A Phylogeny-Based Global Nomenclature System and Automated Annotation Tool for H1 Hemagglutinin Genes from Swine Influenza A Viruses


Virus and Prion Research Unit, National Animal Disease Center, USDA-ARS, Ames, Iowa, USA; Bioinformatics Institute, University of Auckland, Auckland, New Zealand; Department of Zoology, University of Cambridge, Cambridge, United Kingdom; J. Craig Venter Institute, La Jolla, California, USA; Department of Pathology, University of California, San Diego, California, USA; Laboratory of Virology, Faculty of Veterinary Medicine, Ghent University, Merelbeke, Belgium; Animal and Plant Health Agency, Weybridge, United Kingdom; National Veterinary Services Laboratories, USDA-APHIS, Ames, Iowa, USA; ANSES, Ploufragan-Plouzané Laboratory, Swine Virology Immunology Unit, Ploufragan, France; Division of Transboundary Animal Disease, National Institute of Animal Health, National Agriculture and Food Research Organization, Ibaraki, Japan; Canadian Food Inspection Agency, National Centre for Foreign Animal Disease, Winnipeg, Manitoba, Canada; Embrapa Swine and Poultry, Animal Health and Genetic Laboratory, Concórdia, SC, Brazil; Instituto de Patobiología, CICVyA INTA, Hurlingham, Buenos Aires, Argentina; Influenza Division, National Center for Immunization and Respiratory Diseases, Centers for Disease Control and Prevention, Atlanta, Georgia, USA; Department of Infectious Diseases, St. Jude Children’s Research Hospital, Memphis, Tennessee, USA.

ABSTRACT The H1 subtype of influenza A viruses (IAVs) has been circulating in swine since the 1918 human influenza pandemic. Over time, and aided by further introductions from nonswine hosts, swine H1 viruses have diversified into three genetic lineages. Due to limited global data, these H1 lineages were named based on colloquial context, leading to a proliferation of inconsistent regional naming conventions. In this study, we propose rigorous phylogenetic criteria to establish a globally consistent nomenclature of swine H1 virus hemagglutinin (HA) evolution. These criteria applied to a data set of 7,070 H1 HA sequences led to 28 distinct clades as the basis for the nomenclature. We developed and implemented a web-accessible annotation tool that can assign these biologically informative categories to new sequence data. The annotation tool assigned the combined data set of 7,070 H1 sequences to the correct clade more than 99% of the time. Our analyses indicated that 87% of the swine H1 viruses from 2010 to the present had HAs that belonged to 7 contemporaneous cocirculating clades. Our nomenclature and web-accessible classification tool provide an accurate method for researchers, diagnosticians, and health officials to assign clade designations to HA sequences. We developed and implemented a web-accessible annotation tool that can assign these biologically informative categories to new sequence data. The annotation tool assigned the combined data set of 7,070 H1 sequences to the correct clade more than 99% of the time. Our analyses indicated that 87% of the swine H1 viruses from 2010 to the present had HAs that belonged to 7 contemporaneous cocirculating clades. Our nomenclature and web-accessible classification tool provide an accurate method for researchers, diagnosticians, and health officials to assign clade designations to HA sequences. The tool can be updated readily to track evolving nomenclature as new clades emerge, ensuring continued relevance. A common global nomenclature facilitates comparisons of IAVs infecting humans and pigs, within and between regions, and can provide insight into the diversity of swine H1 influenza virus and its impact on vaccine strain selection, diagnostic reagents, and test performance, thereby simplifying communication of such data.

IMPORTANCE A fundamental goal in the biological sciences is the definition of groups of organisms based on evolutionary history and the naming of those groups. For influenza A viruses (IAVs) in swine, understanding the hemagglutinin (HA) genetic lineage of a circulating strain aids in vaccine antigen selection and allows for inferences about vaccine efficacy. Previous reporting of H1 virus HA in swine relied...
on colloquial names, frequently with incriminating and stigmatizing geographic toponyms, making comparisons between studies challenging. To overcome this, we developed an adaptable nomenclature using measurable criteria for historical and contemporary evolutionary patterns of H1 global swine IAVs. We also developed a web-accessible tool that classifies viruses according to this nomenclature. This classification system will aid agricultural production and pandemic preparedness through the identification of important changes in swine IAVs and provides terminology enabling discussion of swine IAVs in a common context among animal and human health initiatives.

**KEYWORDS:** H1N1, H1N2, influenza A virus, molecular epidemiology, nomenclature, swine, virus evolution

Influenza A virus (IAV) is one of the most important respiratory pathogens of swine. Infection causes significant financial losses through decreased production, increased vaccination and treatment cost, and increased mortality through interactions with bacterial and other viral infections (1–3). Additionally, swine IAV is a significant zoonotic pathogen with public health relevance; due to the susceptibility of swine to transient infection with IAVs from different species, novel reassorted and potentially pandemic viruses might emerge in swine and spill over to humans (4). Thus, insights into patterns of swine IAV genetic diversity allow identification of novel viral lineages, provide criteria for rational intervention in swine agriculture, and facilitate public health pandemic preparedness.

