CHARACTERIZING INTRA-HOST DIVERSITY OF INFLUENZA VIRUS AND MODELING TRANSMISSION NETWORKS DURING NATURAL EPIDEMICS

by

Timothy Song

B.S. Biology and Computer Science, The College of New Jersey, 2011

Submitted to the Graduate Faculty of

School of Medicine in partial fulfillment

of the requirements for the degree of

Masters of Science

University of Pittsburgh

2015

UNIVERSITY OF PITTSBURGH

SCHOOL OF MEDICINE

This thesis was presented

by

Timothy Song

It was defended on

August 12, 2015

and approved by

Roni Rosenfeld, Professor, Department of Computer Science, Carnegie Mellon University

Nathan Clark, Professor, Department of Computational & Systems Biology, University of Pittsburgh

Elodie Ghedin, Professor, Department of Biology, New York University

CHARACTERIZING INTRA-HOST DIVERSITY OF INFLUENZA VIRUS AND MODELING TRANSMISSION NETWORKS DURING NATURAL EPIDEMICS

Timothy Song, M.S.

University of Pittsburgh, 2015

Influenza A viruses are characterized by high genetic diversity due to error-prone replication, large population sizes, and strong natural selection. While most of what we know about influenza evolution has come from population scale epidemiological studies based on the analysis of a limited number of consensus sequences, these are limiting for outbreak investigations. The analysis of virus genetic diversity present in an infected host provides a richer genetic fingerprint with which to infer host-to-host virus transmission. Despite the use of animal models to characterize extent of intra-host diversity and what proportion of this diversity that is transmitted between individuals, less is known about these key evolutionary parameters in human populations. To quantify and characterize influenza virus variants that can achieve sustainable transmission in new hosts, we used household donor/recipient pairs of infected individuals from a Hong Kong community during the first wave of the 2009 H1N1 pandemic when seasonal H3N2 was also co-circulating. While the same variants were often found in multiple members of the community during the epidemic, the relative frequencies of variants fluctuate, with patterns of genetic variation more similar within than between households. We estimated the effective population size of influenza A virus across these donor/recipient pairs to be in the range of 100-200 contributing members, which enabled the transmission of multiple virus lineages among individuals, including antigenic variants.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF EQUATIONS

PREFACE

This work is to be submitted to Nature Genetics.

1.0 INTRODUCTION

Influenza A virus is a single stranded RNA virus that infects millions of people every year worldwide. Its error-prone RNA-dependent RNA polymerase leads to high genetic diversity. Even with the availability of vast clinical data, questions remain on what is the intra-host virus genetic diversity, what constitutes an effective dose and what is the genetic bottleneck at transmission.

The advent of high throughput sequencing, such as Next Generation Sequencing (NGS), has enabled the rapid genomic characterization of clinical samples from influenza-infected individuals. We developed a pipeline to analyze high throughput data from Illumina HiSeq, taking into consideration the high read coverage (up to 6000x), PCR errors, and sequence specific errors to identify real mutations due to the error prone polymerase of the RNA virus. Using this approach, we characterized the intra-host genetic diversity of influenza virus populations collected from nasal swabs of individuals infected during the 2009 H1N1 pandemic in Hong Kong where seasonal H3N2 was also co-circulating. Samples were collected from index cases and their household contacts; some individuals were sampled at two different time points. By sequencing the viruses present in these samples, we reconstructed the virus population structure over time and after transmission events. By looking at virus genetic data beyond the consensus, we were able to identify multiple strains within individuals and circulating during the epidemic, and observed that certain strain frequencies fluctuated over time.

Our goal was to determine how intra-host viral evolution influences inter-host viral transmission in a natural environment. We compared variants that are shared between hosts with those that arise by *de novo* mutation. We then identified relationships between samples, observed potential transmission links, and estimated the effective number of virions transmitted that are contributing members to the infection in contact cases.

1.1 WHY STUDY INFLUENZA?

Influenza A viruses are a consistent threat and a burden to human health. These pathogens cause respiratory tract infections, and in severe cases cause high morbidity and mortality. Most people with influenza-like illness, including cough and fever within 48 hours of symptom onset, are likely to have influenza (1). In the United States alone, 36,000 people die annually of influenza A virus infections (2). The 2009 H1N1 pandemic is thought to have infected 0.01% of the world population and resulted in 284,000 deaths worldwide (3). The phylogeny of influenza A is marked by antigenic cluster jumps, which have occurred on average every 3 years, and typically correspond to occurrences of vaccine failure (2).

Influenza viruses are of the family *Orthomyxoviridae* and are single-stranded, negative sense RNA viruses. Type A has greater genetic diversity than types B and C viruses, and infects the widest range of host species, including birds, swine, horses, and humans (2). The total length of the viral genome is around 13 Kb and has eight distinct segments encoding 10-11 proteins (4). The segmented genome can undergo reassortment, which occurs when two or more viruses infect the same cell and the resulting new viral particles contain RNA segments from each of the

"parental" viruses. This can provide an evolutionary advantage because segments of the virus can reassort to create antigenically novel strains, potentially leading to pandemics (4).

The two surface proteins, hemagglutinin (HA) and neuraminidase (NA), exhibit greater amino acid variability than other proteins. The HA binds to cell surface receptors and allow the virus to penetrate into the cytoplasm while the NA enables budding of new virions from the infected cell. There are 16 HA and 9 NA subtypes. Within a single subtype (such as H3N2) there can be multiple and diverse viral lineages co-circulating—including antigenic variants. Co-infection of cells can lead to reassortment and contribute to intra-host diversity (5).

1.2 INTRA-HOST DIVERSITY

We refer to intra-host diversity as the genetic variation of the virus population within the infected host. Intra-host diversity is due to the error-prone RNA-dependent RNA polymerase of the virus and to the host's immune status. This diversity allows the viruses to transition into new genetic space after being exposed to selective pressures, such as host immunity or antiviral treatment (6). In molecular epidemiology studies of influenza, intra-host diversity is overlooked and most of the focus is on the consensus sequence (i.e. the genetic average of individual variants in a population). What we then observe is a stark contrast between the vibrant mutant spectrum of influenza diversity and a single static consensus sequence of influenza. This becomes important because even in rapidly mutating populations, the emerging variant is detectable before the mutation becomes fixed (7). Longitudinal studies of influenza show that the mutational spectrum of influenza can change considerably over time (8). The new variants and phenotypes may persist without changing the consensus sequence and in some cases fixation may never occur (7).

However, the immune status of the host (naïve, previously exposed or vaccinated) can lead to different patterns of sequence diversity, and variants of the former may have antigenic significance. Identifying variants that become fixed are potential clues as to the presence of variants of interest that may exist as minor populations.

1.3 THE TRANSMISSION BOTTLENECK

The transmission of influenza can occur by direct contact, aerosol or droplet transmission. A recent study of infected individuals demonstrates that in infected patients, as many as 10^5 viral copies can be excreted over a 30 minute period by aerosol (9). In addition, aerosol administration to volunteers found that the minimal infectious dose can be fewer than 10 virions (10). Asymptomatic infected hosts can also be infectious (11). Infectious influenza virions can originate from the upper respiratory tract and may be the source for direct and airborne transmission events (12). There are several processes that can affect the bottleneck during transmission. A low infectious dose could severely limit the number of particles transmitted and cause a founder effect, which would reveal very low genetic diversity immediately after transmission (13). Another factor could be selective pressures of the host, where diversity is diminished as natural selection would eliminate viruses not fit enough for establishment of infection (13).

2.0 QUANTIFYING INFLUENZA VIRUS DIVERSITY AND TRANSMISSION IN HUMANS

2.1 INTRODUCTION

Influenza A viruses are characterized by high genetic diversity due to error-prone replication, large population sizes, and strong natural selection (14-16). While most of what we have learned about influenza evolution has come from population level epidemiological studies based on the analysis of consensus sequences (17) they are limiting for outbreak investigations. The analysis of virus genetic diversity present in an infected host provides a richer genetic fingerprint with which to infer virus transmission from host to host (18-22). Despite attempts to characterize intra-host diversity and the transmission bottleneck of the influenza A virus in various animal models (19, 23) it is still not well understood for human populations (24). To characterize patterns of viral evolution at a finer-scale, we performed deep sequencing on nasopharyngeal swabs collected from index cases with confirmed influenza along with their household contacts.

