Infection, Genetics and Evolution 73 (2019) 295–305

Contents lists available at ScienceDirect

Infection, Genetics and Evolution

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

Research paper

Evaluating an automated clustering approach in a perspective of ongoing surveillance of porcine reproductive and respiratory syndrome virus (PRRSV) field strains

Marie-Ève Lambert[a,b,*], Julie Arsenault[a,b], Pascal Audet[a,b], Benjamin Delisle[a,b], Sylvie D'Allaire[a,b]

[a] Laboratoire d'épidémiologie et de médecine porcine (LEMP), Faculty of Veterinary Medicine, Université de Montréal, St. Hyacinthe, Quebec, Canada
[b] Swine and Poultry Infectious Diseases Research Center (CRIPA), Faculty of Veterinary Medicine, Université de Montréal, St. Hyacinthe, Quebec, Canada

ARTICLE INFO

Keywords:
PRRS
ORF5
Classification
Surveillance
Phylogeny

ABSTRACT

Porcine reproductive and respiratory syndrome virus (PRRSV) has a major economic impact on the swine industry. The important genetic diversity needs to be considered for disease management. In this regard, information on the circulating endemic strains and their dispersal patterns through ongoing surveillance is beneficial. The objective of this project was to classify Quebec PRRSV ORF5 sequences in genetic clusters and evaluate stability of clustering results over a three-year period using an in-house automated clustering system. Phylogeny based on maximum likelihood (ML) was first inferred on 3661 sequences collected in 1998–2013 (Run 1). Then, sequences collected between January 2014 and September 2016 were sequentially added into 11 consecutive runs, each one covering a three-month period. For each run, detection of clusters, which were defined as groups of ≥15 sequences having a ≥70% rapid bootstrap support (RBS) value, was automated in Python. Cluster stability was described for each cluster and run based on the number of sequences, RBS value, maximum pairwise distance and agreement in sequence assignment to a specific cluster. First and last run identified 29 and 33 clusters, respectively. In the last run, about 77% of the sequences were classified by the system. Most clusters were stable through time, with sequences attributed to one cluster in Run 1 staying in the same cluster for the 11 remaining runs. However, some initial groups were further subdivided into subgroups with time, which is important for monitoring since one specific wild-type cluster increased from 0% in 2007 to 45% of all sequences in 2016. This automated classification system will be integrated into ongoing surveillance activities, to facilitate communication and decision-making for stakeholders of the swine industry.

1. Introduction

Porcine reproductive and respiratory syndrome (PRRS) is one of the most costly diseases for the swine industry due to major reproductive and respiratory problems (Holtkamp et al., 2013). The causative agent is a single-stranded RNA enveloped virus of 15,000 base pairs (bp) encoding for at least ten functional open reading frames (ORFs) (Rahe and Murtaugh, 2017). This virus evolves mainly through punctual mutations, but also recombination (Kappes and Faaberg, 2015). An important genetic diversity was found among genotype 2 (North American type) strains in Quebec, Canada, which is consistent with the high genetic diversity reported worldwide (Delisle et al., 2012; Lambert et al., 2012; Lambert et al., 2018; Shi et al., 2010; Stadejek et al., 2013). This important heterogeneity increases the risk of clinical disease, considering the absence of complete protection following a heterologous challenge (Diaz et al., 2012). On the other hand, viral diversity can be used advantageously to gain information for disease control. For example, at the herd level, viral sequences can be compared to investigate potential sources of infection and hypothesize on pathways of virus transmission for better targeted control measures (Lambert et al., 2012; Larochelle et al., 2003). At regional level, genetically related strains can be grouped to rapidly appraise spatiotemporal patterns which would be useful to implement collective measures, especially for area-based regional control and elimination (ARC&E) project (Arruda

Abbreviations: PRRSV, Porcine reproductive and respiratory syndrome virus; ORF, Open reading frame; ML, Maximum likelihood; RBS, Rapid bootstrap support; LEMP, Laboratoire d'épidémiologie et de médecine porcine

* Corresponding author at: Laboratoire d'épidémiologie et de médecine porcine (LEMP), Faculté de médecine vétérinaire, Université de Montréal, 3200 Sicotte, St. Hyacinthe, Quebec J2S 2M2, Canada.
E-mail addresses: marie-eve.lambert@umontreal.ca (M.-È. Lambert), julie.arsenault@umontreal.ca (J. Arsenault), pascal.audet@umontreal.ca (P. Audet), benjamin.delisle@umontreal.ca (B. Delisle), sylvie.dallaire@umontreal.ca (S. D'Allaire).

https://doi.org/10.1016/j.meegid.2019.04.014
Received 23 November 2018; Received in revised form 6 April 2019; Accepted 18 April 2019
Available online 27 April 2019

et al., 2017). Through the years, this integration of molecular data in epidemiological studies and surveillance programs has increased rapidly (Alvarez et al., 2016; Vasylyeva et al., 2016)

Different techniques can be used to estimate the genetic relatedness between PRRSV strains. RFLP patterns were initially used to discriminate vaccine from wild-type PRRSV strains (Cai et al., 2002; Wesley et al., 1998). However, subsequent studies demonstrated that RFLP patterns should be used with caution for epidemiological investigations, as they do not sufficiently discriminate between wild-type strains (Brar et al., 2011). Most phylogenetic studies on PRRSV were based on sequencing data from a single gene, mainly ORF5 gene (GP5) as it represents an immunologically important genome segment associated with virus neutralization (Ostrowski et al., 2002; Plagemann, 2004; Plagemann et al., 2002; Popescu et al., 2017) and exhibits a considerable level of diversity. More recently, whole genome sequencing (WGS) has been developed (Zhang et al., 2017). However, the availability and lower cost of ORF5 compared to WGS have allowed the construction through time of large ORF5 sequence databases worldwide used for surveillance and research purposes.

Different methods can be used to group PRRSV ORF5 sequences into genetic clusters. Most studies on PRRSV relied on hierarchical distance-based methods such as neighbour-joining (NJ) or model-based inference such as maximum likelihood (ML) for building up a tree topology as the basis for further classification (An et al., 2007; Delisle et al., 2012; Wang et al., 2008; Yoon et al., 2008; Yoshii et al., 2005). More recently, Bayesian phylogenetic inferences gained in popularity to investigate PRRSV evolutionary process on smaller datasets or subsample of larger datasets or to examine more thoroughly a monophyletic cluster previously identified using ML inference (Alkhamis et al., 2016; Shi et al., 2010). In most of the previously published studies, clusters were defined by visual appraisal of the tree topology, with or without assessment of confidence in tree branches, such as bootstrap, aLRT, or posterior probability (An et al., 2007; Delisle et al., 2012; Wang et al., 2008; Yoon et al., 2008; Yoshii et al., 2005). Other researchers have used average pairwise genetic distance thresholds as a criterion to form lineages and sub-lineages (Shi et al., 2010).

In Quebec, collective control measures against PRRS at the provincial level are a priority. Monitoring changes in viral subpopulations would help to guide these actions, for example by providing information on how control zones should be delineated to limit the dispersal of genetically divergent strains. For ongoing surveillance of endemic strains, genetic clusters would need to be updated several times a year with the addition of new sequences from field submissions. Thus an automate system is needed to reveal major clusters within a large phylogenetic tree inferred on several thousands of sequences without having to identify them visually and to manually extract them. The system would also have to compare results between subsequent classification runs to describe changes in size of genetic clusters through time. Moreover, to provide constant information on circulating strains to stakeholders of the swine industry, the stability of clustering results obtained by the system is necessary to ensure that once detected, major clusters are still identified in subsequent classification runs and that sequences belonging to a cluster remain classified in the same cluster through time. This would avoid confusion when clustering results are to be transferred to end users for field applications.

The main objective of this research project was to automatically classify Quebec PRRSV ORF5 sequences into genetic clusters and evaluate the stability of the clustering results over a three-year period, in a perspective of ongoing surveillance of endemic strains.