The global genetic diversity of swine IAV H1 during the last century is a result of the establishment of IAVs from other species in swine populations and subsequent evolution via antigenic shift and drift (5–8). Broadly, there is continual cocirculation of two dominant H1 subtypes (H1N1 and H1N2), within which there are three major lineages resulting from the separate introductions of genetically and antigenically distinct viruses (9, 10). The first endemic swine IAV lineage originated from the 1918 Spanish flu pandemic, leading to the viruses currently classified as “classical-swine” H1N1 (11). In the late 1990s, the classical-swine viruses reassorted their internal genes with those of a lineage of triple-reassortant H3N2 lineage viruses, leading to a spurt of diversification of the hemagglutinin (HA) genes and new genetic H1 clades within the classical lineage (12–15), including the H1N1 pandemic 2009 viruses (H1N1pdm09) (7, 16). The second endemic swine IAV lineage resulted from the spillover of H1 viruses from wild birds in Europe with subsequent export to Asia. Viruses from this lineage are referred to as Eurasian avian-like (10, 17–19). The third endemic swine IAV lineage resulted from repeated human seasonal IAVs spilling into swine herds and subsequent evolution in pigs. These viruses were first recognized in Europe in the 1990s (20), with independent introductions occurring in North American (21, 22) and South American (23) swine herds.

Within these three major lineages, numerous genetic clades of HA have evolved within specific geographical regions, and naming of these clades has been according to regional systems (Table 1). For example, in the United States, a nomenclature system that grouped viruses into one of seven HA H1 clades using Greek letters was adopted (22, 24, 25). In Europe, the European Surveillance Network for Influenza in Pigs (ESNIP) defined four major HA H1 clades, based on host and/or regional introduction history (26). Contemporary HA H1 genes in Europe have been classified as avian-like swine H1avN1 lineage, human-like reassortant swine H1huN2 lineage, or H1N1pdm09 lineage; additionally, classical-swine H1N1 viruses were transiently identified in the 1970s and 1980s. Similarly, IAV in Asia reflects the regional introduction and subsequent evolution and cocirculation of multiple genetic clades of classical-swine H1N1, avian-like H1N1, and human seasonal-like H1N1 and H1N2 viruses (6, 27, 28). However, swine move frequently within and sporadically between countries, and clades of originally geographically restricted viruses can be dispersed globally, rendering geographical and regional clade names uninformative. Importantly, current clade descriptors are di-
Nomenclature of Swine Influenza H1 Viruses

TABLE 1 List of colloquial names (if available) for swine influenza A virus H1 hemagglutinin clades and associated laboratory studies

<table>
<thead>
<tr>
<th>Clade</th>
<th>Colloquial name</th>
<th>Distribution</th>
<th>Reference(s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Classical swine lineage</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1A.1</td>
<td>α-H1</td>
<td>Canada, China, Hong Kong, Italy, Japan, Mexico, Thailand, United Kingdom, USA</td>
<td>22</td>
</tr>
<tr>
<td>1A.1.1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1A.1.2</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1A.1.3</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1A.2</td>
<td>β-H1</td>
<td>Mexico, South Korea, USA</td>
<td>22</td>
</tr>
<tr>
<td>1A.3</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1A.3.1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1A.3.2</td>
<td>γ-2-H1</td>
<td>Mexico, USA</td>
<td>24</td>
</tr>
<tr>
<td>1A.3.3</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1A.3.3.1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1A.3.3.2</td>
<td>H1N1pdm09</td>
<td>37 countries</td>
<td>4, 16</td>
</tr>
<tr>
<td>1A.3.3.3</td>
<td>γ-H1</td>
<td>South Korea, USA</td>
<td>22</td>
</tr>
<tr>
<td>Human seasonal lineage</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1B.1</td>
<td>European human-like reassortant H1 huN2 (derived from A/swine/Scotland/410440/94)</td>
<td>Ireland, United Kingdom</td>
<td>26</td>
</tr>
<tr>
<td>1B.1.1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1B.1.2</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1B.1.2.1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1B.1.2.2</td>
<td>A/swine/Italy/4675/2003</td>
<td>Italy</td>
<td>26</td>
</tr>
<tr>
<td>1B.1.2.3</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1B.2</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1B.2.1</td>
<td>δ-2</td>
<td>USA</td>
<td>21</td>
</tr>
<tr>
<td>1B.2.2</td>
<td>δ-1</td>
<td>Argentina, Brazil, Canada, United Kingdom, USA</td>
<td>21</td>
</tr>
<tr>
<td>1B.2.2.1</td>
<td>δ-1a</td>
<td>USA</td>
<td></td>
</tr>
<tr>
<td>1B.2.2.2</td>
<td>δ-1b</td>
<td>USA</td>
<td></td>
</tr>
<tr>
<td>Eurasian avian lineage</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1C.1</td>
<td>Avian-like swine H1 avN1 (derived from A/swine/Arnsberg/6554/1979 and A/swine/Belgium/WVL1/1979)</td>
<td>Belgium, Canada, France, Germany, Hong Kong, Ireland, Italy, Netherlands, Spain, United Kingdom</td>
<td>15, 61</td>
</tr>
<tr>
<td>1C.2</td>
<td>Avian-like swine H1 avN1 (derived from A/swine/Ille et Vilaine/1455/1999)</td>
<td>Belgium, Denmark, Finland, Germany, Italy, Mexico, Netherlands, Poland, Sweden</td>
<td>15, 61</td>
</tr>
<tr>
<td>1C.2.1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1C.2.2</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1C.2.3</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
tool matched expertly curated, manual classification of the sequences >99% of the time. This tool will be released on the Influenza Research Database (IRD) at http://www.fludb.org (30, 31) to facilitate the adoption of the unified nomenclature.

