance, it is known that canine parvovirus type-2 (CPV-2) is divided into 3 antigenic types. Two new antigenic types of CPV-2, type 2a (CPV-2a) and type 2b (CPV-2b), emerged and have virtually replaced the original CPV-2 strain worldwide [48]. Another antigenic type, CPV type 2c (CPV-2c) emerged in the 2000s and spread globally [49, 50]. The most relevant changes between these 3 types are in residue 426 of the VP2 protein. Types CPV-2a, -2b, and -2c presented Asn (N), Asp (D) and Glu (E) at this position, respectively. These non-synonymous changes confer somewhat different antigenicity to these 3 types, although they have cross-reactivity each other [51]. Thus, recent CPV vaccines contain all three

antigenic types for better protection.

Antigenicity changes resulted from non-synonymous mutation may impede diagnosis using antigen/antibody reaction. Nucleocapsid (N) protein is considered to play an important role in inducing cell-mediated immunity [52]. Because of these features, N protein is commonly used as a target for diagnosis and vaccine development [53, 54]. Porcine respiratory and reproductive virus (PRRSV), which causes severe problems in the swine industry, can be classified into two genogroups, type 1 (European) and type 2 (North American), based on the N gene [55]. These two genogroups present different N protein antigenicities; therefore, recent commercial ELISA kits include recombinant N proteins of both European and North American PRRSVs [56]. In the immunological diagnosis of PEDV, commercial PEDV ELISA kits have been showing poor performance, and the results of neutralizing assays using cell culture sometimes mismatch with those of the ELISA assays. In fact, Chang et al. recently reported that the antibodies induced by the G2b PEDV strain poorly reacted with a commercial N-based ELISA kit, which showed a sensitivity of 37% [57], which may be the result of antigenicity differences between the genogroups.

Generally, virus isolation is considered to be a gold standard for viral diagnosis. However, this method is not easily applied for routine diagnosis. Virus isolation based on cell culture requires skillful techniques as well as increased labor and time [58]. Also, virus isolation might fail due to the presence of antibodies in extracellular fluids which inhibit viral replications in vitro [58, 59]. Compared to virus isolation, PCR assay is a very useful method due to its easy, rapid, sensitive, and inexpensive features [60, 61]. However, genetic mutation can result in potential mismatches and false-negative results

[62, 63]. For example, primer and template mismatches have been reported to hamper proper diagnosis of several viruses including feline calicivirus, porcine respiratory and reproductive virus, influenza virus, respiratory syncytial virus, dengue virus and hepatitis B virus [64].

Chapter I

Genetic and phylogenetic analysis of sever acute respiratory syndrome coronavirus type 2 (SARS-CoV-2)

Abstract

The S glycoprotein of coronaviruses is important for viral entry and pathogenesis with most variable sequences. Therefore, we analyzed the S gene sequences of SARS-CoV-2 to better understand the antigenicity and immunogenicity of this virus in this study. In phylogenetic analysis, two subtypes (SARS-CoV-2a and -b) were confirmed within SARS-CoV-2 strains. These two subtypes were divided by a novel non-synonymous mutation of D614G. This may play a crucial role in the evolution of SARS-CoV-2 to evade the host immune system. The region containing this mutation point was confirmed as a B-cell epitope located in the S1 domain, and SARS-CoV-2b strains exhibited severe reduced antigenic indexes compared to SARS-CoV-2a in this area. This may allow these two subtypes to have different antigenicity. If the two subtypes have different serological characteristics, a vaccine for both subtypes will be more effective to prevent COVID-19. Thus, further study is urgently required to confirm the antigenicity of these two subtypes.

Keywords: COVID-19; SARS-CoV-2; spike protein; antigenicity

1. Introduction

Coronavirus (CoV) is a class of genetically diverse RNA viruses found in a wide range of hosts including reptiles, birds, and mammals. Most pathogenic CoVs usually cause respiratory and intestinal symptoms in animals [8, 65-68]. Over the past 20 years, a few novel beta coronaviruses originated from bats have been transmitted to humans and caused severe respiratory syndrome. SARS-CoV and MERS-CoV were first introduced to humans in 2002 and 2012, respectively [8]. Recently, a new novel beta-coronavirus named SARS-CoV-2 first broke out in Wuhan city, China. The first case of this virus was reported in December 2019 and this virus has subsequently spread explosively worldwide and severely threatened human health [8, 69].

CoV is generally composed of four major structural proteins: nucleocapsid protein (N), membrane (M), envelope (E), and spike glycoprotein (S). Among these proteins, the S glycoprotein plays crucial roles in viral entry and pathogenesis as its widely exposed structure forms large petal-shaped spikes on the surface of the virion [70]. Mutations in the spike glycoprotein can allow novel coronavirus strains to infect humans and spread pandemically [71]. Therefore, S gene encoding S glycoprotein has widely been used for molecular analysis of coronaviruses due to the significant features of the S glycoprotein a_ecting the antigenicity and immunogenicity [66, 67, 72-75]. Thus far, there has been little data comparing and analyzing S gene sequences within SARS-CoV-2. Generally, several types of coronavirus are divided into subtypes depending on amino acid mutations in S gene sequences, and molecular analysis based on the S gene can provide insights into antigenicity, immunogenicity, or evolutionary trends [66, 67, 72, 76]. Thus,

we analyzed the S gene sequences of SARS-CoV-2 to better understand this virus in this study.

2. Material and Methods

2.1. Sample collection and Phylogenetic analysis

For phylogenetic analysis based on the S gene, 144 sequences of SARS-CoV-2 that globally originated from several countries (China, USA, Italy, Spain, Japan, Vietnam, Taiwan, and Pakistan) were retrieved from GenBank. Using IQ-TREE v1.6.12 [77], the genetic relationships between SARS-CoV-2 were inferred by the maximum likelihood (ML) method. The "-m MFP" option was invoked to help select the data best-fit amino acid substitution model. The branch support values were estimated by ultrafast bootstrap approximation [78] implemented in IQ-TREE [77] via the "-bb 1000" option. The reconstructed phylogenies were displayed and midpoint rooted by FigTree v1.4.3.

2.2. Epitope prediction and antigenic index analysis on S gene sequences

B-cell epitopes on the S1 subunit were predicted by BepiPred-2.0 [79], the Chou & Fasman method [80], the Kolaskar and Tongaonkar method [81] and Parker's Hydrophilicity [82]. Subsequently, antigenic indexes of each amino acid in this region were calculated by the Jameson–Wolf method [83].

3. Results

3.1. Phylogenetic analysis on S gene of SARS-CoV-2

In the ML tree, completely divided clades were identified among the analyzed SARS-CoV-2 strains (Figure 29A). Interestingly, only one reliable non-synonymous change was found to distinguish between subtypes A and B in this study. SARS-Cov-2a and -2b strains consistently exhibited Ala (D) and Gly (G) at the amino acid sequence position 614, respectively (Figure 29B).

3.2. Epitope prediction of S protein

Regions between amino acids 614 and 621 were equally identified as a B cell epitope by all four methods (Figure 29C). The predicted B-cell epitope including amino acid 614 was located in a relatively well-exposed part of the S1 subunit in the 3D-view structure (Figure 29D); this B-cell epitope sequence was identified within the sequence corresponding to the SD-1/-2 domain (Figure 29E).

3.3. Antigenic index analysis on the epitope of S1 subunit

The antigenic indexes of each amino acid in this region (amino acids 613–621) were calculated by the Jameson–Wolf method. When an antigen index was >0.5, it was believed to be a reliable position as an epitope [83]. The results of the antigenic index analysis showed severely reduced indexes of amino acids 615–617 in the SARS-CoV-2b strains compared to SARS-CoV-2a; it is predicted that the change of D614G affects the antigenicity of this region (Figure 29F).

4. Discussion

In the phylogenetic tree, only one reliable non-synonymous change was found to distinguish between subtypes A and B in this study. SARS-Cov-2a and -2b strains consistently exhibited Ala (D) and Gly (G) at the amino acid sequence position 614, respectively. The virus consistently evolves to evade the host immune system with non-synonymous mutations that are so-called positive selection. More evolved viruses are better able to survive, thus such viruses will likely be dominant within the group [84]. SARS-CoV-2a includes the China strains confirmed in 2019, but SARS-CoV-2b only includes the USA strains confirmed after 2020. In addition, the USA was one of the latest countries to experience a COVID-19 outbreak. Although there was only a few months' difference, SARS-CoV-2b may be a more evolved form. If the mutation of D614G plays a crucial role in the positive selection process, SARS-CoV-2b will be the dominant type of SARS-CoV-2 in the future. More long-term tracking will be required to validate this assumption.

The S glycoprotein is important for viral entry and pathogenesis with the most variable sequences in the coronavirus genomes. Human beta-coronavirus S proteins are cleaved into S1 and S2 subunits by host proteases [85]. The S1 subunit forming a globular shape is responsible for receptor binding [86] while the S2 subunit forming a rod shape mediates membrane fusion [87]. More specifically, the S1 subunit is composed of two major domains (S1-NTD and S1-CTD) and two sub-domains (SD-1 and SD-2). One or both of the major domains is potentially responsible for binding host-receptors, and the sub-domains that are complex folding of elements may allow receptor-induced

conformational changes [74, 87, 88]. Thus, mutations within the S1 region are associated with changes in antigenicity and viral pathogenicity [89].

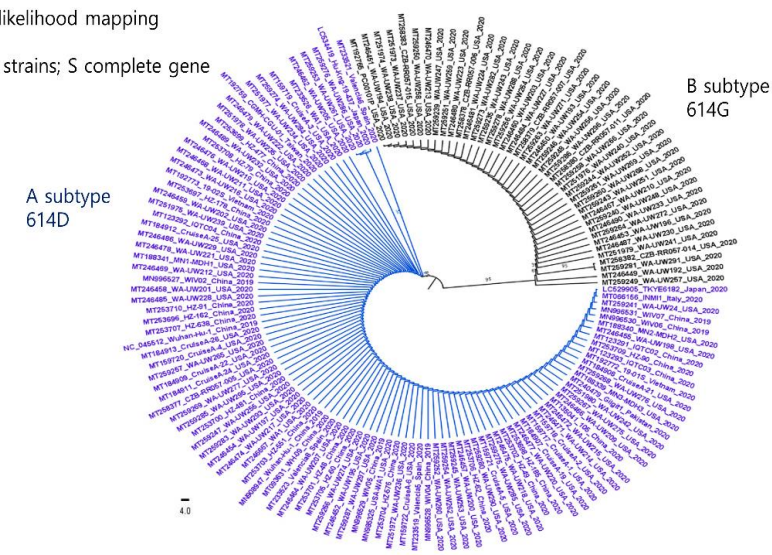
In fact, the S1 subunit contains numerous major and minor neutralizing antibody epitopes [19, 90], thus it is difficult to investigate all putative epitopes to discover which can play a crucial role on the antigenicity and immunogenicity of viruses. In this situation, inferences considering both aspects of virus evolution and epitope analysis can be helpful to investigate which epitopes really play a crucial role. In the prediction of epitopes of S protein, regions between amino acids 614 and 621 were equally identified as a B cell epitope by all four methods. The predicted B-cell epitope including amino acid 614 was located in a relatively well-exposed part of the S1 subunit in the 3D-view structure; this B-cell epitope sequence was identified within the sequence corresponding to the SD-1/-2 domain.

The results of the antigenic index analysis showed severely reduced indexes of amino acids 615–617 in the SARS-CoV-2b strains compared to SARS-CoV-2a; it is predicted that the change of D614G affects the antigenicity of this region. Since no amino acid changes were found in this area other than the change of D614G, it is believed that this amino acid change alters the conformation of these immunogenic determinants; consequently, this region is expected to no longer act as a B-cell epitope in SARS-CoV-2b. B-cell plays a major role in recognizing pathogens and stimulating adaptive immunity in the immune response against virus infection. Thus, the elimination of B cell epitopes will likely reduce immunogenicity by hampering the immune cell recognition of the virus [91, 92].