2. Material and methods

2.1. Sequence collection

A total of 4995 ORF5 PRRSV sequences were obtained from the database of the Laboratoire d'épidémiologie et de médecine porcine (LEMP) of the Université de Montréal. These sequences were collected between January 1, 1998, and September 30, 2016, from Quebec swine production sites through field submissions from regular veterinary services, active surveillance (e.g. ARC&E) or LEMP research projects. Samples were submitted by veterinarians to the molecular diagnostic laboratory of the Faculty of Veterinary Medicine (FVM) of the Université de Montréal or to two private laboratories. RNA extraction, RT-PCR and sequencing were performed according to routine protocol of each laboratory and sequence data were transferred to the LEMP database. Four commercial vaccine sequences were also added to the dataset: Ingelvac PRRS® MLV (Boehringer Ingelheim Vetmedica Inc., St. Joseph, Missouri, USA), Ingelvac PRRS® ATP (Boehringer Ingelheim Vetmedica Inc., St. Joseph, Missouri, USA), Fostera® PRRS (Zoetis, Florham Park, New Jersey, USA) and Prime Pac™ PRRS+ (Merck Animal Health, Summit, NJ). Sequences were examined for the presence of unusual characters other than known IUB symbols and, when present, were replaced by N character.

2.2. Detection of recombinants

Detection of recombinant signals was first performed on the entire sequence dataset (n = 4995). Considering the file size limit for RDP analysis, the dataset was split according to year in three sets of 2000 sequences each, the middle one overlapping the others. Detection of potential recombinant sequences was carried out by using an exploratory primary scan for mosaic signals with RDP, Geneconv and MaxChi detection methods implemented in RDP4 version 4.79 software. These three methods were also used in addition to Bootscan, SisScan, Chimaera and 3-Seq for a secondary scan. Default settings for each method were used throughout the procedure using 0.05 *p*-value with a Bonferroni correction for multiple testing. Sequences identified by at least one primary method were considered as significant recombinants and excluded from the dataset.

2.3. Automated classification system

An automated classification system was applied on the sequence dataset and then, the stability of genetic clusters through time was assessed on a three-year period to mimic an ongoing process of classification. More specifically, following a starting tree leading to a first classification (Run 1, 1998–2013), sequences sampled between 2014 and 2016 were gradually introduced in the dataset into 11 consecutive runs, each addition including all sequences sampled in a three-month period.

2.3.1. First classification run (Run 1)

2.3.1.1. Selection and alignment of dataset. A first dataset was created by selecting sequences submitted between January 1998 and December 31, 2013 (n = 3661). The dataset was sorted in increasing order of sampling date and a multiple alignment was performed using Clustal Omega with default settings (Sievers et al., 2011). Clustal Omega was chosen based on its efficiency at minimizing the number of gaps in different dataset sizes with default open gap values as well as its capability to handle a large number of sequences in a timely manner (Lambert et al., 2019b). As the software retains for its analysis only the first sequence listed among a particular group of sequences having all 100% similarity, this sorting ensures that all sequences selected in a run would be present in subsequent runs. The aligned dataset entered the classification pipeline as shown in Fig. 1.

2.3.1.2. Phylogeny inference. A ML phylogeny was inferred using RAxML (Pthreads AVX version 8.2.8) based on a GTR gamma evolutionary model. The phylogeny was computed on 1000 randomized stepwise Maximum Parsimony starting trees for the final search of the best-scoring ML tree. To determine branch support on the Run 1 best tree, 1000 rapid bootstraps were run. A mid-pointed root

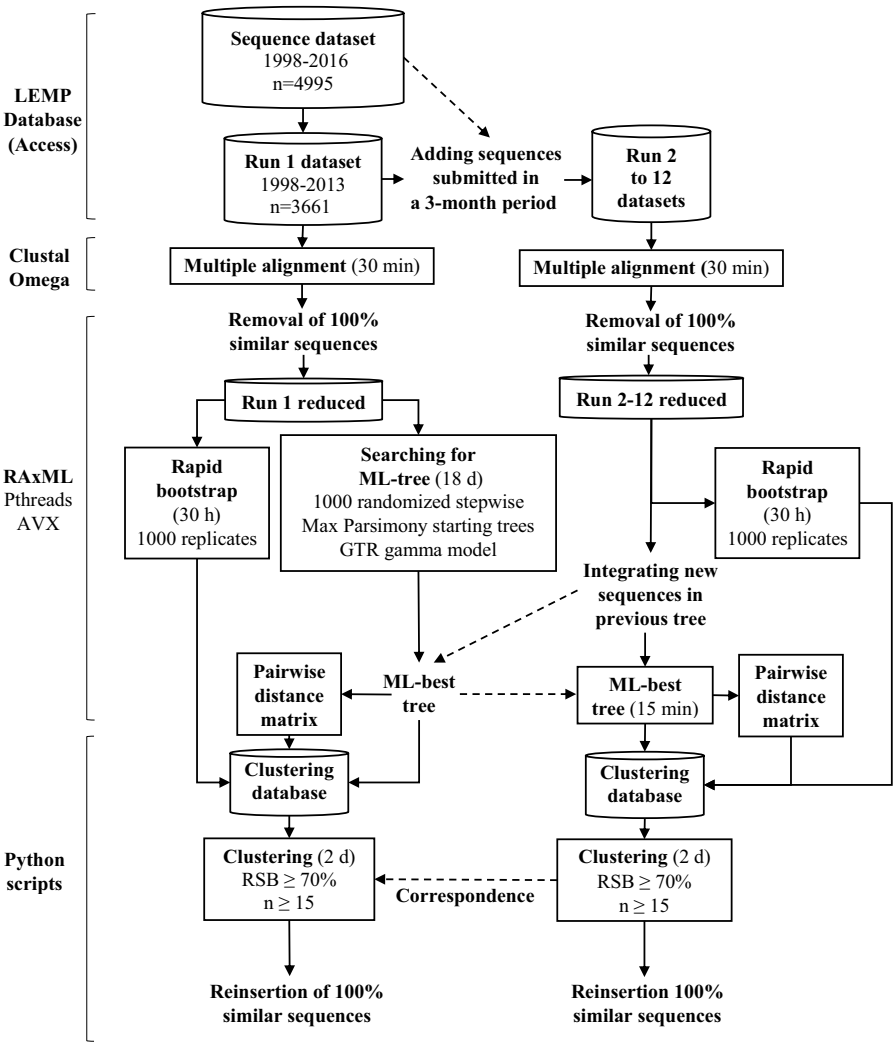

Fig. 1. Description of the pipeline used to classify Quebec PRRSV sequences. The first step consisted of a classification of sequences submitted between January 1, 1998, and December 31, 2013 (Run 1, n = 3661). Next steps consisted of consecutive runs performed after adding sequences submitted in the following three-month periods to the dataset used in the previous classification run.

was placed on the tree for visual representation. Information regarding the tree branching pattern and RBS values were exported into a Newick partition file. In addition, a pairwise evolutionary distance matrix of all pairs of sequences was computed using a RAxML function (RAxML v8.2.X Manual, https://sco.h-its.org/exelixis/resource/download/NewManual.pdf).

2.3.1.3. Sequence clustering. Hierarchical structure of the nodes was extracted from the Newick partition file and the list of 100% similar sequences previously removed (Section 2.3.1.1) were stored into a relational database for downstream reinsertion within clustering steps. Based on the pairwise evolutionary distance matrix, the maximum distance observed among all pairs of sequences included in each node and were registered in the database.

Sequences from the Run 1 phylogeny (1998–2013, n = 3661) were classified into genetic clusters using several algorithms written in Python. The first step was to identify all genetic clusters, which were defined as a group of ≥15 sequences linked to a branch of the tree with RBS value of ≥70%. To avoid multiple subdivisions of genetic cluster due to consecutive RBS values ≥70% near the tips of the tree, a limit was set on evolutionary distance ≤0.025 beyond which further clusters were not considered (Fig. 2).

The second step of the clustering process was to keep track of the hierarchical structure among the different clusters identified, linking major ones to their minor imbricated cluster(s). Major clusters, which were by definition closer to the mid-point root and mutually exclusive, were first identified. Then minor clusters formed by node closer to each tip of the tree were automatically linked to the corresponding major cluster (Fig. 2).

In a third step, 100% similar sequence(s) with at least one other sequence initially dropped from the Run 1 dataset for phylogeny inference were added into corresponding cluster. Sequences not comprised into any major cluster were considered as unclassified by the automated system.

