Inferring the inter-host transmission of influenza A virus using patterns of intra-host genetic variation

J. Conrad Stack1, Pablo R. Murcia2, Bryan T. Grenfell3,4, James L. N. Wood5 and Edward C. Holmes1,4

1Center for Infectious Disease Dynamics, Department of Biology, The Pennsylvania State University, University Park, PA 16802, USA
2Medical Research Council – University of Glasgow Centre for Virus Research, Institute of Infection, Inflammation and Immunity, College of Medical, Veterinary and Life Sciences, University of Glasgow, Glasgow G12 8TA, UK
3Department of Ecology and Evolutionary Biology, Princeton University, Princeton, NJ 08544, USA
4Fogarty International Center, National Institutes of Health, Bethesda, MD 20850, USA
5Disease Dynamics Unit, Department of Veterinary Medicine, University of Cambridge, Cambridge CB3 0ES, UK

Influenza A viruses (IAVs) cause acute, highly transmissible infections in a wide range of animal species. Understanding how these viruses are transmitted within and between susceptible host populations is critical to the development of effective control strategies. While viral gene sequences have been used to make inferences about IAV transmission dynamics at the epidemiological scale, their utility in accurately determining patterns of inter-host transmission in the short-term—i.e. who infected whom—has not been strongly established. Herein, we use intra-host sequence data from the viral HA1 (hemagglutinin) gene domain from two transmission studies employing different IAV subtypes in their natural hosts—H3N8 in horses and H1N1 in pigs—to determine how well these data recapitulate the known pattern of inter-host transmission. Although no mutations were fixed over the course of either experimental transmission chain, we show that some minor, transient alleles can provide evidence of host-to-host transmission and, importantly, can be distinguished from those that cannot.

1. Introduction

The substantial and recurring impact of influenza viruses (IAVs) on humans and other animal species makes understanding their evolutionary patterns and processes a research priority. In their natural reservoir, wild waterfowl, IAV infections usually result in relatively mild disease. However, IAVs are also able to spill-over and become established in new host species, including humans, pigs and horses. The large and dynamic host range of IAVs reflects their ability to adapt rapidly to new host environments, itself a function of high rates of mutation [1] and large peak population sizes within hosts [2–4]. Despite this diversity, the decreasing cost of genome sequencing has been highly beneficial to the study of IAV evolution and epidemiology, allowing exploration of the link between patterns of transmission within and between host populations and among species. The design of IAV vaccines may also be improved by a more accurate picture of global viral diversity and patterns of spread [5–7].

Although gene sequence data have greatly enhanced the study of viral evolution at the epidemiological scale [8–10], establishing the relationship between the evolutionary and transmission dynamics of IAV at more localized scales has proved more difficult. Most epidemiological studies use viral population consensus sequences, which represent the dominant nucleotide at every position, and where little variation is observed among sequences sampled from hosts linked by direct transmission. Although the analysis of consensus sequences is valuable for many aspects of molecular epidemiology, that nucleotide fixation events (i.e. changes to the consensus sequence) are not expected to
occur over such short time spans mean that they are of little value at the scale of individual transmission events. Indeed, at smaller scales, substantial genetic variation is only observed in intra-host viral populations [9]. Although the deep sequencing of intra-host viral genetic variation is likely to largely capture transient, deleterious variants which differ little from the consensus sequence [11], it is possible that this diversity is sufficient to recover precise patterns of inter-host transmission.

To date, little has been done to explore how intra-host viral sequence data might be integrated into the study of inter-host transmission dynamics, including reconstructing the exact pathway of host-to-host transmission [12,13]. To address this issue, we conducted a novel analysis of two previously published datasets, involving the H3N8 subtype in horses (Equus caballus) [2] and the Eurasian H1N1 subtype in swine (Sus scrofa) [4,14]. As both studies used experimentally controlled host transmission patterns they provide an excellent starting point for better understanding the relationship between intra- and inter-host viral evolution.

Equine influenza is a common, acute, highly transmissible and typically non-fatal upper respiratory tract infection [15]. The H3N8 subtype was first identified in 1963 and has since become endemic to North America and Europe and epidemic in other regions. Although routine vaccination has greatly reduced the economic losses associated with equine influenza, and H3N8 in horses is characterized by lower rates of antigenic change than those of IAVs in other mammals [16], large-scale outbreaks continue to occur [15,17].

