nature genetics

Quantifying influenza virus diversity and transmission in humans

Leo L M Poon[1,14], Timothy Song[2,3,14], Roni Rosenfeld[4], Xudong Lin[5], Matthew B Rogers[2,13], Bin Zhou[3], Robert Sebra[6], Rebecca A Halpin[5], Yi Guan[1], Alan Twaddle[3], Jay V DePasse[7], Timothy B Stockwell[5], David E Wentworth[5,13], Edward C Holmes[8,9], Benjamin Greenbaum[10], Joseph S M Peiris[1], Benjamin J Cowling[11,15] & Elodie Ghedin[3,12,15]

Influenza A virus is characterized by high genetic diversity[1–3]. However, most of what is known about influenza evolution has come from consensus sequences sampled at the epidemiological scale[4] that only represent the dominant virus lineage within each infected host. Less is known about the extent of within-host virus diversity and what proportion of this diversity is transmitted between individuals[5]. To characterize virus variants that achieve sustainable transmission in new hosts, we examined within-host virus genetic diversity in household donor-recipient pairs from the first wave of the 2009 H1N1 pandemic when seasonal H3N2 was co-circulating. Although the same variants were found in multiple members of the community, the relative frequencies of variants fluctuated, with patterns of genetic variation more similar within than between households. We estimated the effective population size of influenza A virus across donor-recipient pairs to be approximately 100–200 contributing members, which enabled the transmission of multiple lineages, including antigenic variants.

We have previously shown that pandemic H1N1 and seasonal H3N2 viruses—both present during the first wave of the H1N1 pandemic in Hong Kong[6]—have similar transmission potential in household settings and that antigenic variants of H3N2 co-circulated with clades of H1N1/2009 (refs. 6,7). In other parts of the world and during the same time period, the unseasonal transmission of H3N2 was observed along with the transmission of pandemic H1N1 virus[8]. To characterize patterns of viral evolution at a finer scale and, in particular, the extent of virus genetic diversity that was transmitted among hosts, we performed whole-genome deep sequencing on nasopharyngeal swabs collected from index cases with confirmed influenza along with their household contacts. Notably, the household epidemiological information enabled us to assign donor-recipient pairs in suspected transmission events with relatively high confidence, to compare these with unrelated pairs and to estimate spatiotemporal transmission chains.

The virus sample set was collected in July and August of 2009 from 84 individuals (67 index patients and 17 other household members) living in Hong Kong; 16 patients were sampled twice, 2–4 d apart. We estimated within-host virus diversity for each sample by mapping polymorphic sites onto the consensus genome assemblies to generate a list of single-nucleotide variants (SNVs, or minor variants) present at a frequency of at least 3%. Within-host diversity was measured by Shannon entropy, H, assuming site independence. Mean within-host diversity was significantly higher (Wilcoxon rank-sum test, $P = 1.89 \times 10^{-12}$) for H3N2 ($H = 33$) than for H1N1/2009 ($H = 13$) viruses. There was no significant Pearson correlation between high within-host virus diversity and high viral titer[7] ($r = -0.3$ for H1N1 and $r = -0.16$ for H3N2) for most of the genes, with the exception of PA and M for H1N1/2009 (**Supplementary Table 1**).

Phylogenetic analysis clustered whole-genome consensus sequences by household for each group of patients diagnosed as infected with either H3N2 (**Fig. 1**) or H1N1/2009 (**Supplementary Fig. 1**). Comparisons of the phylogenetic trees from each gene showed no evidence for reassortment within this population during the time-frame of the study (data not shown). Three antigenic sublineages of H3N2 (A/Brisbane/10/2007-like, A/Victoria/208/2009-like and A/Perth/16/2009-like) and three clades of H1N1/2009 (clades 3, 6 and 7) circulated in this population[6]. Despite the relatively small population size in the study, one case of mixed-subtype infection was

[1]Public Health Laboratory Sciences, School of Public Health, The University of Hong Kong, Hong Kong, China. [2]Department of Computational and Systems Biology, University of Pittsburgh School of Medicine, Pittsburgh, Pennsylvania, USA. [3]Center for Genomics and Systems Biology, Department of Biology, New York University, New York, New York, USA. [4]School of Computer Sciences, Carnegie Mellon University, Pittsburgh, Pennsylvania, USA. [5]J. Craig Venter Institute, Rockville, Maryland, USA. [6]Department of Genetics and Genomic Sciences, Icahn School of Medicine at Mount Sinai, New York, New York, USA. [7]Pittsburgh Supercomputer Center, Carnegie Mellon University, Pittsburgh, Pennsylvania, USA. [8]Marie Bashir Institute for Infectious Diseases and Biosecurity, Charles Perkins Centre, School of Biological Sciences, The University of Sydney, Sydney, New South Wales, Australia. [9]Sydney Medical School, The University of Sydney, Sydney, New South Wales, Australia. [10]Tisch Cancer Institute, Departments of Medicine, Hematology and Medical Oncology, and Pathology, Icahn School of Medicine at Mount Sinai, New York, New York, USA. [11]Epidemiology and Biostatistics, School of Public Health, The University of Hong Kong, Hong Kong, China. [12]College of Global Public Health, New York University, New York, New York, USA. [13]Present addresses: Children's Hospital of Pittsburgh, University of Pittsburgh School of Medicine, Pittsburgh, Pennsylvania, USA (M.B.R.) and Influenza Division, Centers for Disease Control and Prevention, Atlanta, Georgia, USA (D.E.W.). [14]These authors contributed equally to this work. [15]These authors jointly supervised this work. Correspondence should be addressed to E.G. (elodie.ghedin@nyu.edu) or B.J.C. (bcowling@hku.hk).

Received 18 August 2015; accepted 7 December 2015; published online 4 January 2016; doi:10.1038/ng.3479

