Global molecular genetic analysis of porcine circovirus type 2 (PCV2) sequences confirms the presence of four main PCV2 genotypes and reveals a rapid increase of PCV2d

Chao-Ting Xiao,1 Patrick G. Halbur1 and Tanja Opriessnig1,2

1Department of Veterinary Diagnostic and Production Animal Medicine, College of Veterinary Medicine, Iowa State University, Ames, IA, USA
2Roslin Institute and The Royal (Dick) School of Veterinary Studies, University of Edinburgh, Midlothian, UK

The oldest porcine circovirus type 2 (PCV2) sequence dates back to 1962 and is among several hundreds of publicly available PCV2 sequences. Despite this resource, few studies have investigated the global genetic diversity of PCV2. To evaluate the phylogenetic relationship of PCV2 strains, 1680 PCV2 open reading frame 2 (ORF2) sequences were compared and analysed by methods of neighbour-joining, maximum-likelihood, Bayesian inference and network analysis. Four distinct clades were consistently identified and included PCV2a, PCV2b, PCV2c and PCV2d; the $\rho$-distance between PCV2d and PCV2b was $0.055 \pm 0.008$, larger than the PCV2 genotype-definition cut-off of 0.035, supporting PCV2d as an independent genotype. Among the 1680 sequences, 278–285 (16.5–17 %) were classified as PCV2a, 1007–1058 (59.9–63 %) as PCV2b, three (0.2 %) as PCV2c and 322–323 (19.2 %) as PCV2d, with the remaining 12–78 sequences (0.7–4.6 %) classified as intermediate clades or strains by the various methods. Classification of strains to genotypes differed based on the number of sequences used for the analysis, indicating that sample size is important when determining classification and assessing PCV2 trends and shifts. PCV2d was initially identified in 1999 in samples collected in Switzerland, now appears to be widespread in China and has been present in North America since 2012. During 2012–2013, 37 % of all investigated PCV2 sequences from US pigs were classified as PCV2d and overall data analysis suggests an ongoing genotype shift from PCV2b towards PCV2d. The present analyses indicate that PCV2d emerged approximately 20 years ago.

The genome of PCV2 contains 1766–1768 nt (Guo et al., 2010; Hamel et al., 1998; Meehan et al., 1998) with at least four ORFs. ORF1 encodes two proteins associated with replication designated Rep and Rep’ (Cheung, 2003; Mankertz et al., 1998), ORF2 encodes the capsid protein (Nawagitgul et al., 2000), ORF3 encodes a protein which is thought to play a role in apoptosis (Liu et al., 2005) and ORF4 encodes a newly discovered protein with a role in suppressing caspase activity and regulation of CD4$^+$ and CD8$^+$ T lymphocytes (He et al., 2013).

Since the late 1990s, PCV2 has emerged as a major pig pathogen around the globe. Although PCV2 is a small DNA virus and as such would be expected to be rather conserved, PCV2 isolates display substantial genetic variations (Beach & Meng, 2012; Patterson & Opriessnig, 2010; Segales et al., 2013). The rate of nucleotide substitution for PCV2 has been estimated to be in the order of $1.2 \times 10^{-3}$

---

The GenBank/EMBL/DDBJ accession numbers for the PCV2 ORF2 sequences identified are KM190983–KM191128.

Six supplementary tables and two supplementary figures are available with the online Supplementary Material.
substitutions per site and year, which is the highest recorded substitution rate for a single-stranded DNA virus, allowing PCV2 to maintain evolutionary dynamics closer to single-stranded RNA viruses than to double-stranded DNA viruses (Firth et al., 2009).

Due to the rapidly increasing numbers of reported PCV2 strains, a unified nomenclature for PCV2, using small letters to indicate subgenotypes, was proposed by European and North American research groups in 2008 (Segalés et al., 2008). PCV2 strains were subdivided into phylogenetic groups or genotypes based on the proportion of different nucleotide sites in ORF2 between two sequences (p-distance) by using a distance cut-off of 0.035 (Segalés et al., 2008) and three main genotypes were identified: PCV2a, PCV2b and PCV2c. PCV2a can now be subdivided into four clusters (2A to 2D) (Olvera et al., 2007) and, until 2000, was the dominant PCV2 genotype in the global pig population. PCV2b can now be subdivided into three clusters (1A to 1C) (Olvera et al., 2007) and has been present in Europe and Asia since 1997 (Dupont et al., 2008); however, it was not detected in North America until 2005 (Beach & Meng, 2012). PCV2c was identified in archived swine serum samples collected in 1980, 1987 and 1990 from Denmark (Dupont et al., 2008) and has not been detected in other samples or locations since that time.