Swine influenza manifests similar symptoms in pigs as H3N8 does in horses and is the cause of major economic disruption [18]. It is caused by three different IAV subtypes, including two distinct strains of H1N1, one of which is an Eurasian (or ‘avian-like’) strain first observed in the late 1970s [18]. Unlike horses, domestic pigs play a central role in the global transmission and evolution of contemporary IAVs [19]. The use of pigs as livestock brings them into frequent contact with potential carriers of IAV from human and other domestic, notably poultry, populations. Indeed, swine IAVs have been the progenitors of several large human influenza epidemics [20].

The two studies used here implemented experimental viral transmission and generated relatively large datasets of intra-host viral HA1 (haemagglutinin) domain sequences. In both cases, experimental infections were initiated and allowed to progress naturally through small naive populations (approx. 10 individuals). Contacts between infected and susceptible hosts were carefully controlled to greatly reduce the number of possible transmission networks, to the point where the route was effectively fixed. The viral population of each host was sequenced at least once over the 2–6 days following infectious contact (or mechanical inoculation). This sampling protocol resulted in temporally and spatially stratified sequence datasets whose intra-host sampling depth ranged from 6 to 154 sequences. Importantly, although both datasets showed very little change at the level of the consensus sequence with no fixation events, there was substantial intra-host genetic variation.

The central aim of our study was to determine whether patterns of intra-host genetic variation can retrace the pathway of inter-host viral transmission, even in cases where the population consensus sequences have remained unchanged. To this end, we used comparative analyses to establish whether minor viral alleles (i.e. those present beneath the consensus) were shared between hosts in a manner consistent with the known pattern of host-to-host contact.

2. Methods

(a) Datasets

More details about the two datasets examined here can be found in the original publications [2,4,14]. Briefly, each equine H3N8 and swine H1N1 virus was first grown in eggs and then inoculated into a pair of naive animals. These initial inoculated hosts were then exposed to a predetermined pair of susceptible naive hosts and were removed once this new set showed clinical signs of influenza infection. This process was repeated until all hosts had been exposed. Nasal swabs were collected from hosts between 2 and 6 days after mechanical inoculation or infectious contact. Swabs were subject to RNA extraction and the HA1 subunit gene (comprising 903 and 939 nucleotides for the equine and swine datasets, respectively), was amplified by RT-PCR, cloned and sequenced. The general statistics for the sequence datasets of both studies are shown in table 1, while the schematic of the host contact networks are presented in figure 1.

Table 1. Summary statistics for the two datasets used in this study. $D_{est}$ is a measure of subpopulation differentiation. Asterisks represent the level of statistical significance based on 1000 permutation tests (**0.001; *0.01).

<table>
<thead>
<tr>
<th></th>
<th>equine H3N8</th>
<th>swine H1N1</th>
</tr>
</thead>
<tbody>
<tr>
<td>no. sequences</td>
<td>2366</td>
<td>868</td>
</tr>
<tr>
<td>no. animals</td>
<td>11</td>
<td>10</td>
</tr>
<tr>
<td>time span (days)</td>
<td>14</td>
<td>14</td>
</tr>
<tr>
<td>no. viral populations</td>
<td>30</td>
<td>21</td>
</tr>
<tr>
<td>range of sampling depth</td>
<td>44 – 154</td>
<td>6 – 77</td>
</tr>
<tr>
<td>no. unique mutations</td>
<td>244</td>
<td>328</td>
</tr>
<tr>
<td>no. unique alleles</td>
<td>251</td>
<td>331</td>
</tr>
<tr>
<td>minor alleles (%)</td>
<td>14.58</td>
<td>56.11</td>
</tr>
<tr>
<td>mean sequence diversity</td>
<td>0.00038</td>
<td>0.00184</td>
</tr>
<tr>
<td>$D_{est}$ by viral population</td>
<td>0.0058***</td>
<td>0.2983***</td>
</tr>
<tr>
<td>$D_{est}$ by host</td>
<td>0.0013**</td>
<td>0.2939***</td>
</tr>
<tr>
<td>$D_{est}$ by day</td>
<td>0.0004</td>
<td>0.2051***</td>
</tr>
</tbody>
</table>

(b) Sequence analysis

In both studies, the population consensus sequence was identical to the strains used to inoculate the first infected animals—A/Equine/Newmarket/1/1993 (H3N8) and A/swine/England/453/2006 (H1N1)—and was the dominant allele. All mutations from the consensus were tallied and their genomic locations recorded, including whether they fell within hypothesized epitope regions [21,22].

