Title: Implementation of a PRRSV Strain Database (Renewal of NPB project #06-127)"
NPB #07-125

Investigator: Kay S. Faaberg

Institution: National Animal Disease Center, ARS, USDA

Date Submitted: January 5, 2009

Industry Summary: The PRRSV strain database (http://prrsvdb.org) has been supported by the National Pork Board for four years. The database (PRRSVdb; 9,531 sequences) is available for queries in order to obtain detailed information about PRRSV sequences (percent similarity to vaccines or other database isolates, ORF5 RFLP analysis, year and state of isolation, phylogenetic relationships, GenBank submission numbers, relevant clinical information if provided by the respective veterinary diagnostic laboratory (VDL)). The project directly addresses the NPB directive to implement a National PRRSV Sequence Database and has relevance to a number of PRRS Initiative efforts such as surveying potentially new field strains for vaccines, examining sequence conservation among various isolates for improvement of diagnostics, epitope evaluation and assessment of virus spread locally, nationally and internationally. The database is comprised presently of four veterinary diagnostic laboratories sequenced isolates and those independently deposited in GenBank. In addition, the software interface of the database has been updated to provide additional flexibility to the PRRSVdb. This flex-based web interface provides improved phylogeny viewing and user-uploaded sequence analysis, either dependent or independent of the database. This research is currently the result of several sources: the National Pork Board, the National Animal Disease Center-USDA-ARS (Kay Faaberg), the Minnesota Veterinary Diagnostic Laboratory (MVDL; James Collins and the Molecular Diagnostic Section), South Dakota Animal Disease Research & Diagnostic Laboratory (Jane Christopher-Hennings, Travis Clement), Hong Kong University (Frederick Leung, Hon Chung Chau) and the Manitoba Veterinary Services Branch (Andre Hamel) with software engineers (Trevor Wennblom, John Crow) formerly located at the Center for Biomedical Research Informatics (CBRI) of the University of Minnesota. Since the CBRI was eliminated, Trevor Wennblom moved to the Minnesota Supercomputing Institute and has worked off hours on the PRRSV Database and John Crow (Co-PI) relocated to the National Center for Genome Resources (NCGR) in Santa Fe, New Mexico. At present, the PRRSV Database has not received additional funding and thus has been forced to stop development. John Crow is currently paying monthly fees for database maintenance and online publication to Slicehost (http://www.slicehost.com), a database warehouse management firm, out of his personal funds at a rate of $38/month.

Scientific Abstract: The PRRSV database now contains 9531 unique sequences with accessions from the Minnesota Veterinary Diagnostic Laboratory, the South Dakota Animal Disease Research & Diagnostic Laboratory, Hong Kong University and the Manitoba Veterinary Services Branch as well as several hundred PRRSV sequences incorporated from GenBank. Field surveillance data has been added to new submissions in cases when the VDLs have provided it, we continue to improve the online submission tool and have boosted the query capability of the database. A stable novel web-page format was unveiled this past year, using software tools that had recently become available. Finally, the database maintenance and viewing was migrated to Slicehost, an independent database warehouse.

Introduction: This report details the improvements made during the 2007-2008 cycle to the initial NPB grant entitled "Implementation of a PRRSV Strain Database" (NPB project #04-118). The merit of the research proposed was indicated by the PRRSV research community's desire to have all PRRSV nucleotide sequences and tools to investigate PRRS disease available in one easy-access website. In the past year, we have extended the diagnostic laboratories to include not only PRRSV sequencing centers located at the Minnesota Veterinary Diagnostic Laboratory (MVDL; James Collins), South Dakota Animal Disease Research & Diagnostic Laboratory (SDADRDL; Jane Christopher-Hennings) and the Manitoba Veterinary Services Branch (Andre Hamel), but also Hong Kong University (HKU; Frederick Leung). We had hoped to add the Iowa State University Diagnostic Laboratory as well (ISUVDL; Kyoung-Jun Yoon), but after repeated attempts by the PI and the Database Software Manager to contact the Yoon laboratory, we still have not received any sequences from ISUVDL. In addition, we successfully transferred the PRRSV database (PRRSVdb) to a new flex-based interface that shows the MySQL Database on a more user-friendly web page. As a result, the appearance of the database is entirely new, but the supportive data is maintained in the same way behind the interface. Lastly, to provide for continued access to the PRRSVdb after the University of Minnesota declined to provide continued support of the database research, the database maintenance and viewing was migrated to Slicehost, an independent database warehouse.

Objectives:

Objective 1. Incorporation of PRRSV sequences generated by the South Dakota Animal Disease Research & Diagnostic Laboratory (SDADRDL), the Iowa State University Veterinary Diagnostic Laboratory (ISUVDL) and Hong Kong University (HKU).