We have previously shown that pandemic H1N1 and seasonal H3N2 viruses—both present during the first wave of the H1N1 pandemic in Hong Kong (25)—have similar transmission potential in household settings, and that different antigenic variants of H3N2 co-circulated with different clades of H1N1/2009 (25, 26). In other parts of the world, and during the same time period, the unseasonal transmission of H3N2 was observed along with pandemic

H1N1 virus (27). To quantify and characterize influenza virus variants that can achieve sustainable transmission, we used household donor/recipient pairs of infected individuals from this Hong Kong community. To characterize patterns of viral evolution at a finer-scale, we performed deep sequencing on nasopharyngeal swabs collected from index cases with confirmed influenza along with their household contacts. We captured whole genome data and genetic diversity of the virus population within each infected patient. The household epidemiological information enables us to assign with relatively high confidence donor/recipient pairs in suspected transmission events and compare with unrelated pairs, all while estimating spatio-temporal transmission chains. We estimated the effective population size that enabled the transmission of multiple virus lineages among individuals, including antigenic variants.

2.2 MATERIALS AND METHODS

2.2.1 Sample collection

Retrospective pooled specimens of nasal and throat swabs studied in a previous household influenza transmission investigations (28, 29) were subjected to next generation sequencing. This dataset comprises 102 virus samples (55 H1N1/2009 and 47 H3N2) collected from 86 individuals in Hong Kong over July and August 2009. There were multiple home visits and 16 individuals were sampled twice on 2 or 3 household visits (visit 1, V1; visit 2, V2; visit 3, V3), 2-4 days apart.

2.2.2 Sample preparation, sequencing and variant calls.

Multi-segment reverse-transcription PCR (M-RT-PCR) (28) was used to amplify influenza-specific segments from total RNA, followed by sequence independent single primer amplification (SISPA) (29). Each RNA sample was subjected to 2 rounds of M-RT-PCR and these 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-ends sequencing on different lanes of a HiSeq2000 sequencer (Illumina). Potential SISPA PCR duplicate reads were removed with the JCVI ELVIRA package [http://sourceforge.net/projects/elvira/]. SISPA barcoded reads were demultiplexed with JCVI DNA Barcode Deconvolution software [http://sourceforge.net/projects/deconvolver/]. CLC Bio software was used to map barcode-trimmed reads to a reference genome and to remove low quality reads. Minor variants were identified using the JCVI ELVIRA package.

2.2.3 Variant analysis

Minor variants were identified using the JCVI ELVIRA package, which applies statistical tests to minimize false positive single nucleotide variants (SNV) calls that can be caused by sequence specific errors (SSE) that may occur in Illumina platforms (30). This involves observing the forward and reverse reads of a SNV call; based on a binomial distribution cumulative probability, we calculate the p-values. If both p-values are within a Bonferroni-corrected significance level (alpha = .05), the SNV call is accepted. A minimum minor allele frequency of 3% was used as the threshold; this cutoff was based on the same control sample that was sequenced in two different sequence runs, and then examining concordance (SNV found in both

samples) and discordance (SNV found in only one of 2 samples) for different frequency thresholds. At 3%, 16/17 sites were concordant, while at 4% 14/14 sites were concordant. We chose the lower cut-off to gain more information, even if the error was higher.

2.2.4 Quantification of intra-host diversity

We used Shannon entropy to quantify the intra-host diversity of each sample through the relative frequencies of each single nucleotide variant 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 = 1.27E-06).

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

Equation 1 - Shannon entropy

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

2.2.5 Genetic distance across samples

We compare each sample against every other sample (all-versus-all pairwise comparison) at each variant nucleotide position using an L1-norm:

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

Equation 2 - L1-norm distance

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). p and q are vectors containing the relative frequencies of the different variant nucleotides observed (these are analogous to "alleles").

Between two 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 (CDS); N is the length of the CDS.

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

Equation 3 - Dissimilarity distance

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

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

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

Equation 4 - L2-norm measure

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

The third method we used was the Jensen-Shannon divergence (JSD). The JSD modifies the Kullback-Leibler divergence so that the resulting output is symmetric and will always have a finite value:

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

Equation 5 – Kullback-Leibler divergence

The JSD is calculated by:

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

Equation 6 - Jensen-Shannon divergence

where

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

Equation 7 - JSD probability measure (M)

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

2.2.6 Estimating the virus effective population size

We used a modified version of the Wright-Fisher idealized population model (31) to estimate the effective population size of influenza A virus from the shared SNVs in our donor/recipient pairs. This model assumes the population does not grow or shrink, there are discrete generations, that every generation is "replaced" by offspring, and that each of the variant sites is independent. We then calculate a variance effective size, the size of a Wright-Fisher population with the same variance,

$$N_i = \frac{E[p_j]E[q_j]}{2var(\Delta)}$$

Equation 8 -- Modified Wright-Fisher idealized population model

where N is the effective population size for a given nucleotide position i, q is the major variant frequency of a donor j, and p is the minor variant frequency of j. For variants that were shared by all donors for a given strain with a frequency greater than 0.01 (1%), we calculated the change in variant frequency between donor and recipients for all pairs,

$$\Delta = p_j - p'_j$$

Equation 9 - Delta of donor and recipient frequencies

with p'_j being the minor variant frequency of the recipient. The variance in this quantity appears in the effective size formula. For H1N1, the size of j is 8 unique donor-recipient pairs with 21 shared variants. The equivalent values for H3N2 are j of 6 unique donor/recipient pairs with 81 shared variants.

To confirm the scale of our estimates, we employed a second method that utilizes Kullback-Leibler divergence to measure Ebola virus transmission (32). The aim of this approach is to measure the distance from a true probability distribution, *q*, to a target probability distribution, *p*, which are our donor and recipient populations, respectively, and use 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 ($\widehat{N}$) on infectious dose size.

$$\widehat{N} = \frac{\mathrm{s}}{2\sum_i^S KL(q_i|p_i)} < N_{eff}$$

Equation 10 - Modified Kullback-Leibler divergence estimate

The number of shared variants between donor and recipient is represented by *s*. A variant has to be shared by both donor and recipient to be included. $KL(q_i|p_i)$ 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 on both the donor and recipient. We calculated this for each donor/recipient pair for H1N1/2009 and H3N2.

2.2.7 Phylogenetic analyses

All eight influenza A coding sequences were concatenated into an alignment of 13,392 nucleotides (nt) for H1N1/2009, and 13,425 nt for H3N2. Coding sequences were concatenated in the order of the segment number on which they were encoded (PB2-PB1-PA-HA-NP-NA-M1-M2-NS1-NS2). All isolates were included except for 781_V1(0), which appeared to be a reassorted isolate, encoding genes related to both H1N1 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 based on their position in widely sampled single gene phylogenies (data not shown). Two additional taxa—A/Brisbane/10/2007 and A/Nanjing/1/200—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 (33) using the GTR nucleotide substitution model, with among-site rate variation modeled using a discrete gamma distribution using four rate categories. Bootstrap support values were generated using 1,000 fast bootstrap replicates, and represented as percentages on nodes (values below 50% not shown).

2.2.8 Haplotype reconstruction by single molecule sequencing

SNVs identified by Illumina sequencing were phased into haplotypes by SMRT sequencing on the PacBio platform for 6 of our donor/recipient pairs using the viral isolates (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)). DNA library preparation and sequencing were performed according to the manufacturer's instructions and reflect the P6-C4 sequencing enzyme and chemistry, using 4-hour movie collection parameters. Each barcoded influenza M-RTPCR 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 2-3 samples per SMRTbell library preparation as follows. The barcoded amplicon pools were then re-purified using a 1.8X AMPure XP purification step (1.8X AMPure beads added, by volume, to each sample in 200 µL EB, vortexed for 10 minutes at 2,000 rpm, followed by two washes with 70% alcohol and finally diluted in EB). This AMPure XP purification step assures removal of any damaged fragments and/or biological contaminant. After purification, ~100 ng of each of the purified, unsheared samples was taken into end-repair, which was incubated at 25ºC for 5 minutes, followed by the second 1.8X Ampure XP purification step. Next, 0.75 µM of Blunt Adapter was added to the cDNA, followed by 1X 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 adapters to the DNA amplicons. This solution was incubated at 25ºC overnight, followed by a 65ºC 10-minute ligase denaturation step. After ligation, the library was treated with an exonuclease cocktail to remove un-ligated DNA fragments using a solution of 1.81 U/µL Exo III 18 and 0.18 U/µL Exo VII, then incubated at 37ºC for 1 hour. Two additional 1.8X Ampure XP purifications steps were performed to remove

any adapter 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 ratio of 10:1 polymerase to SMRTbell at 0.5 nM for 4 hours at 30ºC and then held at 4ºC until ready for magbead loading, prior to sequencing. The magnetic bead-loading step was conducted at 4ºC for 60-minutes per manufacturer's guidelines. The magbead-loaded, polymerase-bound, SMRTbell libraries were placed onto the RSII machine at a sequencing concentration of 50 pM and configured for a 240-minute 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 a 10-pass, 95% single molecule CCS filter criteria to yield ~23,000 – 25,000 post-filtered reads per SMRTcell for each of the pooled sample sets. Continuous long read data with 21-26 single-molecule passes, and ~99.2% accuracy was generated and produced filtered CCS FASTA and FASTQ files were generated for variant calling, after completing the RS_ReadsOfInsert.1 pipeline version 2.3.0.