Except where stated, all data analyses were carried out in the R software environment [23] supplemented with a number of add-on packages [24–29]. The genetic distance between each pair of viral sequences was computed as the raw number of differences between them (i.e. the p-distance). This metric is appropriate given the low sequence diversity (table 1). The overall differentiation in the sequence datasets was measured by $D_{est}$ [30]. Calculating this population-level statistic requires that all alleles are subdivided into groups based on some arbitrary
permutations for each subdivision) were used to assess whether within- and between-group comparisons. Bootstrap tests (1000 transmission network, these alleles were then categorized into minor alleles present in more than one host and are referred to as ‘I’ in equine H3N8 and ‘IV’ in swine H1N1, all other (i.e. minor) alleles contained a set of one or more mutations. Individual viral sequences were assigned to alleles using the haplotype function from the pegas R package, excluding nucleotide positions with missing data. By default, this assigned a unique roman numeral to each allele. Except for the two dominant alleles (‘I’ for equine H3N8 and ‘IV’ in swine H1N1), all other (i.e. minor) alleles contained a set of one or more mutations. Allele networks were calculated from *p*-distances under the assumption of an infinite sites model. Minor alleles were further categorized according to the following ‘explanatory’ characteristics: (i) their frequency across the whole study, (ii,iii) the number of animals and viral populations in which they occurred, (iv) the number of descendants they potentially gave rise to (see below), (v) the number of mutations separating them from the consensus sequence, (vi) their betweenness centrality in the allele network, and whether or not any of their nucleotide mutations, (vii) occurred in epitope regions, (viii) induced a non-synonymous change, or (ix) induced a premature stop codon (table 2). The potential descendants of an allele were defined as all other alleles whose set of mutations was both larger than and a superset of its set. The betweenness centrality of an allele is related to this; it is a graph-theory measure that represents how ‘central’ an allele is in the allele network based on the number of times it occurs in the set of shortest paths connecting all node pairs.

To determine the importance of each explanatory characteristic, they were compared against a binary ‘response’ characteristic: that is, whether or not a given allele was shared between hosts known to have had direct contact (table 2, response). To determine which explanatory variables could most accurately predicted the response variable, classification tree analyses were carried out on the shared minor allele characteristic datasets from both the equine H3N8 and swine H1N1 studies, as well as on a combination of the two. Importance scores for each explanatory variable (table 2) were calculated and are proportional to the reduction in the misclassification rate which each variable affects at a particular internal node, summed over all internal nodes [28]. The relative rank of a variable’s importance score provides an indication of how strongly it is associated with the response variable. These scores are not comparable between datasets. Each classification tree was constructed using a binary recursive partitioning technique [31] in which data observations are recursively split into two groups based on a series of classification functions (i.e. internal nodes or splits), until some stopping criteria are met. Classification functions were only estimated at nodes that contained 10 per cent or more of the observations [32], and proposed splits were only accepted if they improved the tree fit by a factor of 0.001 as gauged by a cross-validation procedure (n = 1000). Every allele characteristic was considered as a candidate (competitor) for each internal node and was ranked according to the amount it reduced the Gini impurity of a

![Figure 1. Schematic of the networks of known contacts between hosts (red lines) and sampling dates (black dots) for the (a) equine H3N8 and (b) swine H1N1 experimental transmission studies. Y-axis labels indicate the unique identifiers randomly assigned to each animal in either study. The x-axis indicates the study day (day 0 is the day in which animals were mechanically inoculated). Red dots indicate the day on which that individual animal came into contact with an infected host or was mechanically inoculated. Where multiple red lines radiate from a single red dot there is some ambiguity as to which animal transmitted to which when three (in the equine study) or four (in the swine study) animals were housed together. For example, on day 3 of the swine study, pigs 113 and 115, having shown symptoms of infection, were housed with pigs 116 and 104.](http://rspb.royalsocietypublishing.org/Downloaded from http://rspb.royalsocietypublishing.org/)

