IRD US swine H1 hemagglutinin clade classification tool
June 24, 2015

This document describes the IRD's US swine HA(H1) clade classification tool, which was developed in conjunction with Tavis Anderson and other swine influenza experts at the USDA. The tool is an adaptation of the tool used for HA(H5) classification. See the original SOP for H5 classification for more details. Note that this tool classifies HA(H1) sequences regardless of host species and regardless of NA subtype.

Goal

The goal of the swine HA(H1) clade classification tool is to assign a standard USDA clade to HA(H1) sequences whose HA belongs to one of the swine H1 clades recognized in the US, by finding that clade in [1,2] to which the query sequence is most closely related. When the query sequence is from a virus outside of the recognized US swine H1 clades, the classification tool will indicate this fact.

Algorithm

We first generated a phylogenetic tree with bootstrap support from all HA(H1) sequences regardless of host and NA subtype. From this tree, we carefully selected representatives of each clade, with particular emphasis on finding the oldest and youngest viruses in the clade. The final selection defines a reference tree of ~100 H1 sequences representing all of the currently defined US swine H1 clades, together with representatives of other H1 clades. An aligned HA(H5) sequence from A/chicken/Scotland/1959 was included as the outgroup for this phylogenetic tree.

The details of the algorithm for assigning a clade are the same as the details for assignment of HA(H5) clade. Briefly, the query sequence is aligned against the alignment of the sequences from the reference tree, then attached to the phylogeny of these sequences using pplacer [3]. Key to our procedure for clade assignment is that the tree of representative sequences does not change, and therefore acts as a "scaffold" upon which the query sequence is hung.

Output

When a query sequence is placed unequivocally within the bounds of a single clade, the classification assigned is that of the clade. When classification of a query sequence is less certain, a designation that includes the suffix "-like" is used to indicate its nearest relation.

The accuracy of clade assignments made by the IRD clade annotation tool has been evaluated by running it against the USDA classifications for all sequences not in the IRD reference set. The tool has been found to be highly accurate: the IRD tool predicts the correct clade more than 99% of the time.

Limitations

When applied to non-HA sequences, the IRD swine H1 Clade Classification Tool returns unpredictable and likely erroneous results.

Inaccuracies can occur when sequences are faulty, or are very short (< 300 nucleotides).

Authorship

The updated and original versions of the tool were designed by Dr. Catherine Macken, a member of the Influenza Research Database (IRD) Team, currently ar the University of Auckland, New Zealand.

References

[1] Anderson, TK, Nelson, MI, Kitikoon, P, Swenson, SL, Korslund, JA, and Vincent, AL. (2013). Population dynamics of co-circulating swine influenza A viruses in the United States from 2009 to 2012. *Influenza and Other Respiratory Viruses* **7(S4)**: 42–51.

[2] Anderson, TK, Campbell, BA, Nelson, MI, Lewis, NS, Janas-Martindale, A, Killian, ML, and 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.

[3] Matsen, FA, Kodner, RB and Armbrust, EV (2010). pplacer: linear time maximum-likelihood and Bayesian phylogenetic placement of sequences onto a fixed reference tree. *BMC Biomathematics* **11**:538