When reflecting the above results, SARS-CoV-2b may have reduced immunogenicity

compared to SARS-CoV-2a. If so, it can permit persistent or recurrent infection of SARS-CoV-2b while evading immune cell recognition. In addition, the mutation of the S1 domain may induce different antigenicity and viral pathogenicity between the two subtypes. Different virus subtypes will likely have somewhat different serological features depending on their antigenicity, although there may be some cross-reactivity. In addition, a certain subtype can serologically cover other serotypes [75, 93]. Thus, this point should be considered to create a new SARS-CoV-2 vaccine. Indeed, if the two serotypes have different serological characteristics, a vaccine that includes both subtypes will be more effective at preventing COVID-19, particularly when developing a killed vaccine that has a narrow protection range compared to a live vaccine. This study was confined to investigating the phylogenetic and genetic features of SARS-CoV-2 due to limited information. Therefore, further study on the cross-reactivity between these two subtypes is required to validate our assumption. Viruses have continually evolved through genetic mutations to evade host immune systems in the long history of the fight between humans and viruses. SARS-CoV-2, which has only recently been introduced in humans, will continue to evolve for survival in the current situation, in which this virus has already become a pandemic. To respond properly against this virus, continuous surveillance of this virus' adaptation to evade host immune systems is important in the future.

(A). Maximum likelihood mapping



(B)

(C). Predicted of B cell epitope;
Reference;GenBank: NC045512, Wuhan-Hu-1, China, 2019

BepiPred-2.0

Threshold: 0.5

```
...EEEE.........EEEEEEE....EEEEEEEEEEEEEEEEEEE
TPGTNTSNQVAVLYQDVNCTEVPVAIHADQLTPTWRVYSTGSNVFQT
-600-------610-------620-------630-------640----
```

Predicted residue scores:

Chou & Fasman

Threshold: 1.0

Position	Residue	Start	End	Peptide	Score
614	D	611	617	LYQDVNC	1.06
615	V	612	618	YQDVNCT	1.113
616	N	613	619	QDVNCTE	1.056

Kolaskar & Tongaonkar

Threshold: 1.041

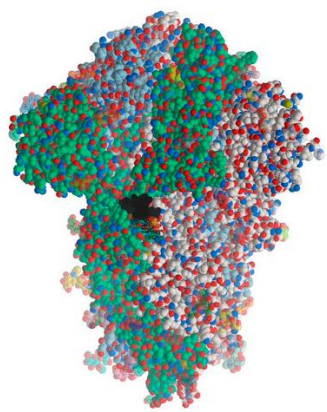
Position	Residue	Start	End	Peptide	Score
613	Q	610	616	VLYQDVN	1.119
614	D	611	617	LYQDVNC	1.123
615	V	612	618	YQDVNCT	1.075
616	N	613	619	QDVNCTE	1.03
617	C	614	620	DVNCTEV	1.083
618	T	615	621	VNCTEVP	1.111
619	E	616	622	NCTEVPV	1.111
620	V	617	623	CTEVPVA	1.152
621	P	618	624	TEVPVAI	1.115

Parker Hydrophilicity

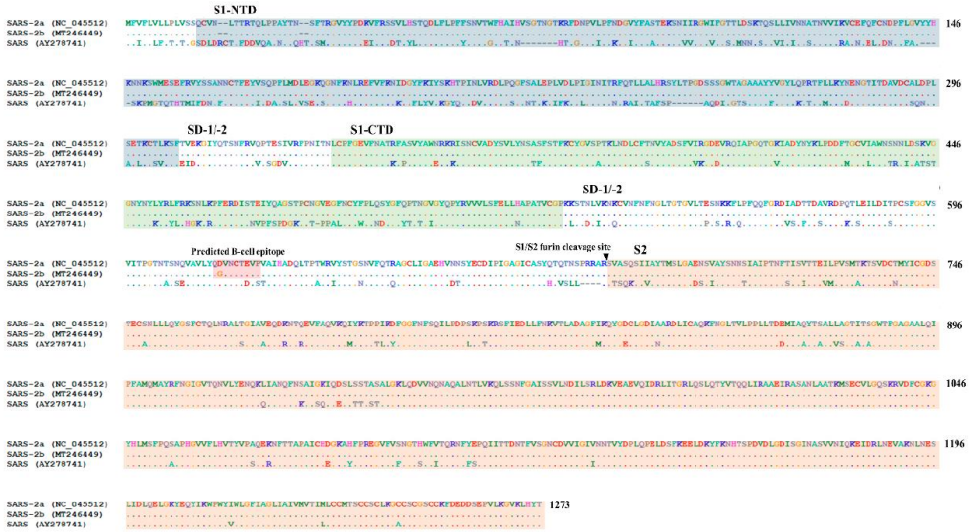
Threshold: 1.238

Position	Residue	Start	End	Peptide	Score
614	D	611	617	LYQDVNC	1.371
615	V	612	618	YQDVNCT	3.429
616	N	613	619	QDVNCTE	4.814
617	C	614	620	DVNCTEV	3.429
618	T	615	621	VNCTEVP	2.3
619	E	616	622	NCTEVPV	2.3
620	V	617	623	CTEVPVA	1.6

(D). COVID-19 3D protein
Reference; PDB 6VXX



(E). SARS-CoV-2 S gene sequence



(F). Jameson–Wolf antigenic index

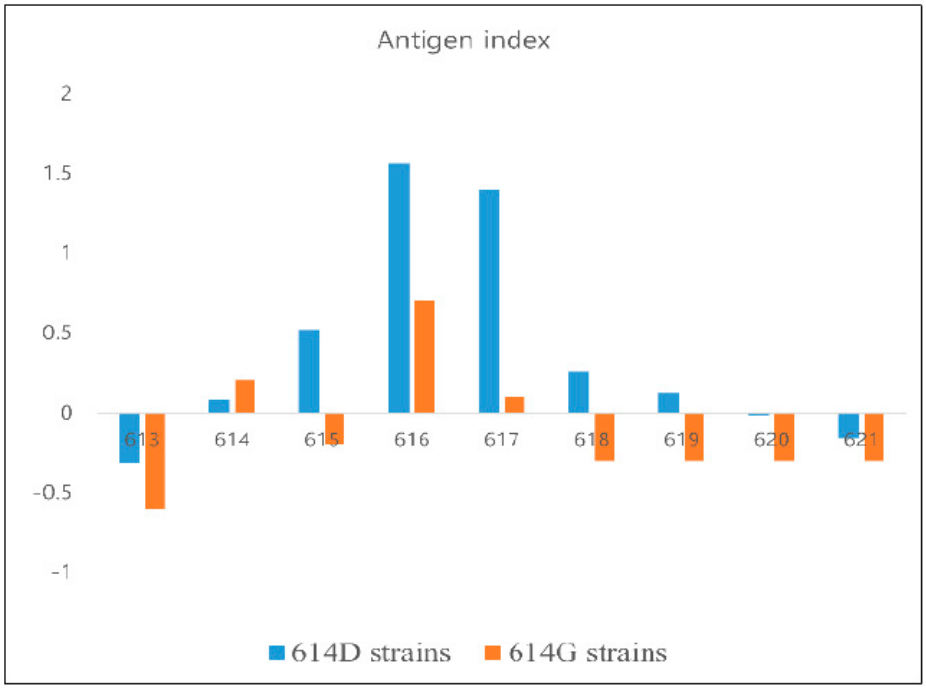


Figure 6. Characterization of the S complete gene in SARS-CoV-2. (**A**) Phylogenetic analysis of SARS-CoV-2 strains based on the S gene. The phylogenetic trees were reconstructed from 144 sequences of SARS-CoV-2 collected globally. Thus, two subtypes (SARS-CoV-2a and -2b) were completely divided. (**B**) Alignment of SARS-CoV sequences including the aa 614 position are highlighted in gray. A novel reliable non-synonymous mutation was identified to distinguish the A and B subtypes. SARS-Cov-2a and -2b strains consistently exhibited Ala (**D**) and Gly (G) at the amino acid sequence position 614, respectively. (**C**) Identification of B-cell epitopes in the adjacent area with aa 614. The B-cell epitope was predicted by BepiPred-2.0 [79], the Chou & Fasman method [80], the Kolaskar and Tongaonkar method [81], and Parker's Hydrophilicity [82]. The 614–621 region was predicted to consist of epitopes. (**D**) The 3D-structure of SARS-CoV-2 Spike protein by Mol soft Mol Browser 3.8–5 according to the original publication from the National Center for Biotechnology Information (NCBI): PDB;6VXX. The predicted B-cell epitope (aa 613–620) highlighted in black color was located at a relatively well-exposed part. (**E**) Sequence alignment of SARS-CoV-1 and -2. The S1 subunit is responsible for receptor binding and the S2 subunit mediates membrane fusion. The S1 subunit consists of two major domains capable of binding to host receptors: an amino (N)-terminal domain (NTD) and a carboxy (C)-terminal domain (CTD) and two sub-domains that may allow receptor-induced conformational changes: SD-1 and SD-2. (**F**) The antigenic index of each amino acid constituting this region (amino acids 613–621) by the Jameson–Wolf method [83].

Chapter II

Molecular characterization of porcine epidemic diarrhea virus and its new genetic classification based on the nucleocapsid gene

Abstract

Porcine epidemic diarrhea virus (PEDV) causes continuous, significant damage to the swine industry worldwide. By RT-PCR-based methods, this study demonstrated the ongoing presence of PEDV in pigs of all ages in Korea at the average detection rate of 9.92%. By the application of Bayesian phylogenetic analysis, it was found that the nucleocapsid (N) gene of PEDV could evolve at similar rates to the spike (S) gene at the order of 10^{-4} substitutions/site/year. Based on branching patterns of PEDV strains, three main N gene-base genogroups (N1, N2, and N3) and two sub-genogroups (N3a, N3b) were proposed in this study. By analyzing the antigenic index, possible antigenic differences also emerged in both the spike and nucleocapsid proteins between the three genogroups. The antigenic indexes of genogroup N3 strains were significantly lower compared with those of genogroups N1 and N2 strains in the B-cell epitope of the nucleocapsid protein. Similarly, significantly lower antigenic indexes in some parts of the B-cell epitope sequences of the spike protein (COE, S1D, and 2C10) were also identified. PEDV mutants derived from genetic mutations of the S and N genes may cause severe damage to swine farms by evading established host immunities.

Keywords: PEDV, phylogenetic analysis, antigenicity, S gene, N gene

1. Introduction

Porcine epidemic diarrhea virus (PEDV) is an enveloped, single-stranded RNA virus belonging to the family *Coronaviridae*, subfamily *Coronavirinae*, genus *Alphacoronavirus* and subgenus *Pedacovirus*. PEDV is one of the major pathogens causing acute enteritis disease, which is characterized by vomiting and watery diarrhea and commonly leads to high rates of mortality and morbidity in suckling piglets [10]. The disease was first reported in the UK in 1971, and the prototype virus—designated as PEDV CV777—was subsequently identified in Belgium [21]. Since the 1980s, PEDV has been widespread throughout Asia, where it has been regarded as an endemic disease for many years [94, 95]. In the late 2010s, new and highly pathogenic strains were reported in China. These new strains were pathologically more critical than the classic strains, resulting in morbidities of 80%–100% and mortality rates of 50%–100% in infected suckling piglets [11]. In May 2013, these new highly pathogenic strains moved from China to the USA and rapidly spread across the country, massively impacting the swine industry; they affected more than 4,000 farms, accounting for more than 7 million piglets within the year [26, 27]. Subsequently, these strains have become pandemic [28].

PEDV has an approximately 28 kb long genome and consists of seven open reading frames (ORF), encoding non-structural or structural proteins [96]. ORF1ab and ORF3 genes encode non-structural proteins. ORF1a codes for the large polyprotein PP1a, while ORF1b is always expressed with PP1a as the fusion protein PP1a/b through ribosomal frameshifting. PP1a and PP1a/b are further processed into 16 non-structural proteins (nsp1 to nsp16). ORF3 codes an accessory protein that is likely to be an additional non-

structural protein (Song and Park 2012). The envelope (E), membrane (M), spike (S), and nucleocapsid (N) genes encode four major structural proteins [97]. The N protein, which is the most abundant protein in the virus particle, provides the structural basis for the helical nucleocapsid surrounding the virus genome [52, 98]. The M and E proteins form a viral envelope by assembly. The M protein is the most abundant component and the E protein is less abundant in the viral envelope [99]. The S protein is very exposed and forms large petal-shaped spikes on the surface of the virus [100].