2.3 RESULTS

2.3.1 Intra-host diversity of Hong Kong samples

The virus sample set was collected in July and August 2009 from 86 individuals (67 index patients and 17 other household members) living in Hong Kong; 16 patients were sampled twice, 2-4 days apart. We estimated intra-host virus diversity for each sample by mapping polymorphic sites onto the consensus genome assemblies to generate a list of single SNVs (or minor variants) present at a frequency of at least 3%. Intra-host diversity was calculated as the Shannon entropy, *H*, by summing the entropies for each such site, assuming site independence. Mean intra-host diversity was significantly higher (Wilcoxon rank-sum test p = 1.89e-12) for H3N2 (H = 33) than H1N1/2009 (H = 13). There was no significant Pearson correlation between high intra-host virus diversity and high viral titer [13] (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 (**Table 1**).

Table 1 - Pearson's correlation between quantitative viral loads (qPCR) and variant counts

strain	passage	segment	p value	# of samples
H1N1/2009	P0	PB2	0.68	53
H1N1/2009	P0	PB1	0.16	53
H1N1/2009	P0	PA	0.02	53
H1N1/2009	P0	HA	0.15	53
H1N1/2009	P0	NP	0.16	53
H1N1/2009	P0	NA	0.80	53
H1N1/2009	P0	MP	0.02	53
H1N1/2009	P0	NS	0.81	53
H3N2	P0	PB2	0.25	45
H3N2	P0	PB1	0.60	45
H3N2	P0	PA	0.23	45
H3N2	P0	HA	0.65	45
H3N2	P0	NP	0.43	45

H3N2	P0	NA	0.39	45
H3N2	P0	MP	0.73	45
H3N2	P0	NS	0.64	45

Data indicate that there is correlation between genetic diversity and viral load only for the M and PA segments in the nasopharyngeal swabs. P0 = nasopharyngeal swabs, no passage. Significance is $p < 0.05$.
qPCR data available at http://web.hku.hk/~bcowling/influenza/HK_H1N1_study.htm

2.3.1.1 Phylogenetic analysis

Phylogenetic analysis clustered whole genome consensus sequences by household for each group of patients diagnosed as being infected with either H1N1/2009 (**Figure 1**) or H3N2 (**Figure 2**). Phylogenetic analyses of each gene individually provided no evidence for reassortment within this population during the timeframe of the study (data not shown).

Three clades of H1N1/2009 (clades 3, 6 and 7) and three antigenic sublineages of H3N2 (A/Brisbane/10/2007-like, A/Victoria/2008/2009-like, and A/Perth/16/2009-like) circulated in this population (34). Despite the relatively small population size, one case of mixed subtype infection was observed, indicating that dual infection with seasonal and pandemic strains may not be a rare event (35).

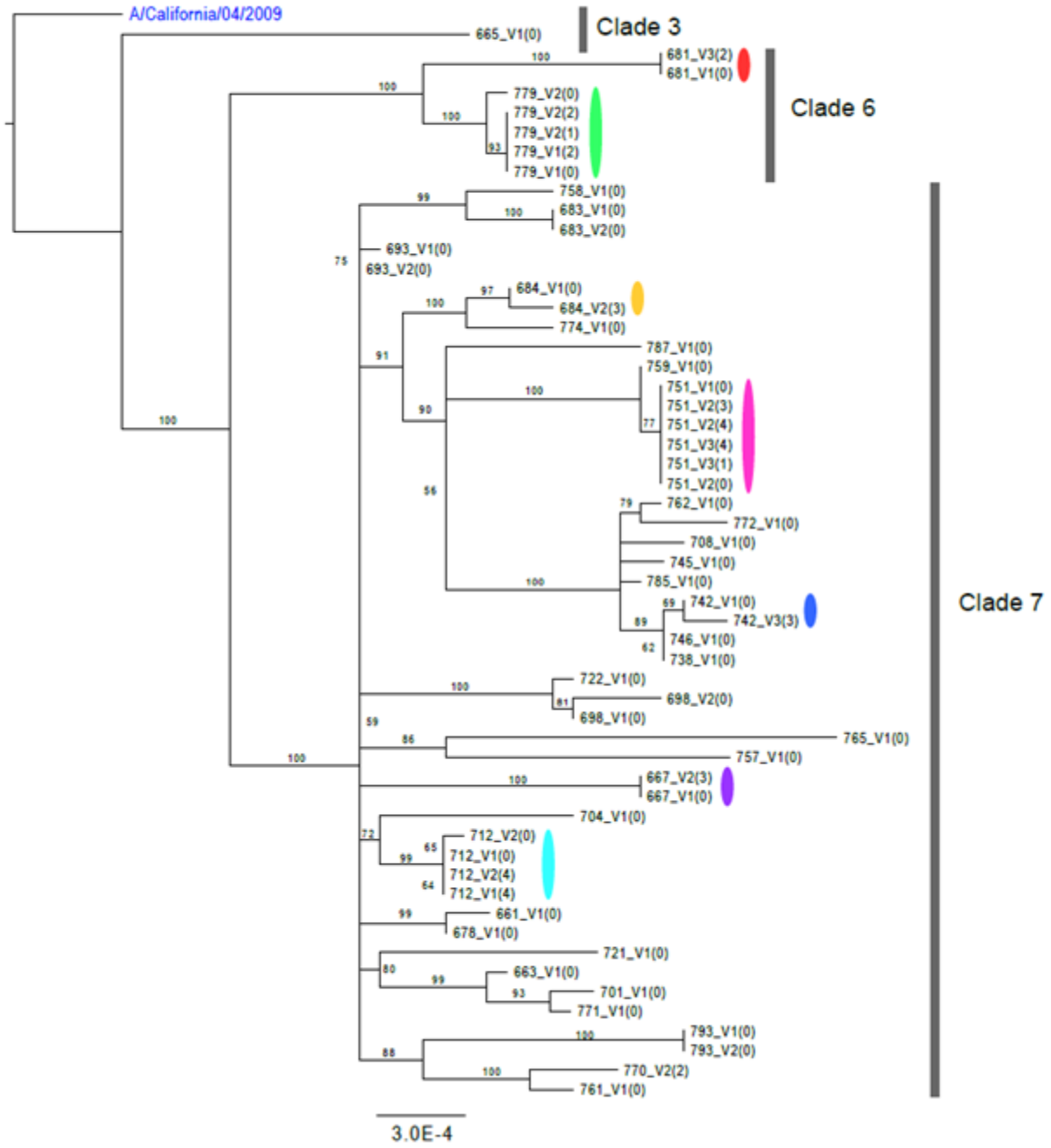

Figure 1 - Maximum likelihood phylogenies of concatenated genomes for H1N1/2009.