(c) Characterization of minor alleles

Individual viral sequences were assigned to alleles using the haplotype function from the pegas R package, excluding nucleotide positions with missing data. By default, each allele was assigned a unique roman numeral to each allele. Except for the two dominant alleles (‘I’ for equine H3N8 and ‘IV’ in swine H1N1), all other (i.e. minor) alleles contained a set of one or more mutations. Allele networks were calculated from *p*-distances under the assumption of an infinite sites model. Minor alleles were further categorized according to the following ‘explanatory’ characteristics: (i) their frequency across the whole study, (ii,iii) the number of animals and viral populations in which they occurred, (iv) the number of descendants they potentially gave rise to (see below), (v) the number of mutations separating them from the consensus sequence, (vi) their betweenness centrality in the allele network, and whether or not any of their nucleotide mutations, (vii) occurred in epitope regions, (viii) induced a non-synonymous change, or (ix) induced a premature stop codon (table 2). The potential descendants of an allele were defined as all other alleles whose set of mutations was both larger than and a superset of its set. The betweenness centrality of an allele is related to this; it is a graph-theory measure that represents how ‘central’ an allele is in the allele network based on the number of times it occurs in the set of shortest paths connecting all node pairs.

To determine the importance of each explanatory characteristic, they were compared against a binary ‘response’ characteristic: that is, whether or not a given allele was shared between hosts known to have had direct contact (table 2, response). To determine which explanatory variables could most accurately predicted the response variable, classification tree analyses were carried out on the shared minor allele characteristic datasets from both the equine H3N8 and swine H1N1 studies, as well as on a combination of the two. Importance scores for each explanatory variable (table 2) were calculated and are proportional to the reduction in the misclassification rate which each variable affects at a particular internal node, summed over all internal nodes [28]. The relative rank of a variable’s importance score provides an indication of how strongly it is associated with the response variable. These scores are not comparable between datasets. Each classification tree was constructed using a binary recursive partitioning technique [31] in which data observations are recursively split into two groups based on a series of classification functions (i.e. internal nodes or splits), until some stopping criteria are met. Classification functions were only estimated at nodes that contained 10 per cent or more of the observations [32], and proposed splits were only accepted if they improved the tree fit by a factor of 0.001 as gauged by a cross-validation procedure (n = 1000). Every allele characteristic was considered as a candidate (competitor) for each internal node and was ranked according to the amount it reduced the Gini impurity of a

![Figure 1. Schematic of the networks of known contacts between hosts (red lines) and sampling dates (black dots) for the (a) equine H3N8 and (b) swine H1N1 experimental transmission studies. Y-axis labels indicate the unique identifiers randomly assigned to each animal in either study. The x-axis indicates the study day (day 0 is the day in which animals were mechanically inoculated). Red dots indicate the day on which that individual animal came into contact with an infected host or was mechanically inoculated. Where multiple red lines radiate from a single red dot there is some ambiguity as to which animal transmitted to which when three (in the equine study) or four (in the swine study) animals were housed together. For example, on day 3 of the swine study, pigs 113 and 115, having shown symptoms of infection, were housed with pigs 116 and 104.](http://rspb.royalsocietypublishing.org/Downloaded from http://rspb.royalsocietypublishing.org/)
node (which is equivalent to the expected error rate) [33]. To avoid model over-fitting, leaf nodes (i.e. unsplit internal nodes) were constrained to contain a minimum of two observations. The accuracy of each classification tree is represented by its confusion matrix, which shows the predicted values of the binary response variable against its actual value in the dataset (electronic supplementary material, table S1). More details on this analytical procedure are given in the electronic supplementary material.

Finally, to determine whether associations between the response variable and explanatory characteristics might be missed due to inconsistent sampling, we evaluated the statistical association between the instance of each allele and the size of the viral population. Accordingly, for a given allele, viral population sizes were divided into two groups based on whether or not they contained a specific allele. One-tailed t-tests of population means were then used to determine if the size of viral populations that contained a specific allele were significantly larger than those that did not (electronic supplementary material, tables S2 and S3), with all p-values adjusted for multiple comparisons [34].

(d) Inference of the transmission network

Shared alleles judged most probably to be indicative of host-to-host transmission events using the classification tree for the combined dataset (because their characteristics were significantly associated with known host contact—see §3) were used to estimate the known transmission network. Classification trees can be biased in their selection of explanatory variables at each internal node, known transmission network. Classification trees can be biased in their selection of explanatory variables at each internal node, and associated summary statistics. Boxes are arranged to show how alleles were progressively segregated based on the hosts they were observed in. Box colours match with those assigned to the alleles displayed in figures 3 and 4.

**Figure 2.** Flow diagram of the allele analysis (a) equine H3N8; (b) swine H1N1 and associated summary statistics. Boxes are arranged to show how alleles were progressively segregated based on the hosts they were observed in. Box colours match with those assigned to the alleles displayed in figures 3 and 4.

