RESEARCH REPORT
SWINE HEALTH

Title: United States Swine Pathogen Database – **NPB #16-222**

Investigator: Dr. Kay Faaberg, Ph.D.

Institution: USDA-ARS National Animal Disease Center

Date Submitted: February 20, 2018

Industry Summary

Veterinarians have long needed a centralized database to compare and contrast the pathogen nucleotide sequences of their clinical samples to those seen concurrently in the field throughout the United States. In addition, meaningful molecular epidemiology studies require regular access to current nucleotide sequence data to provide criteria for updating vaccine composition and the identification of emerging viral threats. Presently, the majority of genetic sequence data derived from pathogens infecting US swine are not readily available to researchers outside of veterinary diagnostic laboratories. To overcome this problem, the National Pork Board granted the first year of funding (NPB #16-222) for a two-year project that would create a combined database for secure housing of all nucleotide sequences produced in the diagnostic laboratories of key universities. The funded NPB research has now established collaborations between the major swine diagnostic laboratories (South Dakota State University and Kansas State University initially, and then once the database is fully established, the University of Minnesota and Iowa State University) and the USDA-ARS National Animal Disease Center in order to generate a centralized sequence and related clinical information repository. Genetic sequences from diagnostic submissions [e.g., porcine reproductive and respiratory syndrome virus (PRRSV); Seneca virus A (SVA); porcine epidemic diarrhea virus (PEDV); others] are currently housed in a private centralized sequence database with an interactive public website that we have under development (United States Swine Pathogen Database). The submitting laboratories and other interested researchers will have access to all of the information stored on our database to view, query, and download genomic and other bibliographic information for molecular epidemiological analyses. The ultimate goal of this resource is to provide a resource for the swine health community to facilitate the development of vaccines, diagnostics and therapeutics to combat endemic (e.g., PRRSV) and emerging viruses (e.g., SVA).

Contact

Dr. Kay Faaberg, kay.faaberg@ars.usda.gov
Dr. Tavis Anderson, tavis.anderson@ars.usda.gov

Keywords

Database, epidemiology, evolution, PRRSV, PEDV, SVA, swine

Scientific Abstract

Veterinary diagnostic laboratories derive partial nucleotide sequences of thousands of isolates of PRRSV, SVA, PEDV and other coronaviruses annually, and with the advent of next generation sequencing, near full-length genomes are also rapidly produced. Presently, the sequence data are only released to the client, as the samples are associated with sensitive information. At the same time, however, this information is critical and can provide objective criteria for: 1) vaccine design; 2) determining when and how fast pathogens are spreading across the landscape; and 3) identifying transmission hotspots. In tandem with the USDA Agriculture Research Service Big Data initiative, we have generated a centralized nucleotide sequence relational database housed at the National Animal Disease Center. We have implemented the Tripal toolkit, using Drupal for Content Management, and the Chado relational database schema. Hosting is via a BlueHost cloud service with resource scaling, dedicated support for the prevention of data theft and control of database vulnerabilities, and the service is ~2X faster for general database query tasks and ~6X faster for more complex analyses. Each genetic sequence housed in the database contains at a minimum four core data items: genomic information; the date of collection; the US State the collection was made; and a unique identifier. Additionally, custom curation and annotation pipelines have determined PRRSV genotype (Type 1 or 2), the location of open reading frames and non-structural proteins, generated amino acid sequences, and identified putative frame shifts. This repository is currently private but will be publicly accessible in the future. The resource will ultimately provide researchers timely access to sequences discovered by highly qualified veterinary diagnosticians, allowing for biological data mining and epidemiological studies. The result of this effort will be a better understanding concerning the appearance of novel viruses in the United States, how these novel isolates are moving through the US and abroad, and discovering new patterns of biological consequence.