RESULTS

Global genetic diversity and swine H1 clade designations. Substantial genetic diversity was demonstrated in H1 viruses circulating in swine over the past 5 years (2010 to present) and among geographic regions (Fig. 1 and 2). Three major first-order H1 lineages continued to circulate in pigs (Fig. 1; also see Fig. S1 in the supplemental material): the 1A classical lineage, viruses related to the 1918 human influenza pan-
demic; the 1B human seasonal lineage, the result of multiple human-to-swine transmission episodes of human seasonal H1 strains over decades; and the 1C Eurasian avian lineage, arising from an introduction from wild birds into pigs in the 1970s. The majority (~87%) of the viruses from 2010 to the present were placed into seven clades. The numerically dominant clades reflected intensive surveillance in the United States (24, 25), investigator sequencing efforts in Canada (e.g., references 32 and 33), and the rapid dissemination of the 2009 H1N1 pandemic virus (H1N1pdm09) across global swine populations (7, 16). Similarly, coordinated surveillance in Europe (26, 34) and Asia (6) captured two primary clades of 1C Eurasian avian lineage currently circulating in the two continents.

**Clade designations for 1A (classical) swine lineage.** The 1A (classical) lineage contained 1,889 viruses from 34 countries collected from 2010 to the present (Fig. 1 and 2). According to our nomenclature rules, we refined the classification of 1A viruses into three second-order divisions, each of which corresponds to earlier, regional classifica-
tions (25): 1A.1 viruses, similar to classical "α-H1" viruses (n = 120, 7 countries); 1A.2 viruses, similar to "β-H1" viruses (n = 77, from United States, South Korea, and Mexico); and 1A.3 viruses, similar to "γ-H1" viruses (n = 1,692, 34 countries). Applying our nomenclature rules again leads to further subdivision of these second-order divisions into third- and fourth-order divisions. The only clade within the 1A lineage that had wide global distribution was that of the 2009 H1N1 pandemic clade (1A.3.3.2) with 541 viruses in 34 countries. Some clades within the 1A lineage were geographically constrained: 1A.1.2 viruses were detected only in Thailand (n = 13); 1A.1.3, 1A.3.3, and 1A.3.3.1 were restricted to China and Hong Kong (n = 28); and 1A.3.1 was restricted to Mexico (n = 14). The remaining clades were more diffuse, potentially reflecting the dissemination of viruses with agricultural trade (see reference 35). For example, the 1A.1.1 clade was detected in Canada, the United States, and Taiwan (n = 82); the 1A.3.2 clade was detected in Mexico and the United States (analogous to "γ-2 H1" viruses [24]); and the 1A.3.3.3 viruses were isolated in South Korea and the United States (analogous to "γ-H1" viruses).

The 1A classical lineage within- and between-clade average pairwise distances (APDs) are presented in Table 2. Each clade had an APD of >7% from other clades and an APD of <7% within the clade, although some minor exceptions were made when all other clade-defining criteria were met and mitigating circumstances supported the exception. Within-clade exceptions were made for the first-order 1A.1 (APD, 7.8%) and the extensive 1A.1.1 second-order clade (APD, 9.5%) that represented multiple monophyletic clades of viruses that individually did not meet our criteria for further division based on the number of recent sequences. The exception to the >7% distance between-clade threshold was associated with clades nested within 1A.3.3 (1A.3.3.1, 1A.3.3.2, and 1A.3.3.3); however, the 1A.3.3 clade required these additional third-order divisions to differentiate the H1N1 pandemic viruses (1A.3.3.2) from North American "γ-H1" viruses (1A.3.3.3) and a geographically isolated clade of viruses from China (1A.3.3.1).

Clade designations for 1B (human seasonal) swine lineage. The 1B (human seasonal) lineage contained 1,447 viruses from 13 countries collected from 2010 to the present (Fig. 1 and 2). Applying our nomenclature rules led to two second-order divisions corresponding to established clades: the 1B.1 viruses, related to a reassortant H1N2 virus that emerged in Great Britain in 1994 (n = 132, 7 European countries) (20), and 1B.2 viruses, related to the "δ-1 H1" and "δ-2 H1" viruses (n = 1,315, 6 countries) (22). We defined two third-order 1B.1 clades: the 1B.1.1 (n = 24) viruses, circulating predominantly in the United Kingdom, with one virus collected in France, and 1B.1.2 (n = 108) viruses circulating in continental Europe. The fourth-order divisions of 1B.1.2 reflect geographic boundaries: 1B.1.2.1 (n = 24) from Belgium, Germany, Italy, and Netherlands; 1B.1.2.2 (n = 21) from Italy; and 1B.1.2.3 (n = 54) from France.

**TABLE 2** Average percent pairwise nucleotide distances within and between H1 1A (classical lineage) phylogenetic clades