**Table 2.** Allele characteristics (i–ix), their data type, and their relative importance in the classification trees built using the equine H3N8 dataset, the swine H1N1 dataset and both datasets combined. Variable importance scores are calculated using information derived from the decision tree building process and high scores do not necessarily indicate that a variable is present in the final tree. All values for the number of mutations (characteristic v) were 1.0 and hence had no discriminating capability.

<table>
<thead>
<tr>
<th>ID</th>
<th>description</th>
<th>data type</th>
<th>usage</th>
<th>equine H3N8</th>
<th>swine H1N1</th>
<th>combined</th>
</tr>
</thead>
<tbody>
<tr>
<td>response</td>
<td>shared between connected animals</td>
<td>binary</td>
<td>response</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>i</td>
<td>frequency in whole study</td>
<td>numeric</td>
<td>explanatory</td>
<td>5.49</td>
<td>6.76</td>
<td>7.04</td>
</tr>
<tr>
<td>ii</td>
<td>no. animals</td>
<td>cardinal</td>
<td>explanatory</td>
<td>3.66</td>
<td>4.96</td>
<td>5.53</td>
</tr>
<tr>
<td>iii</td>
<td>no. viral populations</td>
<td>cardinal</td>
<td>explanatory</td>
<td>3.70</td>
<td>7.76</td>
<td>9.85</td>
</tr>
<tr>
<td>iv</td>
<td>no. potential descendants</td>
<td>cardinal</td>
<td>explanatory</td>
<td>3.53</td>
<td>4.49</td>
<td>8.28</td>
</tr>
<tr>
<td>v</td>
<td>no. mutations</td>
<td>cardinal</td>
<td>explanatory</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>vi</td>
<td>betweenness centrality</td>
<td>binary</td>
<td>explanatory</td>
<td>1.94</td>
<td>2.34</td>
<td>3.17</td>
</tr>
<tr>
<td>vii</td>
<td>has mutation in epitope</td>
<td>binary</td>
<td>explanatory</td>
<td>0.93</td>
<td>6.02</td>
<td>5.33</td>
</tr>
<tr>
<td>viii</td>
<td>has non-synonymous mutation</td>
<td>binary</td>
<td>explanatory</td>
<td>0.55</td>
<td>4.70</td>
<td>1.99</td>
</tr>
<tr>
<td>ix</td>
<td>has premature stop codon</td>
<td>binary</td>
<td>explanatory</td>
<td>0.56</td>
<td>0.94</td>
<td>1.52</td>
</tr>
</tbody>
</table>

The accuracy of each classification tree is represented by its confusion matrix, which shows the predicted values of the binary response variable against its actual value in the dataset (electronic supplementary material, table S1). More details on this analytical procedure are given in the electronic supplementary material.

Finally, to determine whether associations between the response variable and explanatory characteristics might be missed due to inconsistent sampling, we evaluated the statistical association between the instance of each allele and the size of the viral population. Accordingly, for a given allele, viral population sizes were divided into two groups based on whether or not they contained a specific allele. One-tailed t-tests of population means were then used to determine if the size of viral populations that contained a specific allele were significantly larger than those that did not (electronic supplementary material, tables S2 and S3), with all p-values adjusted for multiple comparisons [34].

(d) Inference of the transmission network

Shared alleles judged most probably to be indicative of host-to-host transmission events using the classification tree for the combined dataset (because their characteristics were significantly associated with known host contact—see §3) were used to estimate the known transmission network. Classification trees can be biased in their selection of explanatory variables at each internal node, favouring those with larger numbers of possible values [35]. This bias was not expected to affect in our analyses because of the small number of observations and the relatively low variance in variable ranges. Indeed, additional tree-based predictors robust to this bias [35] predicted the same minor alleles as our original
classification trees (see the electronic supplementary material). Shared minor alleles with stop codon mutations were also included in estimations of the transmission network because they appeared to be transmitted between individuals (see §3). Edges were drawn between host animals on the basis of shared alleles and proximity of sampling days on which the shared alleles were observed. In the simplest case, when alleles were observed in only two hosts, an edge was drawn connecting them. For shared alleles observed in three or more hosts, hosts were first ordered by sampling date and then serially connected with edges. Additional edges were then added between those hosts not previously connected if their sample dates were 1 or fewer days apart.

