Preventive Veterinary Medicine 123 (2016) 175–184

Contents lists available at ScienceDirect

Preventive Veterinary Medicine

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

Genomic and evolutionary inferences between American and global strains of porcine epidemic diarrhea virus

Matthew C. Jarvis[a], Ham Ching Lam[a], Yan Zhang[b], Leyi Wang[b], Richard A. Hesse[c], Ben M. Hause[c], Anastasia Vlasova[d], Qiuhong Wang[d], Jianqiang Zhang[e], Martha I. Nelson[f], Michael P. Murtaugh[g], Douglas Marthaler[a,*]

[a] Department of Veterinary Population Medicine, College of Veterinary Medicine, University of Minnesota, St. Paul, MN, USA
[b] Animal Disease Diagnostic Laboratory, Ohio Department of Agriculture, Reynoldsburg, OH, USA
[c] Kansas State Veterinary Diagnostic Laboratory and Department of Diagnostic Medicine and Pathobiology, Kansas State University, Manhattan, KS, USA
[d] Food Animal Health Research Program, Department of Veterinary Preventive Medicine, Ohio Agricultural Research and Development Center, College of Food Agriculture and Environmental Sciences, The Ohio State University, Wooster, OH, USA
[e] Department of Veterinary Diagnostic and Production Animal Medicine, College of Veterinary Medicine, Iowa State University, Ames, IA, USA
[f] Division of International Epidemiology and Population Studies, Fogarty International Center, National Institutes of Health, Bethesda, MD, USA
[g] Department of Veterinary and Biomedical Sciences, College of Veterinary Medicine, University of Minnesota, St. Paul, MN, USA

ARTICLE INFO

Article history:
Received 2 April 2015
Received in revised form
29 September 2015
Accepted 22 October 2015

Keywords:
Porcine epidemic diarrhea virus
Molecular analysis
Bayesian analysis
Complete genome

ABSTRACT

Porcine epidemic diarrhea virus (PEDV) has caused severe economic losses both recently in the United States (US) and historically throughout Europe and Asia. Traditionally, analysis of the spike gene has been used to determine phylogenetic relationships between PEDV strains. We determined the complete genomes of 93 PEDV field samples from US swine and analyzed the data in conjunction with complete genome sequences available from GenBank ($n = 126$) to determine the most variable genomic areas. Our results indicate high levels of variation within the ORF1 and spike regions while the C-terminal domains of structural genes were highly conserved. Analysis of the Receptor Binding Domains in the spike gene revealed a limited number of amino acid substitutions in US strains compared to Asian strains. Phylogenetic analysis of the complete genome sequence data revealed high rates of recombination, resulting in differing evolutionary patterns in phylogenies inferred for the spike region versus whole genomes. These finding suggest that significant genetic events outside of the spike region have contributed to the evolution of PEDV.

1. Introduction

Porcine epidemic diarrhea virus (PEDV) causes diarrhea, vomiting, and dehydration, leading to high mortality (up to 100%) in suckling piglets. PEDV was first discovered in the United Kingdom in 1971, and later was found in Belgium, Hungary, France, Italy, and the Czech Republic (Chasey and Cartwright, 1978; Fan et al., 2012). In 1986, PEDV was first reported in China, and proceeded to spread throughout Asia (Cui, 1990; Song and Park, 2012). In late 2010, a "variant" PEDV strain with increased pathogenesis compared to the classical strains was reported in China, and is regarded as the first of the pandemic strains (Sun et al., 2012). Highly pathogenic PEDV was first detected in the United States (US) in April 2013, and rapidly spread to 31 states, as well as Mexico and Canada (Oka et al., 2014; Ojkic et al., 2015). In December 2013, a second PEDV strain OH851, initially called "US variant" (later renamed S-INDEL), was detected in the US, which shared the same insertions and deletions in the S1 region as the classical PEDV strain, CV777 (Wang et al., 2014; EFSA AHAW Panel, 2014). The S-INDEL strains were reported with lower virulence in the field and were later identified in June 2013 upon a retroactive study (Vlasova et al., 2014). To date, the highly virulent PEDV has been reported in Colombia, Dominican Republic, Japan, Peru, Germany, Portugal, South Korea, and Ukraine (EFSA AHAW Panel, 2014; Lee and Lee, 2014; Hanke et al., 2015; Murakami et al., 2015; Mesquita et al., 2015).

* Corresponding author at: Veterinary Population Medicine, College of Veterinary Medicine, University of Minnesota, 1333 Gortner Ave, Saint Paul, MN 55108, USA.
E-mail address: marth027@umn.edu (D. Marthaler).

http://dx.doi.org/10.1016/j.prevetmed.2015.10.020

PEDV is a single-stranded, positive sense RNA virus belonging to the family *Coronaviridae*, genus *Alphacoronavirus*. The PEDV genome is approximately 28 kb in length and roughly two-thirds of the genome consists of open reading frame (ORF) 1, which encodes 16 non-structural proteins (nsps) (Lai et al., 2013). These nsps play important roles in viral replication, post-translational processing, and immune evasion (Lai et al., 2013). The virus produces various structural proteins, including spike, membrane, and nucleocapsid (Lai et al., 2013). The spike protein is crucial to cell attachment and infection, and the envelope is an integral membrane protein, aiding in membrane fusion while the nucleocapsid protein is necessary for genomic packaging (Hagemeijer and de Haan, 2015). In addition, the PEDV genome includes ORF3, located between the spike and membrane genes, that encodes an ion channel protein possibly associated with PEDV pathogenesis (Park et al., 2008; Wang et al., 2012).