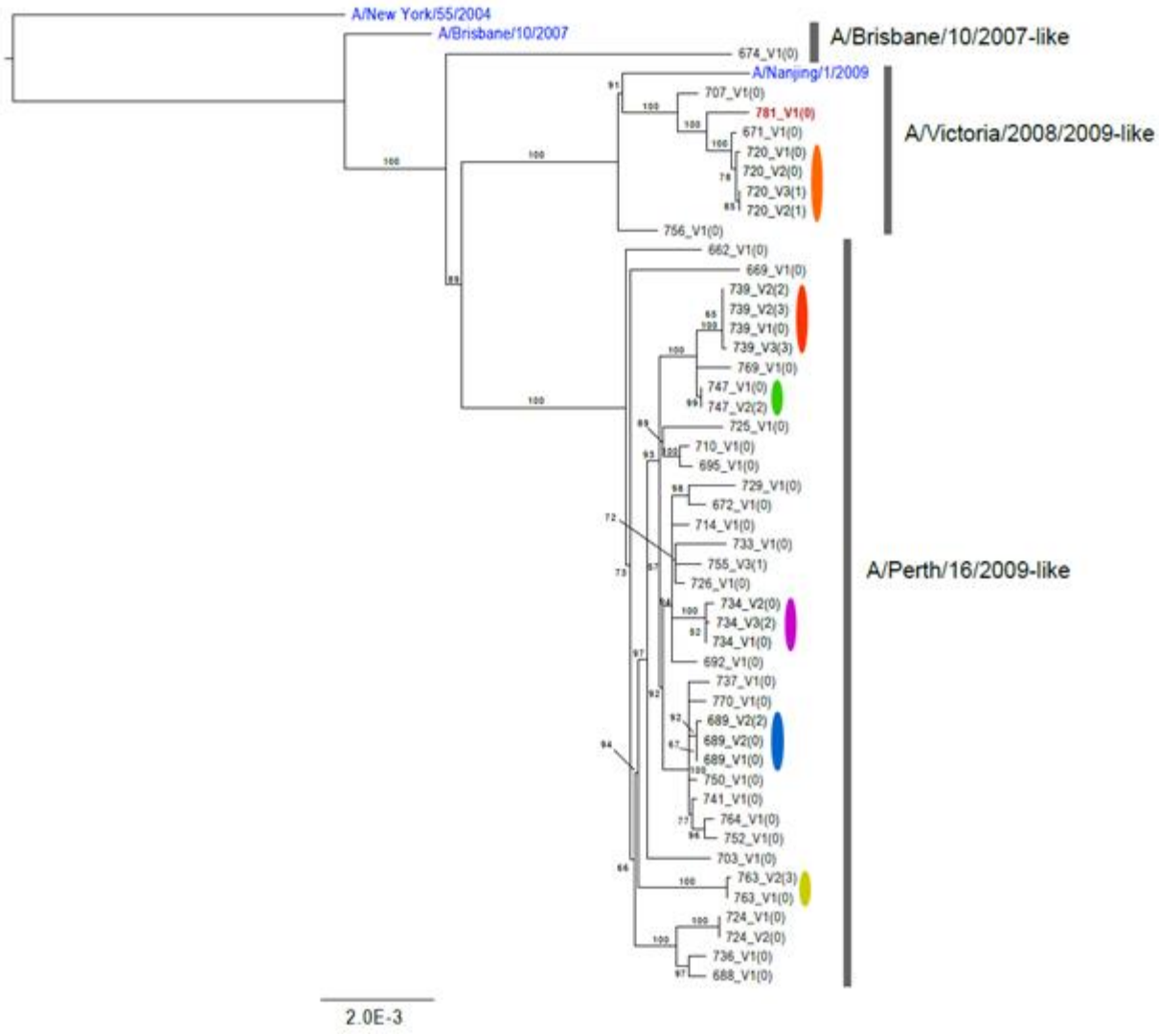

Figure 2 - Maximum likelihood phylogenies of concatenated genomes for H3N2

M1/M2 and NS1/NS2 genes were represented as one segment for each covering the sequence between the first ATG to the last stop codon. Bootstrap support values are shown as percentages on nodes. Values below 50% were treated as equivocal and not shown on the figure. Public sequences downloaded from GenBank for use as out groups, or included within the diversity of the samples, are colored in blue. One patient, 781_V1(0), was shown to also be infected with H1N1/2009 clade 7 after having been diagnosed with H3N2 strain A/Victoria/2008/2009-like. Only the HA and NA from the H1N1/2009 could be unambiguously assembled from this individual (accession CY115455 and CY115458), while a whole genome was assembled for the H3N2. Note that scales are different for both trees. Households with more than one member are colored.

2.3.2 Shared SNVs and haplotype phasing

We compared SNVs across samples to determine if minor variants were shared within and between households.

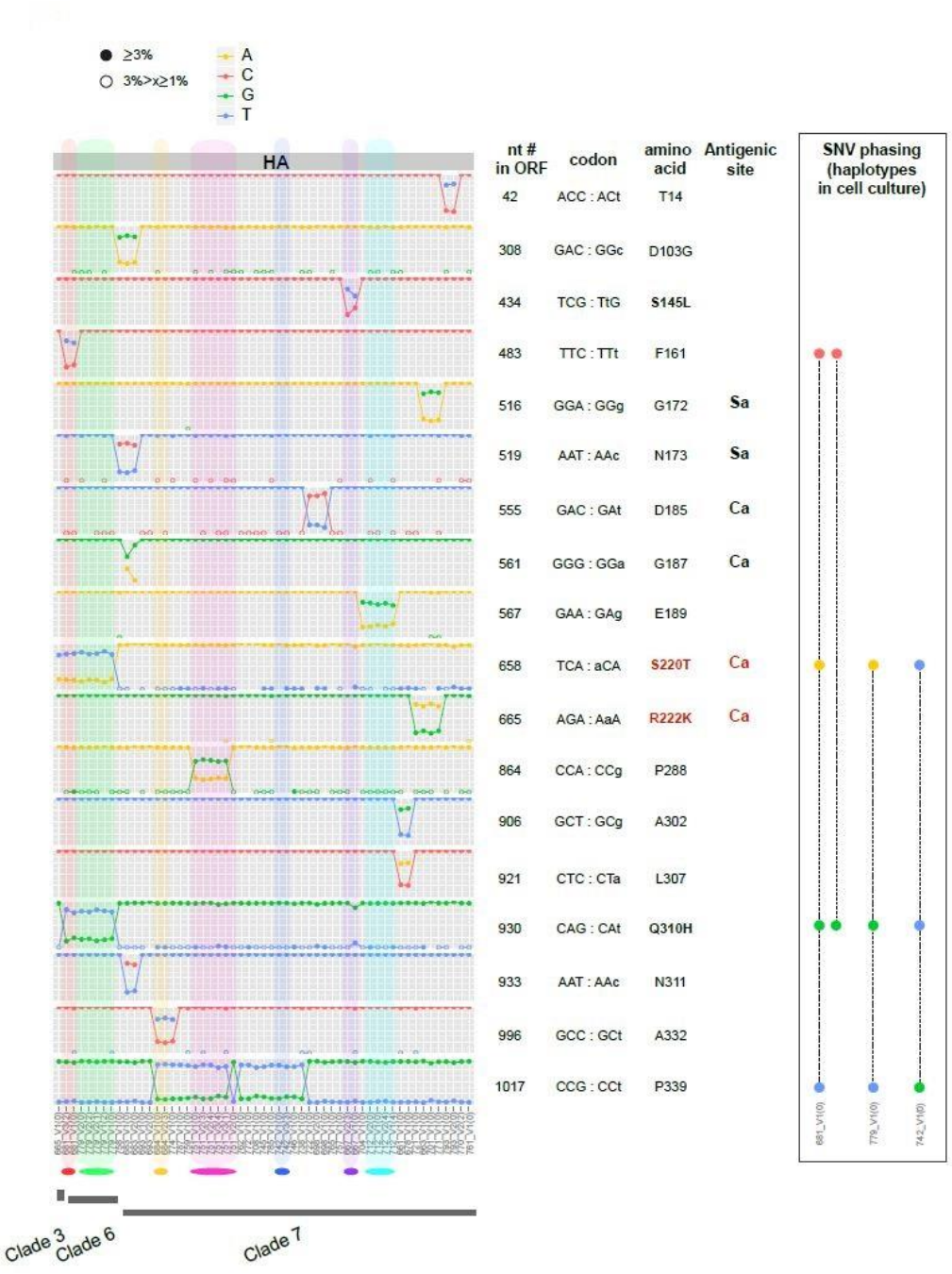

Figure 3 - Comparison of HA minor variant frequencies across households in H1N1/2009

Figure 4 - Comparison of HA minor variant frequencies across households in H3N2

Only polymorphic sites located in the HA1 domain are represented. The amino acid positions were numbered according to the first methionine (start codon) of the protein (and not according to the HA1 numbering schema). The x-axis lists samples by position on the phylogenetic trees in Fig. 1 or Fig. 2; households with more than one member are colored. The y-axis displays nucleotide frequencies with graph lines corresponding to 0, 25%, 75% and 100% frequency. ORF = open reading frame; Antigenic site = previously identified as corresponding to antigenic sites. Text in red highlights non-silent mutations located in antigenic sites. Closed circles represent minor variants found at a frequency 3% and higher, while open circles correspond to frequencies equal or higher than 1%, but below 3%. Boxes show how minor variant nucleotides are phased on the same molecules, representing haplotypes. These were determined from single molecule sequencing of cell culture viruses for 6 household 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).

For both H1N1/2009 (**Figure 3**) and H3N2 (**Figure 4)** we observed multiple positions in HA—including potential antigenic sites—where the minor variant nucleotide in one clade or strain became the major nucleotide in another, with evidence of mixed infection at many other sites across the genome (see Appendix). To confirm these findings observed from the clinical specimens, we phased the SNVs into haplotypes by single molecule sequencing for 12 of the cell culture samples from 6 different households.

Notably, although the consensus sequence points towards the sample belonging to one strain, the patient is often infected with two or more strains; many of these variants could be detected in multiple families. This, along with the haplotype information, suggests that a number of the SNVs are not *de novo* mutations that occurred in the household's index patient, but are shared across the community as a whole. We see a similar phenomenon when looking at global consensus sequences across seasons. Using human 2008 H3 sequences as a reference, we observed a shift of nucleotide frequency at some positions in subsequent seasons of H3N2 epidemics (**Figure 5**).