2.3.2. Subsequent classification runs (Runs 2 to 12)

2.3.2.1. Selection and alignment of datasets. Sequences submitted between January 2014 and September 2016 inclusively (n = 1059) were added into 11 consecutive runs (Run 2 to 12), each one comprising sequences sampled in a trimester. Time intervals were defined as follows: January 1 to March 31, April 1 to June 30, July 1 to September 30, October 1 to December 31 until September 30, 2016 and sequences were added to the previous run dataset. Each resulting dataset was sorted, aligned and entered in the classification pipeline as described for the first run.

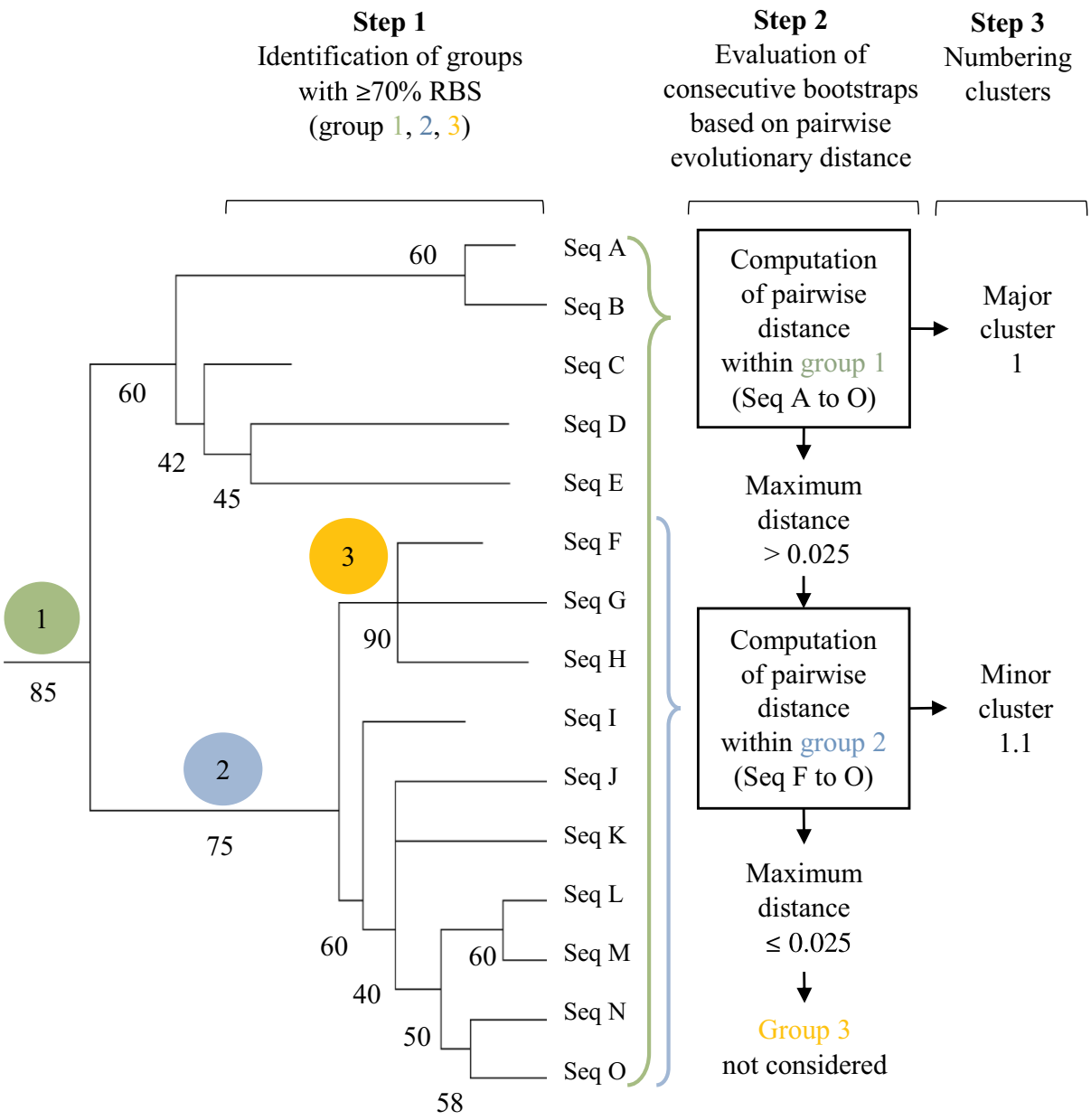

Fig. 2. Evaluation of consecutive rapid bootstrap support (RBS) values ≥70% for sequence clustering. The figure shows the phylogenetic affiliation reconstructed for 15 sequences (Seq A to Seq O) clustered in 3 different groups having ≥70% RBS values. Groups with ≥70% RBS values were not considered if already inserted into another group located closer to the mid-point root and gathering sequences having less than < 0.025 pairwise evolutionary distance with each other.

2.3.2.2. Phylogeny inference and clustering. Dataset for Run 2 was integrated into the Run 1 ML-tree using parameters estimated for the Run 1. Then, dataset for Run 3 was integrated into Run 2 ML-tree and so on until Run 12 was completed. For each run, sequence clustering was done using the same approach as the one described for Run 1.

2.4. Correspondence of genetic clusters through time

Algorithms were developed to establish correspondence of genetic clusters between the different runs (Figs. 1 and 3). Correspondence was determined by comparing the position of each sequence ID of a particular cluster from one run to their location in the next run. A correspondence was then assigned to the cluster in this latter run gathering the highest number of sequences. The same process was applied for all clusters found in the tree (Fig. 3).

2.5. Automating the classification process

The entire process which involved selection and alignment of datasets, phylogeny inference, clustering, correspondence of genetic clusters through time and general statistics regarding the genetic clusters was pipelined and run in a Linux environment (Ubuntu 14.04 LTS) on a Dell Precision T7610 workstation with 10 Intel Xeon Processor E5–2670 @ 2.5 GHz, 128 GB of RAM (DDR3) and 2 TB HD. All computer resources were solely attributed to the classification process. Computation time necessary for inferring of a complete ML-tree (Run 1), and also for adding sequences to an existing tree (Run 2 to Run 12) was noted.

Code lines to execute the different tasks involved in the pipeline consisted of several scripts written in Bash, Python, MySQL or VBA. Furthermore, several file formats had to be managed by the scripts (fasta, Newick, csv, txt). An example of input datasets, scripts and output files are provided in ClassificationExample.zip using a