Introduction

The National Center for Biotechnology Information (NCBI) sequence repository GenBank is the unequivocal master database of all scientific genetic data. However, GenBank is too large to easily facilitate precise molecular biological inquiries, does not mine all data to make sure it is genetically accurate, and concentrates its efforts into successfully storing genomic data that is uploaded by researchers. In addition, swine veterinary diagnostic laboratories are in service to their clients and generally do not have the funds or time to spend on uploading sequences to GenBank, which demands extensive annotation of each nucleotide sequence including identifying start and stop sites for translation, the type of virus, potential open reading fames, and the inclusion of data useful for epidemiologic studies (e.g., age of pig, collection location). Thus, swine disease researchers, outside of diagnostic laboratories, have no method with which to identify novel isolates, the length of time a nucleotide sequence is seen in the swine population, where pathogens emerged, re-emergence of an identical pathogen seen previously, and other knowledge which may be applied to improve animal health.

To address this problem for PRRSV, the National Pork Board funded the development of the porcine reproductive and respiratory syndrome virus database (prrsvdb) from 2005-2008. The prrsvdb archived over 13,000 PRRSV ORF5 sequences from both Type 1 (European) and Type 2 (North American) isolates from predominantly the UMN and SDSU VDLs, and deposited over 8200 unique sequence submissions to GenBank. These sequences included many index PRRSV derived from early 1990 field isolates and was used heavily by molecular epidemiologists and other researchers worldwide (e.g., (1-3)). Unfortunately, long term support for the database was elusive and it was terminated in 2009 with the unpublished data archived, but not accessible to the research community.

With the advent of a national Big Data initiative by the USDA, the Agricultural Research Service has developed the basic infrastructure to establish and support a database that hosts swine pathogen genomic data. This new database will, at conclusion, incorporate nucleotide sequences from the major veterinary diagnostic laboratories for swine health and disease – South Dakota State University,

Kansas State University, University of Minnesota and Iowa State University. This will be achieved by establishing and maintaining a swine pathogen genetic repository (United States Swine Pathogen Database) that operates as a web-based, curated, stable, relational database as part of the next-generation ARS SCINet. The core function of this database is to collect, store, annotate and view and query genomic data. The database has first tackled the collection, curation, and annotation of PRRSV (the most sequences and extremely variable swine pathogen) with standard pipelines developed that will allow SVA, PEDV and other coronaviruses to be incorporated in the future. Given alternate resources for swine influenza A virus (fludb.org), the addition of these sequence data is a long-term goal that will result in a comprehensive database that covers all major pathogens infecting US swine.

Objectives

The major objectives of the project were:

A. Establish and maintain a web-based, curated, stable relational database of swine pathogens.
B. Centralize and standardize genomic data from regional veterinary diagnostic laboratories.
C. Display, query, and download annotated genomic data facilitating research needs

Materials & Methods

Infrastructure for a web-based, curated relational database of swine pathogens

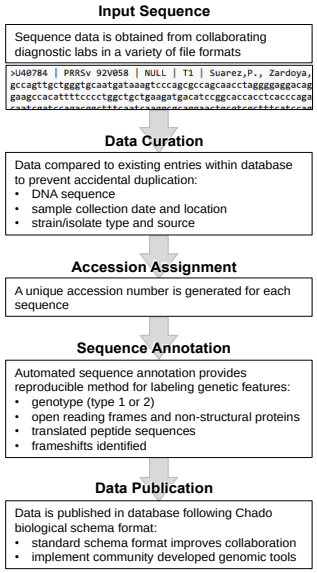
The Agricultural Research Service has built a next-generation science network within Internet2 infrastructure named SCINet. It is a dedicated scientific research network for data computing, and connects six core ARS locations (Ames IA, Stoneville MS, Albany CA, Beltsville MD, Clay Center NE, Fort Collins CO) with other research organizations and individuals, and serves as a research data conduit with high-performance computing capabilities. The high-performance computing (HPC) component of SCINet consists of 1600 compute cores with 16 terabytes (TB) of total RAM, 64 TB of total local storage, and 2 petabytes (PB) of raw, shared but secure and privately managed storage. In addition, SCINet and the HPC are integrated within Amazon Web Services, allowing the system to use Amazon EC2 computing and secure data archiving facilities.