Figure 5 - Nucleotide usage frequency at positions with transmissible variants in human viral HA genes

Full-length Human H1N1/2009 (2009-2013, N=9870; upper panel) and H3N2 (2008-2013, N=4587; lower panel) HA sequences were downloaded from GenBank. Using the HK data set to select sites where minor variants were shared within or between households, we summarized the frequency of these polymorphic sites across different years for each subtype. Data were further subdivided into lineages (H1N1: clade 6 and clade 7; H3N2: A/Perth/16/2009 and A/Victoria/208/2009). Boxplots show the median of the frequency; the bottom and top of each box represent the first and third quartiles. The length of the whiskers is defined as a function of the inner quartile range and they extend to the most extreme data point within the 75%-25% data range. Outliers are marked by black dots.

This phenomenon is more pronounced for variants from the A/Victoria/208/2009-like lineage, in sharp contrast to the decreasing trend observed for the A/Perth/16/2009-like lineage. However, we did not see such a trend in pandemic H1N1 after the 2009 season. Additionally, frequency variations in H1N1/2009 are far less common than in H3N2. One should note that the A/Victoria/208/2009-like virus replaced the A/Perth/16/2009-like virus as the dominant lineage in recent years, leading to a change of vaccine strain from A/Perth/16/2009-like virus to A/Victoria/208/2009-like virus in 2012. In contrast, pandemic H1N1 virus is antigenically stable

and there was no change of vaccine strain after its introduction in humans in 2009. Overall, our data indicate that some synonymous/non-synonymous mutations could be transmitted between individuals at low frequency levels. Genetic distance between samples

Since each virus sample collected in our study will contain *de novo* mutations and/or 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 pairwise comparison for each variant nucleotide position using an L1-norm. We grouped pairwise comparisons by longitudinal pairs (same individual, sampled at two different visits), transmission pairs (within households), and across household pairs (**Figure 6**).

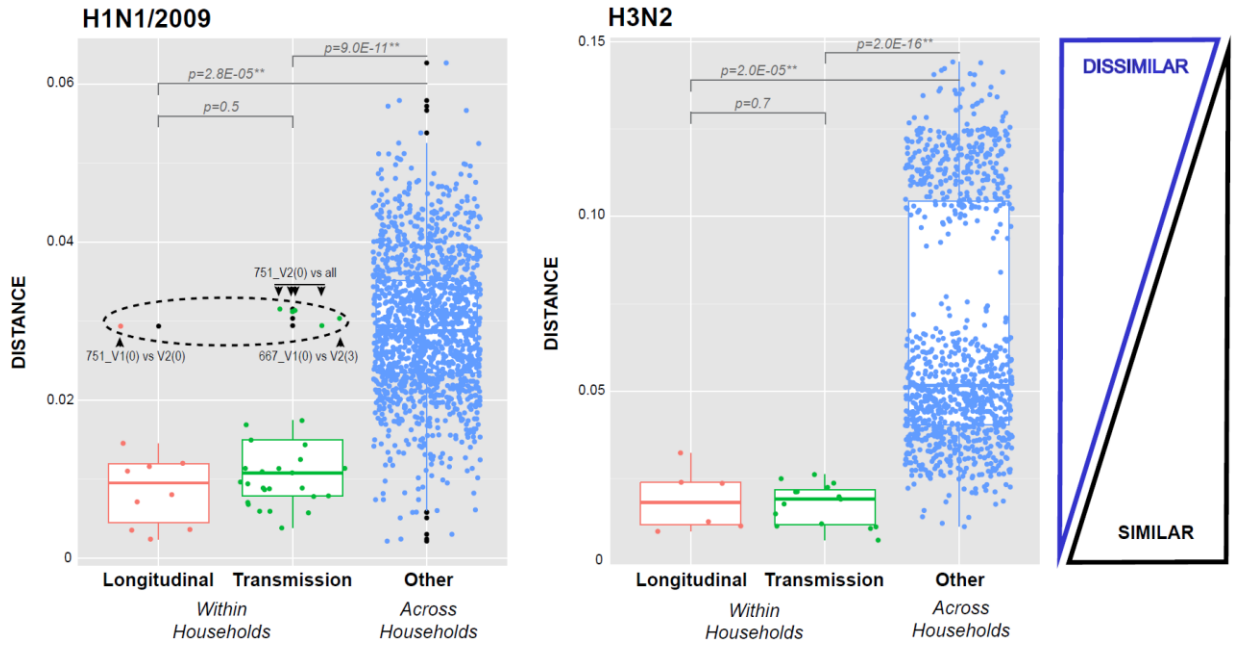

Figure 6 - Box-plots of L1-norm pairwise genetic distance within and across households

We use the L1-norm values obtained from the variant nucleotide analysis across all genes to compare overall genetic distance of longitudinal pairs (there are 16 individuals in 12 households who have been sampled at two different time points, 2-3 days apart) and transmission pairs (there are 13 households where at least 2 members have been sampled, with a total of 22 predicted

donor and recipient pairs within households, and 22 more when including more than one time point per individual), compared to all other comparisons across households (every other possible sample pair combination). The boxplots 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 marked 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 was then used in within household pairwise comparisons (H1N1/2009, Transmission), the visit 2 sample appeared clearly as an outlier. The pairwise genetic distance between the index case in household 667 (667_V1(0)) and its other household member (667_V2(3)) also appeared as an outlier pair.

We determined that the median L1 genetic distances between household pairs or longitudinal pairs are significantly closer than any random pairing, while within household median genetic distance is not significantly different than that observed for longitudinal pairs, indicating minor variants and their proportions can be used to infer inter-host transmission, even if a number of these correspond to co-infecting variants that are shared with individuals across households. Interestingly, for H1N1/2009 we see a few "within household" pairs that are outliers (**Figure 6**, *dashed circle*), further evidence of mixed infection. For example, the visit 2 sample for the index case of household 751 (751_V2(0)) has multiple polymorphic major sites as compared to the other samples from the same household, including its visit 1 sample (751_V1(0)). Similarly, for the index case of household 667 (667_V1(0)), SNV frequencies are different when compared to the contact case (667_V2(3)). These also demonstrate that some minor variants can occasionally become dominant after a single transmission.

2.3.3 Transmission network

After excluding outliers and considering only a single sample (visit 1) per individual, there were 21 viable "within household" transmission pairs. To select other potential epidemic links within the community, outside of the household transmissions, we used the transmission and

longitudinal pairs to identify outliers and determine a threshold of maximum genetic distance (after excluding outliers) (**Figure 6**). Each pair was epidemiologically linked to a short transmission chain (see below). Using the consensus sequences, we inferred transmission networks across the population using a parsimony and graph-based algorithm (36, 37).