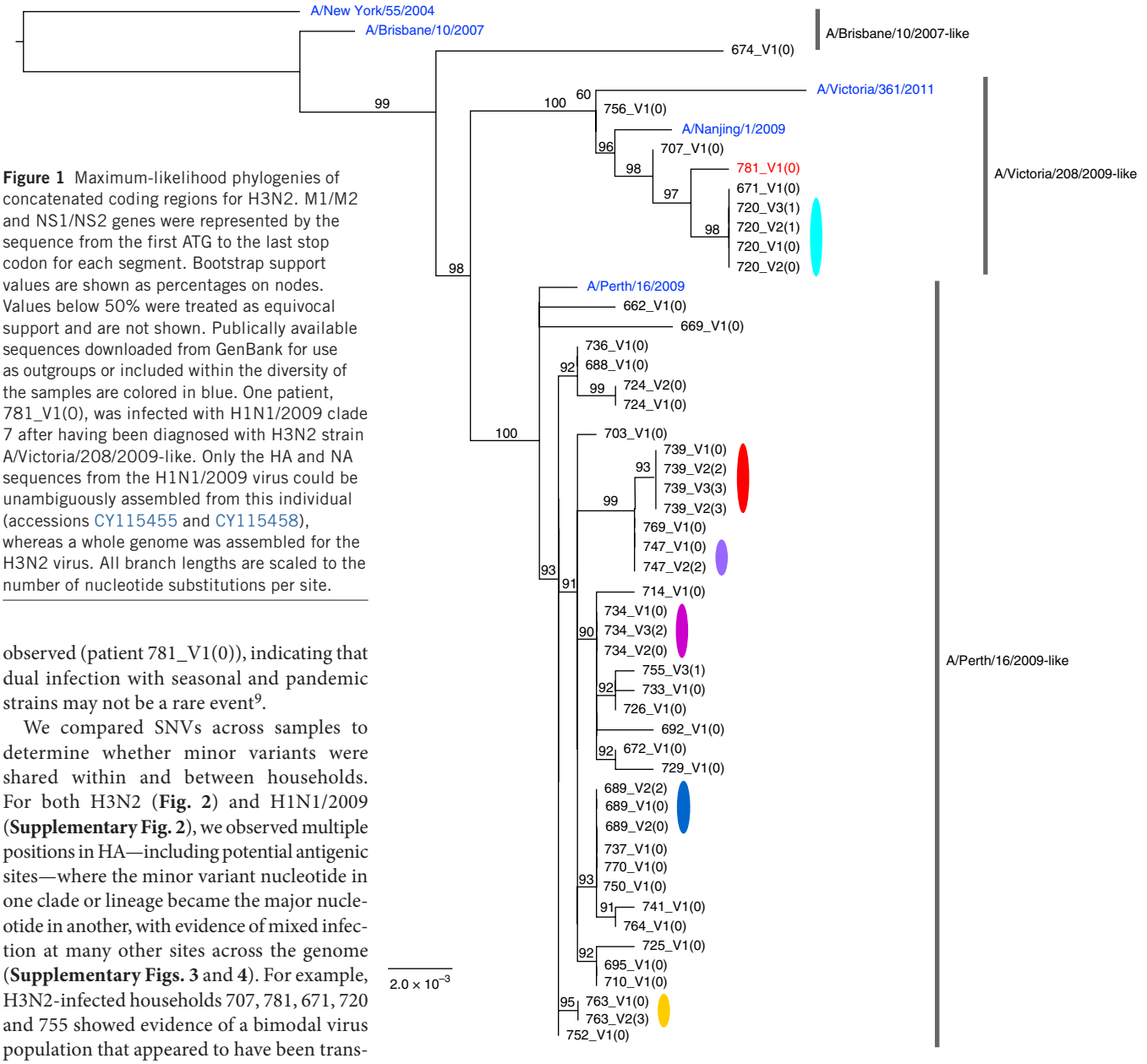

Figure 1 Maximum-likelihood phylogenies of concatenated coding regions for H3N2. M1/M2 and NS1/NS2 genes were represented by the sequence from the first ATG to the last stop codon for each segment. Bootstrap support values are shown as percentages on nodes. Values below 50% were treated as equivocal support and are not shown. Publically available sequences downloaded from GenBank for use as outgroups or included within the diversity of the samples are colored in blue. One patient, 781_V1(0), was infected with H1N1/2009 clade 7 after having been diagnosed with H3N2 strain A/Victoria/208/2009-like. Only the HA and NA sequences from the H1N1/2009 virus could be unambiguously assembled from this individual (accessions CY115455 and CY115458), whereas a whole genome was assembled for the H3N2 virus. All branch lengths are scaled to the number of nucleotide substitutions per site.

observed (patient 781_V1(0)), indicating that dual infection with seasonal and pandemic strains may not be a rare event[9].

We compared SNVs across samples to determine whether minor variants were shared within and between households. For both H3N2 (**Fig. 2**) and H1N1/2009 (**Supplementary Fig. 2**), we observed multiple positions in HA—including potential antigenic sites—where the minor variant nucleotide in one clade or lineage became the major nucleotide in another, with evidence of mixed infection at many other sites across the genome (**Supplementary Figs. 3** and **4**). For example, H3N2-infected households 707, 781, 671, 720 and 755 showed evidence of a bimodal virus population that appeared to have been transmitted intact in multiple transmission events. Overall, we tentatively estimate that approximately 66% of the H3N2-infected patients and 40% of the H1N1/2009-infected patients likely harbored mixed-lineage infections (**Supplementary Table 2**). To confirm these findings from the clinical specimens, we phased the SNVs into haplotypes by single-molecule sequencing for 12 of the cell culture samples from six different households (**Fig. 2**, **Supplementary Fig. 2** and **Supplementary Tables 3–8**). Notably, although the dominant haplotype indicated that each sample belonged to one major lineage, patients often carried a minor haplotype that resembled a separate lineage. This suggests that a number of the SNVs are not only *de novo* mutations that occurred in the index patient from a household but are also shared across the community as a whole. We saw a similar sharing of variant nucleotides when looking at global consensus sequences across seasons. Using HA consensus sequence data available in GenBank and human 2008 H3 sequences as a reference, we observed an increase in the frequency of specific nucleotides at some variant positions in subsequent seasons of H3N2 epidemics (**Supplementary Fig. 5**). This phenomenon was more pronounced for variants from the A/Victoria/208/2009-like lineage, in marked contrast to the reduced frequency of variant nucleotides observed for the A/Perth/16/2009-like lineage. No such increase was observed in pandemic H1N1 after the 2009 season. Additionally, frequency variations in H1N1/2009 were far less common than in H3N2. It is important to note that the A/Victoria/208/2009-like virus has replaced the A/Perth/16/2009-like virus as the dominant lineage in recent years, leading in 2012 to a change of vaccine strain from A/Perth/16/2009-like virus to A/Victoria/361/2011-like virus (a phylogenetic subgroup of A/Victoria/208/2009). In contrast, pandemic H1N1 virus is antigenically stable, and there has been no change of vaccine strain since its introduction in humans in 2009. Overall, these data indicate that some viral lineages can be transmitted between individuals below current surveillance thresholds.

Because each virus sample collected will contain *de novo* mutations and potentially represent a mixed infection, we determined the similarity of the viral populations across the data set. To this end, we calculated the genetic distance between samples by performing an all-versus-all