Construction of a curated relational database of swine pathogens

The US Swine Pathogen Database was built, developed, and is currently maintained by the ORISE scientist (Blake Inderski) funded by the National Pork Board and ARS with guidance and assistance from Drs. Tavis Anderson and Kay Faaberg. The database has three components: the first for data integration, submission, cleaning, and curation (staging database); the second is a locally stored and archived read-only database containing the highly curated data (production database); and the third is the user-interface and query database. The staging database is populated by data downloaded from public data sources (e.g., NCBI Genbank), the original prrsvdb, and is in the process of incorporating data submitted by our veterinary diagnostic laboratory partners.

Downloaded and submitted data is initially quarantined where it undergoes a series of automated data curation steps and data validation (Figure 1). This ensures data integrity and quality before it is released to the production database (e.g., data that does not fulfill minimum requirements is not be accepted, but is maintained in the staging database). This process involves custom python scripts that: 1) standardize input data; 2) filter and identify low quality data; 3) anonymize the submission through the random assignment of a unique identifier; and 4) annotate the sequence (Figure 1). These steps also take advantage of tools within Tripal and Chado (4, 5). Specifically, the database uses the Chado relational database schema and framework for the storage of genomic sequence data, its annotation with user-provided descriptive information (e.g., date of collection, age of pig, notes and comments on clinical signs), and inferred information such as frameshifts and location of open reading frames. The Chado system is flexible and allows the storage, integration, and presentation of genome and epidemiologic data. This database has also developed a custom search tool using PHP, Javascript, and CSS that allows users to search, select, and download data for research purposes.

A.



B.

Unique ID	GenBank Accession	Author/s	Strain or Isolate Name	Sample Source	Sample Collection Date	Sample Collection Location	Genotype	Sequence Length	Translation Type/s
GB:0000000	U40784	Suarez,P.; Zardoya,R.	PRRSv 92V0				T1	257	ORF7_NT1
GB:0000000	U40783	Suarez,P.; Zardoya,R.	PRRSv AV30				T1	257	ORF7_NT1
GB:0000000	U40782	Suarez,P.;	PRRSv 8D				T1	257	ORF7_NT1

Figure 1. Conceptual model describing the (A) automated pipeline implemented in the US Swine Pathogen Database that takes raw sequence data to (B) fully anonymized and annotated virus sequence record in the relational database.

Our approach in construction and curation of this relational database is to use open source tools that are technologically advanced but have a track record of successful implementation in the construction of genomic databases. The current location of the database is within secured ARS SCINet infrastructure, with the public facing website in development on BlueHost cloud servers: the integration of the database with ARS infrastructure and with cloud services guarantees scalability, availability, and extensibility as the database grows.

Results

A. Establish and maintain a web-based, curated, stable relational database of swine pathogens.

The US Swine Pathogen Database was developed using Tripal (4), a construction toolkit for online genome databases. We used the modular relational database schema Chado (5) to store the sequence and associated metadata. The database employs the use of controlled vocabularies such as the Sequence Ontology (6), and Gene Ontology (7): this approach ensures that each component of a sequence record has a clear definition and relationships to all other components in the record. The database is currently implemented on a UNIX operating system (Ubuntu 16.04 LTS), an Apache web server (http://httpd.apache.org/), and a PostgreSQL database server (http://postgres.org). In addition, our Tripal site development is based upon Drupal (http://drupal.org), an extensible modular Content Management System.

The strength of our framework is that that it ensures data standardization and is nested within a larger user community ensuring continued relevance and support.