Among these structural proteins, the S protein plays important roles in virus infection and the induction of neutralizing antibodies [24] and shows substantial genetic diversity [97]. Because of these features, genetic analyses based on the S gene have been commonly used to investigate PEDV evolution [27, 28, 67]. The PEDV has diverged into several subgroups based on the genetic diversification of the S gene. So far, two genotypes (G1 and G2) have been identified by S gene phylogenetic analysis. Each genogroup can be further divided into two subtypes, G1a/b and G2a/b, respectively [101, 102]. Recently, a third subtype (G2c) was identified within the G2 genogroup [28]. N protein is involved in several biological and immunological activities of the virus, including viral nucleolar localization, host cycle ER stress, S-phase prolongation, inhibition of interferon-β production [103-105], and induction of abundant antibodies [52, 98]. Despite the significant features of the N protein, there has been less investigation on the N protein compared to the S protein. In order to understand better about the evolutionary aspects of PEDV, this study applied multiple bioinformatic tools to investigate the genetic diversity of PEDV.

2. Material and Methods

2.1. Sample collection, PEDV detection by PCR, and complete sequencing

Six hundred and seventy-two fecal samples were collected from 83 farms in 2017 and 235 fecal samples from 33 farms in 2018. Rectal swabs were randomly collected from pigs of all stages (suckling, weaned, growing, finishing, gilt, sow) in swine farms dispersed throughout 9 provinces of South Korea. RNA extraction from the samples was performed using an RNA/DNA Extraction kit (Invitrogen, Carlsbad, CA, USA) according to the manufacturer's instructions and the extracted RNA samples were stored at −70°C. The RNA was converted into cDNA with a commercial kit (RNA to cDNA EcoDry Premix, Clontech, Otsu, Japan), following the manufacturer's protocol. PEDV in the collected fecal samples was detected by PEDV-PCR (Chung, Van Giap Nguyen et al. 2015). Commercial kits of Median diagnostic (Korea) were used to detect Porcine reproductive and respiratory syndrome virus (cat. NS-PRR-11), Porcine parvovirus (cat. NS-ABO-11) and Japanese encephalitis virus (cat. NS-ABO-12). The other three enteric pathogens were detected by specific primers and PCR conditions reported previously, such as: Porcine deltacoronavirus [106], Transmissible gastroenteritis virus, and Porcine group A rotavirus [107].

Full-length genome sequencing was conducted with 26 overlapping primer pairs using positive samples from swine farms that were severely damaged by porcine epidemic diarrhea [108], and 7 strains (Y178, S6, S10, S12, S14, S97 and S100) were fully sequenced. Information relating to the 7 strains is provided in Table 1. The full-length genome sequences of the 7 strains were registered in GenBank (accession numbers

MH891584–MH891590).

2.2. Genetic analysis of recombination

For the genetic analyses, 72 other previously registered complete genome sequences were retrieved from GenBank, and the sequences originated from Asia (China, Korea, Japan), America (USA), and Europe (Germany, Belgium) from 1978 to 2018. RDP v4.51 program (Martin, Murrell et al. 2015) with 5 default algorithms (RDP, GENECONV, MaxChi, SiScan and Bootscan) was applied to identify recombinants in the alignment of the S and N genes. The general options were: “sequences are linear”, “highest acceptable p-value = 0.05” and “Bonferroni correction” = true. Upon detection, new recombination-free alignments were created by the option of “save alignment with recombinant regions removed”. Four datasets with identical number of sequences (n = 79, seven obtained in this study) were generated for subsequent analyses: complete S gene (D1), S gene without recombinant regions (D2), complete N gene (D3) and N gene without recombinant regions (D4).

Table 1. Information of PEDV positive samples in this study

Sample Number	Sample collection*			Farm information			Other pathogens*** (TGEV, RotaV, PDCoV, PRRSV, PPV, JEV)	GeneBank
	Stage	Collected Date	Province	Farm name	No. of sows	Vaccinated**		
Y178	Suckling	21-Feb-2017	Gyeongnam	HA	1500	PRRSV (MLV), PPV (VLP)	Negative	MH891588
S6	Suckling	09-Jan-2018	Chungnam	DB	200	PRRSV (MLV), PPV (VLP)	Negative	MH891589
S10	Suckling	06-Feb-2018	Gyeonggido	WD	200	PPV (VLP), PRRSV (MLV), PEDV (killed)	Negative	MH891590
S12	Suckling	06-Feb-2018	Gyeonggido	WT	200	PPV (VLP), PRRSV (MLV), PEDV (killed)	Negative	MH891584
S14	Suckling	07-Feb-2018	Gyeonggido	SW	400	PPV (VLP), PRRSV (MLV), JEV (MLV)	Negative	MH891585
S97	Suckling	17-Apr-2018	Gyeonggido	NG	400	PPV (VLP), PRRSV (MLV), PEDV (Killed)	Negative	MH891586
S100	Suckling	19-Apr-2018	Gyeonggido	YG	700	PPV (VLP), PRRSV (MLV), JEV (MLV), PEDV (Killed)	Negative	MH891587

* Suckling piglets in the investigated farms suffered from severe watery diarrhea and high mortality

* MLV: modified live vaccine, VLP: virus like particle

*** TGEV (Transmissible gastroenteritis virus), RotaV (Rotavirus group A), PDCoV (Porcine deltacoronavirus), PRRSV (Porcine reproductive and respiratory syndrome virus), PPV (Porcine parvovirus), JEV (Japanese encephalitis virus)

2.3. Bayesian phylogenetic analysis

BEAST package v2.6.1 [109] which is available at the CIPRES Science Gateway [110] was used to infer the phylogenetic relationships between sequences and co-estimate the substitution rates from the above mentioned D1-D4 datasets. The genetic classification of PEDV based on the S gene followed previous publication [28]. For the model of nucleotide substitution, bModelTest tool [111] implemented in BEAST 2 was selected which helps to infer the most appropriate substitution model. For molecular clock model, four models of strict clock, uncorrelated lognormal and exponential relaxed-clock [112] and random local clock [113] were specified. For tree prior, three coalescent models implemented in BEAST 2 were tested, including coalescent constant population, coalescent exponential population and coalescent Bayesian skyline plot [114]. Each analysis was run for 100 million chains, sampling every 10 000 generations. The output log files were analyzed in Tracer v1.7.1 [115] to assess the convergence (effective sample size > 100). Path sampling analyses [116] were also performed to select the best fit molecular clock and tree prior models for each dataset. For that analysis, the number of path steps were 100, and the length of each chain were one million iterations. The nucleotide substitution rates and phylogenetic tree of each D1 - D4 dataset were inferred from the data best-fit combining models. The phylogenic trees were summarized with TreeAnnotator v2.6.1 to produce the maximum clade credibility tree, which was displayed using FigTree v1.4.3.

2.4. Pairwise genetic distance (p-Distance) analysis

p-distances within each dataset were calculated using MEGA V. 7 software [117]. The option for gaps/missing data treatment was specified as “partial deletion”. The obtained results were displayed in a frequency distribution histogram of *p*-distance. Basically, a lower genetic relationship between two sequences indicated a higher *p*-distance, and well-bounded areas with peaks in the histogram indicated the presence of different genetic clusters [118].

2.5. Inferring ancestral amino acid changes

The Baseml program implemented in package PAML 4.9j [119] was used to reconstruct amino acid changes on the evolutionary path of PEDV based on the N gene. The input tree topology for that analysis was the maximum clade credibility tree inferred by BEAST 2 under the data best-fit combining models. Non-synonymous substitutions that occurred on the given branches of a phylogeny were annotated by the treesub program (available online: https://github.com/tamuri/treesub).

2.6. Amino acids and antigenic index analysis of N gene

Amino acid sequences deduced from nucleotide sequences of the N gene were aligned and comparatively analyzed to investigate the non-synonymous changes according to genetically distinct clusters. Subsequently, antigenic index analysis of the complete amino acid sequences of the N gene was performed to evaluate the possible antigenic variation of N proteins. The antigenic index of each amino acid was calculated by the Jameson-Wolf method [83], and the calculated indexes of each strain were compared.

2.7. Antigenic index analysis of B-cell epitopes in Korean PEDV strains

Antigenic index analyses were performed to evaluate the antigenic variation of S proteins within the Korean PEDV strains. The antigenic index of each amino acid was calculated by the Jameson-Wolf method for 3 previously identified S protein neutralizing epitopes, COE (within the $S1^B$ region) [120], S1D [121], and 2C10 [122]. Subsequently, the calculated indexes of the Korean strains were compared.

3. RESULTS

3.1. The detection of PEDV in Korea from 2017 to 2018

The detection rate of PEDV from 2017 to 2018 was 9.92% (90/907). Specifically, the positive rate in 2017 and 2018 was 8.63% (58/672) and 13.62% (32/235), respectively. The positive rate in 2018 had somewhat increased compared with in 2017. From the detection rates of each growth stage, the highest rate was seen in the suckling stage (Table 2). Geographically, the province with a higher concentration of swine farms showed higher positive samples of PEDV (Figure 7).

Table 2. Detection results of PEDV according to each stage from 2017 to 2018

2017 year / stage*	Suckling	Weaned	Growing	Finishing	Gilt	Sow	Total
Number of samples	325	162	42	37	43	63	672
Positive samples	33	15	2	3	2	3	58
%	10.15	9.26	4.76	8.11	4.65	4.76	8.63
2018 year / stage*	Suckling	Weaned	Grower	Finisher	Gilt	Sow	Total
Number of samples	76	58	34	20	18	29	235
Positive samples	16	8	2	2	2	2	32
%	21.05	13.79	5.88	10.00	11.11	6.90	13.62
Total / stage*	Suckling	Weaned	Grower	Finisher	Gilt	Sow	Total
Number of samples	401	220	76	57	61	92	907
Positive samples	49	23	4	5	4	5	90
%	12.22	10.45	5.26	8.77	6.56	5.43	9.92

* Samples were sorted into 6 stages: female (gilt and sow), suckling (<21 d), weaned (21– 60 d), growing (60–90 d); and finishing (≥ 90 d).