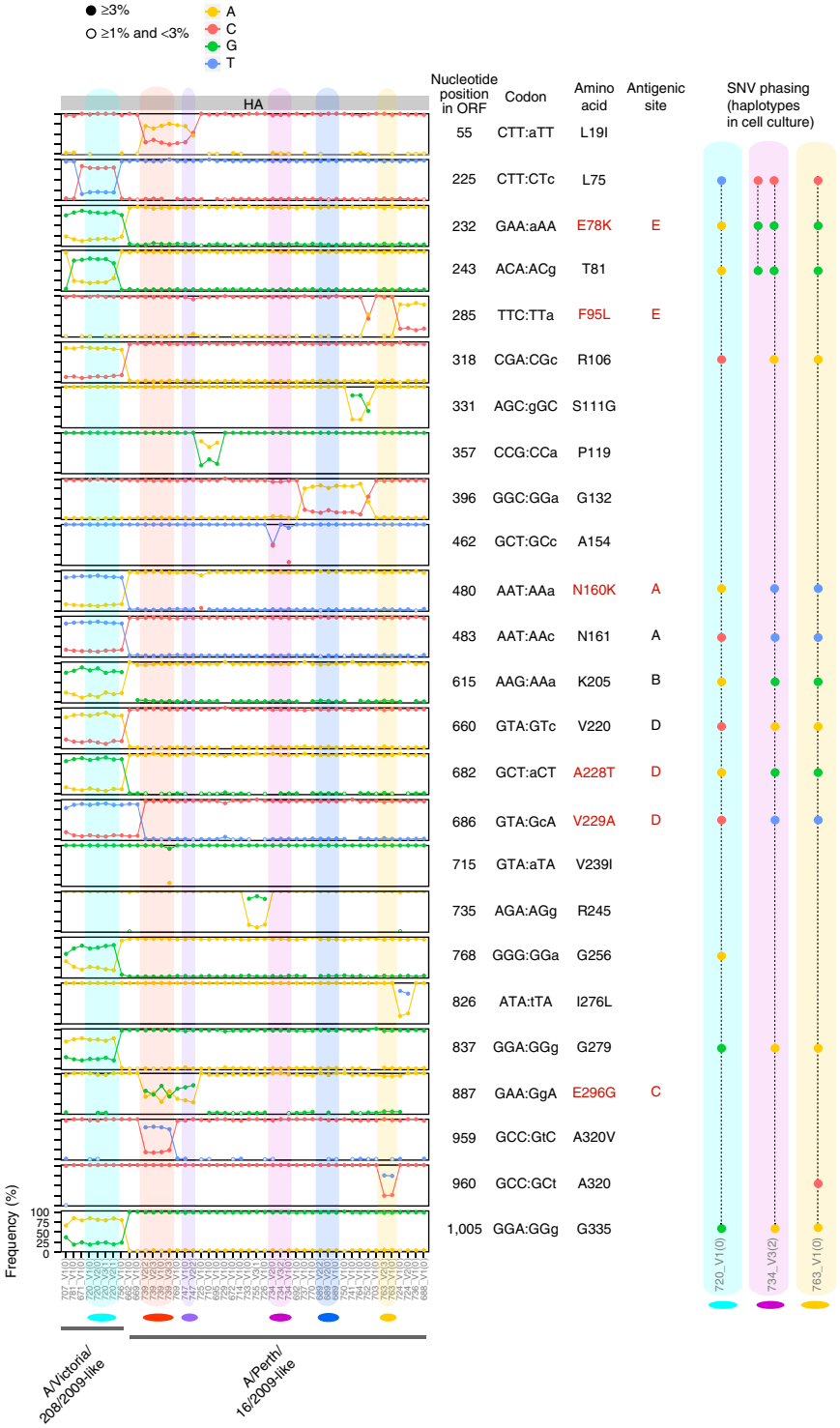

Figure 2 Comparison of HA minor variant frequencies across households. Only polymorphic sites located in the HA1 domain are represented. Amino acid positions are numbered according to the first methionine (start codon) of the protein (and not according to the HA1 numbering schema). Site information for all segments is available in **Supplementary Figure 3**. The *x* axis lists samples by position on the phylogenetic trees in **Figure 1**; households with more than one member are shown on a colored background. The *y* axis displays nucleotide frequencies, with graph lines corresponding to 0%, 25%, 75% and 100% frequency. Antigenic sites correspond to previously identified antigenic sites[16–19]. Text in red highlights nonsynonymous alterations located in antigenic sites. Filled circles represent minor variants found at a frequency of 3% or higher, whereas open circles correspond to frequencies equal to or greater than 1% but below 3%. Lowercase letters in the codon column highlight the minor variant sites. SNV phasing shows how minor variant nucleotides are phased on the same molecules, representing haplotypes. These were determined from single-molecule sequencing of viruses in cell culture for three household pairs: 720_V1(0) and 720_V2(1) (**Supplementary Table 6**), 734_V1(0) and 734_V3(2) (**Supplementary Table 7**), and 763_V1(0) and 763_V2(3) (**Supplementary Table 8**).

pairwise comparison for each variant nucleotide position using the sum of the absolute value of distances between nucleotide frequencies, that is, L1-norm (Online Methods). We grouped pairwise comparisons into longitudinal pairs (from the same individual sampled at two different visits), within-household pairs and across-household pairs (**Fig. 3**). We determined that the median L1 genetic distances for within-household or longitudinal pairs were significantly shorter than those for any random pairing (Student's *t* test, $P << 0.01$). This indicates that minor variants and their proportions can be used to infer transmission between hosts, even if a number of these correspond to variants from coinfecting viruses that are shared by individuals across households. Interestingly, for H1N1/2009, we saw a number of within-household pairs that were outliers (**Fig. 3**, dashed circle), providing further evidence of mixed infection. For example, variants present at a minor frequency in most of the samples from household 751 had become dominant in the sample from the second visit for the index case (751_V2(0)) (**Supplementary Figs. 2** and **4**). Although random sampling effects will influence mutational frequencies, such a profound increase in frequency is compatible with a selective advantage in this patient.

After excluding outliers and considering only a single sample (from the first visit) for each individual, there were 21 viable within-household transmission pairs. To select other potential epidemic links within the community, we used the transmission and longitudinal pairs to identify outliers and determine a threshold of maximum genetic distance for a transmission event (after excluding outliers) (**Fig. 3**). Each pair was epidemiologically linked to a short transmission chain. Using consensus sequences, we first inferred transmission networks across the population using a parsimony- and graph-based algorithm[10,11]. We then used minor variant data to highlight potential localized outbreaks (**Fig. 4**) with cross-region links (between Hong Kong Island, Kowloon and New Territories). This network agrees with the fact that there is a high volume of population flow within Hong Kong each day, allowing ample opportunity for influenza transmission across regions.