Data curation and annotation

The US Swine Pathogen Database, to handle the large amount of RNA sequences generated during this project, used an automated data analysis pipeline to ascertain the presence of potential sequence artifacts or other possible contributors to poor data quality prior to inclusion in the production database (Figure 1). The pipeline flags sequences when autocuration fails and these data are reported to the developer who can use them to correct data prior to final submission to the database. During the curation process, the pipeline aligns each sequence to a consensus whole genome reference sequence from NCBI Genbank; aligned sequences are then assessed for quality (i.e., insertions, deletions, complete or incomplete gene, identification of ambiguities and poor quality sequence data), open reading frame discovery, functional annotation, and general genetic lineage (i.e., Type 1 vs Type 2 Porcine Reproductive and Respiratory Syndrome Virus). If a sequence has quality issues, and the provenance of the sequence is known, these may be corrected. Importantly, each sequence receives an anonymous accession number and only the submitting diagnostic lab maintains the link between the original diagnostic lab report and the database record. More than 42,000 PRRSV sequences have been curated, annotated, and incorporated into the production database (Figure 2).

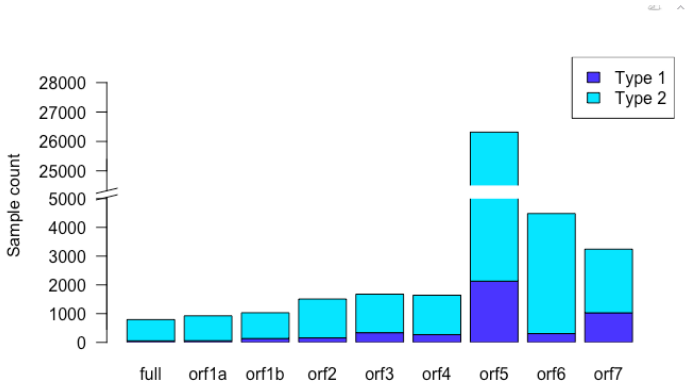


Figure 2. Porcine reproductive and respiratory syndrome virus sequence information currently housed in the US Swine Pathogen Database. Following annotation by the custom pipeline, the data has been binned by whether it is a complete genome (full), the presence of a complete open reading frame, and colored as either Type 1 PRRSV (European origin) or to Type 2 PRRSV (North American origin).

Custom search interface and public availability

A custom search tool was developed within the US Swine Pathogen database using PHP with JavaScript and Cascading Style Sheets (CSS) (Figure 3).

The database is currently housed on private secure servers within ARS SCINet. Following additional testing, a public-facing website that will allow user access will be released (provisionally at swinepathogendb.org) in fall of 2018. The public-facing website is hosted on cloud servers allowing for reliable and scalable service as the database continues to grow.

B. Centralize and standardize genomic data from regional veterinary diagnostic laboratories.

This objective was to have started about three months ago and continue into the second year, but we were delayed in identifying a suitable candidate for the software production of the database. Therefore, this research aim is in its initial stages. The first two laboratories we are working with are the Animal Disease Research and Diagnostic Laboratory of South Dakota State University and the Kansas State Veterinary Diagnostic Laboratory. We opted to work with these two laboratories first because they were most enthusiastic and their data, while immense, would not be quite as cumbersome as the other two laboratories. We plan to approach the other two laboratories when the sequence downloading process has been streamlined. Presently, we have established a secure web based data sharing portal within ARS SCINet, restricting access to five employees of SDSU: SDSU will begin downloading data very soon. KSU will begin this process shortly thereafter.