In 2009 two new PCV2 genotypes, PCV2d and PCV2e, were identified based on analysis of 40 Chinese PCV2 sequences collected between 2004 and 2008 and 56 PCV2 sequences obtained from GenBank (Wang et al., 2009). However, when the analysis was repeated by adding more than 200 ORF2 or complete PCV2 genome sequences from GenBank, only PCV2a, PCV2b and PCV2c could be differentiated (Cortey et al., 2011). Moreover, PCV2 isolates that initially fell into the proposed new genotype PCV2d were nearly identical to PCV2b subtype 1C strains, while PCV2e isolates clustered together with other PCV2a sequences (Cortey et al., 2011; Olvera et al., 2007). In 2010, several unique PCV2 strains with an ORF2 length of 705 nt instead of 702, resulting from a mutation in the stop codon, and 234 amino acids instead of 233 were identified in Asia (Guo et al., 2010). These strains showed a close relationship with PCV2b subtype 1C and PCV2d and were subsequently designated PCV2d (Guo et al., 2010). Later the same research group renamed these strains as PCV2 mutants (Guo et al., 2012). Similar viruses were also detected in North America in 2012 and in South America in 2014 and designated mutant PCV2b (mPCV2b) (Opriessnig et al., 2013; Salgado et al., 2014; Xiao et al., 2012).

Phylogenetic analysis is known to be a powerful tool and is now widely used for investigating the evolution of viruses. However, a frustrating aspect of this type of analysis is the large number of different inference methods that can be utilized and the lack of uniformity in the use of these methods by different experts (Huelsenbeck et al., 2001). Generally, there are four main groups of statistical methods for reconstructing phylogenetic trees from molecular data: (1) parsimony methods, (2) distance methods such as neighbour-joining (NJ), (3) likelihood methods such as maximum-likelihood (ML) and (4) Bayesian inference (BI). ML and BI frequently outperform parsimony and distance methods (Huelsenbeck et al., 2001; Olvera et al., 2010) and have become the standard for phylogenetic analyses (Olvera et al., 2010). BI is based on a quantity called ‘the posterior probability of a tree’. This is a combination of the prior probability of phylogeny and the likelihood by Bayes’ theorem and is recognized to be efficient, especially when analysing large phylogenetic trees or using complex evolutionary models, when attempting to detect footprints of natural selection in DNA sequences and when dealing with uncertainty (Huelsenbeck et al., 2001). Simultaneously with phylogenetic trees, phylogenetic networks have been adopted by evolutionary biologists and are used to analyse data that cannot be represented accurately by a tree, including gene lateral transmission and recombination events (Morrison, 2010). The network analysis is considered to complement trees and to increase robustness and importance of results (Bapteste et al., 2013).

It is recognized that PCV2 isolates undergo recombination (Olvera et al., 2007). Although recombination events can occur in ORF2 (Cheung, 2009), they mainly occur within the ORF1 gene. Therefore the ORF2 gene is more suitable as a phylogenetic and epidemiological marker; that is, a phylogenetic analysis based on PCV2 ORF2 is thought to be similar to a whole PCV2 genome analysis (Olvera et al., 2007). In the present study, to investigate the molecular genetic relationship between PCV2 strains on a global basis, with emphasis on recent US strains, we analysed available classic and emergent ORF2 sequences deposited in GenBank until 2014, together with ORF2 sequences obtained from US pigs between 2011 and 2013 with NJ, ML, BI and network analysis. The results provide a global view of the genetic divergence of PCV2 and indicate a high prevalence of the emerging PCV2d in the USA, possibly supporting an ongoing genotype shift.

RESULTS

Phylogenetic analysis and genotype identification

Four major global PCV2 clades defined as PCV2a, PCV2b, PCV2c and PCV2d were consistently delimited by the methods of NJ, ML, Markov chain Monte Carlo algorithms (MCMC) and median-joining (MJ) networks (Fig. 1, Table 1). Of the 1680 sequences, 278–285 (16.5–17 %) were classified as PCV2a, 1007–1058 (59.9–63 %) as PCV2b, three (0.2 %) as PCV2c and 322–323 (19.2 %) as PCV2d.