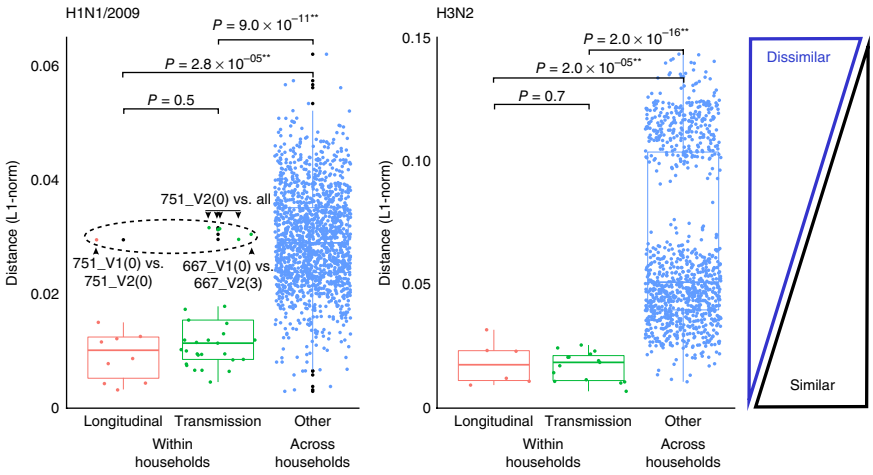

Figure 3 Box plots of L1-norm pairwise genetic distance within and across households. We used the L1-norm values obtained from the variant nucleotide analysis across all genes to compare the overall genetic distance of longitudinal pairs and transmission pairs to every other possible sample pair combination. Each dot represents the genetic distance between a unique pair. The longitudinal pairs are represented by 16 individuals in 12 households who were sampled at two different time points, 2–3 d apart. The transmission pairs are from 13 households where at least two members were sampled; there is a total of 22 predicted donor-recipient pairs within households and 22 more when including more than one time point per individual. The box plots show the median of the distances; the bottom and top of each box represent the first and third quartiles. The lengths of the whiskers extend to 1.5 times the interquartile range. Outliers are represented by black dots. The dashed black circle in the H1N1/2009 plot marks the outliers. One of the H1N1/2009 pairs—household 751, index case (0), visit 1 and visit 2: 751_V1(0) and 751_V2(0)—had a pairwise genetic distance that was above the expected threshold (H1N1/2009, longitudinal). When each of these samples was then used in within-household pairwise comparisons (H1N1/2009, transmission), the sample from the second visit clearly appeared to be an outlier. The pairwise genetic distance between the index case in household 667 (667_V1(0)) and the other member of this household (667_V2(3)) also appeared to be an outlier. Double asterisks indicate statistical significance.

To further explore shared virus populations within households, we compared minor variants at each position in donor (index case) and recipient transmission pair samples. Most variants found in the donor were shared by the potential recipient (**Fig. 5**, colored dots). The frequency of shared variants was much lower in pairs of unrelated samples (**Fig. 5**, black dots), although we found more shared variants in H3N2

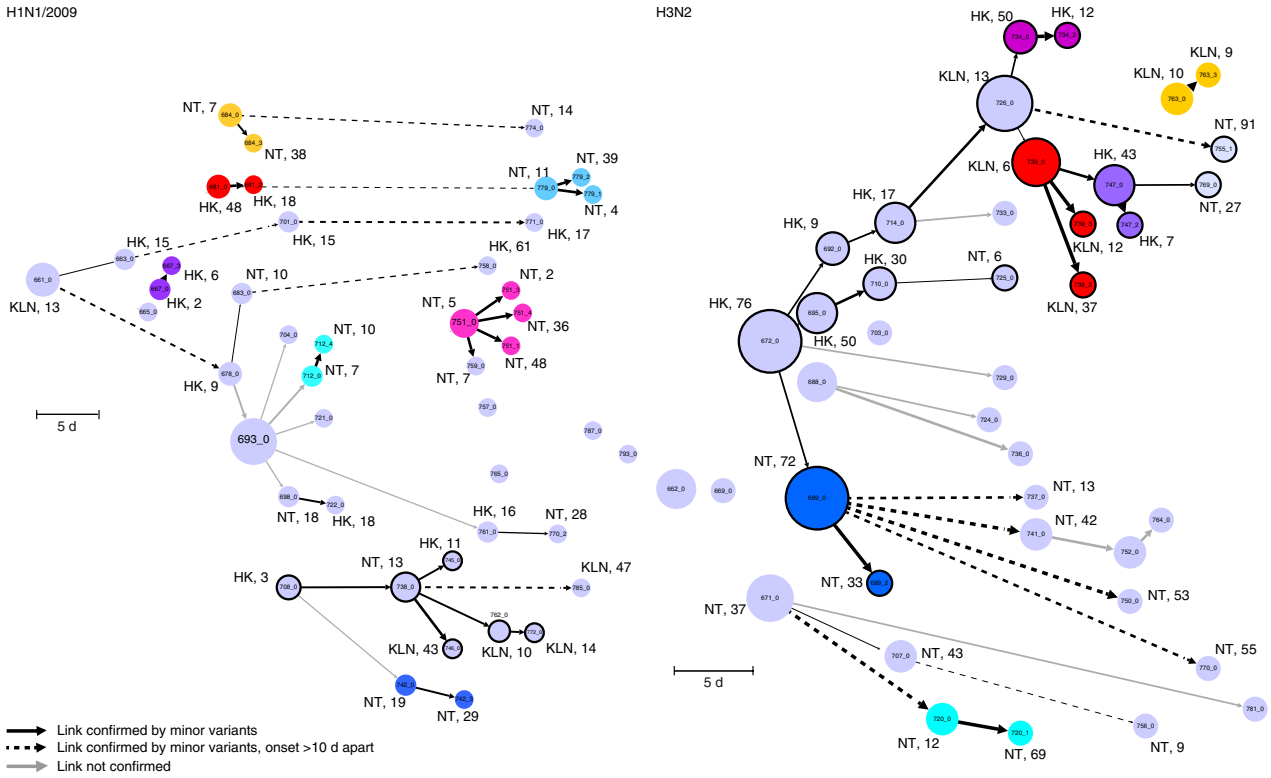

Figure 4 Reconstruction of potential transmission pathways of H1N1/2009 and H3N2 outbreaks. Transmission networks are inferred from consensus whole-genome sequences and date of onset. Each sample is a node on the graph, and the directed edges indicate putative ancestries and transmissions. Time is represented on the *x* axis and indicates the number of days since the first date of onset. A unique color is assigned to each household with more than one member sampled. The size of each node is determined by the number of outgoing edges, known as the out degree. A dashed line indicates a putative transmission link between samples with dates of onset separated by >10 d. The weight of each edge is inversely proportional to the number of nucleotide differences between the two samples connected (the thicker the edge, the smaller the number of differences). Nucleotide differences were separated into quartiles. H1N1/2009: 0–2 nt, 3–6 nt, 7–15 nt and 16–28 nt. H3N2: 0–5 nt, 6–9 nt, 10–19 nt and 20–45 nt. Links were confirmed by genetic distances (L1-norm method) and normalized for edge weights. Circles with thick black edges are nodes within a chain of transmission with more than two individuals. Locality and age of the patient (in years) are indicated for a number of the nodes. HK, Hong Kong; NT, New Territories; KLN, Kowloon.