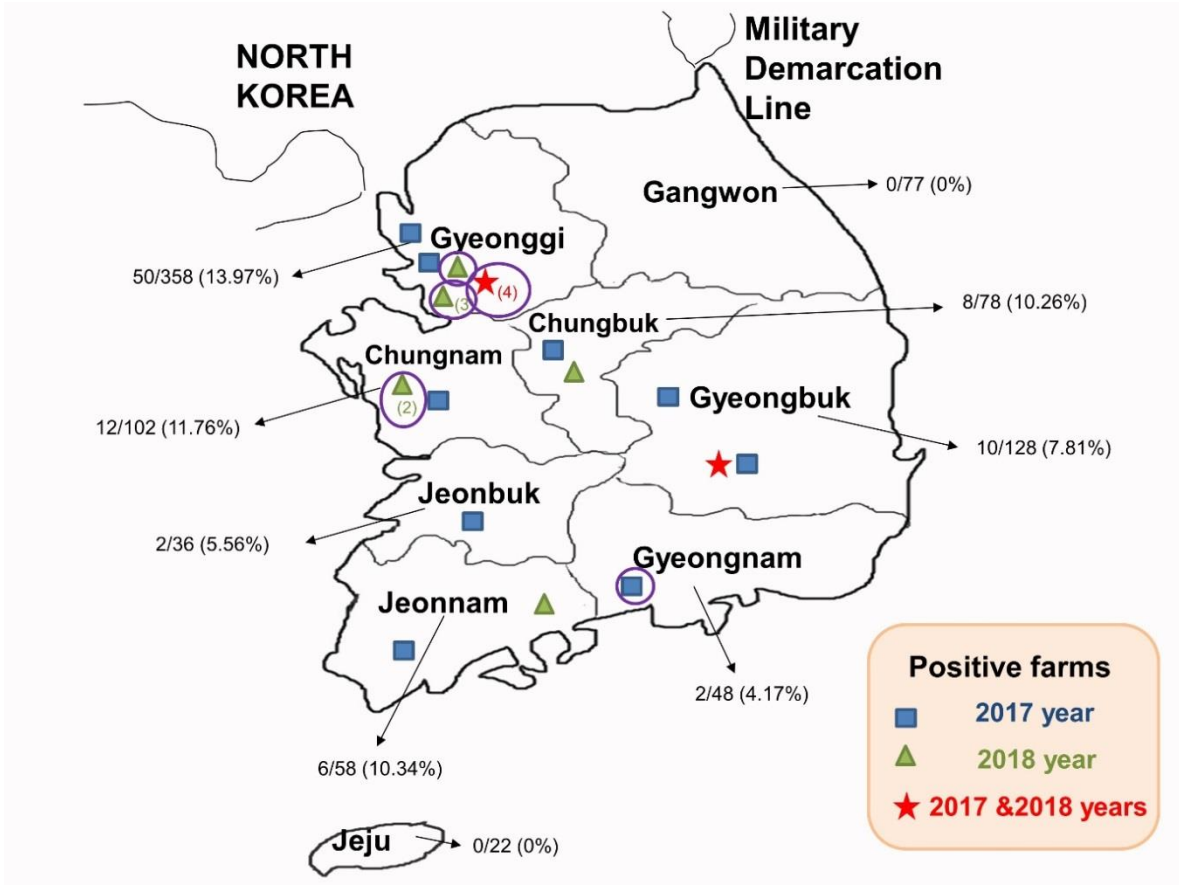


Figure 7. Distribution of total PEDV-positive swine farms in nine provinces of South Korea from 2017 to 2018. Marks (square, triangle, and star) indicated the locations of the swine farms. If the swine farms were located in the adjacent spot, the number of these swine farms was indicated in parentheses below the marks. The sample numbers (positive samples/total samples) and detection rate of each province are indicated by black arrows. The purple circles indicate the PEDV-positive farms where seven Korean PEDV strains sequenced in this study originated.

3.2. Phylogenetic analysis of global PEDV strains

In the phylogenetic trees inferred from the D1–D2 datasets of the S gene (Figure 8), the *PEDV* strains were classified into five sub-genogroups (G1a, G1b, G2a, G2b, and G2c), which were previously designated [28]. However, the relationships between sub-genogroups differed. The D1 dataset contains the complete S gene supported for the clusters of (G1a, G1b) and (G2c, (G2a, G2b)). The D2 dataset contains the S gene without recombinant regions supported for the different clusters of (G1b, G2c) and (G2b, G2a). In both datasets, histograms of pairwise *p*-distances exhibited a discrete distribution betwen sub-genogroups. That was in agreement with the tree topologies and posterior support values at the nodes to each sub-genogroup (Figure 8). In both the D1 and D2 datasets, the seven Korean strains identified in this study (S6, S10, S12, S14, S97, S100, and Y178) were clustered within sub-genogroup G2a.

Supported by high posterior probability values (0.90–1), the phylogenetic trees inferred from the D3–D4 datasets of the N gene (Figure 8) suggested that the classification of *PEDV* strains into four S gene-based sub-genogroups G1a, G1b, G2b, G2a/G2c was more reliable. In both datasets, it was observed that the S gene-based G2a and G2c were not monophyletic (pink branches). Differing from the S gene-based phylogenies (Figure 8), the Ngene-based trees with or without the elimination of recombinant regions had identical topologies of (G1a, (G1b, (G2b, G2a/G2c))). As the result, this study proposed an N-based genotyping of *PEDV* as N1, N2, N3b and N3a, which were equivalent to the S-based genotyping of G1a, G1b, G2b, and G2a/G2c, respectively. That classification was also supported by a clear bimodal distribution of genetic distance between the proposed genogroups N1–N2 and N1–N3a/N3b (inserted histograms, Figure 9).

According to that scheme, the seven Korean strains identified in this study (S6, S10, S12, S14, S97, S100, and Y178) were within sub-genogroup N3a.

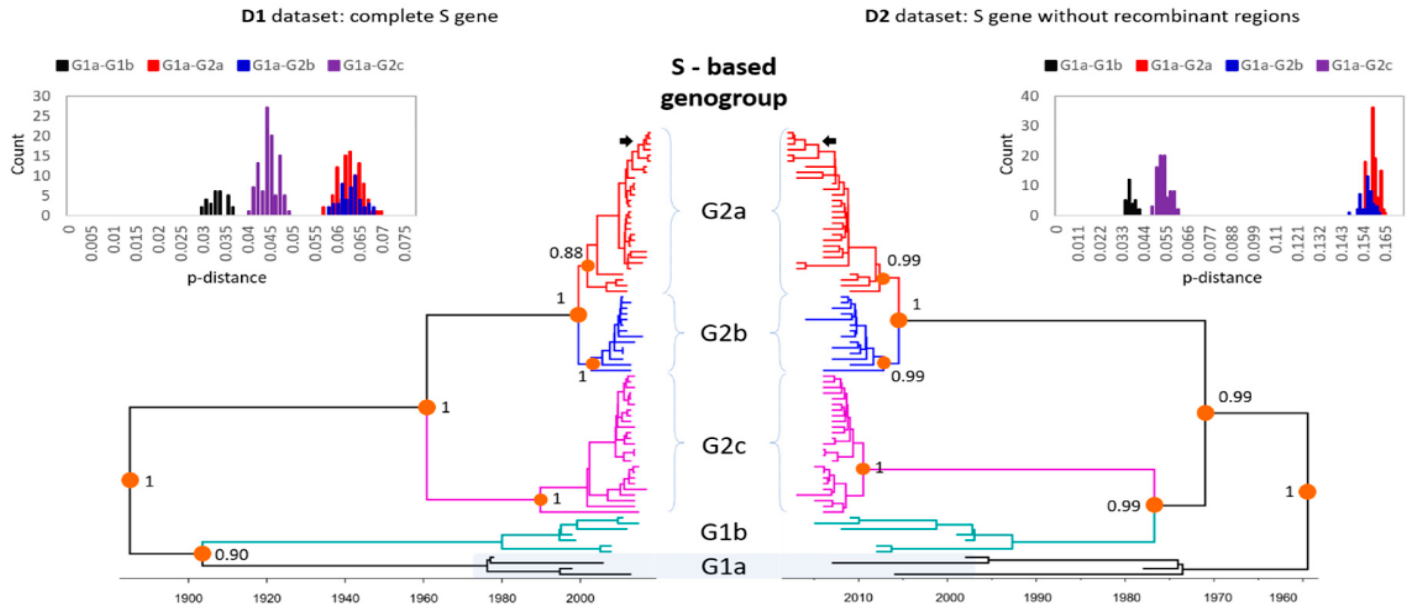


Figure 8. The time-scale maximum clade credibility phylogeny of global PEDV strains based on the S gene. The phylogenetic trees were constructed based on the S gene without removing the recombinant regions (D1) and with the recombinant regions removed (D2). The sub-genogroups were designated as G1a, G1b, G2a, G2b, and G2c. Each sub-genogroup was colored consistently between the D1 and D2 datasets. The inserted histograms of pairwise *p*-distances were between sub-genogroups G1a–G1b, G1a–G2a, G1a–G2b, and G1a–G2c. The *p*-distance is calculated by dividing the number of nucleotide differences by the total number of nucleotides compared. After removing recombinant regions, some sequences might contain large deletions. In other words, the total number of nucleotides became smaller. Thus, the *p*-distance on the right panel was larger than that on the left panel. The Korean strains identified in this study are marked by arrows.

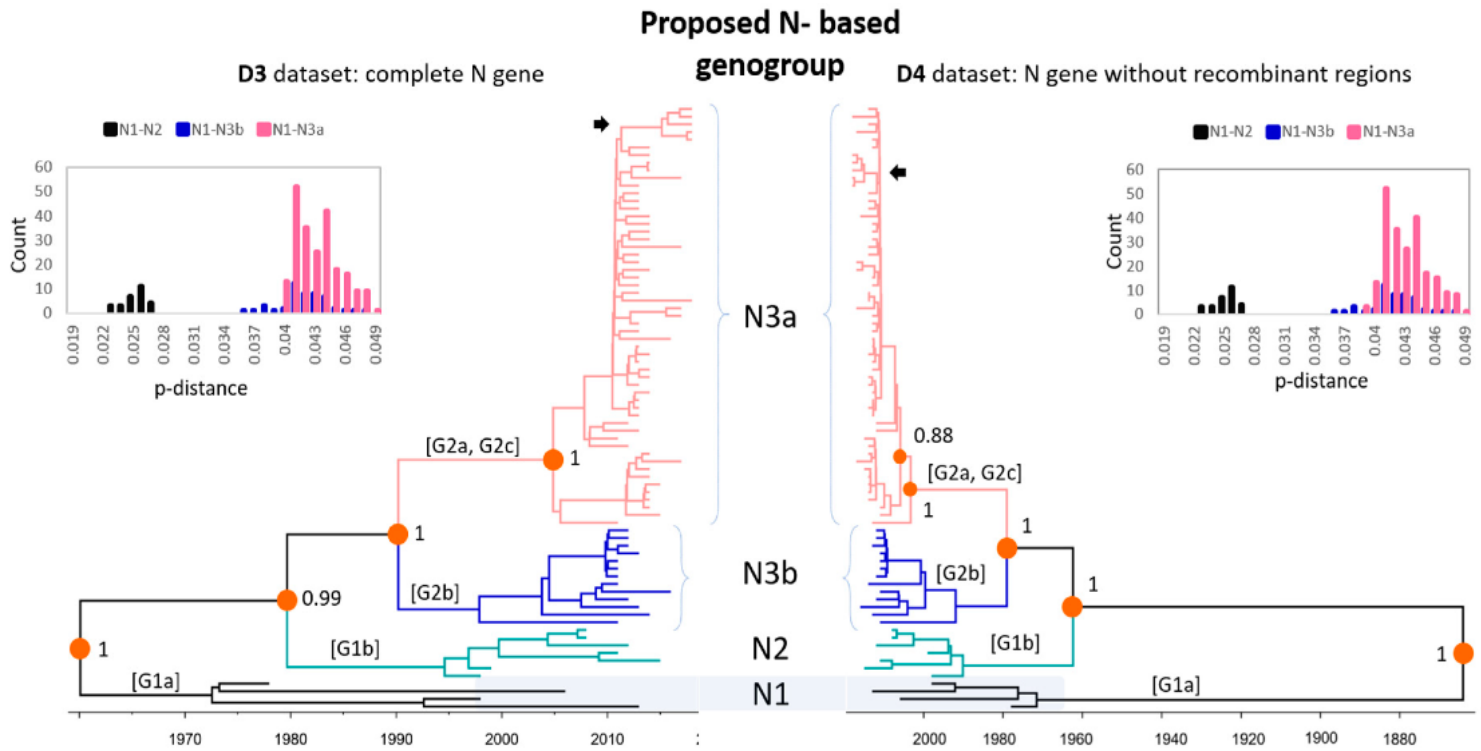


Figure 9. The time-scale maximum clade credibility phylogeny of global PEDV strains based on the nucleocapsid (N) gene. The phylogenetic trees were constructed based on the N gene without removing the recombinant regions (D3) and with the recombinant regions removed (D4). The spike (S) gene-based sub-genogroups were designated as G1a, G1b, G2a, G2b, and G2c. Each sub-genogroup was colored consistently with the D1–D4 datasets. The inserted panels are histograms of pairwise *p*-distances between geno-, sub-genogroups N1–N2 and N1–N3a/N3b. The Korean strains identified in this study are marked by arrows.

3.3. Evolutionary rates of PEDV genes

The estimated mean nucleotide substitutions of S and N genes were at the order of 10^{-4} nucleotide substitutions/ site/ year (Table 3). The substitution rates of complete S and N genes were not significantly different as the 95% highest posterior density (HPD) overlapped (Table 3). Inferring from two datasets with the recombinant regions were removed, the S gene showed substantial higher substitution rates than the N gene because of non-overlapping 95% HPD (6.18×10^{-4} - 10.02×10^{-4} vs. 2.12×10^{-4} - 4.52×10^{-4}, respectively).