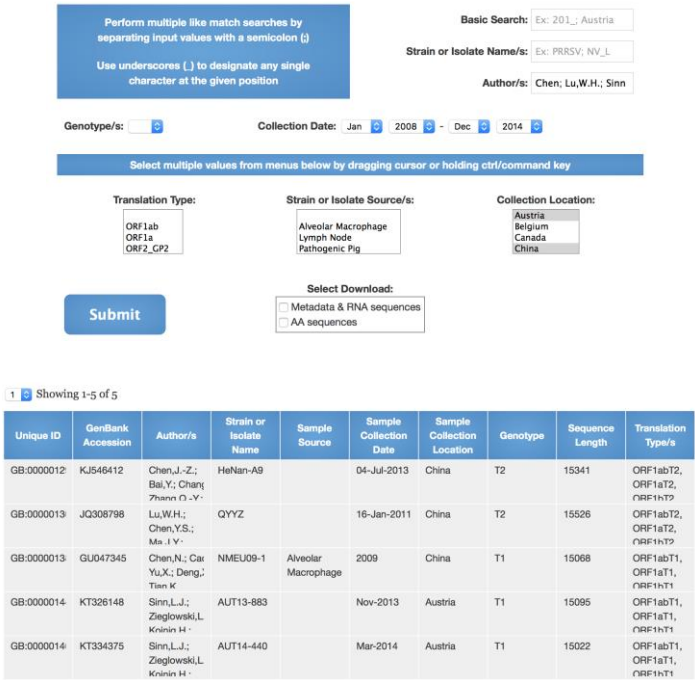


Figure 3. Custom search tool developed for the Swine Pathogen Database. All data fields are searchable with a minimum required content for a sequence to be included in the production database being: sequence information; collection date; collection location; and a unique identifier. Additional information is developed through annotation pipelines.

C. Display, query, and download annotated genomic data facilitating research needs

While the US Swine pathogen database currently has access to over 40,000 nucleotide sequences at present and they could be displayed queried and downloaded, none of the present day sequences and related data have been added. This aim will be implemented during Year 2, funded by ARS.

Discussion

The diversity of viruses that currently circulate in swine continues to increase: this poses a challenge for producers, as established control measures are unlikely to work. In addition, novel viruses appear to periodically emerge in swine populations (e.g., SVA, PEDV) which can lead to establishment and persistence of antigenically distinct viruses that do not have appropriate vaccines because there is limited information from which to derive rational polyvalent formulations. Thus, given that genetic makeup of viruses continually changes, monitoring the patterns of genetic change of viruses in swine is needed to identify possible emerging threats, and also help control endemic viruses. Our database, once released to the public, will aid in improving agricultural production and preparedness for endemic and novel swine pathogens by allowing for the identification of important changes in virus evolution and providing a benchmark from which to measure success of intervention strategies. These interventions include timely vaccine and diagnostic updates for use in swine, as well as factors to prevent infection and transmission, such as changes in production practices or farm management.

The database and curation pipeline has provided a recoding system that will standardize data from the major veterinary diagnostic labs. Further, the search interface allows the dissemination of these data in standard format (e.g., FASTA sequence file) based upon user queries. The queries are based on: virus; gene; date of collection; location of collection; and if additional metadata are available (e.g., isolate source). This is simple, but because the interface allows researchers to screen data rapidly it has the power to provide information for vaccine design, determining when and how fast pathogens are spreading across the landscape, and identifying transmission hotspots at a coarse spatial scale. All these analyses require large-scale genomic information, the date the virus was collected, and the location (state) of collection and these data have not been available, or have been addressed using small – and perhaps inappropriate – datasets. Thus, the utility of this resource, through the integration of surveillance and diagnostic samples will provide considerable power to molecular epidemiological studies conducted in the future.

The major benefit of these data is that genomic information of thousands of sequences may allow for the identification of critical amino acid substitutions associated with particular genetic clades of viruses. Understanding these critical substitutions can be used inform vaccine updates and composition (8-10). Second, through the standardized availability of date of collection on all data, users may create time-scaled phylogenies allowing the time of emergence for novel viral isolates to be determined, and evolutionary rate of change may be calculated. Third, this database will facilitate the analysis of viral spread across the landscape using the state level geographical information provided by the diagnostic laboratories (e.g., 11, 12). Through comparative analysis of the localities of the sequences and the reconstructed phylogeny, the migration history of virus isolate can be traced. This provides critical information for determining where a virus came from, where it may be going, and how fast it may be moving there. The database will also be used to assess where on each pathogen genome nucleotide and amino acid changes are more rapidly occurring, where insertions and deletions are evolving, and where other key translational events are located, all of importance to molecular virology.

