Data in Brief 8 (2016) 549–552

Contents lists available at ScienceDirect

Data in Brief

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

Data Article

Phylodynamic analysis of porcine circovirus type 2: Methodological approach and datasets

Giovanni Franzo [a,*], Martì Cortey [b], Joaquim Segalés [b], Joseph Hughes [c], Michele Drigo [a]

[a] Department of Animal Medicine, Production and Health (MAPS), University of Padua, Viale dell'Università 16, 35020 Legnaro (PD), Italy
[b] Centre de Recerca en Sanitat Animal (CReSA), UAB-IRTA, Barcelona, Spain
[c] MRC-University of Glasgow Centre for Virus Research, Glasgow, Scotland, UK

ARTICLE INFO

Article history:
Received 2 May 2016
Received in revised form
30 May 2016
Accepted 7 June 2016
Available online 15 June 2016

ABSTRACT

Since its first description, PCV2 has emerged as one of the most economically relevant diseases for the swine industry. Despite the introduction of vaccines effective in controlling clinical syndromes, PCV2 spread was not prevented and some potential evidences of vaccine immuno escape have recently been reported ("Complete genome sequence of a novel porcine circovirus type 2b variant present in cases of vaccine failures in the United States" (Xiao and Halbur, 2012) [1], "Genetic and antigenic characterization of a newly emerging porcine circovirus type 2b mutant first isolated in cases of vaccine failure in Korea" (Seo et al., 2014) [2]). In this article, we used a collection of PCV2 full genomes, provided in the present manuscript, and several phylogentic, phylodynamic and bioinformatic methods to investigate different aspects of PCV2 epidemiology, history and evolution (more thoroughly described in **"PHYLODYNAMIC ANALYSIS of PORCINE CIRCOVIRUS TYPE 2 REVEALS GLOBAL WAVES of EMERGING GENOTYPES and the CIRCULATION of RECOMBINANT FORMS"** [3]**.** The methodological approaches used to consistently detect recombiantion events and estimate population dymanics and spreading patterns of rapidly evolving ssDNA viruses are herein reported. Programs used are described and original scripts have been provided. Ensembled databases used are also made available. These consist of a broad collection of complete genome sequences (i.e. 843 sequences; 63 complete genomes of PCV2a, 310 of PCV2b, 4 of PCV2c, 217 of

DOI of original article: http://dx.doi.org/10.1016/j.ympev.2016.04.028
* Corresponding author.
E-mail addresses: giovanni.franzo@unipd.it (G. Franzo), marti.cortey@irta.cat (M. Cortey).

http://dx.doi.org/10.1016/j.dib.2016.06.005

PCV2d, 64 of CRF01, 140 of CRF02 and 45 of CRF03.), divided in differnt ORF (i.e. ORF1, ORF2 and intergenic regions), of PCV2 genotypes and major Circulating Recombinat Forms (CRF) properly annotated with respective collection data and country. Globally, all of these data can be used as a starting point for further studies and for classification purpose.

Specifications Table

Subject area	Biology, Genetics and Genomics
More specific subject area	Phylogenetics and Phylogenomics
Type of data	Excel file
How data was acquired	Sequence data were downloaded from GenBank, manually checked and annotated. Analysis were performed using state of art freely available programs for phylogeny, population dynamics and selective pressure analysis.
Data format	Raw, filtered
Experimental factors	PCV2 complete genome sequences were downloaded from Genbank and annotated with the respective collection country and data. Sequences have been aligned and the consistency of the alignment was checked. All sequences were scanned for recombination and subdivided in genotypes or recombinant forms. Databases generated in this way were used for further analysis.
Experimental features	PCV2 sequences download, quality check and annotation. Sequence alignment, recombination analysis, Coalescent based analysis of population parameters and reconstruction of viral spreading patterns. Analysis of selective pressure acting on different coding regions.
Data source location	n/a
Data accessibility	Data are within the article

Value of the data

- Most extensive collection of PCV2 full genome sequences with available metadata.
- Proper annotation linking genetic data to country of origin and collection data
- Full description of several approaches used to analyze different aspects of viral evolution
- Datasets suitable for further evolutionary studies and for PCV2 classification purpose
- Standardized approach that can be used for follow-up studies on PCV2 evolution.