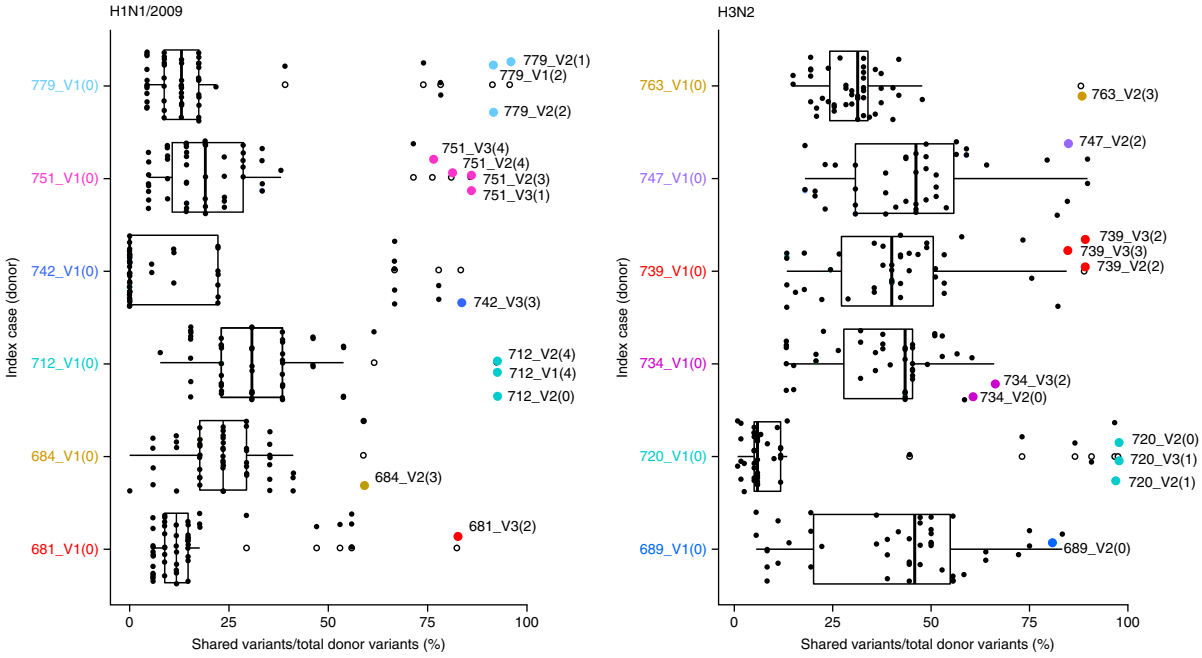

Figure 5 Box plots comparing frequencies for shared variants within and across households. We compared variant frequencies for shared variants in index cases and other members of the same household (colored dots) or any other sample (black dots). Boxes indicate interquartile range, and unfilled circles represent outliers. Household members tend to share most of the variants found in the index case. The index case from each H1N1/2009 household is compared to 54 other samples; the index case from each H3N2 household is compared to 46 other samples.

than in H1N1/2009 pairs. We observed that the relative frequency of variants in the recipient was more often similar to that found in the donor than to that found in any other individual (Wilcoxon signed-rank test, $P < 0.05$), which implies the lack of a substantial genetic bottleneck at transmission. This in turn suggests that shared variants found in the recipient are not the result of *de novo* mutations but are more likely present in viruses that transmit between hosts and replicate.

From the household transmission pairs, we estimated the probability that multiple variants are transmitted between hosts. In particular, polymorphic sites with variants only detected in the donor and ones with variants detected in both the donor and recipient were selected to determine the probability of transmission as a function of variant frequency. Accordingly, for H1N1/2009, a donor variant found at a frequency of 10% has a 64% chance of being transmitted to the recipient; for H3N2, a donor variant at a frequency of 10% has an 86% chance of transmission (**Fig. 6**). Because of the limited sample size, it was not possible to determine with confidence the probability of transmission for variants present at frequencies below 10%.

To infer the sizes of the virus populations before and after transmission that are able to generate productive progeny, we estimated the effective population size, N_e, by modifying a version of the Wright-Fisher idealized population model for our data. Specifically, for donor-recipient pairs, we took the frequency of the shared minor variants, p, and the frequency of the major nucleotide at that position, q; we then calculated the variance of the difference in donor-recipient frequencies to obtain a variance effective size. Through these calculations, we obtained a mean of 192 viral particles (median = 124, mean s.d. range = 114–276) for H1N1/2009 and a mean of 248 viral particles (median = 138, mean s.d. range = 47–457) for H3N2. To confirm the scale of our estimates, we used a different method based on the Kullback-Leibler divergence (KLD)[12] (Online Methods). This method gave a mean of 90 viral particles (median = 80, mean s.d. = 55) for H1N1/2009 and a mean of 114 viral particles (median = 121, mean s.d. = 55) for H3N2. To estimate how many haplotypes would be present within these replicating populations, we used the phased SNV and reconstructed haplotype data and observed an average of three haplotypes for H1N1/2009 and five haplotypes for H3N2 transmitted across donor-recipient pairs (**Supplementary Tables 3–8**). The sample size was too small for the difference between H1N1/2009 and H3N2 to be significant. It is, however, theoretically possible that H3N2 has a higher effective population size than H1N1/2009 because

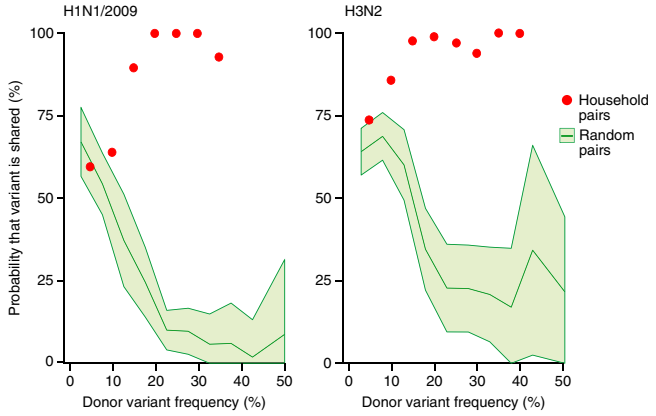

Figure 6 Probability of variant transmission as a function of the relative frequency of the minor variant. Variants that were only detected in the donor and ones that were shared by the donor and recipient were used in determining the probability of transmission. Household pairs (red dots) are comparisons between members of the same household. Each data point represents the proportion of shared variants over the total number of variants found in a window size of 10%. Random pairs (green shaded area) are 30 random donor-recipient pairs resampled 100 times to obtain an estimate of standard deviation.

the virus has been circulating in the human population since 1968, such that there is greater background genetic diversity and hence a greater diversity of lineages that can be transmitted among hosts. Crucially, these effective population size and haplotype estimates suggest that multiple variants can routinely be transmitted between individuals, such that any transmission bottlenecks are fairly loose and a relatively small number of viral particles can initiate a productive infection with a number of variant strains that are co-transmitted.

In sum, we have analyzed minor variant dynamics in the transmission of influenza A virus within and across households during an epidemic and used this information to determine potential transmission events. The information on minor variants shared by donors and recipients in transmission pairs was then used to estimate the number of viral particles that are able to infect and replicate in the recipient. The approach taken here could help define how prior immunity or other host factors, as well as virus subtype and strain, may affect transmission dose, of which our estimates of effective population size likely capture the lower bounds. Indeed, this analysis showed the transmission of multiple variants, both from mixed infections and within-host *de novo* haplotypes, indicating a relatively loose transmission bottleneck. Notably, the data for shared variants also suggest that there was a single coinfection or superinfection event involving two genetically distinct viruses during this epidemic, with the bimodal virus population then being transmitted intact in multiple subsequent transmission events. This finding is unsurprising in light of recent observations that natural selection can act on pools of virus variants linked by their colocalization in the same cell[13]. In addition, this demonstrates that there are likely more cases of mixed lineages within infected patients than can be captured with standard consensus-based diagnostic assays. Such coinfections will facilitate the occurrence of reassortment and may help explain the frequent detection of reassortants between seasonal H3 viruses[14]. Although similar observations have been made in animal studies[11,15], to our knowledge, this is the first demonstration for influenza A virus in humans. Characterizing the genetic information of transmitted virions allows a better understanding of influenza virus transmission in humans and provides more accurate information for modeling epidemics and for disease control strategies.