3.4. Amino acids and antigenic index analysis of N gene sequences

Several consistent changes allowed the differentiation of the genogroups and sub-genogroups based on the N gene (Figure 10). Genogroup N1 differed from genogroups N2 and N3 by 7 non-synonymous substitutions (A84G, K205N, M216V, P381L, Q395L, N398H, and V408A). Unique changes leading to branches (N2, (N3a, N3b)) were A142T, H242L, Q397L, and E400D. Genogroup N2 was further characterized by two non-synonymous substitutions (A145T, K380I). Finally, the main branch leading to sub-genogroup N3a displayed five changes (K123N, M216V, R241K, K252R, and N255S).

In the antigenic index analysis of N gene sequences, significant reductions were found in amino acid positions 122–126 according to the genogroups (Figure 11). Genogroup N3 exhibited lower antigenic indexes, below 0.5 (the cut-off value), compared with genogroups N1 and N2. These amino acid sequences were located in a B-cell epitope sequence of the N protein, which is one of the PEDV-specific epitopes (amino acids 18–133 and 252–262) previously identified [123].

3.5. Antigenic index analysis of S protein B-cell epitopes

When analyzing the Korean PEDV strains, newly identified in this study, based on the three major B-cell epitope sequences of the S protein (COE, S1D, and 2C10), the six strains exhibited significantly lower antigenic indexes in some parts of the B-cell epitope sequences (Figure 12). The strains S6, S10, S12, and S100 exhibited lower antigenic indexes, below the cut-off value of 0.5, in amino acid positions 623–627 of the COE region, which contained the non-synonymous change K623N. The S97 strain exhibited lower antigenic indexes, below the cut-off value of 0.5, in amino acid positions 634–638 in the COE region, and the non-synonymous change E635V was observed in this region. The Y178 strain also showed antigenic indexes below 0.5 at amino acids 764–771 of the S1D region, which contained the non-synonymous change S768F. It is likely that each of the mentioned non-synonymous changes affected the antigenic indexes of the adjacent regions.

Table 3. Estimated nucleotide substitution rates for S and N genes.

Dataset	Data best fit molecular clock*	Data best fit tree prior	Geometric mean rate (x 10^{-4})**	95% HPD interval (x 10^{-4})***
D1: Complete S gene	RLC	Constant	5.23	3.81 - 6.53
D2: S gene without recombinant regions	RLC	Exponential	7.98	6.18 - 10.02
D3: Complete N gene	RLC	Constant	6.58	4.34 - 9.03
D4: N gene without recombinant regions	RLC	BSP	3.24	2.12 - 4.52

* Random local clock model (RLC)

** The geometric mean nucleotide substitution rate (substitutions/site/year) was inferred from the data best fit molecular clock and coalescent tree prior of Constant population size, Exponential population size and Bayesian skyline plot (BSP)

*** Highest posterior density (HPD)

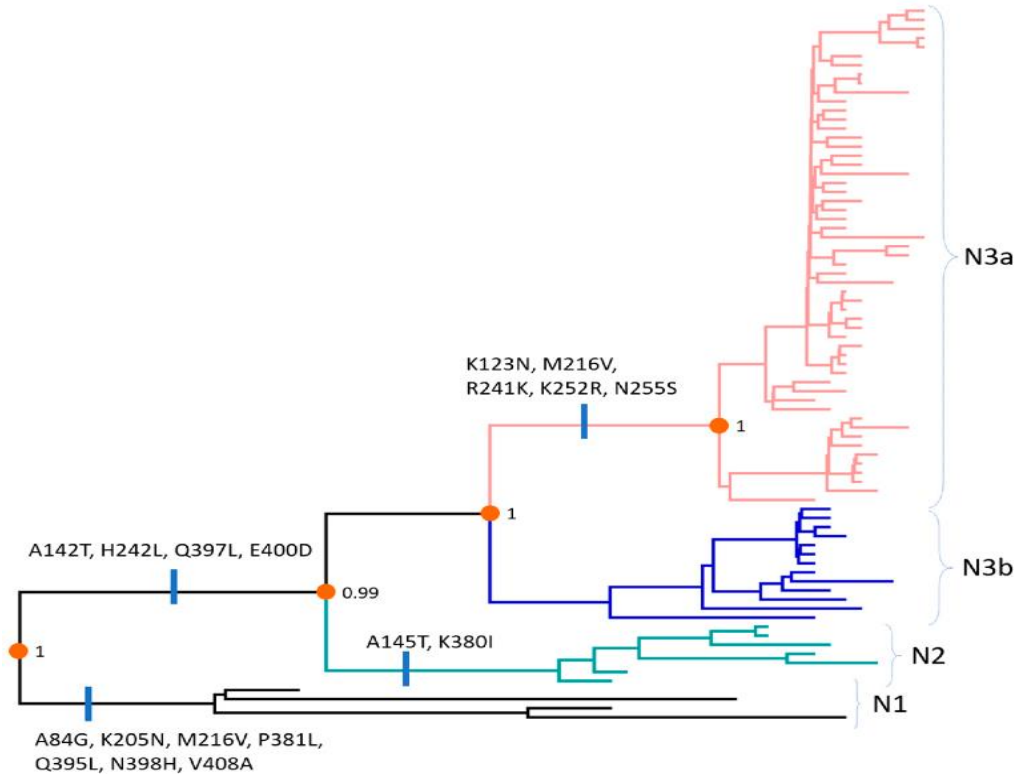


Figure 10. The maximum clade credibility tree based on the N gene with reconstructed non-synonymous substitutions were mapped to the branches of the phylogeny. For clarity, posterior values were shown for main separating nodes.

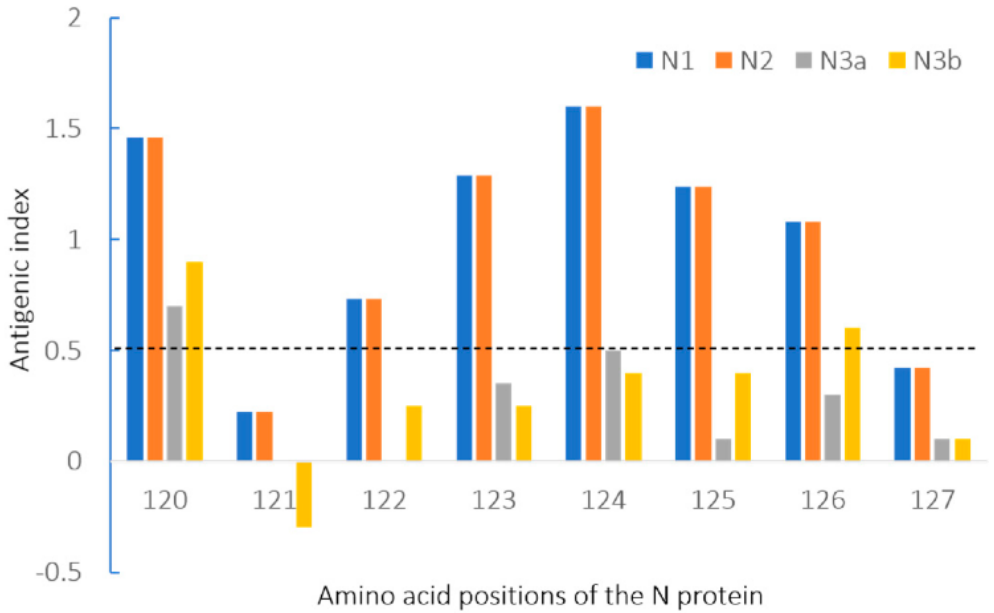


Figure 11. Antigenic index analysis of N gene sequences in PEDV strains. Genogroup N3 exhibited lower antigenic indexes, below 0.5 (cut-off value), compared with genogroups N1 and N2. This region is in a B-cell epitope sequence of the N protein, which is one of the PEDV-specific epitopes (amino acids 18–133 and 252–262).

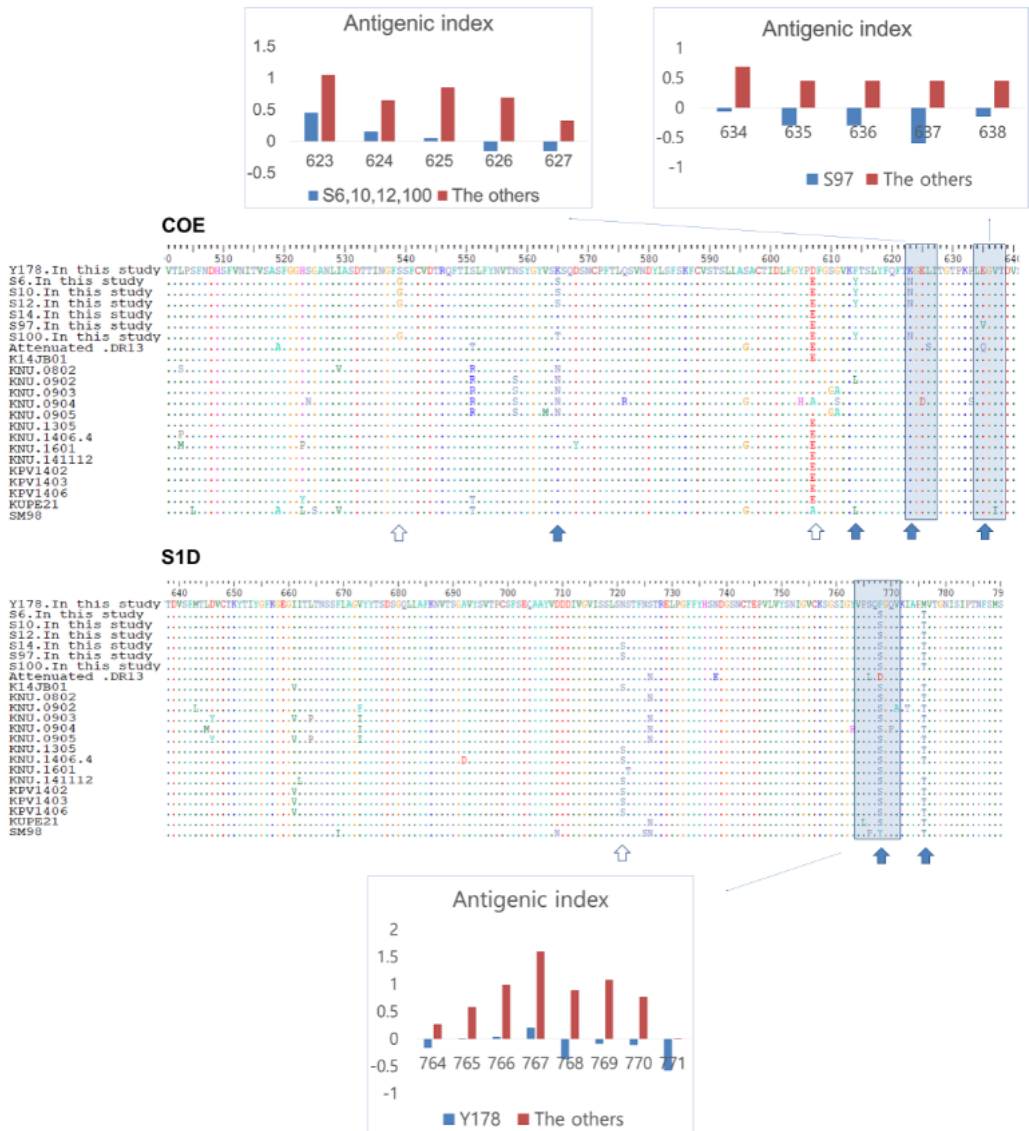


Figure 12. Antigenic index analysis of S protein B-cell epitopes in Korean PEDV strains. Strains S6, S10, S12, and S100 exhibited lower antigenic indexes, below the cut-off value of 0.5, at amino acids 623–627 of the COE region. Strain S97 exhibited reduced antigenic indexes, below 0.5, at amino acids 634–638 of the COE region. Strain Y178 exhibited reduced antigenic indexes, below 0.5, at amino acids 764–771 in the S1D region.