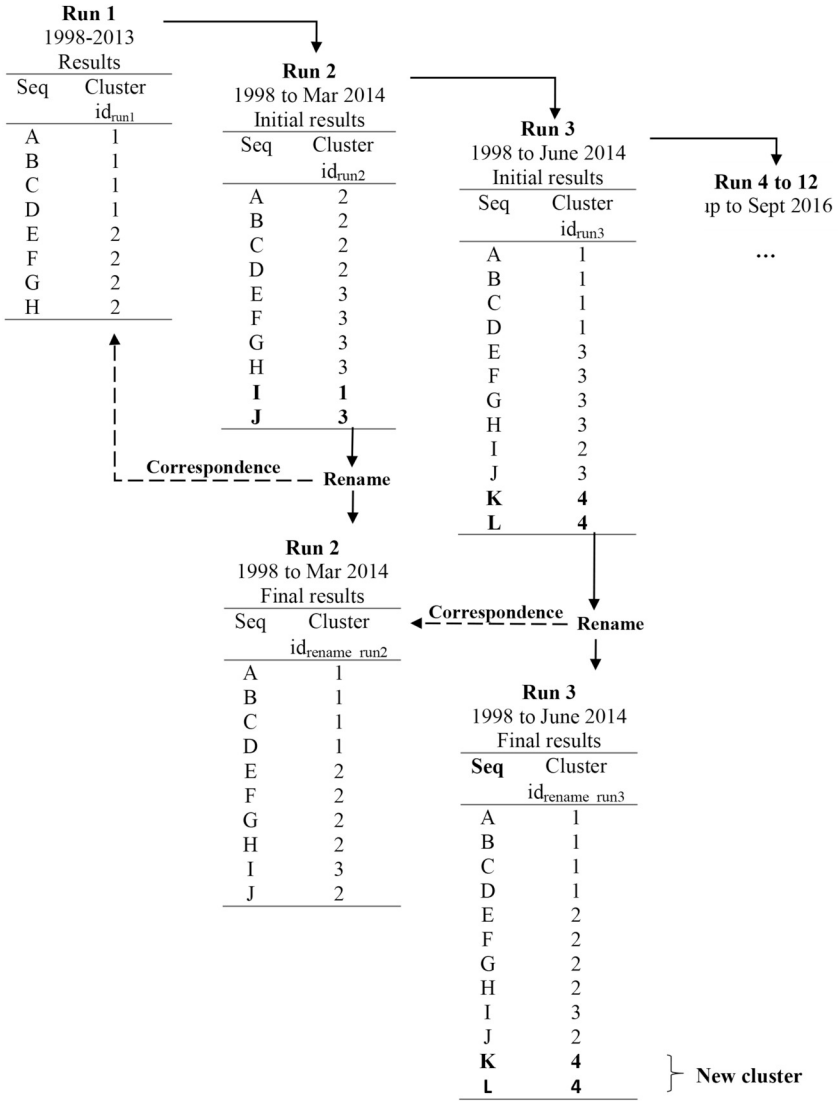

Fig. 3. Workflow correspondence search and renaming of genetic clusters obtained between classification runs. New sequences added for the next run in the system are marked in bold. A correspondence was determined by comparing the position of each sequence ID of a particular cluster from one run to their location in next run. For example, each sequence of a particular cluster in Run 1 was compared to their location in Run 2. The correspondence was assigned to the cluster in Run 2 that gathered the highest number of sequences from a cluster of the last run and so on for each run.

subsample of 440 sequences randomly selected from Run 12 dataset (n = 4289). Instructions regarding the example are also provided in the supplementary material 1 (1_SupplementaryMaterial_Instructions.doc).

2.6. Description of the dataset, clusters and stability of major clusters throughout the runs

The datasets used for the different runs were described according to several characteristics, among them number of sequences, dates, and number of clusters (see Table 1 for more details). Results of each run were only described in terms of major clusters without considering minor clusters. Intracluster diversity was calculated using mean and maximum pairwise evolutionary distance among sequences in a cluster. Stability of a cluster through different runs was defined as the capability of the system to assign a particular sequence to the same major cluster through different runs (Table 2). After having established the correspondence of clusters between runs (Section 2.4), the agreement of sequence assignment was computed for each major cluster and each run except the first one. For example, for a specific cluster of Run 2, it was defined as the number of sequences member of a particular cluster into Run 1 that are also assigned to the same cluster in Run 2 over total number of sequences in Run 1 cluster, and so on.

Sequences belonging to the different genetic clusters identified in final run (Run 12) were extracted from the aligned Run 12 dataset. For each pair of clusters, each sequence belonging to one cluster was compared to each other individual sequences from the other cluster. The similarity between each pair of these sequences was computed as the sum of matches over the total number of aligned bases, expressed as a percentage. Average (min-max) similarity among pairs of clusters was calculated to describe intercluster similarity.

3. Results

For the first run with sequences from 1998 to 2013 (n = 3661), 409 sequences were temporarily excluded because of a 100% pairwise genetic similarity observed with at least one other sequence. The remaining 3252 sequences were used to build the tree. The approximate time of the different steps involved in the first classification process was: 30 min for multiple alignment, 18 days for ML-tree phylogeny, 30 h for estimation of RBS values, 2 days for clustering algorithms as described in Section 2.3.1. Table 1 provides a description of the 12 datasets and general results on clustering for each of them. An average

Table 1
Description of datasets and general results on clustering for Quebec PRRSV ORF5 sequences for each run (1 to 12)[a].

Dataset characteristics and clustering results	Runs of classification											
	1	2	3	4	5	6	7	8	9	10	11	12
Dataset characteristics												
Beginning year	1998	1998	1998	1998	1998	1998	1998	1998	1998	1998	1998	1998
Ending year	2013	2014	2014	2014	2014	2015	2015	2015	2015	2016	2016	2016
Month-day of last sample	12–31	03–31	06–30	09–30	12–31	03–31	06–30	09–30	12–31	03–31	06–30	09–30
Number of sequences	3661	3784	3890	3982	4133	4234	4332	4402	4582	4717	4858	4958
Number of newly added sequences[b]	–	123	106	92	151	101	98	70	180	135	141	100
Number of 100% similar sequences removed[c]	409	429	448	462	486	510	532	553	586	623	650	669
Number of sequences in the ML-tree	3252	3355	3442	3520	3647	3724	3800	3849	3996	4094	4208	4289
Clustering results												
Number of classified sequences	2212	2259	2362	2450	2553	2854	2728	2775	2919	3028	3172	3218
% of classified sequences in ML-tree[d]	68.0	67.3	68.6	69.6	70.0	76.6	71.8	72.1	73.0	74.0	75.4	75.0
Number of major clusters[e]	29	27	29	29	31	31	32	32	32	33	34	33
≥40 sequences	10	10	10	9	10	10	9	9	9	9	9	10
20–39 sequences	9	8	8	9	8	8	9	10	11	11	12	10
< 20 sequences	10	9	11	11	13	13	14	13	12	13	13	13

[a] Run 1 included sequences from the province of Quebec submitted between 1998 and 2013. Runs 2 to 12 represented further classification performed every three-months to simulate ongoing classification of newly submitted sequences.
[b] Difference in number of sequences with previous run.
[c] Number of sequences removed for further phylogenetic analyses because of having 100% pairwise similarity with at least one other sequence in the dataset.
[d] Number of sequences attributed to one of the major clusters over the number of sequences in the ML-tree in the run.
[e] Larger genetic clusters that are closer to the mid-point rooting and mutually exclusive. Each of them may also include minor clusters (not shown).

(min-max) of 118 (70–180) newly submitted sequences per run were added to the classification system.

According to the different runs, the proportion of classified sequences before reintroducing 100% similar sequences was between 67 and 77% (Table 1). Major clusters were almost equally distributed among the clusters of < 20, 20–40 and > 40 sequences, respectively.

Table 2 shows results on genetic clusters identified by the automated system for the different runs according to their number of sequences, RBS value, maximum evolutionary distance observed among pairs of sequences, and percentage of sequences classified into the same cluster on consecutive runs. All major clusters identified using the RBS threshold were over the 0.025 maximum pairwise distance, except cluster 36 identified in runs 6 and 7. Agreement of sequence assignment to a cluster between runs was 100% for all clusters. The maximum pairwise genetic diversity observed between sequences within a specific cluster ranged from 0.03 (clusters 3 and 8) to 0.29 (cluster 25).

For Run 1, a total of 29 clusters respecting both criteria (sequences ≥15, RBS ≥70%) were observed. According to these criteria, most of them (n = 27) remained stable for subsequent runs; ten with sequence addition, and 17 without. For the two remaining clusters (5, 21), having 58 and 25 sequences respectively, RBS values decreased under the threshold in at least one of the subsequent runs, even though no sequence was added or displaced to another cluster. Six new clusters were identified between Runs 4 and 10 (clusters 32, 33, 35, 36, 37, 38) and detected up to Run 12. These viral subpopulations were already present in the earlier dataset but did not fulfill the two criteria to be identified as clusters by the automated system. Three other clusters (cluster 30, 31, 34) emerged in Runs 3 and 6, but RBS went under the 70% threshold for one or several subsequent runs.

Once reinserting in the final run (Run 12) sequences that were 100% similar, 3802 of 4958 (77%) were classified (Table 1). A total of 115 clusters were detected by the clustering algorithm; 33 were major clusters (Table 1) and others were considered as minor clusters embedded into major ones (Fig. 2). Average intercluster similarity observed between the 33 major clusters was 88.5% and detailed results are provided in Supplementary Tables S1 and S2. A timeline of major clusters that once detected by the system, have persisted in subsequent runs until the final one is provided in Fig. 4. Each of the 32 clusters is represented by a colour bar showing year(s) for which at least one sequence of the cluster was submitted into the LEMP-DB. Cluster 25 included a total of 1117 vaccine-like sequences (MLV, ATP, Fostera) which represented 23% of the submitted sequences (1998–2016, n = 4958). This latter cluster was then subdivided into two minor clusters, MLV-like (n = 902, RBS = 85) and ATP-like (n = 204, RBS = 100). Fostera-like sequences also grouped together (RBS = 87), but in insufficient number to be detected by the system (< 15). The remaining major clusters represented wild-type sequences and seven clusters (29, 18, 11, 14, 9, 19, 27) contained > 70 sequences each. Cluster 29 represented from 0 to 45% of sequence submissions between 2007 and 2016. In the last year (2016), 13 clusters were observed among submitted sequences, and cluster 29 was the most prevalent wild-type cluster in Quebec (n = 1308), with several nested minor clusters (results not shown). Six out of 32 wild-type clusters were not observed since 2010 (e.g. Cluster 1, 3, 15, 16, 20, 26). The mean (min-max) number of years with sequences submitted to the LEMP-DB among wild-type groupings was 8 (3–15) years. The highest number of different clusters per year (n = 22), including vaccine-like grouping was observed in 2006.