By using the network analysis, besides the main PCV2a branch, two minor clades were identified which were also classified as PCV2a as the sequences clustered with PCV2a by the methods of NJ and ML (Fig. 1d, Table S1, available for download).
Fig. 1. Evolutionary analysis based on 1680 PCV2 ORF2 sequences. The intermediate (IM) strains and clades that could not be included into the major clades are indicated. The trees were rooted by a PCV1 ORF2 sequence. (a) Neighbour-joining tree based on the $p$-distance model. (b) Maximum-likelihood analysis base on the General Time Reversible (GTR) model with rates of gamma distributed and proportion of invariant sites ($I^* + I$). (c) Maximum clade credibility tree inferred by Bayesian inference with the model of GTR + $I^* + I$. (d) Network analysis indicates the relationship of the defined haplotypes. The red dots indicate the haplotypes that are connected to the neighbour genotypes. The additional coloured dots indicate the haplotypes belonging to different IM clades as defined by other methods.
in the online Supplementary Material). One connected to PCV2d (Table S1) and showed the closest relationship with PCV2d strain S98-305/2 (JX512855) identified in 1998 in Switzerland. The other minor clade connected to the intermediate (IM) 3 clade (described in the intermediate strain section) and corresponded to strains previously reported as PCV2e, PCV2a–2f or recombinants (Table S1).

The PCV2d clade included not only the recently identified PCV2 mutant strains but also strains previously designated PCV2b subtype 1C (Table 2). Notably, from the present phylogenetic data, the PCV2d strains can be traced back to 1998 in Switzerland (JX512855). Furthermore, with the exception of three strains located in the higher order branches (JX512855, FJ644927 and JF827599), two subclades could be recognized within the PCV2d group by the NJ and ML methods (Fig. 2). One subclade includes strains identified between 1999 and 2011 (designated PCV2d-1), while the other subclade includes more recent strains identified between 2006 and 2014 (designated PCV2d-2) (Table S2). Specifically, PCV2d-1 contains 58 strains identified in 11 countries including Belgium, China, Germany, India, Japan, The Netherlands, Romania, Serbia, South Korea, Switzerland and Vietnam. In contrast, PCV2d-2 contains 261 strains identified in China, Romania, South Korea, the USA and Vietnam. The ORF2 amino acid length does not correlate with the PCV2d genotype definition as some PCV2d strains do not encode the additional amino acid in the terminal ORF2 region and have only 233 amino acids instead of 234 (Table 2).

Interestingly, all of the PCV2d strains identified in the USA cluster within PCV2d-2 and show high nucleotide identities of 99.3–100% with a 2008 PCV2d strain from China (HM038017). The present data also indicate that PCV2d-2 is currently the predominant PCV2d subgenotype while PCV2d-1 is a possible ancestor of PCV2d-2.

Table 1. Clustering results of the 1680 PCV2 strains based on the nucleotide sequence analysis of ORF2 using neighbour-joining, maximum-likelihood, Bayesian inference and network analysis methods

<table>
<thead>
<tr>
<th>Method</th>
<th>Number of strains in each of the IM clades</th>
<th>Single IM strains</th>
<th>PCV2a no. (%)</th>
<th>PCV2b no. (%)</th>
<th>PCV2c no. (%)</th>
<th>PCV2d no. (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>IM1</td>
<td>IM2</td>
<td>IM3</td>
<td>IM4</td>
<td>IM5</td>
<td>IM6</td>
</tr>
<tr>
<td>NJ</td>
<td>47</td>
<td>9</td>
<td>5</td>
<td>3</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>ML</td>
<td>47</td>
<td>9</td>
<td>5</td>
<td>3</td>
<td>4</td>
<td>3</td>
</tr>
<tr>
<td>BI</td>
<td>49</td>
<td>9</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>8</td>
</tr>
<tr>
<td>Overall</td>
<td>47–49</td>
<td>9</td>
<td>3–5</td>
<td>3</td>
<td>4</td>
<td>8</td>
</tr>
</tbody>
</table>

–, Not identified.
BI, Bayesian inference; IM, identified intermediate; ML, maximum-likelihood; NJ, neighbour-joining.

Table 2. Location and date of identification of selected PCV2 strains now classified into the PCV2d clade