| Clade | APD (%) | Within clade | 1A.1 | 1A.1.1 | 1A.1.2 | 1A.1.3 | 1A.2 | 1A.3 | 1A.3.1 | 1A.3.2 | 1A.3.3 | 1A.3.3.1 | 1A.3.3.2 | 1A.3.3.3 | 1A.3.3.3.1 | 1A.3.3.2 | 1A.3.3.3.3 | 1A.3.3.3.2 |
|-------|---------|--------------|------|--------|--------|--------|-----|------|--------|--------|-------|----------|----------|----------|-----------|----------|----------|
| 1A.1  | 7.8     |              |      |        |        |        |     |      |        |        |       |          |          |          |            |          |          |
| 1A.1.1| 9.5     | 11.5         |      |        |        |        |     |      |        |        |       |          |          |          |            |          |          |
| 1A.1.2| 5.4     | 11.4         | 13.8 |        |        |        |     |      |        |        |       |          |          |          |            |          |          |
| 1A.1.3| 3.0     | 8.3          | 11.8 | 10.1   |        |        |     |      |        |        |       |          |          |          |            |          |          |
| 1A.2  | 3.8     | 10.8         | 14.0 | 12.5   | 8.7    |        |     |      |        |        |       |          |          |          |            |          |          |
| 1A.3  | 4.2     | 10.8         | 13.7 | 12.2   | 8.7    | 10.6   |     |      |        |        |       |          |          |          |            |          |          |
| 1A.3.1| 1.4     | 11.2         | 13.7 | 12.7   | 9.1    | 12.1   | 11.4 |      |        |        |       |          |          |          |            |          |          |
| 1A.3.2| 5.3     | 11.1         | 14.1 | 12.7   | 9.0    | 11.3   | 10.0 | 11.2 |        |        |       |          |          |          |            |          |          |
| 1A.3.3| 5.1     | 10.1         | 13.4 | 11.6   | 7.6    | 10.2   | 9.7  | 10.2 | 8.1    |        |       |          |          |          |            |          |          |
| 1A.3.3.1| 2.7 | 10.5          | 13.5 | 12.1   | 7.9    | 10.1   | 10.0 | 10.3 | 8.3    | 5.0    |       |          |          |          |            |          |          |
| 1A.3.3.2| 1.9 | 12.0          | 14.4 | 13.5   | 9.7    | 11.1   | 11.3 | 12.0 | 9.8    | 7.5    | 6.5    |          |          |          |            |          |          |
| 1A.3.3.3| 3.2 | 11.1          | 14.1 | 12.5   | 9.0    | 10.7   | 10.9 | 10.9 | 9.0    | 6.8    | 6.5    | 8.3    |          |          |            |          |          |
The 1B.2 clade contained two third-order clades that corresponded to previously described “/H9254-2 H1” (1B.2.1) and “/H9254-1 H1” (1B.2.2) clades. Based on average pairwise distances, and a large number of viruses, the third-order 1B.2.2 clade met the criteria for further subdivision into 1B.2.2.1 (n/H11005360) and 1B.2.2.2 (n/H11005636). In addition to these named subdivisions, the 1B.2 clade from 2010 to the present contained sporadic human-to-swine transmission episodes (n/H110057) in Argentina, Chile (36), China, Mexico, and Vietnam; these spillovers did not warrant the designation of a clade either due to failure to establish in swine populations or due to insufficient numbers to meet our criteria. Similarly, 1B.2.2 (22) included viruses collected from spatially isolated swine populations in Argentina and Brazil (23) and in Mexico that represent human-to-swine transmission episodes, but the number of viruses is too low to be able to confidently infer a separate clade. To link these viruses to their source population and maintain flexibility should additional surveillance detect more samples, we classified these viruses as “Other-Human.”

The 1B.2 clade contained two third-order clades that corresponded to previously described “6-2 H1” (1B.2.1) and “6-1 H1” (1B.2.2) clades. Based on average pairwise distances, and a large number of viruses, the third-order 1B.2.2 clade met the criteria for further subdivision into 1B.2.2.1 (n = 360) and 1B.2.2.2 (n = 636). In addition to these named subdivisions, the 1B.2 clade from 2010 to the present contained sporadic human-to-swine transmission episodes (n = 7) in Argentina, Chile (36), China, Mexico, and Vietnam; these spillovers did not warrant the designation of a clade either due to failure to establish in swine populations or due to insufficient numbers to meet our criteria. Similarly, 1B.2.2 (22) included viruses collected from spatially isolated swine populations in Argentina and Brazil (23) and in Mexico that represent human-to-swine transmission episodes, but the number of viruses is too low to be able to confidently infer a separate clade. To link these viruses to their source population and maintain flexibility should additional surveillance detect more samples, we classified these viruses as “Other-Human.”

The 1B human seasonal lineage within- and between-clade APDs are presented in Table 3. For the most part, each clade had an APD of >7% from other clades and almost all had an APD of <7% within the clade. The within-clade exceptions were the 1B.1 and 1B.2 clades (APD, 9.9% and 7.5%, respectively). The 1B.1 second-order clade (n = 5) had too few representative sequences to calculate genetic distance, and 1B.2 represented multiple monophyletic clades that individually did not meet our criteria for further division. Similarly, the extensive 1B.1.1 clade (APD, 7.8%) did not meet criteria for further splitting. The exception to the between-clade threshold was associated with clades nested within 1B.2.2 (1B.2.2.1 and 1B.2.2.2). These third-order clade designations were made because of the considerable number of viruses in 1B.2.2 (n = 1,016 from 2010 to present), strong bootstrap support (100%), and moderate between-clade support (APDs of 6.4% and 5.8%, respectively).

**Clade designations for 1C (Eurasian avian) swine lineage.** The 1C (Eurasian avian) lineage consisted of 315 viruses from 14 countries collected from 2010 to the present (Fig. 1 and 2). During this time period, we identified two second-order divisions of geographically isolated virus clades: the 1C.1 viruses in the United Kingdom and the 1C.2 viruses in continental Europe and Asia. Within the 1C.2 clade, three third-order divisions emerged: 1C.2.1 (n = 118) in Belgium, Denmark, France, Germany, Italy, Netherlands, Poland, and Spain; 1C.2.2 (n = 25) in France, Germany, Netherlands, Poland, and Spain; and 1C.2.3 (n = 127) in China and South Korea. Avian H1 HA sequences were generally restricted to two monophyletic clades distinct from, but sister to, the 1C swine viruses: these HA sequences were defined as “Other-Avian.” The within- and between-clade APDs are presented in Table 4. For the most part, each clade had an APD of >7% from other clades and an APD of <7% within the clade. The one within-clade exception in this lineage was 1C.2 (APD, 7.9%), which had multiple monophyletic subclades without adequate statistical support to further divide the data.