4. Discussion

An automated classification system was applied on a large PRRSV ORF5 sequence dataset. The stability of the resulting genetic clusters was assessed through time in order to evaluate the applicability of the system for rapid monitoring of endemic strains. Several methodological decisions were made to ensure that the system would provide results at least every three months to the swine industry stakeholders. This three-month period was selected to allow sufficient time to run the classification pipeline on a significant number of new submissions to assess changes in viral populations and to report them. In fact, future reports integrating clustering results will be designed, produced and sent to the swine industry in order to serve as a communication tool.

Clustal Omega was chosen as it was the fastest multiple alignment algorithm among five others evaluated while providing similar results (Lambert et al., 2019b). An approximate method for ML-tree computation (RAxML) was also selected for its capability to manage additional sequences over time. The complex evolutionary model GTR gamma was used, and the parameters were determined precisely as the number of sequences was sufficient. Moreover, this model was likely more representative of the virus evolution than a simpler model, thus

Table 2
Correspondence and assessment of stability of major clusters through time (Runs 1 to 12)[a].

Genetic clusters	Runs of classification											
	1	2	3	4	5	6	7	8	9	10	11	12
Cluster 1												
Nb of sequences	15	15	15	15	15	15	15	15	15	15	15	15
RBS value	98	99	98	98	98	98	98	98	98	99	99	98
Maximum genetic distance[b]	0.076	0.076	0.076	0.076	0.076	0.076	0.076	0.076	0.076	0.076	0.076	0.076
Cluster 2												
Nb of sequences	22	22	22	23	23	23	23	25	26	27	27	27
RBS value	72	73	74	74	74	74	73	74	73	74	73	72
Maximum genetic distance	0.088	0.088	0.088	0.091	0.091	0.091	0.091	0.098	0.098	0.098	0.098	0.098
Cluster 3												
Nb of sequences	18	18	18	18	18	18	18	18	18	18	18	18
RBS value	84	85	82	84	84	84	84	85	82	85	84	86
Maximum genetic distance	0.033	0.033	0.033	0.033	0.033	0.033	0.033	0.033	0.033	0.033	0.033	0.033
Cluster 4												
Nb of sequences	27	28	28	28	28	28	28	28	28	28	28	28
RBS value	97	97	97	97	97	97	97	97	97	97	97	97
Maximum genetic distance	0.086	0.086	0.086	0.086	0.086	0.086	0.086	0.086	0.086	0.086	0.086	0.086
Cluster 5												
Nb of sequences	58	58	58	58	58	58	58	58	58	58	58	58
RBS value	70	67	66	70	69	68	68	64	67	67	72	69
Maximum genetic distance	0.224	–	–	0.224	–	–	–	–	–	–	0.223	–
Cluster 6												
Nb of sequences	20	20	20	20	20	20	20	20	20	20	20	20
RBS value	97	94	95	96	94	96	96	96	96	97	97	96
Maximum genetic distance	0.115	0.115	0.115	0.115	0.115	0.115	0.115	0.115	0.115	0.115	0.115	0.115
Cluster 7												
Nb of sequences	15	15	15	15	15	15	15	15	15	15	15	15
RBS value	95	95	96	95	94	96	95	95	95	95	94	94
Maximum genetic distance	0.053	0.053	0.053	0.053	0.053	0.053	0.053	0.053	0.053	0.053	0.053	0.053
Cluster 8												
Nb of sequences	22	22	23	23	24	25	25	25	26	26	27	26
RBS value	91	92	91	92	77	77	77	81	82	82	83	79
Maximum genetic distance	0.027	0.027	0.027	0.027	0.029	0.035	0.035	0.035	0.035	0.035	0.035	0.035
Cluster 9												
Nb of sequences	84	84	86	87	93	95	96	96	98	98	98	98
RBS value	80	81	82	81	82	82	82	82	81	79	84	84
Maximum genetic distance	0.123	0.123	0.123	0.123	0.123	0.123	0.123	0.123	0.127	0.127	0.127	0.127
Cluster 10												
Nb of sequences	15	15	15	15	15	15	15	15	15	15	15	15
RBS value	80	79	79	80	80	77	80	81	80	80	71	72
Maximum genetic distance	0.065	0.065	0.065	0.065	0.065	0.065	0.065	0.065	0.065	0.065	0.094	0.094
Cluster 11												
Nb of sequences	99	101	105	106	109	110	110	111	112	114	114	114
RBS value	72	74	75	75	75	74	75	73	72	75	74	75
Maximum genetic distance	0.113	0.12	0.123	0.123	0.132	0.132	0.132	0.132	0.14	0.14	0.14	0.14
Cluster 12												
Nb of sequences	49	51	54	58	58	59	60	60	60	61	61	62
RBS value	92	92	91	92	93	92	93	94	93	93	91	92
Maximum genetic distance	0.118	0.117	0.121	0.121	0.121	0.121	0.125	0.125	0.125	0.125	0.125	0.125
Cluster 13												
Nb of sequences	21	21	21	21	21	21	21	21	21	21	21	21
RBS value	70	70	70	70	71	69	70	71	72	70	71	70
Maximum genetic distance	0.115	0.115	0.115	0.115	0.115	–	0.115	0.114	0.114	0.114	0.114	0.114
Cluster 14												
Nb of sequences	50	54	57	59	63	67	72	76	82	84	88	90
RBS value	77	76	74	78	76	74	78	77	78	75	76	76
Maximum genetic distance	0.12	0.12	0.12	0.12	0.12	0.124	0.124	0.132	0.137	0.137	0.137	0.137
Cluster 15												
Nb of sequences	20	20	20	20	20	20	20	20	20	20	20	20
RBS value	100	100	100	100	100	100	99	99	99	99	100	99
Maximum genetic distance	0.038	0.038	0.038	0.038	0.038	0.038	0.038	0.038	0.038	0.038	0.038	0.038
Cluster 16												
Nb of sequences	15	15	15	15	15	15	15	15	15	15	15	15
RBS value	96	96	96	96	97	97	97	96	96	96	96	96
Maximum genetic distance	0.053	0.053	0.053	0.053	0.053	0.053	0.053	0.053	0.053	0.053	0.053	0.053
Cluster 17												
Nb of sequences	16	16	16	16	16	16	16	16	16	16	16	16
RBS value	98	98	98	98	98	98	98	98	98	98	98	98
Maximum genetic distance	0.049	0.049	0.049	0.049	0.049	0.049	0.049	0.049	0.049	0.049	0.049	0.049
Cluster 18												
Nb of sequences	203	206	209	211	214	217	220	222	230	233	235	237
RBS value	76	75	78	75	74	73	77	74	74	75	74	73
Maximum genetic distance	0.202	0.202	0.202	0.202	0.202	0.202	0.202	0.202	0.201	0.202	0.226	0.226

(*continued on next page*)

Table 2 (*continued*)