URLs. DNA Barcode Deconvolution, http://sourceforge.net/projects/deconvolver; Elvira, http://sourceforge.net/projects/elvira/.

METHODS

Methods and any associated references are available in the online version of the paper.

Accession codes. Sequence data have been deposited in the NCBI Nucleotide and Sequence Read Archive (SRA) databases. The accession numbers for the HA and NA genes are listed in the phylogenetic trees. The Illumina raw sequence reads appear in SRA as BioSamples SAMN01095441–SAMN01095495 for H1N1/2009 and SAMN01095144–SAMN01095190 for H3N2. The PacBio raw sequence reads for the 12 viral isolates appear in SRA under experiment accessions SRX1117304, SRX1117319, SRX1117320, SRX1117563–SRX1117566 and SRX1117568–SRX1117572.

Note: Any Supplementary Information and Source Data files are available in the online version of the paper.

ACKNOWLEDGMENTS

T.S. was a predoctoral trainee supported by US National Institutes of Health T32 training grant T32 EB009403 as part of the HHMI-NIBIB Interfaces Initiative. This research was supported with a grant from the Research Grants Council of the Hong Kong Special Administrative Region, China (project T11-705/14N) (L.L.M.P., Y.G., J.S.M.P. and B.J.C.), federal funds from the National Institute of Allergy and Infectious Diseases, US National Institutes of Health, US Department of Health and Human Services, under contract numbers HHS-N272201400006C (L.L.M.P., Y.G. and J.S.M.P.), HHS-N266200700005C (B.J.C.) and HHS-N272200900007C (E.G., X.L., R.A.H., T.B.S. and D.E.W.), the National Institute of General Medical Science, US National Institutes of Health, under award numbers U54 GM088491 (E.G., R.R. and J.V.D.) and U54 GM088558 (B.J.C.), and National Health and Medical Research Council of Australia Fellowship AF30 (E.C.H.). The data for this manuscript were generated and prepared while D.E.W. was employed at the J. Craig Venter Institute. The opinions expressed in this article are the authors' own and do not reflect the views of the Centers for Disease Control and Prevention, the US Department of Health and Human Services or the US government.

AUTHOR CONTRIBUTIONS

All the authors read and approved the manuscript. L.L.M.P. and E.G. conceived and designed the experiments, supervised research, performed analyses and wrote the manuscript. T.S. analyzed the deep sequence data, performed the variant codon and clustering analyses, and wrote the manuscript. B.G. and R.R. supervised research on the inoculum size estimates and wrote the manuscript. X.L., R.A.H., D.E.W., B.Z. and R.S. performed the sample preparation and sequencing. T.B.S., A.T. and J.V.D. performed the bioinformatic analyses. M.B.R. performed phylogenetic analyses. E.C.H. performed phylogenetic analyses and wrote the manuscript. Y.G. and J.S.M.P. conceived and designed the experiments. B.J.C. conceived and designed the experiments and supervised research.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

1. Bush, R.M., Fitch, W.M., Bender, C.A. & Cox, N.J. Positive selection on the H3 hemagglutinin gene of human influenza virus A. *Mol. Biol. Evol.* **16**, 1457–1465 (1999).
2. Drake, J.W. Rates of spontaneous mutation among RNA viruses. *Proc. Natl. Acad. Sci. USA* **90**, 4171–4175 (1993).
3. Drake, J.W. & Holland, J.J. Mutation rates among RNA viruses. *Proc. Natl. Acad. Sci. USA* **96**, 13910–13913 (1999).
4. Viboud, C., Nelson, M.I., Tan, Y. & Holmes, E.C. Contrasting the epidemiological and evolutionary dynamics of influenza spatial transmission. *Phil. Trans. R. Soc. Lond. B* **368**, 20120199 (2013).
5. Fordyce, S.L. *et al.* Genetic diversity among pandemic 2009 influenza viruses isolated from a transmission chain. *Virol. J.* **10**, 116 (2013).
6. Poon, L.L. *et al.* Viral genetic sequence variations in pandemic H1N1/2009 and seasonal H3N2 influenza viruses within an individual, a household and a community. *J. Clin. Virol.* **52**, 146–150 (2011).
7. Cowling, B.J. *et al.* Comparative epidemiology of pandemic and seasonal influenza A in households. *N. Engl. J. Med.* **362**, 2175–2184 (2010).
8. Ghedin, E. *et al.* Unseasonal transmission of H3N2 influenza A virus during the swine-origin H1N1 pandemic. *J. Virol.* **84**, 5715–5718 (2010).
9. Lee, N., Chan, P.K., Lam, W.Y., Szeto, C.C. & Hui, D.S. Co-infection with pandemic H1N1 and seasonal H3N2 influenza viruses. *Ann. Intern. Med.* **152**, 618–619 (2010).
10. Jombart, T., Eggo, R.M., Dodd, P.J. & Balloux, F. Reconstructing disease outbreaks from genetic data: a graph approach. *Heredity (Edinb.)* **106**, 383–390 (2011).
11. Hughes, J. *et al.* Transmission of equine influenza virus during an outbreak is characterized by frequent mixed infections and loose transmission bottlenecks. *PLoS Pathog.* **8**, e1003081 (2012).
12. Emmett, K.J., Lee, A., Khiabanian, H. & Rabadan, R. High-resolution genomic surveillance of 2014 ebolavirus using shared subclonal variants. *PLoS Curr.* http://dx.doi.org/10.1371/currents.outbreaks.c7fd7946ba606c982668a96bcba43c90 (9 February 2015).
13. Combe, M., Garijo, R., Geller, R., Cuevas, J.M. & Sanjuán, R. Single-cell analysis of RNA virus infection identifies multiple genetically diverse viral genomes within single infectious units. *Cell Host Microbe* **18**, 424–432 (2015).
14. Westgeest, K.B. *et al.* Genomewide analysis of reassortment and evolution of human influenza A(H3N2) viruses circulating between 1968 and 2011. *J. Virol.* **88**, 2844–2857 (2014).
15. Varble, A. *et al.* Influenza A virus transmission bottlenecks are defined by infection route and recipient host. *Cell Host Microbe* **16**, 691–700 (2014).
16. Xu, R. *et al.* Structural basis of preexisting immunity to the 2009 H1N1 pandemic influenza virus. *Science* **328**, 357–360 (2010).
17. Kitikoon, P. *et al.* Pathogenicity and transmission in pigs of the novel A(H3N2)v influenza virus isolated from humans and characterization of swine H3N2 viruses isolated in 2010–2011. *J. Virol.* **86**, 6804–6814 (2012).
18. Tharakaraman, K. *et al.* Antigenically intact hemagglutinin in circulating avian and swine influenza viruses and potential for H3N2 pandemic. *Sci. Rep.* **3**, 1822 (2013).
19. Cong, Y. *et al.* Reassortant between human-like H3N2 and avian H5 subtype influenza A viruses in pigs: a potential public health risk. *PLoS One* **5**, e12591 (2010).