### Table 3: Average percent pairwise nucleotide distances within and between H1 1B (human seasonal lineage) phylogenetic clades

<table>
<thead>
<tr>
<th>Clade</th>
<th>APD (%)</th>
<th>Within clade</th>
<th>Between clade:</th>
<th>1B.1</th>
<th>1B.1.1</th>
<th>1B.1.2</th>
<th>1B.1.2.1</th>
<th>1B.1.2.2</th>
<th>1B.1.2.3</th>
<th>1B.2.1</th>
<th>1B.2.2</th>
<th>1B.2.2.1</th>
<th>1B.2.2.2</th>
<th>1B.2.2.2.1</th>
</tr>
</thead>
<tbody>
<tr>
<td>1B.1</td>
<td>9.9</td>
<td></td>
<td>1B.1</td>
<td>12.3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1B.1.1</td>
<td>7.8</td>
<td></td>
<td>1B.1.1</td>
<td>12.2</td>
<td>12.3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1B.1.2</td>
<td>5.3</td>
<td></td>
<td>1B.1.2</td>
<td>12.2</td>
<td>12.1</td>
<td>9.8</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1B.1.2.1</td>
<td>5.3</td>
<td></td>
<td>1B.1.2.1</td>
<td>13.1</td>
<td>12.6</td>
<td>10.1</td>
<td>9.9</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1B.1.2.2</td>
<td>4.4</td>
<td></td>
<td>1B.1.2.2</td>
<td>12.6</td>
<td>13.5</td>
<td>13.7</td>
<td>13.8</td>
<td>14.7</td>
<td>13.4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1B.2</td>
<td>7.5</td>
<td></td>
<td>1B.2</td>
<td>14.5</td>
<td>15.1</td>
<td>15.4</td>
<td>15.7</td>
<td>16.6</td>
<td>15.2</td>
<td>8.8</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1B.2.1</td>
<td>4.2</td>
<td></td>
<td>1B.2.1</td>
<td>14.6</td>
<td>14.2</td>
<td>14.6</td>
<td>14.6</td>
<td>15.4</td>
<td>14.0</td>
<td>8.1</td>
<td>8.7</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1B.2.2</td>
<td>3.5</td>
<td></td>
<td>1B.2.2</td>
<td>13.9</td>
<td>14.2</td>
<td>14.6</td>
<td>14.6</td>
<td>15.4</td>
<td>14.0</td>
<td>8.1</td>
<td>8.7</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1B.2.2.1</td>
<td>2.3</td>
<td></td>
<td>1B.2.2.1</td>
<td>14.6</td>
<td>15.1</td>
<td>15.6</td>
<td>15.8</td>
<td>16.1</td>
<td>15.2</td>
<td>9.7</td>
<td>10.5</td>
<td>6.4</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1B.2.2.2</td>
<td>2.8</td>
<td></td>
<td>1B.2.2.2</td>
<td>14.1</td>
<td>14.5</td>
<td>14.8</td>
<td>15.0</td>
<td>15.3</td>
<td>14.3</td>
<td>9.1</td>
<td>9.9</td>
<td>5.8</td>
<td>5.3</td>
<td></td>
</tr>
</tbody>
</table>
Consistency of proposed classifications. The clades identified by these global phylogenetic analyses and pairwise-distance criteria were consistently segregated by different phylogenetic approaches and with randomly subsampled data sets. While tree topology varied slightly between Bayesian and maximum likelihood methods, the monophyletic grouping and bootstrap support (or posterior probability) were consistent. There were a number of minor discrepancies in our classification (n = 7 or 0.28% of the randomly subsampled 2,528 viruses [see Data Set S1 in the supplemental material]). Of the 7, 1 HA was incorrectly classified (i.e., 1A.2 virus classified to 1A.3), 1 HA was incorrectly assigned to a lower-order division (1A.1 virus was placed in the 1A.1.2 clade), and the remaining 5 viruses were incorrectly assigned to a higher-order division (1A.3.3 classified to 1A.3).

Automated classification of swine H1 hemagglutinin sequences. The representative phylogeny used for classifying global swine sequences contained 239 H1 viruses of predominantly swine origin, with a few H1 viruses from human and avian hosts to represent the diversity of nonswine H1 viruses. The swine viruses were selected to capture the diversity within each of the defined clades. We used this algorithm to classify all sequences in the final data set of 7,070 IAV HA/H1 sequences from swine, avian, and human hosts, described in Materials and Methods. The classifier ascribed the correct clade in all but 41 instances. Of these 41 sequences, three from clade 1A.3.3.1 were incorrectly assigned clade 1A.3.3. The remaining 38 sequences were assigned a “-like” classification very close to the correct value. For example, five 1A.1 swine sequences were assigned the classification “1A.1-like.” One 1C.2.1 swine sequence was assigned the classification “1C.2-like.” Overall, the classifier was highly accurate in correctly capturing the classifications assigned by the earlier expert phylogenetic curation. Thus, this tool will be valuable for rapidly assigning the appropriate, biologically meaningful clade to new viruses not studied in our analyses. Its implementation on the web, through IRD (http://www.fludb.org), will allow classification of novel sequences to be carried out in clinical or diagnostic settings.