4. Discussion

First of all, this study reflected the ongoing circulation of PEDV in Korea at about 10% (Table 1) and its wide distribution (Figure 7). At the same time, obtaining genomic sequences of PEDV from field samples provided good opportunity for studying the genetic evolution of the virus. As a result, this study applied multiple bioinformatics tools to investigate the genetic diversity of PEDV.

Based on phylogenetic analysis of the complete S gene or S gene excluding recombinant regions, global PEDV strains (Figure 8) can be classified into two genogroups, and each genogroup may be further subdivided into two (G1a and G1b) and three (G2a, G2b, and G2c) sub-genogroups. This result was consistent with that of Guo et al. [28]. Additionally, Guo et al. reported that all these different subgroups existed in Korea, with the most prevalent subgroup being G2a, when they analyzed the Korean PEDV strains identified prior to 2016. Similar to the previous data, all Korean field strains identified from 2017 to 2018 in this study were included in subgroup G2a, indicating that the most prevalent Korean subgroup has not changed since 2016.

Genogroups G1 and G2 are known to have different S protein neutralization activities [24]. However, differences within the genogroups have not been investigated. The six Korean strains (S6, S10, S12, S97, S100, and Y178) had significantly reduced antigenic indexes compared with other Korean strains in some parts of COE [120] and S1D [121], which encode the B-cell epitopes of S protein. These antigenic index reductions may induce somewhat different S protein antigenicities, even within the same genogroup. In fact, strains S10, S12, S97, and S100 originated from swine farms that had suckling piglet mortalities of almost 100%. These swine farms had been regularly using killed

vaccines containing the new genogroup (G2) PEDV strain but were seriously damaged by porcine epidemic diarrhea.

In the literature, the N gene had been used to infer the phylogenetic relationships of PEDV strains [124, 125]. It was noteworthy that the N gene showed a similar evolutionary rate to the S gene in this study. This high evolutionary rate implies that the N gene, as well as the S gene, is likely to have high genetic diversity; accordingly, several sub-genogroups could have diverged. Comparing to the S-based phylogenetic topology, the classical strains (G1 strains) presented the same topology in the N-based and S-based analysis. However, there were some differences in the classification of subgroups on the new genotype strains (G2 strains). The N3 genogroup consisting of the G2 strains was divided into only 2 subgroups (N3a and N3b) not following the S-based subgroups (G2a, G2b and G2c). Specifically, the N3b strains were consistent with the G2b strains, but the G2a and G2c strains were grouped into the same subgroup, N3a, in the N-based topology. This classification of the three genogroups (N1, N2, and N3) was also supported by the consistent variation in the antigenic indexes depending on the genogroups. The number of antigenic indexes of the N3 strains significantly decreased compared to those of the N1 and N2 strains within amino acids 122–126, which code for the B-cell epitope of the N protein [126].

PEDV N protein has an important immunological aspect. Abundant antibodies against N protein are induced at the early stages of PEDV infection [52, 98]. Furthermore, N protein is considered to play an important role in inducing cell-mediated immunity [52]. Because of these features, N protein is commonly used as a target for diagnosis and vaccine development [53, 54]. As mentioned above, several sub-genogroups based on the

N gene were identified in this study. This genetic diversity may change their antigenicities according to their geno- or sub-genogroup (Figure 4). In the immunological diagnosis of PEDV, commercial PEDV ELISA kits have been showing poor performance, and the results of neutralizing assays using cell culture sometimes mismatch with those of the ELISA assays. In fact, Chang et al. recently reported that the antibodies induced by the G2b PEDV strain poorly reacted with a commercial N-based ELISA kit, which showed a sensitivity of 37% [57], which may be the result of antigenicity differences between the genogroups. However, further study is required to validate this hypothesis. Indeed, if there are antigenic differences between the genogroups, a combination of N proteins derived from both genogroups would be required for the accurate immunological diagnosis of PEDV.

Overall, this study revealed that PEDV displayed genetic diversity in both S and N genes which resulted in the divergence into different sub-genogroups and altered antigenic indexes. Such PEDV mutants derived from genetic mutations of the S and N genes may cause severe damage to swine farms because of their ability to evade the unprepared host immune systems.

General conclusions

Viruses have continued to fight with host immune through genetic mutation facilitating immune evasion and this strategy of viruses for their survival will continue in the future. Also, genetic variations, which may change structural form of viral proteins by non-synonymous changes, can hamper diagnostic accuracy. Thus, it is very important work to investigate and track critical genetic events along with virus evolution. These efforts can give worthy information and insight to establish appropriate prevent and diagnostic strategy for viruses. In this study, coronaviruses causing sever disease in human and porcine were investigated by genetic and phylogenetic analysis.

1. Chapter I: Severe acute respiratory syndrome virus type 2 (SARS-CoV-2)

The S glycoprotein of coronaviruses is important for viral entry and pathogenesis with most variable sequences. Therefore, we analyzed the S gene sequences of SARS-CoV-2 to better understand the antigenicity and immunogenicity of this virus in this study. In phylogenetic analysis, two subtypes (SARS-CoV-2a and -b) were confirmed within SARS-CoV-2 strains. These two subtypes were divided by a novel non-synonymous mutation of D614G. This may play a crucial role in the evolution of SARS-CoV-2 to evade the host immune system. The region containing this mutation point was confirmed as a B-cell epitope located in the S1 domain, and SARS-CoV-2b strains exhibited severe reduced antigenic indexes compared to SARS-CoV-2a in this area. This may allow these two subtypes to have different antigenicity. If the two subtypes

have different serological characteristics, a vaccine for both subtypes will be more effective to prevent COVID-19. Thus, further study is urgently required to confirm the antigenicity of these two subtypes.

2. Chapter II: Porcine epidemic diarrhea virus (PEDV)

PEDV strains were classified into three genogroups based on nucleocapsid (N) gene (N1, N2 and N3). The antigenic indexes of genogroup N3 strains were significantly lower compared with those of genogroups N1 and N2 strains in the B-cell epitope of the nucleocapsid protein. Indeed, there is different antigenicity between the genogroups based on the N gene, it may affect diagnostic results using commercial ELISA kits based on N1 protein.

PEDV displayed genetic diversity in both S and N genes which resulted in the divergence into different sub-genogroups and altered antigenic indexes. Such PEDV mutants derived from genetic mutations of the S and N genes may cause severe damage to swine farms because of their ability to evade the unprepared host immune systems.

In conclusion, the crucial genetic variations, which may induce immune evasion or diagnostic error, were revealed in the viruses originated from various species. It is expected that these results provide better understanding for preventing viral infection and more precise diagnosis. Also, constant surveillances through genetic analysis should maintain to appropriately respond to virus evolution in the future.

References

1. Koonin, E.V., T.G. Senkevich, and V.V. Dolja, *The ancient Virus World and evolution of cells.* Biology direct, 2006. **1**(1): p. 29.
2. Breitbart, M. and F. Rohwer, *Here a virus, there a virus, everywhere the same virus?* Trends in microbiology, 2005. **13**(6): p. 278-284.
3. Domingo, E., et al., *Basic concepts in RNA virus evolution.* The FASEB Journal, 1996. **10**(8): p. 859-864.
4. Villarreal, L.P., *Viruses and the evolution of life*. 2005: ASM press.
5. Hampson, A.W., *Influenza virus antigens and 'antigenic drift'*, in *Perspectives in medical virology*. 2002, Elsevier. p. 49-85.
6. Earn, D.J., J. Dushoff, and S.A. Levin, *Ecology and evolution of the flu.* Trends in ecology & evolution, 2002. **17**(7): p. 334-340.
7. Boutwell, C.L., et al., *Viral evolution and escape during acute HIV-1 infection.* The Journal of infectious diseases, 2010. **202**(Suppl 2): p. S309.
8. Sun, J., et al., *COVID-19: Epidemiology, Evolution, and Cross-Disciplinary Perspectives.* Trends Mol Med, 2020. **26**(5): p. 483-495.
9. Alanagreh, L., F. Alzoughool, and M. Atoum, *The Human Coronavirus Disease COVID-19: Its Origin, Characteristics, and Insights into Potential Drugs and Its Mechanisms.* Pathogens, 2020. **9**(5).
10. Debouck, P. and M. Pensaert, *Experimental infection of pigs with a new porcine enteric coronavirus, CV 777.* American journal of veterinary research, 1980. **41**(2): p. 219-223.

11. Li, W., et al., *New variants of porcine epidemic diarrhea virus, China, 2011.* Emerging infectious diseases, 2012. **18**(8): p. 1350.

12. Cunningham, C., H. Zhu, and D. Hillis, *Best-fit maximum-likelihood models for phylogenetic inference: empirical tests with known phylogenies.* Evolution, 1998. **52**(4): p. 978-987.

13. Pompei, S., V. Loreto, and F. Tria, *Phylogenetic properties of RNA viruses.* PLoS One, 2012. **7**(9): p. e44849.

14. Nasir, A. and G. Caetano-Anollés, *A phylogenomic data-driven exploration of viral origins and evolution.* Science advances, 2015. **1**(8): p. e1500527.

15. Monchatre-Leroy, E., et al., *Identification of alpha and beta coronavirus in wildlife species in France: bats, rodents, rabbits, and hedgehogs.* Viruses, 2017. **9**(12): p. 364.

16. Jaimes, J.A., et al., *A tale of two viruses: the distinct spike glycoproteins of feline coronaviruses.* Viruses, 2020. **12**(1): p. 83.

17. Luk, H.K., et al., *Molecular epidemiology, evolution and phylogeny of SARS coronavirus.* Infection, Genetics and Evolution, 2019. **71**: p. 21-30.

18. Cui, J., F. Li, and Z.-L. Shi, *Origin and evolution of pathogenic coronaviruses.* Nature Reviews Microbiology, 2019. **17**(3): p. 181-192.

19. Walls, A.C., et al., *Structure, function, and antigenicity of the SARS-CoV-2 spike glycoprotein.* Cell, 2020.

20. Liu, C., A. von Brunn, and D. Zhu, *Cyclophilin A and CD147: novel therapeutic targets for the treatment of COVID-19.* Med Drug Discov, 2020. **7**: p. 100056.

21. Pensaert, M. and P. De Bouck, *A new coronavirus-like particle associated with diarrhea in swine.* Archives of virology, 1978. **58**(3): p. 243-247.

22. Kocherhans, R., et al., *Completion of the porcine epidemic diarrhoea coronavirus (PEDV) genome sequence.* Virus genes, 2001. **23**(2): p. 137-144.

23. Lee, C., *Porcine epidemic diarrhea virus: an emerging and re-emerging epizootic swine virus.* Virology journal, 2015. **12**(1): p. 193.

24. Liu, J., et al., *Neutralization of genotype 2 porcine epidemic diarrhea virus strains by a novel monoclonal antibody.* Virology, 2017. **507**: p. 257-262.

25. Henzel, A., et al., *Genetic and phylogenetic analyses of capsid protein gene in feline calicivirus isolates from Rio Grande do Sul in southern Brazil.* Virus Res, 2012. **163**(2): p. 667-71.

26. Cima, G., *PED virus reinfecting US herds. Virus estimated to have killed 7 million-plus pigs.* Journal of the American Veterinary Medical Association, 2014. **245**(2): p. 166.

27. Vlasova, A.N., et al., *Distinct characteristics and complex evolution of PEDV strains, North America, May 2013–February 2014.* Emerging infectious diseases, 2014. **20**(10): p. 1620.

28. Guo, J., et al., *Evolutionary and genotypic analyses of global porcine epidemic diarrhea virus strains.* Transboundary and emerging diseases, 2019. **66**(1): p. 111-118.

29. Zambon, M.C., *Epidemiology and pathogenesis of influenza*. Journal of Antimicrobial Chemotherapy, 1999. **44**(suppl_2): p. 3-9.

30. Drake, J.W., et al., *Rates of spontaneous mutation.* Genetics, 1998. **148**(4): p. 1667-1686.

31. Steinhauer, D.A., E. Domingo, and J.J. Holland, *Lack of evidence for proofreading mechanisms associated with an RNA virus polymerase*. Gene, 1992. **122**(2): p. 281-288.