The US Swine Pathogen Database uses state of the art network infrastructure and computational power developed and maintained by ARS. It is secure and stable and our work presently integrates two of the major swine veterinary diagnostic laboratories in the USA. Consequently, our goal as we incorporate more pathogens and publish the website is to provide the most up-to-date collection of swine viral pathogens in the world. In addition, it provides an opportunity for new collaborations to emerge between the investigators who have strengths in different but complementary fields. Collectively, reducing the impact of viral pathogens in US swine requires a fundamental knowledge of what viruses are circulating in the population: our proposal has made good progress to achieving this, and will provide a tool for investigators to develop rational and representative vaccines which will reduce viral burdens, decreasing the economic burden of viral disease and improving animal health.

References

1. **Nguyen VG**, **Kim HK**, **Moon HJ**, **Park SJ**, **Chung HC**, **Choi MK**, **Park BK**. 2014. A Bayesian phylogeographical analysis of type 1 porcine reproductive and respiratory syndrome virus (PRRSV). Transbound Emerg Dis **61**:537–545.

2. **Shi M**, **Lam TT-Y**, **Hon C-C**, **Hui RK-H**, **Faaberg KS**, **Wennblom T**, **Murtaugh MP**, **Stadejek T**, **Leung FC-C**. 2010. Molecular epidemiology of PRRSV: a phylogenetic perspective. Virus Research **154**:7–17.

3. **Anderson TK**, **Laegreid WW**, **Cerutti F**, **Osorio FA**, **Nelson EA**, **Christopher-Hennings J**, **Goldberg TL**. 2012. Ranking viruses: measures of positional importance within networks define core viruses for rational polyvalent vaccine development. Bioinformatics **28**:1624–1632.

4. **Ficklin SP**, **Sanderson L-A**, **Cheng C-H**, **Staton ME**, **Lee T**, **Cho I-H**, **Jung S**, **Bett KE**, **Main D**. 2011. Tripal: a construction toolkit for online genome databases. Database **2011**:bar044.

5. **Mungall CJ**, **Emmert DB**, **The FlyBase Consortium**. 2007. A Chado case study: an ontology-based modular schema for representing genome-associated biological information. Bioinformatics **23**:i337–i346.

6. **Eilbeck K**, **Lewis SE**, **Mungall CJ**, **Yandell M**, **Stein L**, **Durbin R**, **Ashburner M**. 2005. The Sequence Ontology: a tool for the unification of genome annotations. Genome Biology **6**:R44.