1. Data

Supplementary data 1 provides a table reporting the accession number of all (i.e. 843) PCV2 complete genomes and PCV2a ORF2 sequences used in Franzo et al. [3]. For each sequence, the country where it has been sampled and the collection data are also reported. The alignments of all major PCV2 genotypes (i.e. PCV2a, PCV2b, PCV2c and PCV2d) and circulating recombinant forms (CRF) are provided in Supplementary data 2 and could be used for comparison purpose and as a starting point for further studies. Finally, Supplementary data 3 provides an R script for ancestral state reconstruction of per-site amino acid sequence using a maximum likelihood approach.

2. Experimental design, materials and methods

2.1. Dataset

A total of 925 PCV2 complete genome sequences with known collection dates and country of origin were downloaded from GenBank (accessed 06/10/2014 – listed in "Supplementary data 1.xls" in the online version of this article) and aligned using the MAFFT method [4]. All poorly aligned sequences and those displaying degenerate nucleotides or indels which caused reading frame alterations, suggesting sequencing errors, were removed from the dataset (898 sequences were maintained) (Supplementary data 1).

2.1.1. Recombination analysis

The whole dataset was tested for recombination using two programs based on different approaches: RDP4 [5] and GARD [6]. When RDP was used, only recombination events detected by more than 2 methods with a significance value lower than 10^{-5} (p-value $< 10^{-5}$) and Bonferroni correction were accepted. The non-recombinant sequences as well as those sharing recombination events were split into separate datasets and expanded to their original size.

2.2. Genotyping and database preparation

The non-recombinant sequences were classified into genotypes PCV2a, PCV2b, PCV2c or PCV2d according to Franzo et al. 2015 [7].

The most appropriate nucleotide substitution model was selected according to the results of the Akaike information criterion (AIC) score calculated using JModel Test 2.1.2 [8]. A phylogenetic tree was reconstructed using the Maximum likelihood (ML) approach implemented in PhyML [9]. The best tree search method included the combination of two branch swapping algorithms: nearest neighbor interchange (NNI) and subtree pruning and regrafting (SPR). The robustness of the monophyly of the taxa subsets was estimated with the fast non-parametric version of the aLRT (Shimodaira–Hasegawa [SH]-aLRT), developed and implemented in PhyML 3.0 [10]. On the basis of the recombination and phylogenetic analyses, sequences were divided into independent datasets, corresponding to different genotypes and CRFs (i.e. those including more than 30 sequences collected in two or more countries). Every dataset was further divided in three regions, namely *ORF1*, *ORF2* and intergenic region (obtained merging together the major and the minor intergenic regions) and a new alignment was generated on each dataset. The coding regions were aligned at the amino acid level and then the nucleotide sequences were back-translated using the MAFFT algorithm implemented in TranslatorX [11]. All these datasets, comprising different gene alignments, are provided in Supplementary data 2. These include 63 complete genomes of PCV2a, 310 of PCV2b, 4 of PCV2c, 217 of PCV2d, 64 of CRF01, 140 of CRF02 and 45 of CRF03. Additionally a dataset of 83 PCV2a ORF2 sequences is provided.

2.3. BEAST and selective pressures analysis

The time to most recent common ancestor (tMRCA), substitution rates, phylogeography and population dynamics were jointly estimated using a Bayesian serial coalescent approach implemented in BEAST 1.8.1 [12].

The selective pressure on the viral proteins was estimated using different methods based on the ratio between non-synonymous and synonymous substitution rates (dN/dS).

Pervasive diversifying/purifying selection was estimated using SLAC, FEL and FUBAR method while episodic diversifying selection was evaluated using MEME [13–15]. The action of selective pressures was compared among different genes using the *dNdSDistributionComparison.bf* implemented in HyPhy [16]. Differences in the site-by-site selection patterns among different genotypes were investigated for each gene using the batch files *CompareSelectivePressure.bf* implemented in the same program. Ancestral state reconstruction of per site amino acid sequence was performed, based on the time scaled phylogenetic trees, using the maximum likelihood approach of the *ape* package implemented in R [17].The corresponding script is provided in Supplementary data 3.