Researchers have explored various regions of the coronavirus (CoV) genome to link specific areas with virulence and host cell attachment. For example, the spike gene codes for a viral attachment protein that can be divided into the S1 (1-789 aa) and S2 (790-1383 aa) regions (Song and Park, 2012). Comparative analysis of transmissible gastroenteritis virus (TGEV), porcine respiratory coronavirus (PRCV), and murine hepatitis virus (MHV) revealed two main antigenic sites in the S1 region: the N-terminal domain (NTD) and the C-terminal receptor binding domain (RBD) (Li et al., 2007). While both domains can influence virus infectivity, such as in TGEV, one domain tends to be central to a CoV's tropism: the NTD is important for MHV tropism, and the RBD is central to PEDV infectivity and virulence (Reguera et al., 2012). The NTD can bind to various sialic acids on the host cell surface (Reguera et al., 2012). The RBD contains residues that bind to the porcine aminopeptidase-N (pAPN), the host receptor utilized by TGEV and PEDV (Delmas et al., 1992).

Since the last large-scale North American PEDV outbreak ended in the spring of 2014, the complete genomes of 93 PEDV strains from the US were sequenced and analyzed to further understand the origin and phylogenetic relationships among the American and global PEDV strains. In-depth nucleotide and amino acid analysis was conducted to identify genes of high diversity. Bayesian analysis was performed to understand the evolution of PEDV and the emergence of different clades within US strains. In addition, the RBD was modeled to visualize the differences between American and Asian strains to better understand how changes in the RBD might affect vaccine efficacy and development.

2. Materials and methods

2.1. Sample acquisition

Samples were routinely submitted to the University of Minnesota Veterinary Diagnostic Laboratory (UMVDL) for pathogen detection. Between January 2014 and December 2014 samples were screened for PEDV by real time RT-PCR (Vlasova et al., 2014). Samples for complete genome sequencing were selected based on the criteria of a high viral concentration from the RT-PCR results and geographical diversity within the US. A total of 83 samples, including fecal ($n = 38$), intestinal homogenate ($n = 21$), fecal swab ($n = 10$), oral fluid ($n = 5$), feedback ($n = 4$), and environmental ($n = 5$) samples were selected for complete genome sequencing using next generation sequencing (NGS) techniques as previously described (GenBank numbers KR265759–KR265834, KR265840–KR265846) (Marthaler et al., 2013; Marthaler et al., 2014). Whole genomic PEDV sequences obtained using NGS techniques were also generously supplied from Iowa State University ($n = 7$, GenBank numbers KM975735–KM975741) and the Ohio Department of Agriculture ($n = 3$, GenBank numbers KP641661–KP641663), using previously described methods (Wang et al., 2014; Chen et al., 2014).

2.2. Nucleotide and amino acid analysis

Using the complete PEDV genome sequences from this study ($n = 93$) and the available PEDV sequences from GenBank ($n = 126$), two nucleotide alignments were created and analyzed to determine the phylogenetic relationships between American and global PEDV sequences: the concatenation of all ORFs (ORF1, S, ORF3, envelope, membrane, and nucleocapsid), and a S1 alignment. Vaccine and cell-passaged strains were excluded from the analysis (Table S1). Nucleotide and amino acid entropy analyses were performed using the MATLAB software (MATLAB v8.0 and Statistics Toolbox v8.1, The MathWorks, Inc., Natick, MA, USA). Threshold values were determined using previously published methods (Shannon, 1948; Litwin and Jores, 1992).

2.3. Recombination and bayesian analysis (BEAST)

Recombination analysis was performed using the Recombination Detection Program (RDP) v4, which uses multiple detection algorithms, including the RDP Method, GENECOV, and MAXCHI, to check for the presence of recombinant sequences in the sequence dataset (Martin et al., 2015). Window size was set to 100 bp. Breakpoints, the presence of major/minor donor sequences, and confidence intervals were used to determine regions that required excision from the alignment, or if entire sequences needed to be removed from the analysis due to multiple recombination events within the sequence. Recombinant sequences were removed only prior to the Bayesian analysis, but remained in the alignments for all entropy analysis and molecular modeling.

Bayesian Markov Chain Monte Carlo (MCMC) approach using BEAST v1.8.1, with a relaxed molecular clock and Bayesian skyline population (BSP) prior, with a general-time reversible nucleotide substitution and gamma distributed among-site rate variation was used to infer time-scaled phylogeny (Drummond et al., 2002, 2005, 2006, 2012Drummond and Rambaut, 2007; Minin et al., 2008; Drummond and Suchard, 2010). The MCMC chain was run for 800 million generations, with sub-sampling every 80,000 iterations. A Maximum clade credibility (MCC) tree was created by discarding the initial 10% of the chains and summarized in TreeAnnotator (v.1.8.0). Key nodes were identified using FigTree (v.1.4.2) to determine time to most recent common ancestor (tMRCA).

2.4. Molecular modeling

The putative pAPN receptor-binding residues were analyzed to determine residue trends between classical and pandemic strains (Reguera et al., 2012). The C-terminal RBD within the S1 region of the spike gene was modeled using the open-source modeling server SWISS-MODEL provided by the Swiss Institute of Bioinformatics (Biasini et al., 2014). Predicted tertiary structure of the PEDV pAPN RBD was modeled using PRCV as a template since a PEDV template was not available. Spike monomer and trimer models were developed using a theoretical SARS-CoV model as a template (Bernini et al., 2004). Illustrations were created using the open-source Java-based molecular viewer Jmol (Herraez, 2006) and the Python-based molecular viewer PyMOL (The PyMOL Molecular Graphics System, Version 1.7.4 Schrödinger, LLC.).