<table>
<thead>
<tr>
<th>Accession number</th>
<th>Country</th>
<th>Collection or submission date</th>
<th>Length of ORF2 (aa)</th>
<th>Health status</th>
<th>Previous classification</th>
</tr>
</thead>
<tbody>
<tr>
<td>JX512856</td>
<td>Switzerland</td>
<td>1999</td>
<td>233</td>
<td>Healthy</td>
<td>PCV2d (Wang et al., 2009), PCV2b-1C (Olvera et al., 2007)</td>
</tr>
<tr>
<td>AY181946</td>
<td>China</td>
<td>2002</td>
<td>234</td>
<td>Respiratory</td>
<td>PCV2d (Wang et al., 2009)</td>
</tr>
<tr>
<td>AY484410</td>
<td>Netherlands</td>
<td>2003</td>
<td>234</td>
<td>Healthy</td>
<td>PCV2b-1C (Olvera et al., 2007)</td>
</tr>
<tr>
<td>AY713470</td>
<td>Germany</td>
<td>2004</td>
<td>234</td>
<td>Unknown</td>
<td>PCV2b-1C (Knell et al., 2005; Olvera et al., 2007)</td>
</tr>
<tr>
<td>EU450616</td>
<td>South Korea</td>
<td>2006</td>
<td>233</td>
<td>Systemic</td>
<td>PCV2b-1C (Kim et al., 2009)</td>
</tr>
<tr>
<td>GU808525</td>
<td>India</td>
<td>2007</td>
<td>234</td>
<td>Unknown</td>
<td>Not available</td>
</tr>
<tr>
<td>EF990645</td>
<td>Belgium</td>
<td>2008</td>
<td>234</td>
<td>Systemic</td>
<td>PCV2b-1C (Lefebvre et al., 2008)</td>
</tr>
<tr>
<td>JN006443</td>
<td>Romania</td>
<td>2008</td>
<td>234</td>
<td>Unknown</td>
<td>PCV2b (Turcitus et al., 2011)</td>
</tr>
<tr>
<td>JX099786</td>
<td>Vietnam</td>
<td>2008</td>
<td>234</td>
<td>Unknown</td>
<td>PCV2b-1C (Huynh et al., 2014)</td>
</tr>
<tr>
<td>AB462384</td>
<td>Japan</td>
<td>2008</td>
<td>234</td>
<td>Unknown</td>
<td>Not available</td>
</tr>
<tr>
<td>HM038017</td>
<td>China</td>
<td>2008</td>
<td>234</td>
<td>Systemic</td>
<td>PCV2d (Guo et al., 2010)</td>
</tr>
<tr>
<td>HQ378157</td>
<td>Serbia</td>
<td>2010</td>
<td>234</td>
<td>Systemic</td>
<td>PCV2b-1C (Savic et al., 2012)</td>
</tr>
<tr>
<td>JX535296</td>
<td>USA</td>
<td>2012</td>
<td>234</td>
<td>Systemic</td>
<td>mPCV2b (Xiao et al., 2012)</td>
</tr>
<tr>
<td>KJ187306</td>
<td>Brazil</td>
<td>2013</td>
<td>234</td>
<td>Systemic</td>
<td>mPCV2b (Salgado et al., 2014)</td>
</tr>
</tbody>
</table>
For the recent US samples, three distinct clades were consistently identified by the methods of NJ, ML, MCMC and the network analysis (Fig. 3). These clades corresponded to PCV2a (12.6%), PCV2b (52.4%) and PCV2d (35%), whereas PCV2c was not identified (Table S3). The earliest US PCV2d identified was obtained from samples collected in April 2012.

**Intermediate (IM) strains and clades**

From the phylogenetic analysis it became evident that 4.6% (78/1680) of the global PCV2 strains clustered into different genotypes depending on the method utilized (Fig. 1, Tables 1 and S4). These strains were designated IM strains and some have been previously reported to be recombinants. If more than three intermediate strains clustered together, they were classified as an IM clade and if fewer than three strains clustered together, they were regarded as a single IM strain. If more than three intermediate strains clustered together, they were classified as an IM clade and if fewer than three strains clustered together, they were regarded as a single IM strain. In the analysis with NJ method, four IM clades (IM1–IM4) and two single IM strains (JN382187, JN382189) were demarcated. For the ML method, one additional IM clade, IM5, was identified. By the MCMC method only IM1, IM4 and IM6 were identified, whereas by the network method, only IM2 and IM3 were identified (Fig. 1, Tables 1 and S4).

Among all identified IM clades, IM1 contained the most strains with 47 sequences identified by NJ and ML and 49 sequences by BI (MCMC), all identified in Asia (Tables 1 and S4). However, using the network analysis all of the IM1 strains clustered within PCV2b. IM2, defined by methods of NJ, ML and network analysis, included nine strains. IM3 included five sequences by NJ and ML and three sequences using the network analysis. IM4 included three sequences by NJ, ML and BI analysis. IM5 included four strains identified by the ML analysis. Eight sequences belonged to IM6 based on BI analysis (Tables 1 and S4).