DISCUSSION
Swine influenza was first observed in 1918, with the ancestral “classical” H1N1 virus isolated from swine in the 1930s. At present, there are three major evolutionary lineages circulating in swine globally, resulting from the 1918 H1N1 human pandemic, human seasonal H1 viruses, and an avian H1 lineage. As lineages of these viruses were established locally, many of them became ecologically isolated, resulting in divergent evolutionary trajectories (15). We identified 3 first-order lineages, 7 second-order divisions, 13 third-order divisions, and 8 fourth-order divisions that sufficiently capture the historical and current genetic diversity of global swine H1 HA influenza viruses. In doing so, we established rational and rigorous criteria for naming such clades. These criteria are flexible enough to adapt to continued within-clade evolution of viruses and allow for the identification and classification of novel lineages should they emerge.

Our primary goal was to classify HA clades that reflected the evolutionary history of swine IAV. To do so, we use three first-order descriptors—the 1A classical lineage derived from the 1918 human pandemic viruses, the 1B human seasonal lineage...
associated with 1990s human-to-swine transmission episodes, and the 1C Eurasian avian lineage associated with viruses introduced to swine in Europe and Asia from wild birds (37). Following this, we identified monophyletic clades in our phylogeny with at least 10 viruses collected over the preceding 5 years: without exception, these clades had statistical support of \( \geq 70\% \) and generally an average pairwise distance of \(< 7\% \) within clade and \( > 7\% \) between clades. When applying these criteria with different data sets, there were minor discrepancies (\( n = 7 \)): this highlights the nondeterministic nature of maximum likelihood phylogenetic approaches. The solution to this problem is to use multiple approaches, to use more comprehensive data sets, to conduct analyses more than once, and to interpret the data conservatively.

To facilitate the adoption of this system, we implemented an automated annotation tool that can rapidly assign these biologically informative clade designations to new, as-yet-unclassified sequence data. Our tool uses maximum likelihood to rapidly classify a query IAV sequence by placing it on a reference phylogeny of just 239 H1 viruses selected from the named, biologically informative clades. When a query sequence is placed within a named clade, this name is assigned to the query. When a query sequence does not fall within a named clade, it is classified by the neighborhood of its placement, using a “-like” annotation. For example, the tool assigns the classification “1B.2.2-like” to viruses ancestral to both the 1B.2.2.1 and 1B.2.2.2 clades but not placed within the 1B.2.2 clade (see Fig. S2 in the supplemental material). These “-like” viruses have insufficient statistical support to assign them to a monophyletic clade, forcing a placement between existing clades. By using our automated classification, sequences collected during surveillance efforts can quickly be classified to known clades or, if receiving a “-like” designation, can be flagged for additional analyses or additional targeted sample collection.

Our goal of achieving tightly structured definitions for statistically supported clades was challenged by the relatively frequent introduction of avian and human IAVs to swine populations (e.g., see references 5 and 38) and the absence of surveillance in large sections of the world, including some with significant swine populations (10). Another challenge was the forced inclusion of viruses with likely specific regional evolutionary histories into a geographically broader classification because of the paucity of sequences from that region. For example, a small cluster of distinct human seasonal viruses in Brazil (23) were classified as 1B.2.2 although they differed from other 1B.2.2 viruses that circulate in different geographic regions. A unique clade designation for this handful of Brazil viruses might be considered if phylogenetic support was \( > 70\% \) and if additional evidence demonstrated continued circulation of this genetic grouping, such as specific hemagglutination inhibition serosurveillance data. These modified criteria (high statistical support and serosurveillance data) may be applied to interspecies spillover events and undersampled regions and allow the creation of further meaningful clade divisions when additional virologic sampling and sequencing are not feasible.

The most readily available approach to limiting IAV transmission within swine populations is through an appropriate vaccination program that protects against currently circulating genetic and antigenic diversity (15, 39, 40). Importantly, our global classification scheme can inform vaccine strain selection: it is not possible to compose a vaccine with all known viral variants (41), and our scheme provides a mechanism for quickly filtering data spatially and temporally, allowing matching to existing vaccines or selection of representative viruses for vaccine research and development. Experimental studies have demonstrated that protection against infection may be correlated with genetic relatedness of the vaccine strain to challenge strains (e.g., see references 42 and 43). However, vaccine efficacy would likely be compromised when considering all clade levels because there are a substantial number of viruses belonging to as many as 18 genetic second- and third-order clades in each continent (e.g., the United States has 12 cocirculating H1 genetic clades), genetic relatedness is not always a good predictor of protection (e.g., see reference 44) because just one or two amino acid mutations in the HA-1 domain may drive a significant reduction in antigenic cross-reactivity (e.g., see
references 21, 45, and 46), and host immune response affects protection (47–49). Despite this challenge, and in lieu of a universal vaccine, our classification system can identify regional patterns of genetic diversity, which can lead to assessment of antigenic diversity relative to other viruses (15). For example, if the widely dispersed 1A.3.2 viruses (n = 541, H1N1pdm09 viruses) are excluded, 84% of the publicly available swine H1 viruses from 2010 to the present belonged to 6 predominant clades: one from the 1A classical lineage, three from the 1B human seasonal lineage, and two from the 1C Eurasian avian lineage. Though there is no centralized system for matching circulating strains with vaccine seeds, these data and the relatively slow antigenic drift of swine viruses (average of 0.39 antigenic units per year for classical-lineage viruses [15]) suggest that a selection of viruses with regional representation would be sufficient for an acceptable vaccine efficacy that reduces clinical burden and limits virus spread.