Figure 3. Allele networks for all sequences in the (a) equine H3N8 and (b) swine H1N1 studies. Blue nodes represent the dominant allele in each host as well as minor alleles that were shared between hosts known to have had direct contact. Light green nodes represent minor alleles that were observed more than once, but never shared between hosts in direct contact. Node sizes are a function of the number of times the each allele was observed in each host and relative only to others within each network. Some nodes were shrunk to improve clarity. Dark dotted lines connect minor alleles that are shared between hosts in a manner consistent with the known host-to-host contacts. Light green dotted lines connect instances of minor alleles shared between hosts which were never in direct contact. Nodes are positioned vertically based on the first day they came into contact with an infected host. Inset. The number of edges connecting shared minor alleles that were consistent or inconsistent with the pattern of host-to-host contacts.

Figure 4. Approximately transmission networks between hosts constructed using specific minor alleles in the (a) equine H3N8 and (b) swine H1N1 studies. Blue nodes represent the overall dominant alleles (consensus strains) specific to each study and also indirectly represent each individual host. Orange nodes represent shared minor alleles that were selected by specific criteria (see text). Node sizes are drawn relative to the number of alleles they represent. Orange lines connecting hosts represent known host-to-host contacts that were accurately inferred. Dotted, grey lines connect hosts that had no direct contact. The thickness of these lines is proportional to the number of minor alleles they share. Faint dotted lines similar in colour connect shared minor allele nodes. Nodes are positioned in top-down chronological order based on when their minor alleles were observed and are positioned horizontally so that the edges between them are apparent. Inset. The number of edges connecting shared minor alleles that were consistent or inconsistent with the pattern of host-to-host contacts.
3. Results

(a) H3N8 equine influenza

There was a small but statistically significant subpopulation differentiation within the equine H3N8 sequences (table 1), apparent when alleles were subdivided both by horse and by viral population (horse $\times$ day), but not by day only. Given the major bifurcation in the transmission network on day 6 (figure 1), this latter result was expected.

Over the course of this experimental study, there was a single dominant allele (representing the consensus sequence) across all viral populations as well as many minor alleles (table 1). Most minor alleles were only observed once (‘singletons’), and differed from each other by one or more single-occurrence mutations (figure 2). Compared with multiple-occurrence mutations, single-occurrence mutations were on average more likely to be non-synonymous and occur more frequently in epitope regions. However, as this pattern could in part reflect PCR/sequencing errors, these singleton alleles were excluded from the analyses of transmission networks in both this and the swine H1N1 study.

Minor alleles observed more than once (‘non-singletons’) contained fewer non-synonymous mutations and, on average, deviated from the dominant allele by only one mutation (figure 2). The number of new alleles rose steadily through the study and only declined in one of the two transmission chains that resulted from the major bifurcation on day 6 (electronic supplementary material, figure S1). The pattern of these alleles over time suggests that little to no genetic variation is preserved at the sub-consensus level, with only a relatively small number of minor alleles being transmitted between hosts. Indeed, specific mutations are observed at most in four of the 11 horses. However, that minor alleles from different horses tended to be proximal in time suggests that they are transmitted between animals rather than generated de novo (electronic supplementary material, figure S1).

To isolate those alleles that might reflect transmitted genetic variation, we excluded all alleles except those that occurred in more than one horse. Shared minor alleles which were observed in two or more horses known to have been in direct, infectious contact (‘direct contact’, figure 2) were more likely to have occurred more frequently, occurred in more horses and more viral populations, and had more potential descendants than those that were only shared between horses never in direct contact, a pattern that is reflected in the importance score of these explanatory characteristics (table 2). That neither the presence of either epitope nor non-synonymous mutations (characteristics vii and viii) were ranked highly indicates that the pattern of connected alleles may be largely the result of stochastic population processes. Notably, the only shared minor allele with a stop codon mutation (characteristic ix) occurred between directly connected horses.

Figure 3a depicts the network of hosts connected using all shared alleles, regardless of their characteristics. Importantly, this estimated network contains edges that are both consistent and inconsistent with the known network of infectious contacts between hosts (figure 1a). To systematically remove inconsistent edges, we only used the shared alleles that were predicted to be consistent by the classification tree for the combined dataset. The combined classification tree (electronic supplementary material, figure S2) was used here (and below) to select shared minor alleles, and from which we could gauge the effectiveness of the same set of characteristics between two different IAV-host systems. In the combined classification tree, the highly ranked explanatory characteristics were overall allele frequency, the number of viral populations and animals it was observed in, the number of potential descendants, and whether or not the allele contained an epitope mutation (characteristics i–iv, vii). In addition to the shared alleles predicted to have been transmitted using the combined classification tree, the only allele with a stop codon mutation was also included. An approximation of the actual transmission network was then carried out using these shared alleles ($n = 6$, electronic supplementary material, table S4). This clarified the transmission network—depicted in figure 4a—by accurately removing most inconsistent edges, while preserving most consistent edges (see the insets in figures 3 and 4).