**ORF2 sequence comparison**

The previously reported signature motifs in PCV2 ORF2, 5'-ACC/AAC/AAA/ATC/TCT/ATA-3' with the corresponding amino acid sequence TNKISI for PCV2a and 5'-TCA/AAC/CCC/GGC/TCT/GTG-3' and SNPRSV for PCV2b (Cheung et al., 2007), were confirmed in the present study. For PCV2d, the motif 5'-TCA/AAC/CCC/CTC/ACT/GTG-3', corresponding to the amino acid sequence SNPLTV, was identified but is also present in PCV2c strains and therefore needs to be used with care. Unique amino acid changes in PCV2d strains are indicated in Fig. S1 with the majority of these changes located in previously described antibody recognition domains (Trible & Rowland, 2012). The significance of these changes requires further investigation.

**Genetic divergence of PCV2**

Of all the available methods, the NJ method has been most commonly utilized to analyse the evolutionary relationship among PCV2 strains, and the NJ p-distance model has been
considered the standard to define PCV2 genotypes (Olvera et al., 2007; Segalés et al., 2008). To analyse the global genetic divergence of PCV2 strains, the \( p \)-distances of the main clades PCV2a, PCV2b, PCV2c and PCV2d and the IM clades 1–4 were calculated (Table 3). Of the four genotypes, PCV2a strains showed the largest genetic divergence with an average distance (± standard error) of 0.041 ± 0.004 and PCV2b strains showed the least genetic divergence (0.009 ± 0.001), while the divergence of PCV2d strains was between PCV2a and PCV2b (0.015 ± 0.002).

For the genetic divergence among genotypes, PCV2c showed the largest divergence (>0.1) compared to the other three genotypes (Table 3), suggesting that PCV2c isolates diverged early from the other PCV2 genotypes.
Table 3. The average p-distances within and between PCV2 clades identified by the neighbour-joining method, which include ORF2 sequences of 282 PCV2a, 1007 PCV2b, three PCV2c, 322 PCV2d and 64 sequences from intermediate clades IM1–IM4

The numbers of base differences per site from averaging over all sequences pairs between clades are shown. Standard error (se) estimates are shown and were obtained by a bootstrap procedure (1000 replicates); the analysis involved 1678 nt sequences. All positions containing gaps and missing data were eliminated. There were a total of 609 positions in the final dataset. Evolutionary analyses were conducted in MEGA5.

<table>
<thead>
<tr>
<th>Genotype</th>
<th>Average p-distance between PCV2 clades (mean ± se)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>PCV2a</td>
</tr>
<tr>
<td>0.041 ± 0.004</td>
<td>0.074 ± 0.009</td>
</tr>
<tr>
<td>0.009 ± 0.001</td>
<td>0.055 ± 0.008</td>
</tr>
<tr>
<td>0.008 ± 0.002</td>
<td>0.010 ± 0.011</td>
</tr>
<tr>
<td>0.015 ± 0.002</td>
<td>0.008 ± 0.007</td>
</tr>
</tbody>
</table>

Genotype PCV2a and genotype PCV2b showed a divergence of 0.074 ± 0.009. Similarly, PCV2d and PCV2b strains showed a genetic divergence of 0.055 ± 0.008. The distances within and among genotypes were also calculated using Kimura two-parameter and Tamura–Nei methods and the results were similar to the p-distance results (data not shown).

Among recent US strains, PCV2d strains showed the least genetic divergence (0.002 ± 0.001) within genotype (Table S5), which may indicate their close genetic relationship to each other due to their short circulation period in North America. On the contrary, and likely due to their long presence in North America, PCV2a strains displayed the highest genetic divergence (0.017 ± 0.003), which was approximately eight times higher compared to PCV2d and three times higher compared to PCV2b strains. Moreover, similar to the global results, PCV2d strains showed the closest relationship to PCV2b strains with an average distance and standard error of 0.059 ± 0.008, which is much larger than the genotype-defining cut-off of 0.035.

Phylogenetic analysis using a small dataset

To ensure maintenance of the clade topology when fewer isolates were included, a small dataset was analysed that included 10 reference genotype strains and 33 representative strains from the intermediate clades (Fig. 4). While four clades corresponding to PCV2a, PCV2b, PCV2c and PCV2d could be clearly delimited, clustering of the IM strains was inconsistent between the different methods as expected (data not shown). Interestingly, different from the results obtained with the complete dataset, two representative strains, AY754020 and AY874169, from the small PCV2a clade connecting to PCV2d by network (Fig. 1d) were clustered with the oldest PCV2d strain (JX512855) from Switzerland (Fig. 4). Furthermore, the dashed line in Fig. 4, which indicates the p-distance PCV2 genotype threshold of 0.035 in the linearized NJ tree, is located in front of the bifurcation of PCV2a and PCV2b, which is definitely incorrect, highlighting the problems associated with inferring distances directly from the linearized tree as demonstrated (Cortey et al., 2011) (Fig. 3a). The function to linearize the evolutionary tree is no longer available in MEGA 6, perhaps because of the imprecision in directly inferring genetic distance from the linearized tree.