7. **Gene Ontology Consortium**, **Blake JA**, **Dolan M**, **Drabkin H**, **Hill DP**, **Li N**, **Sitnikov D**, **Bridges S**, **Burgess S**, **Buza T**, **McCarthy F**, **Peddinti D**, **Pillai L**, **Carbon S**, **Dietze H**, **Ireland A**, **Lewis SE**, **Mungall CJ**, **Gaudet P**, **Chrisholm RL**, **Fey P**, **Kibbe WA**, **Basu S**, **Siegele DA**, **McIntosh BK**, **Renfro DP**, **Zweifel AE**, **Hu JC**, **Brown NH**, **Tweedie S**, **Alam-Faruque Y**, **Apweiler R**, **Auchinchloss A**, **Axelsen K**, **Bely B**, **Blatter MC**, **Bonilla C**, **Bouguerleret L**, **Boutet E**, **Breuza L**, **Bridge A**, **Chan WM**, **Chavali G**, **Coudert E**, **Dimmer E**, **Estreicher A**, **Famiglietti L**, **Feuermann M**, **Gos A**, **Gruaz-Gumowski N**, **Hieta R**, **Hinz C**, **Hulo C**, **Huntley R**, **James J**, **Jungo F**, **Keller G**, **Laiho K**, **Legge D**, **Lemercier P**, **Lieberherr D**, **Magrane M**, **Martin MJ**, **Masson P**, **Mutowo-Muellenet P**, **O'Donovan C**, **Pedruzzi I**, **Pichler K**, **Poggioli D**, **Porras Millán P**, **Poux S**, **Rivoire C**, **Roechert B**, **Sawford T**, **Schneider M**, **Stutz A**, **Sundaram S**, **Tognolli M**, **Xenarios I**, **Foulgar R**, **Lomax J**, **Roncaglia P**, **Khodiyar VK**, **Lovering RC**, **Talmud PJ**, **Chibucos M**, **Giglio MG**, **Chang HY**, **Hunter S**, **McAnulla C**, **Mitchell A**, **Sangrador A**, **Stephan R**, **Harris MA**, **Oliver SG**, **Rutherford K**, **Wood V**, **Bahler J**, **Lock A**, **Kersey PJ**, **McDowall DM**, **Staines DM**, **Dwinell M**, **Shimoyama M**, **Laulederkind S**, **Hayman T**, **Wang SJ**, **Petri V**, **Lowry T**, **D'Eustachio P**, **Matthews L**, **Balakrishnan R**, **Binkley G**, **Cherry JM**, **Costanzo MC**, **Dwight SS**, **Engel SR**, **Fisk DG**, **Hitz BC**, **Hong EL**, **Karra K**, **Miyasato SR**, **Nash RS**, **Park J**, **Skrzypek MS**, **Weng S**, **Wong ED**, **Berardini TZ**, **Huala E**, **Mi H**, **Thomas PD**, **Chan J**, **Kishore R**, **Sternberg P**, **Van Auken K**, **Howe D**, **Westerfield M**. 2013. Gene Ontology annotations and resources. Nucleic Acids Research **41**:D530–5.

8. **Lewis NS**, **Anderson TK**, **Kitikoon P**, **Skepner E**, **Burke DF**, **Vincent AL**. 2014. Substitutions near the Hemagglutinin Receptor-Binding Site Determine the Antigenic Evolution of Influenza A H3N2 Viruses in U.S. Swine. J Virol **88**:4752–4763.

9. **Abente EJ**, **Santos J**, **Lewis NS**, **Gauger PC**, **Stratton J**, **Skepner E**, **Anderson TK**, **Rajão DS**, **Perez DR**, **Vincent AL**. 2016. The Molecular Determinants of Antibody Recognition and Antigenic Drift in the H3 Hemagglutinin of Swine Influenza A Virus. J Virol **90**:8266–8280.

10. **Anderson TK**, **Campbell BA**, **Nelson MI**, **Lewis NS**, **Janas-Martindale A**, **Killian ML**, **Vincent AL**. 2015. Characterization of co-circulating swine influenza A viruses in North America and the identification of a novel H1 genetic clade with antigenic significance. Virus Research **201**:24–31.

11. **Nelson MI**, **Lemey P**, **Tan Y**, **Vincent A**, **Lam TT-Y**, **Detmer S**, **Viboud C**, **Suchard MA**, **Rambaut A**, **Holmes EC**, **Gramer M**. 2011. Spatial dynamics of human-origin H1 influenza A virus in North American swine. PLoS Pathog **7**:e1002077.

12. **Nelson MI**, **Culhane MR**, **Trovão NS**, **Patnayak DP**, **Halpin RA**, **Lin X**, **Shilts MH**, **Das SR**, **Detmer SE**. 2017. The emergence and evolution of influenza A (H1α) viruses in swine in Canada and the United States. J Gen Virol **98**:2663–2675.