Swine IAV evolution is a complex issue at regional and especially at global levels. The emergence and extinction of clades due to ecological and evolutionary processes, along with spillover events from nonswine hosts, have created a nomenclature quagmire. Consequently, we developed a unified system that accounts for the unique evolutionary history of swine IAV that can be periodically updated as viral diversity expands or contracts. The data to create a classification system and the accompanying automated tool rely exclusively on genetic divergence in the HA and do not infer information on viral phenotype. Future modeling and computational tools can build from and adapt this system. For example, the classification of nonswine H1 viruses could follow the process described here for swine H1, leading to a comprehensive, multihost H1 classification scheme. Incorporating data from functional HA studies could refine clade definitions. For example, including studies on antigenic evolution with genetic classification could provide advanced metrics for clade definition, which would facilitate the selection of vaccine strains and inform risk management policies for agricultural and public health.

**MATERIALS AND METHODS**

**Swine influenza A virus hemagglutinin H1 data set.** All available swine IAV hemagglutinin (HA) H1 sequences from viruses in the IRD (30) were downloaded on 7 June 2016. Only H1N1 and H1N2 subtype viruses were included, and these sequences comprised 8,438 worldwide samples. To restrict our analyses to relevant field viruses, we excluded sequences with “lab” or “laboratory” host. Sequences were then aligned with MAFFT v 7.221 (50, 51), with manual correction and curation in Mesquite v 1.36 (52). The aligned sequences underwent a redundancy analysis within the program mothur v 1.36.0 (53), and sequences with 100% identity were removed. Our final filtering step was to remove poor-quality data using two criteria: sequences were removed if >50% of the HA gene sequence was missing and a sequence was removed if it had more than 5 nucleotide base ambiguities. This process resulted in a set of 6,298 nonidentical H1 HA IAV sequences that represent the full extent of published swine H1 HA genetic diversity worldwide. An additional 428 randomly sampled human seasonal H1 HA sequences and 344 nonidentical H1 HA swine IAV sequences that represent the full extent of published swine HA genetic diversity relative to other viruses (15). For example, if the widely dispersed 1A.3.2 viruses (n = 541, H1N1pdm09 viruses) are excluded, 84% of the publicly available swine H1 viruses from 2010 to the present belonged to 6 predominant clades: one from the 1A classical lineage, three from the 1B human seasonal lineage, and two from the 1C Eurasian avian lineage. Though there is no centralized system for matching circulating strains with vaccine seeds, these data and the relatively slow antigenic drift of swine viruses (average of 0.39 antigenic units per year for classical-lineage viruses [15]) suggest that a selection of viruses with regional representation would be sufficient for an acceptable vaccine efficacy that reduces clinical burden and limits virus spread.

**Phylogenetic methods, clade annotation, and clad comparisons.** From these data, a maximum likelihood tree was inferred using RAxML (v8.2.4 [54]) on the CIPRES Science Gateway (55) employing the rapid bootstrap algorithm, a general time-reversible (GTR) model of nucleotide substitution, and F-distributed rate variation among sites. The statistical support for individual branches was estimated by bootstrap analysis with the number of bootstrap replicates determined automatically using an extended majority-rule consensus tree criterion (56). Using this phylogeny, we defined clades using quantifiable criteria that were applied collectively across the entire data set. Clades were defined based on sharing of a common node and monophyly, statistical support greater than 70% at the clade-defining node, and average percent pairwise nucleotide distances between and within clades of >7% and <7%, respectively, with certain minor exceptions (see Results). Given recent, relatively frequent, spillover of nonswine viruses without subsequent onward transmission in swine populations, we required a minimum of 10 viruses between 2010 and the present in a proposed clade before assigning a clade designation. Using this process, we identified three first-order lineages, seven second-order divisions, 13 third-order divisions, and eight fourth-order divisions (Fig. 1; see also Data Set S1 in the supplemental material). Sampling and sequencing in the 1900s and early 2000s were not representative of the relative abundance of different swine IAV clades (see reference 10); consequently, in Results, we restrict comments on abundance and geographical dispersion to just those data from 2010 to the present.

To validate tree topology, branch support, and the subsequent manual clade designations, we created three separate data sets by separating the 6,298 swine H1 sequences into the three first-order lineages and then randomly subsampling viruses from each second-order division. The first data set
Nomenclature of Swine Influenza H1 Viruses

This study was supported by USDA-ARS; by USDA-APHIS; by an NIH-National Institute of Allergy and Infectious Diseases (NIAID) interagency agreement associated with the Center of Research in Influenza Pathogenesis (CRIP), an NIAID-funded Center of Excellence in Influenza Research and Surveillance (CEIRS, HHSN272201400008C); and by the St. Jude Center of Excellence in Influenza Research and Surveillance (CEIRS, HHSN272201400008C); T.K.A. was funded by USDA-ARS SCA agreement no. 58-3625-4-070 and by an appointment to the USDA-ARS Research Participation Program administered by the Oak Ridge Institute for Science and Education (ORISE) through an interagency agreement between the U.S. Department of Energy (DOE) and USDA under contract no. DE-AC05-06OR23100. N.S.L. was funded by USDA-ARS SCA agreement no. 58-3625-2-103F, the EC FP7 award no. 259949, and CRIP (CEIRS, HHSN272201400008C). The research leading to these results has received funding from the European Union’s Seventh Framework Programme for research, technological development, and demonstration under grant agreement no. 259949.

