IRD Highly Pathogenic H5 Clade Classification Tool
January 14, 2015

This document describes the IRD's latest version of its H5 clade classification tool; it is an updated and abbreviated version of the SOP from the original November 2012 release this tool. Note that this version of the tool classifies H5 viruses regardless of NA subtype; it also classifies viruses in the novel clade provisionally labeled 2.3.4.6.

Goal

The goal of the updated clade classification tool is unchanged from that of the original release in November 2012. Briefly, the tool assigns an H5 clade annotation to an HA/H5 sequence (the "query" sequence) by finding that clade in the updated WHO nomenclature to which the query sequence is most closely related. When the query sequence is from a virus outside of the A/goose/Guangdong/1/96 (GsGua)-lineage, the classification tool will return a result that indicates this fact.

Algorithm

We first generated a phylogenetic tree with bootstrap support from all HA/H5 sequences for which a clade is given in the updated WHO nomenclature. 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 ~300 H5 sequences representing all of the currently defined clades of the A/goose/Guangdong/1/96 (GsGua)-lineage, together with representatives of Eurasian H5 sequences from outside the GsGua lineage, and American low-pathogenic H5 viruses. The early high-pathogenic virus A/chicken/Scotland/1959 was included as the outgroup for a phylogenetic tree of these sequences.

The details of the algorithm for assigning a clade have not changed from the first version of this tool. 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 [2]. 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. If a query sequence is attached to a branch that connects two clades that differ in granularity, such as the branch connecting clade 2.3.4 sequences and clade 2.3.4.3 sequences, the query is associated with the less-detailed clade, here "2.3.4-like." If a query sequence lies on a branch that separates two clades of the same granularity, such as 2.3.1 and 2.3.2, then the query is assigned the classification held in common to the two clades, here "2.3-like." The terminology "like" indicates uncertainty in classification to a higher level of granularity. Sequences that are a sister to A/chicken/Scotland/1959 are classified as "Outgroup."

The accuracy of clade assignments made by the IRD clade annotation tool has been evaluated by running it against the updated WHO 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 about 98% of the time.

Limitations

When applied to non-HA sequences, the IRD H5N1 Clade Classification Tool returns unpredictable and likely erroneous results.

Inaccuracies can occur when viruses belong to old clades, whose definition suffer from paucity of data, leading to phylogenetic uncertainty, or have bad sequence information, 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 from the University of Auckland, New Zealand.

References

[1] World Health Organization/World Organization for Animal Health/Food and Agriculture Organization (WHO/OIE/FAO) H5N1 Evolution Working Group (2014). Revised and updated nomenclature for highly pathogenic avian influenza A (H5N1) viruses. Infl. And Other Resp. Viruses. Doi:10.1111/irv.12230

[2] 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