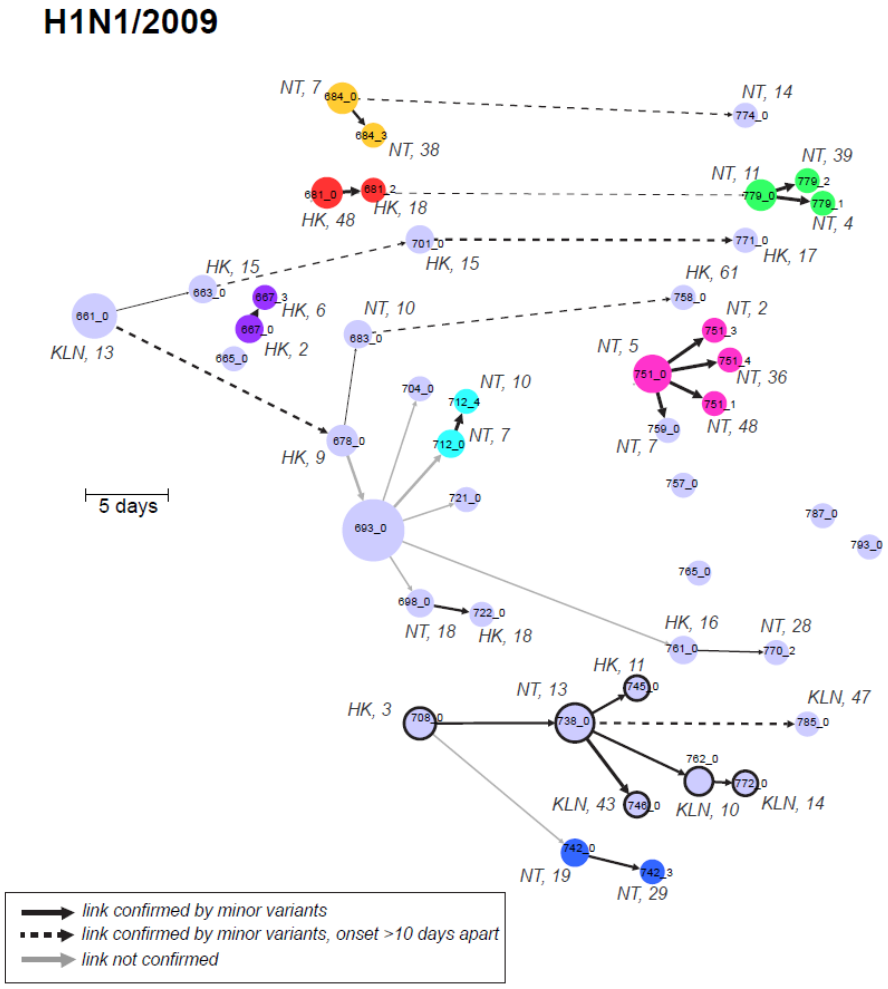

Figure 7 - Reconstruction of potential transmission pathways of H1N1/2009

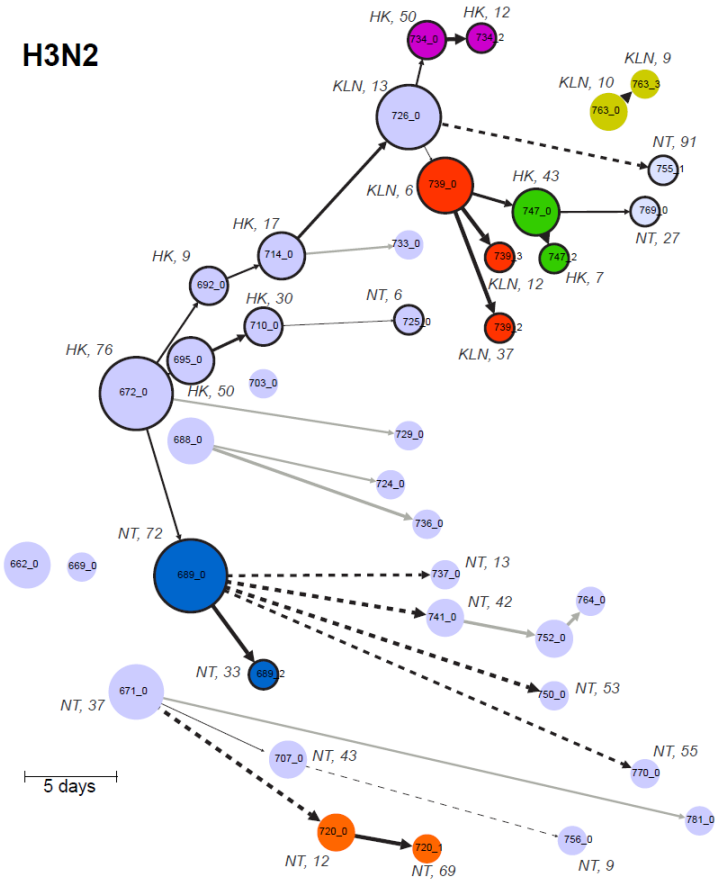

Figure 8 - Reconstruction of potential transmission pathways of H3N2 outbreaks

Transmission networks are inferred from the 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 shows the number of days since the first date of onset. A unique color is assigned to households with more than one member sampled. The size of the node is determined by the number of out degrees. A dashed line indicates a putative transmission link greater than 10 days. The weight of an edge indicates the number of nucleotide differences between two samples (a darker edge = smaller number of differences); Nucleotide differences were separated into quartiles. H1N1/2009: 0-2 nt; 3-6 nt; 7-15 nt; 16-28 nt. H3N2: 0-5 nt; 6-9 nt; 10-19 nt; 20-45 nt. Circles with thick black edges are nodes within a chain of transmission with more than 2 individuals. Locality and age of the patient is indicated for a number of the nodes. HK: Hong Kong; NT: New Territories; KLN: Kowloon.

We then use minor variant data to highlight potential localized outbreaks (**Figure 7 and Figure 8**) with cross-region links (i.e. Hong Kong Island, Kowloon and New Territories). This also 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.

2.3.4 Shared viral populations

To further explore shared virus populations within households, we compared minor variants at each position in donor (index cases) and recipient transmission pair samples.

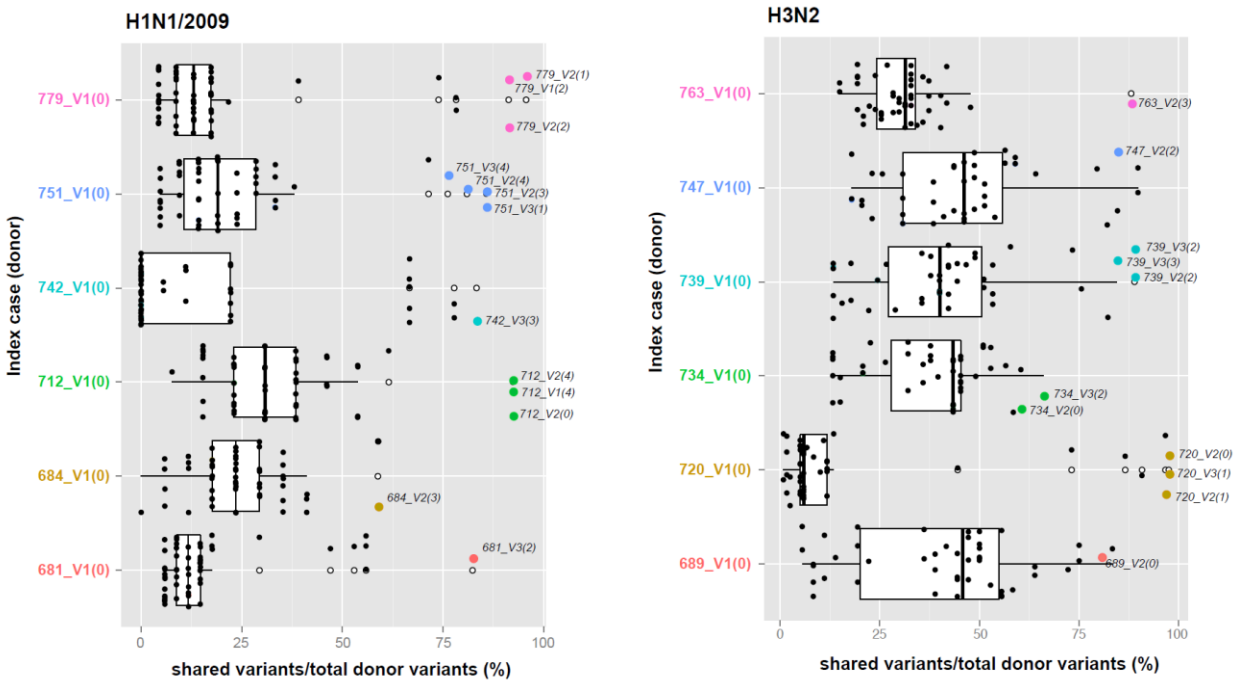

Figure 9 - Box-plots comparing shared variant frequencies within and across households

We compared shared variant frequencies between samples from index cases and their household members (colored dots) or with any other sample (black dots). White boxes indicate interquartile ranges and white dots indicate outliers. Household members tend to share most of the variants found in the index case.

Most variants found in the donor were shared with the potential recipient (**Figure 9**, *colored dots*). The frequency of shared variants is much lower in pairs of unrelated samples (**Figure 9**, *black dots*), although we find more shared variants in H3N2 than in H1N1/2009 pairs. We observe that the relative frequency of variants in the recipient is more often similar to that found in the donor, which is not the case for the same variants found in any other individual (Wilcoxon

signed-rank test, $p < 0.05$). This suggests shared variants found in the recipient are not the result of *de novo* mutation but are more likely present in viruses that transmit and replicate.

2.3.5 Effective population sizes

From the household transmission pairs we estimated the number of variants that can achieve sustainable transmission in new hosts. Polymorphic sites with variants only detected in the donor and those detected in both donor and recipient samples 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 10% has an 86% chance of transmission (**Figure 10**). Because of limited sample size it was not possible to determine with confidence the probability of transmission for variants present at frequencies below 10%.