(b) H1N1 swine influenza

In this case, statistically significant differentiation between subpopulations was detected when viral sequences were subdivided by animal, by sampling day and by viral population (table 1). The significant differentiation by sampling day, not seen in the equine H3N8 study, was expected given the linear rather than bifurcating pattern of host contacts.

Despite relatively shallow intra-host viral sampling from fewer viral populations, the swine H1N1 data on average exhibited an order of magnitude greater sequence diversity than H3N8 in horses (table 1). In addition, the study-wide consensus sequence was not the dominant allele in every viral population, being briefly supplanted roughly halfway through the 2-week study. On average, singleton alleles again comprised both more mutations and higher frequencies of non-synonymous and epitope mutations (figure 2). When the non-singleton alleles were further subdivided based on whether they occurred exclusively in a single host or in multiple hosts, the two resultant groups displayed very different mutation profiles, in contrast to the equine H3N8 study. Non-singleton minor alleles with occurrences in a single host only on average contained more mutations, a greater frequency of non-synonymous mutations, and more mutations in putative epitope regions (figure 2). A similar dichotomy was observed when alleles with occurrences in two or more animals were further separated based on whether or not they were shared between animals with direct, infectious contact. Interestingly, the minor alleles shared between pigs with known contact had significantly higher frequencies of non-synonymous and epitope mutations compared with those that were shared between hosts never in direct contact.

The minor alleles observed in more than one animal were characterized as above. Minor alleles shared between hosts with direct contact were significantly different from those that were not, in that they occurred with greater overall frequency, were observed in more hosts and viral populations, and had more potential descendants, as indicated by their variable importance in the swine H1N1 classification tree (table 2, characteristics i–iv and vi). However, in contrast with the equine H3N8 study, strong associations were observed between the response variable and those allele characteristics that are more suggestive of the action of natural selection; nearly all minor alleles shared between directly connected hosts contained non-synonymous mutations, a higher proportion of which fell in putative epitope regions.
(table 2, characteristics vii and viii; electronic supplementary material, table S5). Finally, one shared minor allele, observed between directly connected hosts, contained a premature stop codon mutation.

Using all shared alleles to connect hosts resulted in a substantial degree of ambiguity regarding which pigs might be linked by transmission (figure 3b), with the estimated transmission network again containing edges that are both consistent and inconsistent with the known network of host-to-host contacts. Hence, predictions from the combined classification tree were again used to select only the subset of shared minor alleles that likely represent actual transmission events. Once again, the only minor allele with a stop codon mutation was included in this set. In this case, the approximated host transmission network constructed using only shared alleles which met at least one of these criteria (n = 5, electronic supplementary material, table S4) again provided a much clearer picture (figure 4b), retaining most edges consistent with the known network of host contacts and eliminating most that were not (see the insets in figures 3 and 4).

4. Discussion

Until recently, the application of phylodynamic methods to the study of IAV was necessarily focused on capturing host–pathogen dynamics at the epidemiological scale, where infections spread among different host populations over the course of months or years [4,7]. With finer-scale intra-host data—such as those used in this study—becoming more accessible, smaller-scale examinations of host-to-host transmission patterns have become possible. However, the evolutionary and population dynamics of viral sequence populations observed over shorter time spans (days or weeks) are not well understood. Using HA1 sequences from two different IAV subtypes sampled from two different host species, we demonstrate here that intra-host viral sequence data collected over relatively short evolutionary time spans from an acute infection do contain a discernible trace of their transmission history even when the population consensus sequence is identical.

A population bottleneck will routinely accompany the inter-host transmission of influenza viruses, such that only a fraction of the overall viral genetic material in an infected host is passed to a susceptible host [12,31]. However, in some systems, such as these two analysed here, it is clear that bottlenecks are broad enough to allow at least some minor genetic variation to be transmitted between hosts. This explains how minor alleles are sometimes shared between hosts and, critically, offers a way to link hosts by transmission using genetic evidence. An alternative explanation is that all minor alleles were generated de novo within each animal separately, although this has already been shown to be statistically unlikely for the equine H3N8 dataset [9]. Indeed, under random de novo generation, we would not expect to observe the statistically significant relationships between certain allele characteristics and known host-to-host contact as we do in both studies (table 2). Whether or not viral genetic sequence data collected over short time spans can be used effectively to reconstruct transmission networks thus becomes a question of how to identify transmitted versus de novo minor alleles, and whether the identified transmitted alleles can be translated into an accurate representation of the actual transmission network.