Transparency document. Supplementary material

Transparency data associated with this article can be found in the online version at http://dx.doi.org/10.1016/j.dib.2016.06.005.

Appendix A. Supplementary material

Supplementary data associated with this article can be found in the online version at http://dx.doi.org/10.1016/j.dib.2016.06.005.

References

[1] C.T. Xiao, P.G. Halbur, T. Opriessnig, Complete genome sequence of a novel porcine circovirus type 2b variant present in cases of vaccine failures in the United States, J. Virol. 86 (22) (2012) 12469-12.
[2] H.W. Seo, C. Park, I. Kang, et al., Genetic and antigenic characterization of a newly emerging porcine circovirus type 2b mutant first isolated in cases of vaccine failure in Korea, Arch. Virol. 159 (11) (2014) 3107–3111.
[3] G. Franzo, M. Cortey, J. Segalés, J. Hughes, M. Drigo, Phylodynamic analysis of Porcine circovirus type 2 reveals global waves of emerging genotypes and the circulation of recombinant forms, Mol. Phylogent Evol. 100 (2016) 269–280 10.1016/j.ympev.2016.04.028.
[4] K. Katoh, D.M. Standley, MAFFT multiple sequence alignment software version 7: improvements in performance and usability, Mol. Biol. Evol. 30 (4) (2013) 772–780.
[5] D.P. Martin, P. Lemey, M. Lott, V. Moulton, D. Posada, P. Lefeuvre, RDP3: a flexible and fast computer program for analyzing recombination, Bioinformatics 26 (19) (2010) 2462–2463.
[6] S.L. Kosakovsky Pond, D. Posada, M.B. Gravenor, C.H. Woelk, S.D. Frost, GARD: a genetic algorithm for recombination detection, Bioinformatics, 22, 3096–3098.
[7] G. Franzo, M. Cortey, A. Olvera, et al., Revisiting the taxonomical classification of porcine circovirus type 2 (PCV2): still a real challenge, Virol. J., 12, 131.
[8] D. Darriba, G.L. Taboada, R. Doallo, D. Posada, JModelTest 2: more models, new heuristics and parallel computing, Nat. Methods. 9 (8) (2012) 772.
[9] S. Guindon, J.- Dufayard, V. Lefort, M. Anisimova, W. Hordijk, O. Gascuel, New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0, Syst. Biol. 59 (3) (2010) 307–321.
[10] M. Anisimova, M. Gil, J.- Dufayard, C. Dessimoz, O. Gascuel, Survey of branch support methods demonstrates accuracy, power, and robustness of fast likelihood-based approximation schemes, Syst. Biol. 60 (5) (2011) 685–699.
[11] F. Abascal, R. Zardoya, M.J. Telford, TranslatorX: multiple alignment of nucleotide sequences guided by amino acid translations, Nucleic Acids Res. 38 (2010) W7–W13.
[12] A.J. Drummond, M.A. Suchard, D. Xie, A. Rambaut, Bayesian phylogenetics with BEAUti and the BEAST 1.7, Mol. Biol. Evol. 29 (8) (2012) 1969.
[13] S.L. Kosakovsky Pond, S.D. Frost, Not so different after all: a comparison of methods for detecting amino acid sites under selection, Mol. Biol. Evol. 22 (5) (2005) 1208–1222.
[14] B. Murrell, S. Moola, A. Mabona, et al., FUBAR: a fast, unconstrained bayesian approximation for inferring selection, Mol. Biol. Evol. 30 (5) (2013) 1196–1205.
[15] B. Murrell, J.O. Wertheim, S. Moola, T. Weighill, K. Scheffler, S.L. Kosakovsky Pond, Detecting individual sites subject to episodic diversifying selection, Plos. Genet. 8 (7) (2012) e1002764.
[16] S.L.K. Pond, HyPhy: hypothesis testing using phylogenies, Bioinformatics 21 (5) (2005) 676.
[17] E. Paradis, J. Claude, K. Strimmer, APE: analyses of phylogenetics and evolution in R language, Bioinformatics, 20, 289–290.