ONLINE METHODS

Sample collection. Retrospective pooled specimens of nasal and throat swabs studied in our previous household influenza transmission investigations[6,7] were subjected to next-generation sequencing by HiSeq 2000 (Illumina). This data set comprises 102 virus samples (55 H1N1/2009 and 47 H3N2) collected from 84 individuals in Hong Kong over July and August of 2009. There were multiple home visits, and 16 individuals were sampled twice on two or three household visits (V1, visit 1; V2, visit 2; V3, visit 3), 2–4 d apart.

Sample preparation and sequencing. Multi-segment RT-PCR (M-RT-PCR)[20] was used to amplify influenza-specific segments from total RNA, followed by sequence-independent, single-primer amplification (SISPA)[21]. Each RNA sample was subjected to two rounds of M-RT-PCR, and the products in turn were amplified by SISPA using different barcodes to control for barcode-specific amplification bias; these technical replicates were then pooled separately for 100-bp paired-end sequencing on different lanes of a HiSeq 2000 sequencer. Potential SISPA PCR duplicate sequence reads were removed with the Elvira package. SISPA barcoded reads were demultiplexed with the DNA Barcode Deconvolution software, and the demultiplexed reads were trimmed of M-RT-PCR primer sequences and low-quality regions. Sequence reads were then *de novo* assembled using CLC Bio's clc_novo_assemble program (Qiagen), and the resulting contigs were used to identify influenza virus reference segment sequences by performing BLASTN searches against complete influenza virus segments available from GenBank. CLC Bio's clc_ref_assemble_long software (version 3.22.55705) was then used to map trimmed reads to the segments of the reference genome.

Phylogenetic analyses. All eight influenza A coding sequences were concatenated into an alignment of 13,425 nucleotides for H3N2 and 13,392 nucleotides for H1N1/2009. Coding sequences were concatenated in the order of the segment number on which they were located (PB2-PB1-PA-HA-NP-NA-M1-M2-NS1-NS2). All isolates were included except for 781_V1(0), which appeared to be a mix of H3N2 and H1N1/2009, with genes related to both the H1N1/2009 and H3N2 strains. Other taxa not included in this study were used as outgroup taxa (A/California/04/2009 and A/New York/55/2004 for H1N1/2009 and H3N2, respectively). These were selected on the basis of their position in widely sampled single-gene phylogenies (data not shown). Two additional taxa—A/Brisbane/10/2007 and A/Nanjing/1/2009—were included in the H3N2 phylogeny to capture the full diversity of this part of the H3N2 tree. Maximum-likelihood phylogenies were generated with RAxML[22] using the GTR nucleotide substitution model, with among-site rate variation modeled using a discrete gamma distribution with four rate categories. Bootstrap support values were generated using 1,000 fast bootstrap replicates and represented as percentages on nodes (values below 50% are not shown).

Variant analysis. Minor variants were identified using the Elvira package, which applies statistical tests to minimize false positive SNV calls that can be caused by sequence-specific errors (SSEs) that may occur on Illumina platforms[23]. This involves observing the forward and reverse reads of an SNV call. On the basis of a binomial distribution cumulative probability, we calculate the *P* values. If both *P* values are within a Bonferroni-corrected significance level ($\alpha = 0.05$), the SNV call is accepted. A minimum minor allele frequency of 3% was used as the threshold with a minimum coverage of 200 reads for a given site (see **Supplementary Table 9** for the average coverage for each sample). This conservative cutoff was selected on the basis of the same control sample that was sequenced in two different sequence runs with examination of concordance (SNV found in both samples) and discordance (SNV found in only one of the two samples) for different frequency thresholds. At 3%, 16 of 17 sites were concordant, whereas at 4% 14 of 14 sites were concordant. We chose the lower cutoff to gain more information, even if the error was higher. As a comparison, at 1%, only 32 of 62 sites were concordant and, at 2%, 16 of 26 sites were concordant.

Quantification of within-host diversity. We used Shannon entropy to quantify the within-host diversity of each sample through the relative frequencies of each SNV using the short-read (Illumina) data. This was done across all segments and assumes that all SNVs are independent of each other. We find that the entropy scores between H1N1/2009 and H3N2 are significantly different from each other ($P = 1.27 \times 10^{-6}$)

$$H(x) = \sum_{i}^{n} P(i) \log_2 P(i)$$

where $P(i)$ is the relative frequency of a variant at position i.

Genetic distance across samples. The genetic distance between samples was estimated using three different methods: L1-norm, L2-norm and the Jensen-Shannon divergence (JSD) measure. For the L1-norm method, we compare each sample against every other sample (all-versus-all pairwise comparison) at each variant nucleotide position

$$d_k(p,q) = \sum_{i=1}^{n} |p_i - q_i|$$

Here d_k is the distance measured at nucleotide position k between two samples, n is the total number of possible nucleotide configurations (A, C, G, T), and p and q are vectors containing the relative frequencies of the different variant nucleotides observed (these are analogous to 'alleles'). For a pair of samples, we observe a nucleotide position of a coding sequence (d_k) and then sum over all positions to obtain D, the distance measured between two samples for a specific coding sequence; N is the length of the coding sequence.

$$D = \sum_{k=1}^{N} d_k$$

This results in a single number that informs us of the distance (or dissimilarity) between two samples for each of the coding sequences. This procedure was repeated across all segments.

We verified our analysis by comparing against two other distance measures. The L2-norm method uses Euclidean distance and follows a similar procedure to the L1-norm method with d_k computed as

$$d_k(p,q) = \sqrt{\sum_{i=1}^{n} (p_i - q_i)^2}$$

D is similarly calculated by summing over all values of d_k.

For the third method, the JSD approach modifies the Kullback-Leibler divergence so that the resulting output is symmetric and will always have a finite value.

$$D_{\mathrm{KL}}(P \,||\, Q) = \sum_{i} \ln\left(\frac{P(i)}{Q(i)}\right) P(i)$$

The JSD is calculated by

$$D_{\mathrm{JSD}}(P \,||\, Q) = \frac{1}{2} D(P \,||\, M) + \frac{1}{2} D(Q \,||\, M)$$

where

$$M = \frac{1}{2}(P + Q)$$

A *t* test was used to score significance between the three methods (data not shown). Because no significance was found, we used the L1-norm method.

Estimating the virus effective population size (N_e). We used a modified version of the Wright-Fisher idealized population model[24] to estimate the effective population size of influenza A virus from the shared SNVs in our donor-recipient pairs. This model assumes that the population does not grow or shrink, that there are discrete generations, that every generation is 'replaced' by offspring and that each of the variant sites is independent (the parameter

values used in the Wright-Fisher calculations can be found in **Supplementary Table 10**). We then calculated a variance effective size, the size of a Wright-Fisher population with the same variance

we focus on the variance effective population size (NeV), defined as the number of breeding individuals in an idealized population that would show the same amount of dispersion of allele frequencies under random genetic drift as the population under consideration (Crow, 1954).

$$N_{e_i}^{v} = \frac{\mathrm{E}[p_j]\mathrm{E}[q_j]}{2\mathrm{var}(\Delta_j)}$$

where $N_{e_i}^{v}$ is the variance effective population size for a given nucleotide position i, q is the major variant frequency of donor j and p is the minor variant frequency of donor j. For variants that were shared by all donors for a given strain with a frequency greater than 1% (we use this less conservative threshold so that we have more sites to include in our estimate and better resolution), we calculated the change in variant frequency between the donor and recipient for all pairs

$$\Delta_j = p_j - p_j'$$

with p_j' being the minor variant frequency of the recipient. The variance in this quantity appears in the effective size formula. For H1N1/2009, the size of j is eight unique donor-recipient pairs with 21 shared variants. For H3N2, the size of j is six unique donor-recipient pairs with 81 shared variants.