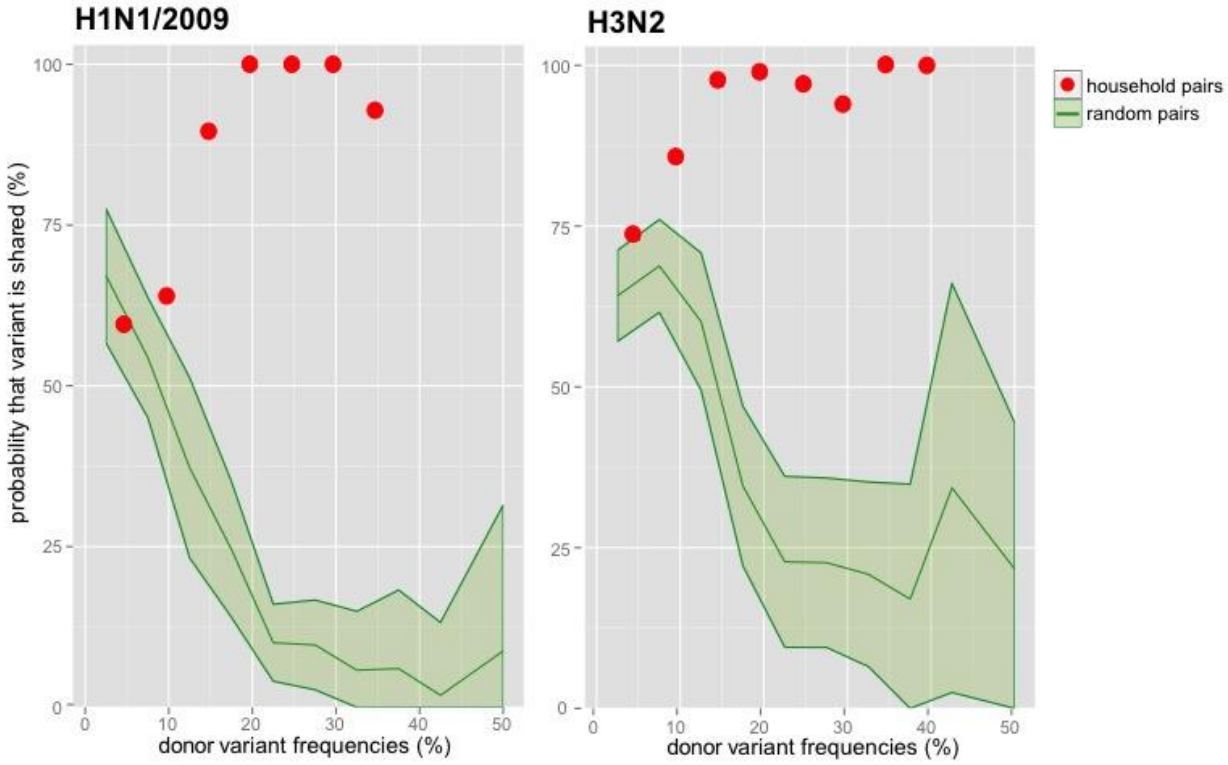

Figure 10 - Probability of variant transmission as a function of relative frequency of the minor variants

Variants that were only detected in the donor and those that were shared between donor and recipient samples were used in determining the probability of transmission. Household pairs (red dots) are comparisons between members of the same household. Each point is 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 get a standard deviation estimate.

To infer the size of the virus population before and after transmission that is able to generate productive progeny, we estimated the effective population size, *Ne*, using a modified version of the Wright-Fisher (WF) idealized population model for our dataset. Specifically, for our donor/recipient pairs we take the frequency of the shared minor variants, *p*; the frequency of the major nucleotide at that position, *q*; and then calculate the variance of the difference in donor/recipient frequencies, to obtain a variance effective size. For this we obtain a mean of 192

viral particles (median 123) for H1N1/2009 and a mean of 248 (median 138) for H3N2. To confirm the scale of our estimates, we utilized a different method based on the Kullback-Leibler divergence (KLD) to estimate an effective size. This gave a mean of 90 (median 80) for H1N1/2009 and a mean of 114 (median 121) for H3N2. To estimate how many haplotypes would be present within these replicating populations, we phased SNVs for the HA segment and reconstructed haplotypes by single molecule sequencing. From this, we observed an average of three haplotypes transmitted across donor/recipient pairs for both H1N1/2009 and H3N2, taking into account phased SNVs for HA. Previous empirical data for H3N2 is of the same order of magnitude as these inferred values (38) and with previous observations that seasonal H3N2 has more co-circulating lineages than pandemic H1N1 (25, 39). Crucially, these effective population and haplotype estimates suggest that multiple variants can be routinely transmitted between individuals, such that any transmission bottlenecks are relatively loose, and that a relatively small number of viral particles can initiate a productive infection with a number of variant strains that are co-transmitted.

2.4 DISCUSSION

We analyzed minor variant dynamics in the transmission of the influenza A virus within and across households during an epidemic. In particular, we used shared minor variant information between donors and recipients in transmission pairs to estimate the number of viral particles that are able to infect and replicate in the recipient, and which revealed the transmission of multiple variants. Our approach could help define how prior immunity or other host factors, as well as virus subtype and strain, may affect transmission dose, which our effective size estimates likely

capture lower bounds on. We also demonstrate that there are likely more cases of mixed strains within infected patients than can be captured with standard consensus-based diagnostic assays. Such co-infections will obviously facilitate the occurrence of reassortment, and may help explain the frequent detection of reassortants between seasonal H3 viruses (40). For some of the co-infected patients we observe potential competition between two strains with different lineages dominating the population found in each individual. Although similar observations have been made in infected animals (37, 41), ours is the first demonstration for influenza A virus in humans. Overall, characterizing the genetic information of transmitted virions allows a better understanding of influenza virus transmission, and provides more accurate information for modeling epidemics and disease control strategies.

3.0 CONCLUSIONS AND FUTURE PERSPECTIVES

This work provides a new analysis framework to better understand the structure of influenza A virus populations within infected hosts and influenza transmission dynamics. Due to the now ubiquitous nature of NGS with deep sequencing and single molecule sequencing capabilities, we can collect detailed information on virus genetic information and evolution in an infection. This can be used in conjunction with epidemiological data to estimate the effective population size at transmission and how prior immunity or other host factors or virus subtype and strain may affect that number. These types of analyses also enable us to look at mixed infections, such as co-infections with two strains of the same influenza A subtype, or two different subtypes—which we observed in one of our Hong Kong patients—and, potentially, different respiratory viruses. These would likely have been ignored in standard consensus-based diagnostic assays.

Future work will include the same type of analyses to characterize transmission networks of school-age children in Pittsburgh schools. Using the variant data information, we can follow minor strains over the course of an epidemic and reconstruct chains of transmission. The next steps will be to link the symptom onset data with our variant analysis to create a transmission model. Preliminary results indicate that there may be multiple variants transmitted. This again suggests that the genetic diversity within an individual plays a much larger role than what is depicted by the consensus sequence alone.

These studies highlight the power of NGS for the fine characterization of viral populations within infected hosts. These data can be overlaid onto epidemiological maps to get better resolution in transmission networks during epidemics.

BIBLIOGRAPHY

1. **Monto AS, Gravenstein S, Elliott M, Colopy M, Schweinle J.** 2000. Clinical signs and symptoms predicting influenza infection. Archives of internal medicine **160:**3243-3247.

2. **Nelson MI, Holmes EC.** 2007. The evolution of epidemic influenza. Nature reviews genetics **8:**196-205.

3. **Dawood FS, Iuliano AD, Reed C, Meltzer MI, Shay DK, Cheng P-Y, Bandaranayake D, Breiman RF, Brooks WA, Buchy P.** 2012. Estimated global mortality associated with the first 12 months of 2009 pandemic influenza A H1N1 virus circulation: a modelling study. The Lancet infectious diseases **12:**687-695.

4. **Hutchinson EC, von Kirchbach JC, Gog JR, Digard P.** 2010. Genome packaging in influenza A virus. Journal of general virology **91:**313-328.

5. **Holmes EC.** 2009. RNA virus genomics: a world of possibilities. The Journal of clinical investigation **119:**2488.

6. **Graci JD, Cameron CE.** 2006. Mechanisms of action of ribavirin against distinct viruses. Reviews in medical virology **16:**37-48.

7. **Bordería AV, Stapleford KA, Vignuzzi M.** 2011. RNA virus population diversity: implications for inter-species transmission. Current opinion in virology **1:**643-648.

8. **Rogers MB, Song T, Sebra R, Greenbaum BD, Hamelin M-E, Fitch A, Twaddle A, Cui L, Holmes EC, Boivin G.** 2015. Intrahost Dynamics of Antiviral Resistance in Influenza A Virus Reflect Complex Patterns of Segment Linkage, Reassortment, and Natural Selection. mBio **6:**e02464-02414.

9. **Milton DK, Fabian MP, Cowling BJ, Grantham ML, McDevitt JJ.** 2013. Influenza virus aerosols in human exhaled breath: particle size, culturability, and effect of surgical masks. PLoS Pathog **9:**e1003205.