Time to the most recent common ancestor (TMRCA) and phylogeographic inference of PCV2d strains

To estimate the rate of evolution and the TMRCA of PCV2d, the initially described strains in each region or country (Table 2), the two PCV2d strains FJ644927 and JF827599 which did not group with PCV2d-1 and PCV2d-2, and representative strains from genotypes PCV2a, PCV2b and PCV2c were utilized (Figs 5 and S2). A total of 29 ORF2 sequences and 24 corresponding full genome sequences were analysed including five ORF2 PCV2d sequences at initial identification in a certain country (EU450616 from South Korea, JN006443 from Romania, EF990645 from Belgium, AB462384 from Japan and HQ378157 from Serbia). The 95 % highest posterior density (HPD) of the estimated parameter values from the posterior analyses and those from the prior analyses did not overlap, suggesting that sufficient temporal structure was present in the data to estimate both the rates of nucleotide substitution and TMRCA. The mean substitution rate for the full genome dataset was 9.77 × 10⁻⁴ nucleotide substitutions per site and year, with a 95 % HPD that ranged from 4.58 × 10⁻⁴ to 1.49 × 10⁻³ substitutions per site and year. The mean substitution rate for the ORF2 dataset was 2.93 × 10⁻³ nucleotide substitutions per site and year, with a 95 % HPD that ranged from 1.71 × 10⁻⁴ to 4.24 × 10⁻³ substitutions per site and year, as described for PCV2a and PCV2b (Firth et al., 2009).

The TMRCA for PCV2d was estimated to be 27 years before present (ybp) (95 % HPD of 17 to 42 ybp) based on
Molecular analysis of PCV2 isolates

**Fig. 4.** Neighbour-joining evolutionary tree constructed by p-distance model based on 43 PCV2 ORF2 sequences. The PCV2 ORF2 sequences include 10 reference sequences and 33 representative strains from the intermediate clades and other PCV2 strains. The tree is drawn to scale, with branch lengths in the same units, number of base differences per site, as those of the evolutionary distances used to infer the phylogenetic tree. The trees were rooted by PCV1 (AY184287). The dashed line indicates the PCV2 genotype-defining threshold of 0.035. The reference strains are indicated by bold font.
the analysis of full genomic sequences, and it was 20 ybp (95% HPD of 15 to 25 ybp) based on the ORF2 dataset. Moreover, based on the present dataset, the TMRCA of PCV2c was estimated to be 37 ybp (95% HPD of 33 to 44 ybp) based on full genome analysis or 34 ybp (95% HPD of 33 to 38 ybp) based on ORF2 sequence analysis. To be consistent, the TMCRA of PCV2a and PCV2b of the selected sequences used in this study were also determined and the results were as reported previously (Firth et al., 2009) (data not shown). The phylogeographic reconstruction based on the present full genomes or ORF2 sequences suggested that PCV2d may have a MRCA location in China (Figs 5 and S2). Two strains, which grouped with PCV2d isolates based on ORF2 analysis and grouped with PCV2b isolates based on full genome analysis, were identified to have recombination events affecting a partial ORF1–ORF2 region (strain JX512855) or ORF2 (strain JF827599).

The relative substitution rates for the three codon positions of ORF2 are 0.674, 0.727 and 1.599 respectively, which indicates that evolution of these sequences is dominated by purifying selection as the third codon position (known as the wobble position) has a much higher rate of evolution (2 times faster) than either of the other codon positions; this is in accordance with previous analyses, where purifying selection was identified in both rep and cap genes (Firth et al., 2009; Olvera et al., 2007).

DISCUSSION

In the present study, 1680 PCV2 ORF2 sequences were analysed by the NJ standard method with the p-distance model together with more outperformed methods such as ML, BI and network analysis. The results confirmed that PCV2 can be divided into four major clades, PCV2a, PCV2b, PCV2c and PCV2d, and that several small or
intermediate clades are located between the major clades. The present study further confirmed that deriving the genetic distance directly from the linearized evolutionary tree can be problematic and can lead to incorrect conclusions as shown in Fig. 4.