Genetic clusters	Runs of classification											
	1	2	3	4	5	6	7	8	9	10	11	12
Cluster 19												
Nb of sequences	83	84	84	85	85	85	85	85	86	87	87	88
RBS value	74	75	75	72	74	74	77	76	76	76	74	75
Maximum genetic distance	0.124	0.12	0.12	0.124	0.124	0.124	0.124	0.124	0.124	0.124	0.124	0.125
Cluster 20												
Nb of sequences	15	15	15	15	15	15	15	15	15	15	15	15
RBS value	100	100	100	100	100	100	100	100	100	100	100	100
Maximum genetic distance	0.061	0.061	0.061	0.061	0.061	0.061	0.061	0.061	0.061	0.061	0.061	0.061
Cluster 21												
Nb of sequences	25	25	25	25	25	25	25	25	25	25	25	25
RBS value	71	69	69	68	69	67	67	68	67	68	67	68
Maximum genetic distance	0.105	–	–	–	–	–	–	–	–	–	–	–
Cluster 22												
Nb of sequences	27	27	27	27	27	27	27	27	27	27	27	27
RBS value	98	99	99	99	99	99	98	99	99	99	99	99
Maximum genetic distance	0.143	0.143	0.143	0.143	0.143	0.143	0.143	0.143	0.143	0.143	0.143	0.143
Cluster 23												
Nb of sequences	17	17	17	17	17	17	17	17	17	17	17	17
RBS value	100	100	100	100	100	100	100	100	100	100	100	100
Maximum genetic distance	0.058	0.058	0.058	0.058	0.058	0.058	0.058	0.058	0.058	0.058	0.058	0.058
Cluster 24												
Nb of sequences	22	22	22	22	22	22	22	22	22	22	22	22
RBS value	95	95	96	95	95	96	96	96	95	95	96	94
Maximum genetic distance	0.158	0.158	0.158	0.158	0.158	0.158	0.158	0.158	0.158	0.158	0.158	0.158
Cluster 25												
Nb of sequences	591	617	644	660	685	701	714	725	762	777	804	826
RBS value	77	76	79	76	74	77	73	74	77	77	78	78
Maximum genetic distance	0.286	0.286	0.285	0.285	0.285	0.285	0.285	0.285	0.285	0.285	0.285	0.285
Cluster 26												
Nb of sequences	17	17	17	17	17	17	17	17	17	17	17	17
RBS value	95	95	95	95	95	94	96	96	95	96	96	95
Maximum genetic distance	0.129	0.129	0.129	0.129	0.129	0.129	0.129	0.129	0.129	0.129	0.129	0.129
Cluster 27												
Nb of sequences	71	71	71	71	71	71	71	71	71	71	71	71
RBS value	85	85	87	86	87	86	86	84	82	86	88	82
Maximum genetic distance	0.186	0.186	0.186	0.186	0.185	0.185	0.186	0.185	0.185	0.186	0.185	0.185
Cluster 28												
Nb of sequences	16	16	16	16	16	16	16	16	16	16	16	16
RBS value	92	93	92	91	91	92	91	91	92	91	90	91
Maximum genetic distance	0.117	0.116	0.116	0.116	0.116	0.116	0.116	0.116	0.116	0.116	0.116	0.116
Cluster 29												
Nb of sequences	559	619	656	696	761	802	843	865	944	1007	1078	1123
RBS value	78	79	79	81	77	76	78	75	78	78	74	78
Maximum genetic distance	0.122	0.123	0.123	0.123	0.123	0.123	0.123	0.123	0.124	0.125	0.135	0.136
Cluster 30												
Nb of sequences	98	98	98	98	98	98	98	98	98	98	98	98
RBS value	68	68	70	68	70	68	68	67	66	68	69	65
Maximum genetic distance	–	–	0.26	–	0.259	–	–	–	–	–	–	–
Cluster 31												
Nb of sequences	18	18	18	18	11	11	11	11	11	11	11	18
RBS value	69	67	72	67	80	79	80	80	79	80	80	68
Maximum genetic distance	–	–	0.075	–	–	–	–	–	–	–	–	–
Cluster 32												
Nb of sequences	14	14	14	16	19	20	20	20	20	20	20	20
RBS value	92	93	91	90	91	89	92	92	90	92	90	92
Maximum genetic distance	–	–	–	0.08	0.083	0.083	0.083	0.083	0.082	0.083	0.083	0.083
Cluster 33												
Nb of sequences	14	14	14	14	15	15	15	15	15	15	15	15
RBS value	74	75	77	76	77	72	76	76	75	75	75	75
Maximum genetic distance	–	–	–	–	0.092	0.092	0.092	0.092	0.092	0.092	0.092	0.092
Cluster 34												
Nb of sequences	341	343	347	354	373	382	389	394	405	414	420	428
RBS value	68	65	67	68	64	70	67	64	64	65	61	62
Maximum genetic distance	–	–	–	–	–	0.170	–	–	–	–	–	–
Cluster 35												
Nb of sequences	12	12	13	14	14	15	16	17	18	18	18	18
RBS value	98	98	98	98	98	98	98	97	98	98	98	98
Maximum genetic distance	–	–	–	–	–	0.087	0.092	0.092	0.092	0.092	0.092	0.092
Cluster 36												
Nb of sequences		2	2	4	12	15	18	20	25	30	34	40
RBS value		100	100	100	100	100	100	100	99	100	99	99
Maximum genetic distance		–	–	–	–	0.015	0.02	0.026	0.026	0.043	0.055	0.059

(*continued on next page*)

Table 2 (*continued*)

Genetic clusters	Runs of classification											
	1	2	3	4	5	6	7	8	9	10	11	12
Cluster 37												
Nb of sequences									14	15	15	15
RBS value									87	86	86	85
Maximum genetic distance									–	0.060	0.060	0.060
Cluster 38												
Nb of sequences						14	17	19	21	23	24	24
RBS value						100	99	98	98	98	98	98
Maximum genetic distance						–	0.039	0.045	0.045	0.045	0.045	0.045

[a] Run1 included sequences from the province of Quebec submitted between 1998 and 2013. Runs 2 to 12 represented further classification performed each additional three-months to simulate ongoing classification of newly submitted sequences.

[b] The maximum pairwise evolutionary genetic distance was not computed when cluster size (≥15 sequences) or RBS value criteria (≥70) were not fulfilled. Hence the cluster was not considered for that specific period.

potentially limiting the need for a change of evolutionary model in the future. Re-estimating model parameters for each run was also considered at first, but since computation reached 30 days for the first run and grew exponentially with the number of sequences, it precluded the re-estimation of the entire tree for each subsequent run. Sequences were rather inserted into the tree of the previous run based on evolutionary parameters estimated from the first tree. Moreover, RBS was used for practical reasons as standard bootstraps were too lengthy, 30 days for 1000 replicates on the baseline tree. Finally, the clustering steps (Section 2.3.1.3) applied on each run were only slightly affected by the number of sequences, insuring that the overall system could be used for larger datasets.

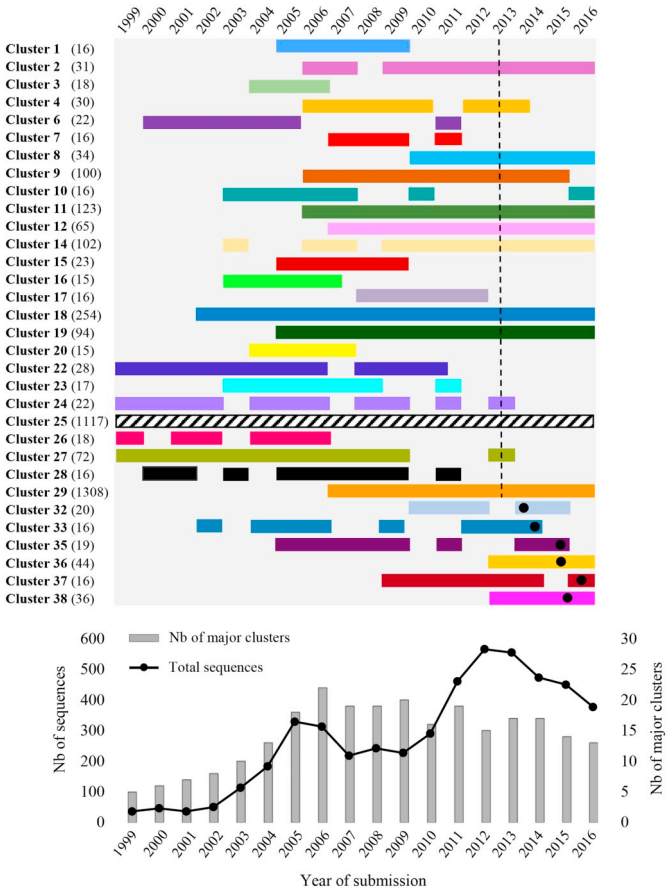

Fig. 4. Gantt chart describing the temporal distribution of sequence submissions to the LEMP-DB according to the 32 major clusters (≥15 sequences, ≥70 RBS) that have persisted since their detection by the classification system. Hatched bar identifies a cluster of vaccine-like sequences (#25). Dashed line indicates clusters detected in the first run performed on sequences gathered in the LEMP-DB up to December 2013. Black dots on the Gantt chart indicate when clusters identified after the first run (n = 6) were first detected. Histogram of the number of major clusters per year with line representing the total number of sequences submitted. The eight sequences submitted in 1998 did not belong to any cluster and are not shown in the graph.