To estimate the variance in the effective population size across all household pairs, we included a standard deviation (SD) parameter defined by

$$\varepsilon = \mathrm{SD}(p_j)$$

which is used in the following modified Wright-Fisher equations

$$N_{e_i}^{v} = \frac{\mathrm{E}[p_j + \varepsilon]\mathrm{E}[q_j - \varepsilon]}{2\mathrm{var}(\Delta_j(\varepsilon,\varepsilon'))}$$

$$N_{e_i}^{v} = \frac{\mathrm{E}[p_j - \varepsilon]\mathrm{E}[q_j + \varepsilon]}{2\mathrm{var}(\Delta_j(\varepsilon,\varepsilon'))}$$

This ensures that $\mathrm{E}[p_j \pm \varepsilon] + \mathrm{E}[q_j \pm \varepsilon] \approx 1$ and captures the mean standard deviation range, and $\Delta_j(\varepsilon,\varepsilon')$ is the change in frequency at the jth site between the donor and recipient.

To confirm the scale of our estimates, we employed a second method that uses Kullback-Leibler divergence, as previously used to measure Ebola virus transmission[12]. This approach measures the distance from a true probability distribution q to a target probability distribution p, which are our donor and recipient populations, respectively, and uses their similarity to estimate the number of times the donor distribution was sampled. As with the Wright-Fisher approach, this assumes independence between variant sites and will consequently return a lower bound estimate ($\hat{N}$) on infectious dose size.

$$\hat{N} = \frac{s}{2\sum_i^S \mathrm{KL}(q_i \mid p_i)} < N_e$$

The number of variants shared by the donor and recipient is represented by s. A variant has to be present in both the donor and recipient to be included. $\mathrm{KL}(q_i|p_i)$ is the Kullback-Leibler divergence from q_i to p_i, where q_i is the set of nucleotide frequencies found in the donor at position i and p_i is the set of nucleotide frequencies found in the recipient at the same site. This value is summed over the variant positions across all segments where a shared variant is discovered for both the donor and recipient. We calculated this for each donor-recipient pair for H1N1/2009 and H3N2.

Haplotype reconstruction by SMRT sequencing. SNVs identified by Illumina sequencing were phased into haplotypes for six of our donor-recipient pairs (H1N1/2009 681_V1(0)/681_V3(2), 742_V1(0)/742_V3(3), 779_V1(0)/779_V2(1); H3N2: 720_V1(0)/720_V2(1), 734_V1(0)/734_V3(2), 763_V1(0)/763_V2(3)) using SMRT sequencing on the PacBio platform (Pacific Biosciences). DNA library preparation and sequencing was performed according to the manufacturer's instructions and reflects the P6-C4 sequencing enzyme and chemistry, using 4-h movie collection parameters. Each barcoded influenza M-RT-PCR cDNA was assessed by Qubit analysis and DNA 12000 Agilent Bioanalyzer gel chip to quantify the mass and size distribution of the double-stranded cDNA present. After quantification, samples were pooled in batches of two or three samples per SMRTbell library preparation. The barcoded amplicon pools were then repurified using a 1.8× AMPure XP purification step to ensure removal of any damaged fragments and/or biological contaminant. After purification, ~100 ng of each of the purified, unsheared samples was used for end repair, with reactions incubated at 25 °C for 5 min, followed by a second 1.8× AMPure XP purification step. Next, 0.75 µM of Blunt Adaptor was added to the cDNA, followed by the addition of 1× template Prep Buffer, 0.05 mM ATP low and 0.75 U/µl T4 ligase to ligate (final volume of 47.5 µl) the SMRTbell adaptors to the DNA amplicons. This solution was incubated at 25 °C overnight, followed by ligase denaturation at 65 °C for 10 min. After ligation, the library was treated with an exonuclease cocktail to remove unligated DNA fragments using a solution of 1.81 U/µl Exo III 18 and 0.18 U/µl Exo VII with reactions incubated at 37 °C for 1 h. Two additional 1.8× AMPure XP purifications steps were performed to remove any adaptor dimer or molecular contamination. Upon completion of library construction, samples were validated using another Agilent Bioanalyzer DNA 12000 gel chip as well as Qubit analysis. For all cases, the yield was sufficient, and primer was annealed to the SMRTbell libraries for sequencing. The polymerase-template complex was then bound to the P6 enzyme using a 10:1 ratio of polymerase to SMRTbell at 0.5 nM for 4 h at 30 °C, and libraries were then held at 4 °C until ready for magbead loading, before sequencing. The magbead-loaded, polymerase-bound SMRTbell libraries were placed on the RSII machine at a sequencing concentration of 50 pM and configured for a 240-min continuous sequencing run to allow for the maximum number of passes for consensus error correction through the reads of insert protocol version 2.3.0. Sequencing was conducted to ample coverage using a single SMRTcell for each of the sample pools, where reads were rigorously filtered using ten-pass, 95% single-molecule CCS filter criteria to yield ~23,000–25,000 reads after filtering per SMRTcell for each of the pooled sample sets. Continuous long-read data with 21–26 single-molecule passes were generated and passed through the RS_ReadsOfInsert.1 pipeline version 2.3.0 using a ~99.2% accuracy cutoff to achieve higher-quality CCS FASTA and FASTQ files for variant calling. Reads were aligned against the same reference genome used for the Illumina data. Alignment was performed with BLASR[25], using the default parameters. Reads that mapped against each segment were retrieved using SAMtools (version 1.2)[26] and converted to FASTA format. We used the variant calls obtained from the Illumina reads and phased them with the PacBio reads to identify linked variants. The GenBank accession code for the H1N1/2009 reference was CY111731 and that for the H3N2 reference was CY106640.

20. Zhou, B. *et al.* Single-reaction genomic amplification accelerates sequencing and vaccine production for classical and Swine origin human influenza a viruses. *J. Virol.* **83**, 10309–10313 (2009).
21. Djikeng, A. *et al.* Viral genome sequencing by random priming methods. *BMC Genomics* **9**, 5 (2008).
22. Stamatakis, A., Hoover, P. & Rougemont, J. A rapid bootstrap algorithm for the RAxML Web servers. *Syst. Biol.* **57**, 758–771 (2008).
23. Nakamura, K. *et al.* Sequence-specific error profile of Illumina sequencers. *Nucleic Acids Res.* **39**, e90 (2011).
24. Charlesworth, B. Fundamental concepts in genetics: effective population size and patterns of molecular evolution and variation. *Nat. Rev. Genet.* **10**, 195–205 (2009).
25. Chaisson, M.J. & Tesler, G. Mapping single molecule sequencing reads using basic local alignment with successive refinement (BLASR): application and theory. *BMC Bioinformatics* **13**, 238 (2012).
26. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).

doi:10.1038/ng.3479