For the six IM clades, IM1 and IM4 were obtained by NJ, ML and BI methods, IM2 and IM3 were identified by NJ, ML, and network methods, IM5 was obtained by the ML method and IM6 was obtained by the BI (MCMC) method. The location of the IM clades in the topology of the trees was method dependent, indicating the overall uncertainty of the phylogenetic status of these strains. It appears reasonable to accept the phylogenetic results if there is agreement by more than two methods and IM5 and IM6 strains could be classified into PCV2d or PCV2a (Table S4). The IM1 clade can be considered active; there are more than 40 of these strains which are limited to Asia but are being reported on an ongoing basis. None of the other IM clades has more than 10 representative strains; all of these clades can be considered inactive. The limited numbers of reported IM2–6 strains perhaps indicates a reduced fitness ultimately leading to extinction of these groups, as reported for some RNA viruses (Elena et al., 2000). Some of the IM strains may be relics of intermediate virus strains during PCV2 evolution or represent artificial recombinants, as sequence amplification artefacts during PCR with Taq polymerase have been reported previously (Martin et al., 2011).

PCV2 was subjected to a drastic global genotype shift from PCV2a to PCV2b around 2003 (Beach & Meng, 2012; Patterson & Opriessnig, 2010) and PCV2b now appears to dominate, with nearly 60% of the sequences deposited in GenBank belonging to the PCV2b genotype. PCV2a strains have the highest genetic divergence (0.041 ± 0.004) within genotype, as expected. PCV2c was further confirmed to be an extinct group of viruses, as no additional PCV2c sequence has been identified since the three PCV2c strains were initially reported (Dupont et al., 2008). In a recent report from China, 68.2% (45/66 samples) of the PCV2 positive samples were PCV2d (Wei et al., 2013), which confirms an earlier report implying PCV2d as the emerging and predominant PCV2 subtype in China (55%; 22/40 samples) (Cai et al., 2012). Since the first identification of PCV2d in the US from a PCV2 vaccinated herd in 2012 (Xiao et al., 2012), the numbers of PCV2d sequences have increased considerably. Besides having been identified in Asia, Europe and North America, PCV2d has been recently detected in pigs in Brazil (Salgado et al., 2014). Taken together, these data strongly suggest that PCV2d not only is present worldwide but also appears to be increasing, perhaps suggesting another global PCV2 genotype shift.

PCV2d can be divided into PCV2d-1 and PCV2d-2 based on NJ and ML analysis. US PCV2d-2 genomes show high identities (>99.9%) to a Chinese PCV2d-2 strain (HM038017) with only one silent mutation in ORF1 (Xiao et al., 2012). Further phylogeographic analysis confirmed that the most likely MRCA location of PCV2d is in China. The estimate of TMRCA suggests that PCV2d originated 20 years ago, similar to PCV2b (Firth et al., 2009), which may indicate that PCV2b and PCV2d bifurcated early and evolved independently in the pig population. The nucleotide sequence identities between the earliest PCV2d strain (JX512856) and recent PCV2d strains (represented by JX535296 identified in 2012) were 96.5% for the full genome, 97.1% for ORF1 and 94.3% for ORF2. Those between the oldest strain (JX512856) and middle-term strains (represented by AY191946 identified in 2002) were 97.2%, 97.4% and 95.7% for full genome, ORF1 and ORF2, respectively, while those between the middle-term strains and recent strains were 98.2%, 98.8% and 97% for full genome, ORF1 and ORF2, respectively. Evolutionarily, this indicates gradual nucleotide substitutions and substantial genetic divergence within PCV2d strains over time, which may have implications for the change of the pathogenicity of the virus, as the earliest JX512856 was identified from a healthy pig with no PCV2-associated disease and many of the recent strains were identified in PCV2 vaccine failure investigations.

Theoretically, natural selection should favour a reduction in mutation rates in static environments to reduce the burden of deleterious mutations (Duffy et al., 2008). The mechanism by which ssDNA viruses can achieve high mutation rates while using host DNA polymerases to replicate is still unknown (Duffy et al., 2008). The observed high substitution rate of PCV2, the emergence of multiple genotypes, the gradual genetic divergence within genotype, such as with PCV2d and genotype shifts, may be a product of various factors including vaccination pressure and natural selection. Particularly for PCV2d, widely used PCV2 vaccination could be important in its emergence. Most pigs in the US are vaccinated against PCV2 at weaning and therefore it can be assumed that the majority of the 2012 and 2013 PCV2d sequences were obtained from vaccinated pigs. However, the vaccination status of PCV2d infected pigs in other geographical regions is less clear and a possible relationship of PCV2 vaccination and occurrence of PCV2d should be investigated further.