All major clusters but two identified in the first run were also detected in subsequent runs, suggesting that the criteria used to define a cluster insured a certain stability of clustering results through time. This stability is important as the purpose of the classification was to give an appraisal of changes in genetic clusters according to time and region for field end users. In addition, all sequences that were classified in a cluster remained in the same cluster in subsequent runs, whatever was the number of sequences added to the system. This latter finding was observed even for clusters with a RBS value slightly lower than the threshold for one or more runs (Clusters 5, 13, 21, 30, 34) which suggests that some clusters below the 70% threshold might also be detectable through time. Stability of clustering results could be partly due to the large dataset used in the baseline tree (> 3000), which is among the largest that has been used for either PRRSV research or surveillance projects worldwide (Balka et al., 2018; Shi et al., 2010). The proportion of classified strains reached 68% in the first run of classification, with an additional 7% until Run 12, indicating that most of the clustering process for major clusters occurred in the first run. When examining major clusters with 15 to 20 sequences identified in Run 1 (10 clusters) or those newly detected between Runs 2 and 12 (6 clusters), the great majority had good support values (≥90%) throughout the subsequent runs, indicating that the system was able to identify stable clusters even with no > 15 sequences. Maybe temporary clusters could be created for groups having borderline thresholds that should be examined in the future. This would certainly lower the number of unclassified strains, which represents a limit of our system.

The number of clusters detected was deemed suitable for ongoing surveillance purposes, which requires a balance between the smallest number of clusters for communication purposes among end users and a sufficient number to allow a within cluster diversity that is phylogenetically meaningful. The classification system revealed major changes in PRRSV sequences submitted by swine veterinarians at the provincial scale over a fifteen-year period. As an example, the first sequences of cluster 29 were submitted in 2007 but since then, it has largely expanded and is now the most important wild-type cluster found in Quebec. The automated system detected several minor clusters that might have emerged at different times within this particular cluster, revealing the importance of a system working at different levels of the tree (major vs. minor clusters) and keeping track of the hierarchical

structure to better monitor changes in PRRSV submitted sequences through time. To better characterize the virus, it would be advisable to apply Bayesian analysis on some particular clusters that would have a manageable size for this type of analysis (Alkhamis et al., 2017; Alkhamis et al., 2016; Baele et al., 2018; Shi et al., 2010). In addition, this system should not be used to determine whether two sequences are similar or not, since intracluster diversity in major and most minor clusters is beyond the 97.5–98% pairwise similarity threshold often used to determine that two sequences are similar (Lambert et al., 2012; Larochelle et al., 2003; Murtaugh, 2012). Other tools have been developed by the research team to investigate sources of contamination in the field or for epidemiological studies (Lambert et al., 2019a).

The time interval between the first sampling date and detection of the cluster by the system was highly variable for the 6 new clusters identified, between 1 and > 10 yrs. Thus, the ability of the system to detect newly emerging strains or new viral introduction is limited. However, since, the system can keep track of all groups, even those not meeting size criterion, it could possibly be used to monitor the size of small groups to enhance detection of emerging cluster. In addition, the system could be adapted to picture unclassified sequences on the overall ML-tree; the minimum pairwise distance observed between a sequence from a particular major cluster and sequences from a reference set may help detect introductions of foreign strains from other provinces or countries (Lambert et al., 2018). Finally, the current pipeline would largely benefit from including phylodynamic approaches that would be better at detecting emerging strains.

This paper documented an automated approach that was successfully applied on a large ML-phylogeny to identify major clusters and to monitor them through time. Hence, the system avoids relying on visual appraisal of a large tree to identify, manually extract and keep track of the hierarchical structure of the genetic clusters through time. It also avoids the need for selecting representative sequences of the tree to be able to compute Bayesian analysis (Shi et al., 2010). The system was designed for a preliminary assessment of genetic diversity within a large dataset, and should be viewed as the first step into phylogenetic analysis, which does not preclude further evaluation on the evolution of viral populations. In fact, further Bayesian analyses should be performed to better characterize the evolutionary rate of particular clusters over time.

5. Conclusion

An automated approach to identify genetic clusters from a large phylogenetic tree was successfully developed to monitor changes in PRRSV sequences submitted by swine veterinarians at the provincial level. The system could be useful for epidemiologic research, surveillance activities and decision-making for different stakeholders of the industry to better organize control activities based on the knowledge of viral populations.

Supplementary data to this article can be found online at https://doi.org/10.1016/j.meegid.2019.04.014.

Funding

This work was supported by the Swine Innovation Porc within the Swine Cluster 2 (project #1343), funded by Agriculture and Agri-Food Canada through the AgriInnovation Program, provincial producer organizations and industry partners. Financial support was also provided by the Éleveurs de Porcs du Québec and Zoetis Canada inc.

Contributions

M-È Lambert, J Arsenault, P Audet, B Delisle and S D'Allaire designed the methodology of the classification system. B Delisle performed the recombinant detection analyses. P Audet coded all programs required for the classification system. P Audet and M-È Lambert performed descriptive statistics. All authors analysed and interpreted data. M-È Lambert wrote the first draft of the manuscript. All authors revised and approved the final manuscript.

Conflicts of interest

The authors declare no conflict of interest.

Acknowledgments

The authors would like to thank all Quebec swine veterinarians (AVIA) and diagnostic laboratories for their willingness to share data.

References

Alkhamis, M.A., Perez, A.M., Murtaugh, M.P., Wang, X., Morrison, R.B., 2016. Applications of bayesian phylodynamic methods in a recent U.S. porcine reproductive and respiratory syndrome virus outbreak. Front. Microbiol. 7, 67. https://doi.org/10.3389/fmicb.2016.00067.

Alkhamis, M.A., Arruda, A.G., Morrison, R.B., Perez, A.M., 2017. Novel approaches for spatial and molecular surveillance of porcine reproductive and respiratory syndrome virus (PRRSv) in the United States. Sci. Rep. 7, 4343. https://doi.org/10.1038/s41598-017-04628-2.

Alvarez, J., Valdes-Donoso, P., Tousignant, S., Alkhamis, M., Morrison, R., Perez, A., 2016. Novel analytic tools for the study of porcine reproductive and respiratory syndrome virus (PRRSv) in endemic settings: lessons learned in the US. Porcine Health Manag. 2, 1–9. https://doi.org/10.1186/s40813-016-0019-0.

An, T.Q., Zhou, Y.J., Liu, G.Q., Tian, Z.J., Li, J., Qiu, H.J., Tong, G.Z., 2007. Genetic diversity and phylogenetic analysis of glycoprotein 5 of PRRSV isolates in mainland China from 1996 to 2006: coexistence of two NA-subgenotypes with great diversity. Vet. Microbiol. 123, 43–52. https://doi.org/10.1016/j.vetmic.2007.02.025.

Arruda, A.G., Friendship, R., Carpenter, J., Hand, K., Ojkic, D., Poljak, Z., 2017. Investigation of the occurrence of porcine reproductive and respiratory virus in swine herds participating in an area regional control and elimination project in Ontario, Canada. Transbound. Emerg. Dis. 64, 89–100. https://doi.org/10.1111/tbed.12343.

Baele, G., Dellicour, S., Suchard, M.A., Lemey, P., Vrancken, B., 2018. Recent advances in computational phylodynamics. Curr. Opin. Virol. 31, 24–32. https://doi.org/10.1016/j.coviro.2018.08.009.

Balka, G., Podgorska, K., Brar, M.S., Balint, A., Cadar, D., Celer, V., Denes, L., Dirbakova, Z., Jedryczko, A., Marton, L., Novosel, D., Petrovic, T., Sirakov, I., Szalay, D., Toplak, I., Leung, F.C., Stadejek, T., 2018. Genetic diversity of PRRSV 1 in Central Eastern Europe in 1994-2014: origin and evolution of the virus in the region. Sci. Rep. 8, 7811. https://doi.org/10.1038/s41598-018-26036-w.

Brar, M.S., Shi, M., Ge, L., Carman, S., Murtaugh, M.P., Leung, F.C., 2011. Porcine reproductive and respiratory syndrome virus in Ontario, Canada 1999 to 2010: genetic diversity and restriction fragment length polymorphisms. J. Gen. Virol. 92, 1391–1397. https://doi.org/10.1099/vir.0.030155-0.