SUPPLEMENTAL MATERIAL

Supplemental material for this article may be found at http://dx.doi.org/10.1128/mSphere.asm.org/00275-16.

Figure S1, TXT file, 1.3 MB.
Figure S2, TIF file, 0.2 MB.
Data Set S1, XLSX file, 0.5 MB.

ACKNOWLEDGMENTS

We gratefully acknowledge the laboratories that deposit swine influenza virus sequences into publicly available databases and the OFFLU network and contributing support staff at all participating organizations and institutions.

This study was supported by USDA-ARS; by USDA-APHIS; by an NIH-National Institute of Allergy and Infectious Diseases (NIAID) interagency agreement associated with the Center of Research in Influenza Pathogenesis (CRIP), an NIAID-funded Center of Excellence in Influenza Research and Surveillance (CEIRS, HHSN272201400008C); and by the St. Jude Center of Excellence in Influenza Research and Surveillance (HHSN266200700005C and HHSN272201400006C). T.K.A. was funded by USDA-ARS SCA agreement no. 58-3625-4-070 and by an appointment to the USDA-ARS Research Participation Program administered by the Oak Ridge Institute for Science and Education (ORISE) through an interagency agreement between the U.S. Department of Energy (DOE) and USDA under contract no. DE-AC05-06OR23100. N.S.L. was funded by USDA-ARS SCA agreement no. 58-3625-2-103F, the EC FP7 award no. 259949, and CRIP (CEIRS, HHSN272201400008C). The research leading to these results has received funding from the European Union’s Seventh Framework Programme for research, technological development, and demonstration under grant agreement no. 259949.

Swine H1 clade classification tool. The H1 gene classification tool is based on a bifurcating scaffold phylogenetic tree inferred using maximum likelihood from 3 to 10 representatives of each well-supported, named clade, to capture the evolutionary relationships among clades. To be included in this representative phylogeny, an H1 sequence was required to be at least 1,600 nucleotides (nt) long but was unrestricted with respect to host species. The classifier uses pplacer (60) to attach a query sequence to a branch in this tree, without reestimating the tree. Thus, the tree of representative sequences acts as a “scaffold” upon which the query sequence is placed. pplacer maximizes the likelihood of the placement by comparing the sequence of the query with the sequences in the tree, given the estimates of the evolutionary parameters underlying the inferred phylogeny. The classifier then assigns a clade to the query based on the clades represented in the local neighborhood of its placement (see Fig. S2 in the supplemental material), as follows: (i) if the query is attached to a terminal branch, then it is assigned the clade of the virus at the tip; (ii) if the query is attached to an internal branch, then it is assigned the clade of the node at the basal end of this branch. Internal nodes are assigned clades according to the rules in a parameter file. Nodes with “-like” classifications fall into internode regions joining subtrees of distinct clades. In our experience, viruses assigned “-like” classifications are often transitional, occurring prior to or during the emergence of a new clade that successfully expands onward. The “-like” designation attempts to capture the position intermediate between older and newer clades.

The classifier is written in perl and is portable, fast, and accurate. Importantly, it is adaptable readily to other clade classification tasks, because it specifies the parameters relevant to a particular application in external files. To date, it has been applied successfully to classification of avian HA/H5 sequences, according to the nomenclature of the WHO/OIE/FAO H5N1 Working Group (29), to distinguishing new pandemic 2009 H1 viruses from earlier seasonal H1 viruses in humans and other hosts, and to classifying U.S. swine H1 HA phylogenetic clades. These three applications have been implemented on IRD (30, 31).

To quantify the within- and between-clade nucleotide distances for the H1 clade designations, the average pairwise distances (APDs) were calculated in MEGA-CC v 7.07 (59) using the p-distance calculation.

**ACKNOWLEDGMENTS**

We gratefully acknowledge the laboratories that deposit swine influenza virus sequences into publicly available databases and the OFFLU network and contributing support staff at all participating organizations and institutions.

This study was supported by USDA-ARS; by USDA-APHIS; by an NIH-National Institute of Allergy and Infectious Diseases (NIAID) interagency agreement associated with the Center of Research in Influenza Pathogenesis (CRIP), an NIAID-funded Center of Excellence in Influenza Research and Surveillance (CEIRS, HHSN272201400008C); and by the St. Jude Center of Excellence in Influenza Research and Surveillance (CEIRS, HHSN272201400008C); T.K.A. was funded by USDA-ARS SCA agreement no. 58-3625-4-070 and by an appointment to the USDA-ARS Research Participation Program administered by the Oak Ridge Institute for Science and Education (ORISE) through an interagency agreement between the U.S. Department of Energy (DOE) and USDA under contract no. DE-AC05-06OR23100. N.S.L. was funded by USDA-ARS SCA agreement no. 58-3625-2-103F, the EC FP7 award no. 259949, and CRIP (CEIRS, HHSN272201400008C). The research leading to these results has received funding from the European Union’s Seventh Framework Programme for research, technological development, and demonstration under grant agreement no. 259949.
REFERENCES


33. Allison M, Deen J, Detmer SE, Gnamer MR, Joo HS, Romagosa A, Volume 1 Issue 6 e00275-16 msphere.asm.org