Notably, the evolutionary patterns observed in both the equine H3N8 and swine H1N1 datasets exhibited important similarities; neither set of intra-host viral sequences preserved any mutations throughout the study periods, the population consensus sequences comprised the majority of alleles in nearly all viral populations, and both datasets contained a large number of minor alleles some of which were observed in multiple viral populations. These minor alleles provided a means to infer specific host-to-host transmission events. There were also a number of common characteristics of the shared minor alleles from both datasets that appear to be predictive of known direct contact. In particular, these alleles occur frequently, are present in a large number of animals and viral populations, have a large number of possible descendants, frequently occur at epitope sites and result in stop codon mutations. The first four of these allele characteristics can be reasonably explained by genetic drift, where non-lethal mutants are passed a few times before being lost stochastically, or are generated de novo in different animals. In contrast, the presence of a stop codon mutation in directly connected hosts is most likely explained by complementation, as described in a variety of RNA viruses [36,37]. Somewhat paradoxically, it is therefore possible that defective viral mutations, such as premature stop codons, are good indicators of the precise patterns of inter-host virus transmission; these mutants seem to occur sporadically enough that they might be useful in teasing out certain host-to-host connections. In addition, although it is unclear whether the same allele characteristics would apply to other host–IAV systems, it is striking that the estimated transmission networks in both datasets connected a majority of hosts and come close to capturing the best possible network. Indeed, most edges in accord with the known network of host-to-host contacts were included in the estimated networks, and those that were not were largely excluded.

However, there are also dissimilarities in the minor allele characteristic associations that highlight which different evolutionary pressures are acting on the two IAV subtypes. In particular, the presence or the absence of non-synonymous mutations and mutations in putative epitope regions (table 2, characteristics vii and viii) were predictive of infectious host contact in the swine H1N1 study, but not in the equine H3N8 study. This, in turn, suggests that the selection pressures acting on H1N1 and H3N8 differ by host species. In addition, only a very small minority of shared minor alleles were non-randomly associated with the size of the viral populations they were observed in relative to those they were not, indicating that this disparity is not caused by inconsistent sampling of viral populations within each study. The dissimilarities in the minor allele characteristic associations could also be in part explained by structural differences in the HA1 proteins of equine H3N8 and swine H1N1 [38]. However, both H3 and H1 have relatively large host ranges, with both successfully infecting wild birds, pigs and humans [34], suggest that they have similar conformational flexibility. Accordingly, dissimilarities between the minor alleles of the equine and swine studies are most likely a reflection of the immunological and physiological differences between the two host species [10,39].

Despite the potential of the approach developed here, certain caveats need to be noted. In particular, there is a
temporal clustering in daily instances of directly connected minor alleles which seems to occur in earlier rather than later hosts. This point is quantitatively borne out for the swine H1N1 dataset by the statistically significant association between directly connected minor alleles and the first day that allele was observed ($p = 0.03$, one-tailed $t$-test of population means). This was noted by the authors of both original studies and explained as the possible consequence of the initial inocula comprising larger and more genetically complex viral populations than those that were naturally transmitted between subsequent hosts. This temporal bias could also reflect adaptive evolution following passage in avian egg cultures during the creation of each inoculation dose, and is clearly a topic that needs to be considered further.

In sum, we have provided an important step towards developing phylodynamic methods that will improve our understanding of the spread and evolution of influenza A virus at the level of individual hosts when population consensus sequences are identical. Additional work is therefore required to establish whether these findings can be generalized to other host–IAV systems, to hosts previously exposed to influenza virus, and to other viral infections.

This work is supported by NIH grant 2 R01 GM080533-06 to E.C.H. P.R.M. is a Wellcome Trust Veterinary Post-doctoral Fellow. J.L.N.W. is supported by the Alborada Trust. B.T.G. and J.L.N.W. were supported by the RAPIDD program of the Science and Technology Directorate, Department of Homeland Security and the Fogarty International Center, National Institutes of Health.

References