32. Roberts, J.D., K. Bebenek, and T.A. Kunkel, *The accuracy of reverse transcriptase from HIV-1*. Science, 1988. **242**(4882): p. 1171-1173.

33. Radding, C.M., *Homologous pairing and strand exchange in genetic recombination.* Annual review of genetics, 1982. **16**(1): p. 405-437.

34. Pérez-Losada, M., et al., *Recombination in viruses: mechanisms, methods of study, and evolutionary consequences.* Infection, Genetics and Evolution, 2015. **30**: p. 296-307.

35. Weller, S.K. and D.M. Coen, *Herpes simplex viruses: mechanisms of DNA replication.* Cold Spring Harbor perspectives in biology, 2012. **4**(9): p. a013011.

36. Johansson, C. and S. Schwartz, *Regulation of human papillomavirus gene expression by splicing and polyadenylation.* Nature reviews Microbiology, 2013. **11**(4): p. 239-251.

37. Kotin, R.M., R.M. Linden, and K.I. Berns, *Characterization of a preferred site on human chromosome 19q for integration of adeno-associated virus DNA by non-homologous recombination.* The EMBO journal, 1992. **11**(13): p. 5071-5078.

38. Ball, L.A., *High-frequency homologous recombination in vaccinia virus DNA.* Journal of virology, 1987. **61**(6): p. 1788-1795.

39. Nagy, P.D. and J.J. Bujarski, *Efficient system of homologous RNA recombination in brome mosaic virus: sequence and structure requirements and accuracy of crossovers.* Journal of Virology, 1995. **69**(1): p. 131-140.

40. Nagy, P.D. and J.J. Bujarski, *Homologous RNA recombination in brome mosaic virus: AU-rich sequences decrease the accuracy of crossovers.* Journal of virology, 1996. **70**(1): p. 415-426.

41. Alejska, M., et al., *Two types of non-homologous RNA recombination in brome mosaic virus.* Acta Biochimica Polonica, 2005. **52**(4): p. 833-844.

42. Baron, S., *Alphaviruses (Togaviridae) and Flaviviruses (Flaviviridae)--Medical Microbiology*. 1996: University of Texas Medical Branch at Galveston.

43. Smith, E.C., N.R. Sexton, and M.R. Denison, *Thinking outside the triangle: replication fidelity of the largest RNA viruses.* Annual Review of Virology, 2014. **1**: p. 111-132.

44. Taubenberger, J.K. and J.C. Kash, *Influenza virus evolution, host adaptation, and pandemic formation.* Cell host & microbe, 2010. **7**(6): p. 440-451.

45. Cuevas, J.M., P. Domingo-Calap, and R. Sanjuán, *The fitness effects of synonymous mutations in DNA and RNA viruses.* Molecular biology and evolution, 2012. **29**(1): p. 17-20.

46. Shields, D.C., et al., *" Silent" sites in Drosophila genes are not neutral: evidence of selection among synonymous codons.* Molecular biology and evolution, 1988. **5**(6): p. 704-716.

47. Gupta, A.M., J. Chakrabarti, and S. Mandal, *Non-synonymous mutations of SARS-CoV-2 leads epitope loss and segregates its variants.* Microbes and infection, 2020.

48. Parrish, C.R., et al., *Natural variation of canine parvovirus.* Science, 1985. **230**(4729): p. 1046-8.

49. Decaro, N., et al., *Genetic analysis of canine parvovirus type 2c.* Virology, 2009. **385**(1): p. 5-10.

50. Hoelzer, K. and C.R. Parrish, *The emergence of parvoviruses of carnivores.* Vet Res, 2010. **41**(6): p. 39.

51. López de Turiso, J.A., et al., *Fine mapping of canine parvovirus B cell epitopes.* J Gen Virol, 1991. **72 (Pt 10)**: p. 2445-56.

52. Saif, L.J., *Coronavirus immunogens.* Veterinary microbiology, 1993. **37**(3-4): p. 285-297.

53. Song, D. and B. Park, *Porcine epidemic diarrhoea virus: a comprehensive review of molecular epidemiology, diagnosis, and vaccines.* Virus genes, 2012. **44**(2): p. 167-175.

54. Hou, X.-L., L.-Y. Yu, and J. Liu, *Development and evaluation of enzyme-linked immunosorbent assay based on recombinant nucleocapsid protein for detection of porcine epidemic diarrhea (PEDV) antibodies.* Veterinary microbiology, 2007. **123**(1-3): p. 86-92.

55. Pesente, P., et al., *Phylogenetic analysis of ORF5 and ORF7 sequences of porcine reproductive and respiratory syndrome virus (PRRSV) from PRRS-positive Italian farms: a showcase for PRRSV epidemiology and its consequences on farm management.* Veterinary microbiology, 2006. **114**(3-4): p. 214-224.

56. Seuberlich, T., et al., *Nucleocapsid protein-based enzyme-linked immunosorbent assay for detection and differentiation of antibodies against European and North American porcine reproductive and respiratory syndrome virus.* Clinical and Diagnostic Laboratory Immunology, 2002. **9**(6): p. 1183-1191.

57. Chang, C.-Y., et al., *Development and comparison of enzyme-linked immunosorbent assays based on recombinant trimeric full-length and*

truncated spike proteins for detecting antibodies against porcine epidemic diarrhea virus. BMC veterinary research, 2019. **15**(1): p. 421.

58. Radford, A.D., et al., *Feline calicivirus infection: ABCD guidelines on prevention and management.* Journal of feline medicine and surgery, 2009. **11**(7): p. 556-564.

59. Gaskell, R., et al., *Isolation of felid herpesvirus I from the trigeminal ganglia of latently infected cats.* Journal of general virology, 1985. **66**(2): p. 391-394.

60. Sykes, J.E., et al., *Detection of feline calicivirus, feline herpesvirus 1 and Chlamydia psittaci mucosal swabs by multiplex RT-PCR/PCR.* Veterinary microbiology, 2001. **81**(2): p. 95-108.

61. Berger, A., et al., *Feline calicivirus and other respiratory pathogens in cats with Feline calicivirus-related symptoms and in clinically healthy cats in Switzerland.* BMC veterinary research, 2015. **11**(1): p. 282.

62. Whiley, D.M. and T.P. Sloots, *Sequence variation in primer targets affects the accuracy of viral quantitative PCR.* Journal of Clinical Virology, 2005. **34**(2): p. 104-107.

63. Kim, S.J., Y.H. Park, and K.T. Park, *Development of a novel reverse transcription PCR and its application to field sample testing for feline calicivirus prevalence in healthy stray cats in Korea.* Journal of Veterinary Science, 2020. **21**(5).

64. Khan, K.A. and P. Cheung, *Presence of mismatches between diagnostic PCR assays and coronavirus SARS-CoV-2 genome.* Royal Society Open Science, 2020. **7**(6): p. 200636.

65. Woo, P.C., et al., *Coronavirus genomics and bioinformatics analysis.* viruses, 2010. **2**(8): p. 1804-1820.

66. Martínez, N., et al., *Molecular and phylogenetic analysis of bovine coronavirus based on the spike glycoprotein gene.* Infection, Genetics and Evolution, 2012. **12**(8): p. 1870-1878.

67. Chung, H.-C., et al., *Molecular characterization of a Korean porcine epidemic diarrhea virus strain NB1.* Canadian Journal of Veterinary Research, 2019. **83**(2): p. 97-103.

68. Li, J., et al., *Game consumption and the 2019 novel coronavirus.* The Lancet Infectious Diseases, 2020. **20**(3): p. 275-276.

69. Zehender, G., et al., *Genomic characterization and phylogenetic analysis of SARS-COV-2 in Italy.* Journal of Medical Virology, 2020.

70. Abraham, S., et al., *Deduced sequence of the bovine coronavirus spike protein and identification of the internal proteolytic cleavage site.* Virology, 1990. **176**(1): p. 296-301.

71. Menachery, V.D., et al., *A SARS-like cluster of circulating bat coronaviruses shows potential for human emergence.* Nature medicine, 2015. **21**(12): p. 1508-1513.

72. Sánchez, C.M., et al., *Genetic evolution and tropism of transmissible gastroenteritis coronaviruses.* Virology, 1992. **190**(1): p. 92-105.

73. Du, L., et al., *Antigenicity and immunogenicity of SARS-CoV S protein receptor-binding domain stably expressed in CHO cells.* Biochemical and biophysical research communications, 2009. **384**(4): p. 486-490.

74. Kirchdoerfer, R.N., et al., *Pre-fusion structure of a human coronavirus spike protein.* Nature, 2016. **531**(7592): p. 118-121.

75. Wang, X., et al., *Immunogenicity and antigenic relationships among spike proteins of porcine epidemic diarrhea virus subtypes G1 and G2.* Archives of virology, 2016. **161**(3): p. 537-547.

76. Ntafis, V., et al., *Canine coronavirus, Greece. Molecular analysis and genetic diversity characterization.* Infection, Genetics and Evolution, 2013. **16**: p. 129-136.

77. Nguyen, L.-T., et al., *IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies.* Molecular biology and evolution, 2015. **32**(1): p. 268-274.

78. Hoang, D.T., et al., *UFBoot2: improving the ultrafast bootstrap approximation.* Molecular biology and evolution, 2018. **35**(2): p. 518-522.

79. Jespersen, M.C., et al., *BepiPred-2.0: improving sequence-based B-cell epitope prediction using conformational epitopes.* Nucleic acids research, 2017. **45**(W1): p. W24-W29.

80. Chou, P.Y. and G.D. Fasman, *Prediction of protein conformation.* Biochemistry, 1974. **13**(2): p. 222-245.

81. Kolaskar, A.S. and P.C. Tongaonkar, *A semi-empirical method for prediction of antigenic determinants on protein antigens.* FEBS letters, 1990. **276**(1-2): p. 172-174.

82. Parker, J., D. Guo, and R. Hodges, *New hydrophilicity scale derived from high-performance liquid chromatography peptide retention data: correlation of predicted surface residues with antigenicity and X-ray-derived accessible sites.* Biochemistry, 1986. **25**(19): p. 5425-5432.

83. Jameson, B. and H. Wolf, *The antigenic index: a novel algorithm for predicting antigenic determinants.* Bioinformatics, 1988. **4**(1): p. 181-186.

84. Pereira, C.A., E.S. Leal, and E.L. Durigon, *Selective regimen shift and demographic growth increase associated with the emergence of high-fitness variants of canine parvovirus.* Infection, Genetics and Evolution, 2007. **7**(3): p. 399-409.

85. Zaki, A.M., et al., *Isolation of a novel coronavirus from a man with pneumonia in Saudi Arabia.* New England Journal of Medicine, 2012. **367**(19): p. 1814-1820.

86. Kubo, H., Y.K. Yamada, and F. Taguchi, *Localization of neutralizing epitopes and the receptor-binding site within the amino-terminal 330 amino acids of the murine coronavirus spike protein.* Journal of Virology, 1994. **68**(9): p. 5403-5410.

87. Luo, Z. and S.R. Weiss, *Roles in cell-to-cell fusion of two conserved hydrophobic regions in the murine coronavirus spike protein.* Virology, 1998. **244**(2): p. 483-494.

88. Li, F., *Structure, function, and evolution of coronavirus spike proteins.* Annual review of virology, 2016. **3**: p. 237-261.

89. Ballesteros, M., C. Sanchez, and L. Enjuanes, *Two amino acid changes at the N-terminus of transmissible gastroenteritis coronavirus spike protein result in the loss of enteric tropism.* Virology, 1997. **227**(2): p. 378-388.

90. Jiang, S., Y. He, and S. Liu, *SARS vaccine development.* Emerging infectious diseases, 2005. **11**(7): p. 1016.

91. Nagata, S. and I. Pastan, *Removal of B cell epitopes as a practical approach for reducing the immunogenicity of foreign protein-based therapeutics.* Advanced drug delivery reviews, 2009. **61**(11): p. 977-985.

92. Lin, Y.-M., et al., *Naturally occurring hepatitis B virus B-cell and T-cell epitope mutants in hepatitis B vaccinated children.* The Scientific World Journal, 2013. **2013**.