10. **Tellier R.** 2009. Aerosol transmission of influenza A virus: a review of new studies. Journal of the Royal Society Interface**:**rsif20090302.

11. **Hsieh Y-H, Tsai C-A, Lin C-Y, Chen J-H, King C-C, Chao D-Y, Cheng K-F.** 2014. Asymptomatic ratio for seasonal H1N1 influenza infection among schoolchildren in Taiwan. BMC infectious diseases **14:**80.

12. **Varble A, Albrecht RA, Backes S, Crumiller M, Bouvier NM, Sachs D, García-Sastre A.** 2014. Influenza A virus transmission bottlenecks are defined by infection route and recipient host. Cell host & microbe **16:**691-700.

13. **Wilker PR, Dinis JM, Starrett G, Imai M, Hatta M, Nelson CW, O'Connor DH, Hughes AL, Neumann G, Kawaoka Y.** 2013. Selection on haemagglutinin imposes a bottleneck during mammalian transmission of reassortant H5N1 influenza viruses. Nature communications **4**.

14. **Drake JW.** 1993. Rates of spontaneous mutation among RNA viruses. Proceedings of the National Academy of Sciences **90:**4171-4175.

15. **Drake JW, Holland JJ.** 1999. Mutation rates among RNA viruses. Proceedings of the National Academy of Sciences **96:**13910-13913.

16. **Bush RM, Fitch WM, Bender CA, Cox NJ.** 1999. Positive selection on the H3 hemagglutinin gene of human influenza virus A. Molecular biology and evolution **16:**1457-1465.

17. **Viboud C, Nelson MI, Tan Y, Holmes EC.** 2013. Contrasting the epidemiological and evolutionary dynamics of influenza spatial transmission. Philosophical Transactions of the Royal Society B: Biological Sciences **368:**20120199.

18. **Ghedin E, Holmes EC, DePasse JV, Pinilla LT, Fitch A, Hamelin M-E, Papenburg J, Boivin G.** 2012. Presence of oseltamivir-resistant pandemic A/H1N1 minor variants before drug therapy with subsequent selection and transmission. Journal of Infectious Diseases **206:**1504-1511.

19. **Murcia PR, Hughes J, Battista P, Lloyd L, Baillie GJ, Ramirez-Gonzalez RH, Ormond D, Oliver K, Elton D, Mumford JA.** 2012. Evolution of an Eurasian avian-like influenza virus in naive and vaccinated pigs. PLoS Pathog **8:**e1002730-e1002730.

20. **Kundu S, Lockwood J, Depledge DP, Chaudhry Y, Aston A, Rao K, Hartley JC, Goodfellow I, Breuer J.** 2013. Next-generation whole genome sequencing identifies the direction of norovirus transmission in linked patients. Clinical infectious diseases**:**cit287.

21. **Simmons HE, Dunham JP, Stack JC, Dickins BJ, Pagan I, Holmes EC, Stephenson AG.** 2012. Deep sequencing reveals persistence of intra-and inter-host genetic diversity in natural and greenhouse populations of zucchini yellow mosaic virus. Journal of General Virology **93:**1831-1840.

22. **Stack JC, Murcia PR, Grenfell BT, Wood JL, Holmes EC.** 2012. Inferring the inter-host transmission of influenza A virus using patterns of intra-host genetic variation. Proceedings of the Royal Society of London B: Biological Sciences**:**rspb20122173.

23. **Hughes J, Allen RC, Baguelin M, Hampson K, Baillie GJ, Elton D, Newton JR, Kellam P, Wood J, Holmes EC.** 2012. Transmission of equine influenza virus during an outbreak is characterized by frequent mixed infections and loose transmission bottlenecks. PLoS Pathog **8:**e1003081.

24. **Fordyce SL, Bragstad K, Pedersen SS, Jensen TG, Gahrn-Hansen B, Daniels R, Hay A, Kampmann M-L, Bruhn CA, Moreno-Mayar JV.** 2013. Genetic diversity among pandemic 2009 influenza viruses isolated from a transmission chain. Virol. J **10:**116.

25. **Poon LL, Chan KH, Chu DK, Fung CC, Cheng CK, Ip DK, Leung GM, Peiris JS, Cowling BJ.** 2011. Viral genetic sequence variations in pandemic H1N1/2009 and seasonal H3N2 influenza viruses within an individual, a household and a community. Journal of clinical virology **52:**146-150.

26. **Cowling BJ, Chan KH, Fang VJ, Lau LL, So HC, Fung RO, Ma ES, Kwong AS, Chan C-W, Tsui WW.** 2010. Comparative epidemiology of pandemic and seasonal influenza A in households. New England Journal of Medicine **362:**2175-2184.

27. **Ghedin E, Wentworth DE, Halpin RA, Lin X, Bera J, DePasse J, Fitch A, Griesemer S, Hine E, Katzel DA.** 2010. Unseasonal transmission of H3N2 influenza A virus during the swine-origin H1N1 pandemic. Journal of virology **84:**5715-5718.

28. **Zhou B, Donnelly ME, Scholes DT, George KS, Hatta M, Kawaoka Y, Wentworth DE.** 2009. Single-reaction genomic amplification accelerates sequencing and vaccine production for classical and Swine origin human influenza a viruses. Journal of virology **83:**10309-10313.

29. **Djikeng A, Halpin R, Kuzmickas R, DePasse J, Feldblyum J, Sengamalay N, Afonso C, Zhang X, Anderson NG, Ghedin E.** 2008. Viral genome sequencing by random priming methods. BMC genomics **9:**5.

30. **Nakamura K, Oshima T, Morimoto T, Ikeda S, Yoshikawa H, Shiwa Y, Ishikawa S, Linak MC, Hirai A, Takahashi H.** 2011. Sequence-specific error profile of Illumina sequencers. Nucleic acids research**:**gkr344.

31. **Charlesworth B.** 2009. Fundamental concepts in genetics: effective population size and patterns of molecular evolution and variation. Nature reviews. Genetics **10:**195-205.

32. **Emmett KJ, Lee A, Khiabanian H, Rabadan R.** 2014. High-resolution genomic surveillance of 2014 ebolavirus using shared subclonal variants. PLoS currents **7**.

33. **Stamatakis A, Hoover P, Rougemont J.** 2008. A rapid bootstrap algorithm for the RAxML web servers. Systematic biology **57:**758-771.

34. **Poon LL, Chan KH, Chu DK, Fung CC, Cheng CK, Ip DK, Leung GM, Peiris JS, Cowling BJ.** 2011. Viral genetic sequence variations in pandemic H1N1/2009 and seasonal H3N2 influenza viruses within an individual, a household and a community.

Journal of clinical virology : the official publication of the Pan American Society for Clinical Virology **52:**146-150.

35. **Lee N, Chan PK, Lam WY, Szeto CC, Hui DS.** 2010. Co-infection with pandemic H1N1 and seasonal H3N2 influenza viruses. Annals of internal medicine **152:**618-619.

36. **Jombart T, Eggo RM, Dodd PJ, Balloux F.** 2011. Reconstructing disease outbreaks from genetic data: a graph approach. Heredity **106:**383-390.

37. **Hughes J, Allen RC, Baguelin M, Hampson K, Baillie GJ, Elton D, Newton JR, Kellam P, Wood JL, Holmes EC, Murcia PR.** 2012. Transmission of equine influenza virus during an outbreak is characterized by frequent mixed infections and loose transmission bottlenecks. PLoS pathogens **8:**e1003081.

38. **Gustin KM, Belser JA, Wadford DA, Pearce MB, Katz JM, Tumpey TM, Maines TR.** 2011. Influenza virus aerosol exposure and analytical system for ferrets. Proceedings of the National Academy of Sciences **108:**8432-8437.

39. **Galiano M, Agapow P-M, Thompson C, Platt S, Underwood A, Ellis J, Myers R, Green J, Zambon M.** 2011. Evolutionary pathways of the pandemic influenza A (H1N1) 2009 in the UK. PloS one **6:**e23779.

40. **Westgeest KB, Russell CA, Lin X, Spronken MI, Bestebroer TM, Bahl J, van Beek R, Skepner E, Halpin RA, de Jong JC, Rimmelzwaan GF, Osterhaus AD, Smith DJ, Wentworth DE, Fouchier RA, de Graaf M.** 2014. Genomewide analysis of reassortment and evolution of human influenza A(H3N2) viruses circulating between 1968 and 2011. Journal of virology **88:**2844-2857.

41. **Varble A, Albrecht RA, Backes S, Crumiller M, Bouvier NM, Sachs D, Garcia-Sastre A, tenOever BR.** 2014. Influenza a virus transmission bottlenecks are defined by infection route and recipient host. Cell host & microbe **16:**691-700.