**METHODS**

**Samples and sequences.** To investigate the molecular genetic relationship of PCV2 strains worldwide, 1537 PCV2 ORF2 gene sequences available through GenBank were downloaded and analysed (GenBank sequences were collected until 17 February 2014). In addition, 143 ORF2 sequences were obtained from January 2011 to December 2013 through the Iowa State University Veterinary Diagnostic Laboratory database and included samples from 13 US states (Iowa, Illinois, Indiana, Kansas, Maryland, Michigan, Missouri, North Carolina, North Dakota, Nebraska, Ohio, Oklahoma and Pennsylvania).

**Sequences analysis.** The sequences were aligned by CLUSTAL W program within software BioEdit (Hall, 1999). Phylogenetic analyses were carried out by BI with the MCMC algorithms with software BEAST 1.7.5 (Drummond et al., 2012), by the methods of NJ and ML with software MEGA5 (Tamura et al., 2011) and by network analysis.
with the MJ algorithm (Bandelt et al., 1999). The haplotypes and haplotype diversity of PCV2 were determined using the software DnSP5.10.1 (Librado & Rozas, 2009). Specifically, the General Time Reversible (GTR) substitution model with a proportion of invariant sites and gamma distributed rate heterogeneity (GTR + I + F) was used by the methods of ML and BI. The p-distance was used to construct the NJ tree with bootstrap replication of 1000 to test the phylogeny. During the analysis of the BI, a relaxed molecular clock with an uncorrelated log-normal distribution of rates and a chain length of 1 × 10^6 generations with sampling every 1000 generation was performed to estimate the posterior probability, using an extended Bayesian skyline plot as a tree prior, and the burn-in was set at 10 % of the sampled states. The trees were assessed by the program Tracer (http://beast.bio.ed.ac.uk/tracer/) and were viewed by FigTree (http://tree.bio.ed.ac.uk/software/figtree/).

To characterize the genetic divergence within and between the genotypes, the distances within and between genotypes were calculated by the methods of p-distance, Kimura two-parameter and Tamura-Nei with the software MEGA5. To detect possible recombination events within the ORF2 gene sequences obtained from the USA, six methods in the Recombination Detection Program (RDP) 4.35 (Martin et al., 2010) were utilized including RDP, GeneConv, BootScan, MaxChi, Chimaera and SiScan. Only putative recombination events detected by more than one method were adopted.

To estimate the rate of evolution and the TMRCA of PCV2 with emphasis on PCV2d, whole genome sequences and ORF2 sequences of representative PCV2 strains were further selected and analysed by BEAST 1.7.5 by a relaxed molecular clock with an uncorrelated log-normal distribution of rates and a Bayesian skyline coalescent model (Drummond et al., 2012). For analysis of full genome sequence, a GTR + I + F model of nucleotide substitution was used. For analysis of ORF2 sequences a Hasegawa–Kishino–Yano model of nucleotide substitution with a proportion of invariant sites and gamma distributed rate heterogeneity (HKY + I + F) was used. To examine the influence of the prior distributions on the nucleotide substitution rate and TMRCA parameters, the MCMC chains were run with the same priors as in the original analysis, but without the inclusion of sequence data. Moreover, the timescaled phylogeny of PCV2d was reconstructed by using a standard continuous-time Markov chain process over discrete sampling locations with the Bayesian stochastic search variable selection (BSSVS) model.

GenBank accession numbers. The PCV2 ORF2 sequences obtained in the present study and also one US sequence obtained in 2007 and two 2011 samples from Mexico included in the global obtained in the present study and also one US sequence obtained (Drummond et al., 2012). For analysis of full genome sequence, a GTR + I + F model of nucleotide substitution was used. For analysis of ORF2 sequences a Hasegawa–Kishino–Yano model of nucleotide substitution with a proportion of invariant sites and gamma distributed rate heterogeneity (HKY + I + F) was used. To examine the influence of the prior distributions on the nucleotide substitution rate and TMRCA parameters, the MCMC chains were run with the same priors as in the original analysis, but without the inclusion of sequence data. Moreover, the timescaled phylogeny of PCV2d was reconstructed by using a standard continuous-time Markov chain process over discrete sampling locations with the Bayesian stochastic search variable selection (BSSVS) model.

GenBank accession numbers. The PCV2 ORF2 sequences obtained in the present study and also one US sequence obtained in 2007 and two 2011 samples from Mexico included in the global analysis have been deposited in GenBank under the accession numbers of KM190983–KM191128.

ACKNOWLEDGEMENTS

We thank Dr Priscilla F. Gerber for assistance with retrieving sequences from GenBank.

REFERENCES