Objective 2. Completion of upgrades to the relational database and website.

Materials and Methods:
The software tools used to run the database include The Basic Local Alignment Search Tool for database specific use (BLAST; http://www.ncbi.nlm.nih.gov/BLAST/) ClustalW (http://www.ebi.ac.uk/clustalw/), Phylip (http://evolution.gs.washington.edu/phylip.html). Protein Family Alignment Annotation Tool (PFAAT; http://pfaat.sourceforge.net/). Native software developed by CBRI includes new software code to result in less errors and better software problem identification, ORF detection software and RFLP analysis. The web framework is Ruby on Rails[TM], but now has been integrated with Flex, an additional framework that helps you build dynamic, interactive rich internet applications. The PRRSV database itself is still based on MySQL software. Note that the rich blend of software from several domains are all open source and were blended to work together by expert software developers, Trevor Wennblom and John Crow.

Results:

Objective 1. Incorporation of PRRSV sequences generated by the South Dakota Animal Disease Research & Diagnostic Laboratory (SDADRDL), the Iowa State University Veterinary Diagnostic Laboratory (ISUVDL) and Hong Kong University (HKU).

The PRRSV database now consists of 9,531 unique nucleotide sequences (>2% nucleotide difference), which have also been submitted to GenBank. Duplicate sequences were noted and electronically "filed" with the isolation date and place. In the que remaining to be added are 560 sequences from the MVDL (Molecular Diagnostic Laboratory), 806 sequences from SDADRDL (Travis Clement) and 131 sequences from HKU (Hon Chung Chau). Imported GenBank sequences total 433. ISUVDL (Kyoung-Jin Yoon) did not submit any sequences, for unknown reasons.

Objective 2. Completion of upgrades to the relational database and website.

    a.   *Flex-based Interface.* Early in 2008, the University of Minnesota merged the Center for Biological Research Informatics (where the database was housed) with another academic unit, the Minnesota Supercomputing Institute (MSI), The MSI refused to support the PRRSVdb, causing concern about the future of the PRRSV Database. In order to provide a protected environment, Co-PI John Crow and Trevor Wennblom moved the database to an inexpensive database warehouse named Slicehost, which also provides online PRRSVdb viewing (www.slicehost.com). Trevor Wennblom released the new interface for working with PRRSV sequence data at prrsvdb.org (below).

By clicking on the text Begin Exploring PRRSVdb in the upper right, the database exploration tablet is shown (below).
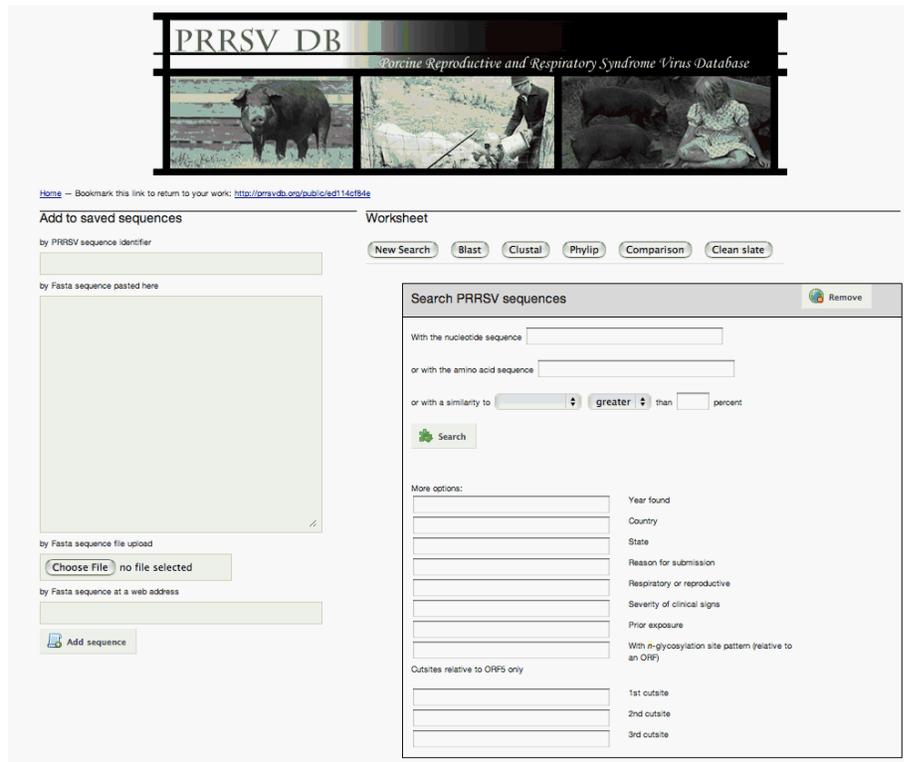
This tablet has one side (left) for loading the sequences you want to work with, and a worksheet side (right) that allows researchers and veterinarians to perform a variety of tasks. On the left side, a user can place a known PRRSV identifier, or they can upload their own sequences in a variety of ways. On the right side, a "Search PRRSV sequences" page is already available, but new blank pages can be brought up by clicking on the New Search button just above this window. As displayed, a user can search by several criteria, one at a time or all at once. If a user wants to assess an input sequence with the database, they can "Blast". After they have assembled the sequences of interest, an alignment can be derived by clicking on "Clustal", which leads to the alignment editors PFAAT and Jalview with their related phylogenetic dendogram (tree) drawing software. An alternative rudimentary tree drawing program is accomplished by the "Phylip" button. A "Percent Identity" grid can also be calculated for all input sequences. Finally, the "Clean slate" option allows users to have a fresh tablet to begin again.