93. Wilson, S., et al., *Vaccination of dogs with canine parvovirus type 2b (CPV-2b) induces neutralising antibody responses to CPV-2a and CPV-2c.* Vaccine, 2014. **32**(42): p. 5420-5424.

94. Kweon, C., et al., *Isolation of porcine epidemic diarrhea virus (PEDV) in Korea.* Korean J Vet Res, 1993. **33**(2): p. 249-54.

95. Chung, H.-C., et al., *Isolation of porcine epidemic diarrhea virus during outbreaks in South Korea, 2013–2014.* Emerging infectious diseases, 2015. **21**(12): p. 2238.

96. Wang, Q., et al., *Emerging and re-emerging coronaviruses in pigs.* Current opinion in virology, 2019. **34**: p. 39-49.

97. Jung, K. and L.J. Saif, *Porcine epidemic diarrhea virus infection: etiology, epidemiology, pathogenesis and immunoprophylaxis.* The Veterinary Journal, 2015. **204**(2): p. 134-143.

98. Lai, M.M. and D. Cavanagh, *The molecular biology of coronaviruses*, in *Advances in virus research*. 1997, Elsevier. p. 1-100.

99. Utiger, A., et al., *Identification of the membrane protein of porcine epidemic diarrhea virus.* Virus Genes, 1995. **10**(2): p. 137-148.

100. Li, W., et al., *Cellular entry of the porcine epidemic diarrhea virus.* Virus research, 2016. **226**: p. 117-127.

101. Van Diep, N., et al., *Molecular characterization of US-like and Asian non-S INDEL strains of porcine epidemic diarrhea virus (PEDV) that circulated in Japan during 2013–2016 and PEDVs collected from recurrent outbreaks.* BMC veterinary research, 2018. **14**(1): p. 96.

102. Su, Y., et al., *Detection and phylogenetic analysis of porcine epidemic diarrhea virus in central China based on the ORF3 gene and the S1 gene.* Virology journal, 2016. **13**(1): p. 192.

103. Ding, Z., et al., *Porcine epidemic diarrhea virus nucleocapsid protein antagonizes beta interferon production by sequestering the interaction between IRF3 and TBK1.* Journal of virology, 2014. **88**(16): p. 8936-8945.

104. Xu, X., et al., *Porcine epidemic diarrhea virus N protein prolongs S-phase cell cycle, induces endoplasmic reticulum stress, and up-regulates interleukin-8 expression.* Veterinary microbiology, 2013. **164**(3-4): p. 212-221.

105. Shi, D., et al., *Molecular characterizations of subcellular localization signals in the nucleocapsid protein of porcine epidemic diarrhea virus.* Viruses, 2014. **6**(3): p. 1253-1273.

106. Chung, H.-C., et al., *New emergence pattern with variant porcine epidemic diarrhea viruses, South Korea, 2012–2015.* Virus research, 2016. **226**: p. 14-19.

107. Song, D.S., et al., *Multiplex reverse transcription-PCR for rapid differential detection of porcine epidemic diarrhea virus, transmissible gastroenteritis virus, and porcine group A rotavirus.* Journal of Veterinary Diagnostic Investigation, 2006. **18**(3): p. 278-281.

108. Gao, Y., et al., *Phylogenetic analysis of porcine epidemic diarrhea virus field strains prevailing recently in China.* Archives of virology, 2013. **158**(3): p. 711-715.

109. Bouckaert, R., et al., *BEAST 2.5: An advanced software platform for Bayesian evolutionary analysis.* PLoS computational biology, 2019. **15**(4): p. e1006650.

110. Miller, M.A., W. Pfeiffer, and T. Schwartz. *Creating the CIPRES Science Gateway for inference of large phylogenetic trees*. in *2010 gateway computing environments workshop (GCE)*. 2010. Ieee.

111. Bouckaert, R.R. and A.J. Drummond, *bModelTest: Bayesian phylogenetic site model averaging and model comparison.* BMC evolutionary biology, 2017. **17**(1): p. 42.

112. Drummond, A.J., et al., *Relaxed phylogenetics and dating with confidence.* PLoS Biol, 2006. **4**(5): p. e88.

113. Drummond, A.J. and M.A. Suchard, *Bayesian random local clocks, or one rate to rule them all.* BMC biology, 2010. **8**(1): p. 1-12.

114. Drummond, A.J., et al., *Bayesian coalescent inference of past population dynamics from molecular sequences.* Molecular biology and evolution, 2005. **22**(5): p. 1185-1192.

115. Rambaut, A., et al., *Posterior summarization in Bayesian phylogenetics using Tracer 1.7.* Systematic biology, 2018. **67**(5): p. 901.

116. Baele, G., et al., *Improving the accuracy of demographic and molecular clock model comparison while accommodating phylogenetic uncertainty.* Molecular biology and evolution, 2012. **29**(9): p. 2157-2167.

117. Kumar, S., et al., *MEGA2: molecular evolutionary genetics analysis software.* Bioinformatics, 2001. **17**(12): p. 1244-5.

118. Hsiao, K.-L., et al., *New phylogenetic groups of torque teno virus identified in eastern Taiwan indigenes.* PloS one, 2016. **11**(2).

119. Yang, Z., *PAML 4: phylogenetic analysis by maximum likelihood.* Mol Biol Evol, 2007. **24**(8): p. 1586-91.

120. Fan, J.-H., et al., *Heterogeneity in membrane protein genes of porcine epidemic diarrhea viruses isolated in China.* Virus genes, 2012. **45**(1): p. 113-117.

121. Sun, D., et al., *Spike protein region (aa 636789) of porcine epidemic diarrhea virus is essential for induction of neutralizing antibodies.* Acta virologica, 2007. **51**(3): p. 149-156.

122. Cruz, D.J.M., C.-J. Kim, and H.-J. Shin, *The GPRLQPY motif located at the carboxy-terminal of the spike protein induces antibodies that neutralize Porcine epidemic diarrhea virus.* Virus research, 2008. **132**(1-2): p. 192-196.

123. Wang, K., et al., *The identification and characterization of two novel epitopes on the nucleocapsid protein of the porcine epidemic diarrhea virus.* Scientific reports, 2016. **6**(1): p. 1-14.

124. Li, Z., et al., *Sequence and phylogenetic analysis of nucleocapsid genes of porcine epidemic diarrhea virus (PEDV) strains in China.* Archives of virology, 2013. **158**(6): p. 1267-1273.

125. Chen, J., et al., *Genetic variation of nucleocapsid genes of porcine epidemic diarrhea virus field strains in China.* Archives of virology, 2013. **158**(6): p. 1397-1401.

126. Wang, K., et al., *The Identification and Characterization of Two Novel Epitopes on the Nucleocapsid Protein of the Porcine Epidemic Diarrhea Virus.* Sci Rep, 2016. **6**: p. 39010.

국문 초록

사람 코로나바이러스 (SARS-CoV-2)와 돼지 코로나바이러스 (PEDV)의 유전학적 분석과 유전적 변이가 바이러스 항원성과 진단에 영향을 미칠 가능성

김 성 재

(지도교수: 박 용 호)

서울대학교 대학원 수의학과

수의미생물학 전공

바이러스는 면역 회피를 유발할 수 있는 유전적 돌연변이를 통해 숙주 면역과 계속해서 싸우고 있으며, 이러한 바이러스의 생존 전략은 앞으로도 계속 될 것이다. 특히, 유전자 변이에 따른 아미노산 서열의 비상동성 (non-synonymous) 변화는 바이러스 에피톱의 항원성을 변화시킬 수 있으며, 이러한 변화는 기존에 개발된 백신의 방어능을 저하시킬 수 있다. 또한, 바이러스의 염기서열의 변이는 현재 일상적으로 사용되는 진단 기술인 중합효소연쇄반응 (PCR)과 효소결합면역흡착분석법 (ELISA)의 진단 정확도를 저해 할 수 있다. 따라서 바이러스가 진화함에 따른 중요한 그들의 유전학적 변화를 조사하고 추적하는 것은 바이러스에 대한 적절한 예방 및 진단 전략을 수립하는 데 매우 큰 도움이 된다. 이 연구에서는 현재 사람과 돼지에서 심각한 문제가 되고 있는 코로나바이러스의 유전적 변이와 그 변이들이 바이러스의 항원성과 진단에 영향을 미칠 가능성에

대해 조사하였다.

첫 번째 장에서는 최근 사람에서 문제가 되고 있는 severe acute respiratory syndrome coronavirus type 2 (SARS-CoV-2)를 분석하였다. 코로나바이러스의 스파이크 (S) 단백질은 바이러스의 세포 내 유입에 결정적인 역할을 하는 표면 단백질이다. 따라서, 이 연구에서는 항원성과 면역학적 특징을 확인하기 위해 SARS-CoV-2의 S 유전자를 분석하였다. S 유전자를 기반으로 한 계통학적 분석에서 SARS-CoV-2 분리주들 사이에 두 개의 유전자 그룹이 존재하는 것을 확인하였다. 이 두 개의 유전자 그룹은 하나의 특이적인 염기서열 변이인 D614G에 의해 나뉘었다. 이 변이는 SARS-CoV-2가 숙주의 면역체계를 회피하는데 결정적인 역할을 할 것으로 생각되었다. D614G 염기서열 변이를 포함하는 S1 domain의 에피톱 부위에 대해 항원 지수 분석을 시행한 결과, SARS-CoV-2b 유전자 그룹이 SARS-CoV-2a 유전자 그룹에 비해 유의적으로 감소한 항원 지수를 보이는 것으로 확인하였다. 따라서, 이 유전적 변이에 의해 두 유전자 그룹간의 항원성 차이가 발생하였을 것으로 생각되었다. 두 유전자 그룹간 항원성 차이가 발생하였다면 두 유전자 그룹을 백신에 포함시키는 것이 COVID-19을 방어하는데 보다 효율적일 것이다. 그러므로, 실제로 두 유전자 그룹간 항원성 차이가 발생하였는지 확인하는 것이 시급하다.

두 번째 장에서는 전 세계 돼지 산업에 지속적이고 심각한 피해를 입히고 있는 돼지유행성설사바이러스 (PEDV)를 분석하였다. 최근 양돈장의 PEDV 유병률은 약 9.92 %로 지속적으로 문제가 되고 있음이 확인되었다. 뉴클레오캡시드 (N) 유전자를 기반으로 베이지안 계통 분석을 진행한 결과, 세 개의 주요 N 유전자 기반 유전자 그룹 (N1, N2 및 N3)과 두 개의 하위 유전자 그룹 (N3a과 N3b) 을 확인하였다. N 단백질에 포함된 에피톱 부분의 항원 지수를 분석한 결과, 유전자 그룹간 항원성에 차이가 있을

것으로 강하게 의심되었다. 에피톱 부위에서 N3 유전자 그룹의 항원 지수는 N1 및 N2 유전자 그룹의 항원 지수에 비해 유의하게 낮았다. 이러한 변화는 N1 단백질을 항원으로 사용하는 ELISA 키트의 진단 결과에 영향을 미칠 것으로 판단되었다. 또한, 최근 확인된 한국 PED 바이러스들의 S 유전자를 분석한 결과, 스파이크 단백질 (COE, S1D 및 2C10)의 B 세포 에피토프 서열의 일부에서 유의적으로 낮은 항원 지수가 확인되었다. 이러한 S 및 N 유전자의 면역학적 주요 부위에 유전적 변이가 발생한 PED 바이러스들은 기존에 확립된 숙주 면역을 회피하여 돼지 농장에 심각한 손상을 줄 수 있기 때문에 지속적인 감시가 필요하다.

이 연구에서 바이러스의 면역 회피나 진단 오류를 유발할 수 있는 중요한 유전적 변이를 현재 심각한 문제가 되고 있는 사람과 돼지 코로나바이러스에서 확인하였다. 이러한 발견은 바이러스 감염 예방에 대한 더 나은 이해와 보다 정확한 진단법을 개발하는데 도움을 줄 것으로 기대한다. 나아가 향후 코로나바이러스 진화에 적절히 대응할 수 있도록 유전자 분석을 통한 지속적인 감시가 유지되어야 한다.

주요어: 코로나바이러스, 유전학적 변이, 항원성, 진단, PEDV, SARS-CoV-2

학번: 2016-30968