Cai, H.Y., Alexander, H., Carman, S., Lloyd, D., Josephson, G., Maxie, M.G., 2002. Restriction fragment length polymorphism of porcine reproductive and respiratory syndrome viruses recovered from Ontario farms, 1998-2000. J. Vet. Diagn. Investig. 14, 343–347. https://doi.org/10.1177/104063870201400415.

Delisle, B., Gagnon, C.A., Lambert, M.E., D'Allaire, S., 2012. Porcine reproductive and respiratory syndrome virus diversity of eastern Canada swine herds in a large sequence dataset reveals two hypervariable regions under positive selection. Infect. Genet. Evol. 12, 1111–1119. https://doi.org/10.1016/j.meegid.2012.03.015.

Diaz, I., Gimeno, M., Darwich, L., Navarro, N., Kuzemtseva, L., Lopez, S., Galindo, I., Segales, J., Martin, M., Pujols, J., Mateu, E., 2012. Characterization of homologous and heterologous adaptive immune responses in porcine reproductive and respiratory syndrome virus infection. Vet. Res. 43, 30. https://doi.org/10.1186/1297-9716-43-30.

Holtkamp, D.J., Kliebenstein, J.B., Neumann, E.J., Zimmerman, J.J., Rotto, H.F., Yoder, T.K., Wang, C., Yeske, P.E., Mowrer, C.L., Haley, C.A., 2013. Assessment of the economic impact of porcine reproductive and respiratory syndrome virus on United States pork producers. J. Swine Health Prod. 21, 72–84.

Kappes, M.A., Faaberg, K.S., 2015. PRRSV structure, replication and recombination: origin of phenotype and genotype diversity. Virology 479-480, 475–486. https://doi.org/10.1016/j.virol.2015.02.012.

Lambert, M.E., Arsenault, J., Poljak, Z., D'Allaire, S., 2012. Correlation among genetic, Euclidean, temporal, and herd ownership distances of porcine reproductive and respiratory syndrome virus strains in Quebec, Canada. BMC Vet. Res. 8, 76. https://doi.org/10.1186/1746-6148-8-76.

Lambert, M.E., Audet, P., Delisle, B., Arsenault, J., D'Allaire, S., 2019a. Porcine reproductive and respiratory syndrome virus: web-based interactive tools to support surveillance and control initiatives. Porcine Health Manag. 5 (10). https://doi.org/10.1186/s40813-019-0117-x.

Lambert, M.E., Arsenault, J., Delisle, B., Audet, P., Poljak, Z., D'Allaire, S., 2019b. Impact of alignment algorithm on the estimation of pairwise genetic similarity of porcine reproductive and respiratory syndrome virus (PRRSV). BMC Vet. Res. 15, 1–10. https://doi.org/10.1186/s12917-019-1890-0.

Lambert, M.E., Delisle, B., Arsenault, J., Poljak, Z., D'Allaire, S., 2018. Diversity of PRRSV strains circulating in Canada. In: 10th European Symposium of Porcine Health

Management, Barcelona, Spain, (p. 441).
Larochelle, R., D'Allaire, S., Magar, R., 2003. Molecular epidemiology of porcine reproductive and respiratory syndrome virus (PRRSV) in Quebec. Virus Res. 96, 3–14. https://doi.org/10.1016/S0168-1702(03)00168-0.
Murtaugh, M.P., 2012. Use and Interpretation of Sequencing in PRRSV Control Programs, Allen D. Leman Swine Conference, Minnesota, United States. pp. 49–55.
Ostrowski, M., Galeota, J.A., Jar, A.M., Platt, K.B., Osorio, F.A., Lopez, O.J., 2002. Identification of neutralizing and nonneutralizing epitopes in the porcine reproductive and respiratory syndrome virus GP5 ectodomain. J. Virol. 76, 4241–4250. https://doi.org/10.1128/JVI.76.9.4241-4250.2002.
Plagemann, P.G., 2004. GP5 ectodomain epitope of porcine reproductive and respiratory syndrome virus, strain Lelystad virus. Virus Res. 102, 225–230. https://doi.org/10.1016/j.virusres.2004.01.031.
Plagemann, P.G., Rowland, R.R., Faaberg, K.S., 2002. The primary neutralization epitope of porcine respiratory and reproductive syndrome virus strain VR-2332 is located in the middle of the GP5 ectodomain. Arch. Virol. 147, 2327–2347. https://doi.org/10.1007/s00705-002-0887-2.
Popescu, L.N., Trible, B.R., Chen, N., Rowland, R.R.R., 2017. GP5 of porcine reproductive and respiratory syndrome virus (PRRSV) as a target for homologous and broadly neutralizing antibodies. Vet. Microbiol. 209, 90–96. https://doi.org/10.1016/j.vetmic.2017.04.016.
Rahe, M.C., Murtaugh, M.P., 2017. Effector mechanisms of humoral immunity to porcine reproductive and respiratory syndrome virus. Vet. Immunol. Immunopathol. 186, 15–18. https://doi.org/10.1016/j.vetimm.2017.02.002.
Shi, M., Lam, T.T., Hon, C.C., Murtaugh, M.P., Davies, P.R., Hui, R.K., Li, J., Wong, L.T., Yip, C.W., Jiang, J.W., Leung, F.C., 2010. Phylogeny-based evolutionary, demographical, and geographical dissection of north American type 2 porcine reproductive and respiratory syndrome viruses. J. Virol. 84, 8700–8711. https://doi.org/10.1128/JVI.02551-09.
Sievers, F., Wilm, A., Dineen, D., Gibson, T.J., Karplus, K., Li, W., Lopez, R., McWilliam, H., Remmert, M., Soding, J., Thompson, J.D., Higgins, D.G., 2011. Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal omega. Mol. Syst. Biol. 7, 539. https://doi.org/10.1038/msb.2011.75.
Stadejek, T., Stankevicius, A., Murtaugh, M.P., Oleksiewicz, M.B., 2013. Molecular evolution of PRRSV in Europe: current state of play. Vet. Microbiol. 165, 21–28. https://doi.org/10.1016/j.vetmic.2013.02.029.
Vasylyeva, T.I., Friedman, S.R., Paraskevis, D., Magiorkinis, G., 2016. Integrating molecular epidemiology and social network analysis to study infectious diseases: towards a socio-molecular era for public health. Infect. Genet. Evol. https://doi.org/10.1016/j.meegid.2016.05.042.
Wang, C., Lee, F., Huang, T.S., Pan, C.H., Jong, M.H., Chao, P.H., 2008. Genetic variation in open reading frame 5 gene of porcine reproductive and respiratory syndrome virus in Taiwan. Vet. Microbiol. 131, 339–347. https://doi.org/10.1016/j.vetmic.2008.04.027.
Wesley, R.D., Mengeling, W.L., Lager, K.M., Clouser, D.F., Landgraf, J.G., Frey, M.L., 1998. Differentiation of a porcine reproductive and respiratory syndrome virus vaccine strain from north American field strains by restriction fragment length polymorphism analysis of ORF 5. J. Vet. Diagn. Investig. 10, 140–144. https://doi.org/10.1177/104063879801000204.
Yoon, S.H., Song, J.Y., Lee, C.H., Choi, E.J., Cho, I.S., Kim, B., 2008. Genetic characterization of the Korean porcine reproductive and respiratory syndrome viruses based on the nucleocapsid protein gene (ORF7) sequences. Arch. Virol. 153, 627–635. https://doi.org/10.1007/s00705-007-0027-0.
Yoshii, M., Kaku, Y., Murakami, Y., Shimizu, M., Kato, K., Ikeda, H., 2005. Genetic variation and geographic distribution of porcine reproductive and respiratory syndrome virus in Japan. Arch. Virol. 150, 2313–2324. https://doi.org/10.1007/s00705-005-0549-2.
Zhang, J., Zheng, Y., Xia, X.Q., Chen, Q., Bade, S.A., Yoon, K.J., Harmon, K.M., Gauger, P.C., Main, R.G., Li, G., 2017. High-throughput whole genome sequencing of porcine reproductive and respiratory syndrome virus from cell culture materials and clinical specimens using next-generation sequencing technology. J. Vet. Diagn. Investig. 29, 41–50. https://doi.org/10.1177/1040638716673404.