Thus, this format allows for additional regions of the genome to be analyzed, and is extremely flexible about the types of input sequences to query (database and/or user defined).

b.   *Allow veterinarians/producers/researchers to temporarily upload their personal sequences.* – Completed, see above description in (a.).

c.   *Modify the database to account for all sequences identified and list when and where the sequence has been detected.* Trevor Wennblom redesigned the database output so that a single particular sequence detected several times in several places will be displayed only once but with multiple dates and multiple locations.

Additional progress:

a.   An online submission tool, limited to location-specific IP domains after approval of PI Faaberg, is now available for direct uploading of nucleotide sequences from the various laboratories (see Figure at right). Alternatively, in the interest of diagnostic laboratory confidentiality issues, a laboratory can request a bank of unique PRRSV ID numbers. In that case, personnel assigned from the laboratory produces an Excel document (spreadsheet) of the PRRSV unique ID numbers and companion nucleotide sequences, year and date of isolation. Going forward, they may elect to submit new sequences using the web submission form and the PRRSV ID numbers given, taking care to inform personnel of specific laboratory confidentiality mandates. After submission, the database manager will check the sequences for potential errors and uniqueness. If the nucleotide sequence is new, the date and location are recorded and then added to the PRRSVdb. If the sequence is not novel, a record of where and when the sequence was detected is added to the original PRRSVdb sequence. After this, the sequence is automatically analyzed by the already developed ORF5 restriction fragment length polymorphism (RFLP) analysis software, N-glycosylation patterns are determined, and novel sequences are uploaded to GenBank with appropriate recognition given to the individual laboratory.

b.   Trevor Wenblom developed a new web/SPARQL/RDF interface (http://prrsvdb.org/db/sparql) for data mining by expert programmers and for development of new databases. The deposited sequence pages now have RDFa (Resource Description Framework Alternative) data embedded in them for third party/computer program recognition. Trevor also worked with personnel at the National Center for Biotechnology Information to correct a problem in their "tbl2asn" program that generates .sqn files for submission to GenBank.

Discussion:

The PRRSV Database has been in existence for four years. In that time, we have expanded the submitted sequences from 600 to over 9500, primarily from the MVDL and SDADRDL. We have also combined, refined and added several unique tools for producers, veterinarians and researchers to mine information obtained from analysis of these sequences. The success of this venture is appreciated by web-viewing statistics. Starting with April of 2008, there were 1,474 unique visitors (different internet addresses), with 88,743 visits, which indicate an average of 60 page views per unique visitor. These numbers verify that the PRRSVdb is a well-used resource.

However, the termination of software support for the PRRSVdb by the University of Minnesota has caused considerable challenges for not only the PI, but also for efforts put forth by many individuals and organizations. First, in order to provide continued database maintenance and online access, the database was moved offsite to Slicehost, a secure database storage facility. In the following months,

we recognized the benefits of this move, principally in the impressive security of Slicehost and yet the ease of which the Database Manager can obtain secure access to this privileged site (the only one with such access) for database upgrades, for a very low cost ($38/month). This monthly fee is being paid by Co-PI John Crow in the interim out of his personal funds. Second, the 2008 NPB proposal was difficult to assemble, yet we believe that the decisions made at the time of submission were indeed correct. We are still awaiting the decision from the NPB's Swine Health Committee on funding of our renewal.

Unless we receive continued support, not only will the PRRSVdb be discontinued but we will lose the technical skill and guidance by the software creator (Trevor Wennblom). John Crow, the Co-PI who has left the University of Minnesota, fortunately wishes to remain associated with our proposed expansion of the PRRSVdb. Although he has moved to the National Center for Genome Resources, his new location provides avenues to explore for additional PRRSVdb resources in the coming years.