1. Supplementary Discussion

1.1 Reassortant variants of H7N9 viruses

Reassortant variants or genotypes of the H7N9 viruses can be defined based on the clade designations of their segments for the H7N9 viruses not involved in co-infections (n=505, Extended Data Table 3). Extended Data Figs. 1 and 2 illustrates the development of H7N9 genotypes from the first wave to the second wave. Of the 48 genotypes identified, 20 were from wave 1 clade (W1) and 28 from wave 2 clades (W2-A, B, C). Only nine genotypes had more than ten isolates and 24 had only a single isolate. Twenty of these genotypes, ten from W2 clade and ten from W2 clades, were found in human isolates. Of these four from W1 clade and two from W2 clades were only found in humans, reflecting limited or no avian surveillance in the locality of these human infections. The genotypes common in humans were also common in chickens.

In the second wave, the Zhejiang viruses in clade W2-A contain 8 genotypes (Fig. 2a, Extended Data Fig. 2). Most of these reassortant variants have their PB2 (28 out of 31) and M (24 out of 34) gene segments from clades other than 1.1 (mainly clade 2). Some of the W2-A viruses that disseminated to Jiangxi have further reassorted to acquire more non-clade 1.1 internal gene segments (e.g. PB1 and PA), mainly from the H9N2 poultry viruses circulating in Jiangxi (Fig. 2a, Extended Data

Fig. 1). Some W2-A viruses (e.g. Ck/ST/4816/14, July 2014) transmitted to Guangdong may have recently reassorted with the local W2-B viruses (Fig. 2a, Extended Data Fig. 1, Extended Data Table 3). In total, 12 reassortant variants (including ZJ, JX and GD viruses) were identified in clade W2-A.

The W2-B viruses were almost exclusively found in Guangdong (Fig. 2), and composed 9 reassortant variants (Extended Data Table 3). Most have four or five gene segments not from clade 1.1 (198 out of 204; Extended Data Fig. 2), primarily from clade 3. Among these, 56 have retained the PA and M gene segments from clade 1.1, while 136 retain only the clade 1.1 M gene.

Most of the clade W2-C viruses had genotypes with four gene segments not from clade 1.1 (Extended Data Fig. 2; n=15 out of 21). Their PB2, PA and M genes were mainly from clade 2, while the NS segment was from clade 3 (Fig. 2a; Extended Data Fig. 1). Seven reassortant variants were identified in clade W2-C.

1.2 Amino acid changes occurring in H7N9 viruses

Some early wave1 viruses (those in the early cluster of the HA phylogeny; available from http://dx.doi.org/10.5061/dryad.5q7kf) possess glutamine at HA residue 235 (numbering based on the complete H7 sequence). This residue was mutated to leucine early in the wave 1 outbreak and has been maintained. Amino acid

substitutions have occurred in predicted antigenic regions A (residue 140), C (residue 285) and E (residue 65) of the HA. R65K and N285D occurred in both wave 1 and wave 2 viruses, while T140A was only observed in wave 2 viruses (see the H7 tree available from http://dx.doi.org/10.5061/dryad.5q7kf). Multiple independent substitutions of R65K have led to a change in the dominant residue from wave 1 (79.4% R) to wave2 (78.3% K) (Extended Data Table 4). For example, K65 was introduced in the wave 1 sequences immediately ancestral to clade W2-B. No bias between human and avian isolates was observed for this site. This codon was also identified as a positively selected site ($p<0.005$; SLAC analysis[38]).

Despite the mutations (R65K, T140A and N285D) that occurred in predicted antigenic regions, haemagglutinin inhibition (HI) assays show that four ferret antisera raised against three wave 1 viruses (two of which are the current vaccine candidates selected by WHO) and one wave 2 virus are reactive against most wave 2 isolates (≤four fold change in titre; Supplementary Data). This suggests that these vaccine candidate strains are still sufficient to protect against the wave 2 viruses in humans.

The neuraminidase proteins of all the H7N9 viruses have a deletion of residues 69-73 in the stalk region, except for A/Shanghai/05/2013 (SH5) and A/Guangdong/03/2013 (GD3), which were isolated in April 2013 and December 2013, respectively. There is no direct phylogenetic relationship between SH5 (very early

wave 1) and GD3 (W2-B clade of wave 2) (see the H7 and N9 trees available from http://dx.doi.org/10.5061/dryad.5q7kf). The amino acid sequences differ in their stalk regions corresponding to the deletion – SH5 has QISNT, GD3 has ETNIT. SH5 may have retained QISNT from the earlier wild bird lineage (e.g. A/northern shoveler/HK/MPL133/2010) while GD3 appears to have duplicated the five amino acids N-terminal to the normally deleted region. The majority of the H7N9 viruses are predicted to be susceptible to neuraminidase inhibition. R294K, which can confer resistance to oseltamivir[14,15], occurred occasionally in the first (1.9%) and second (0.2%) wave viruses, but only in human isolates, which might be a response to treatment.

Residues 191, 559 and 570 in PB2 had lysine, asparagine or methionine, respectively, in the majority (88.3%, 85.5% and 91.0%) of wave 1 viruses examined but have glutamic acid, threonine or isoleucine (99.5%, 99.7% and 87.8%), respectively, in wave 2 viruses. Several residues of the PB1 and PA polymerase proteins show changes in amino acid preference from wave 1 to wave 2 (Extended Data Table 4). In PB1, residues 171, 397 and 525 change from methionine (100%) to valine (53.7%), isoleucine (100%) to methionine (54.2%) and valine (71%) to isoleucine (68.3%), respectively, from wave 1 to wave 2. In PA residues 100 and 394, changes occurred between wave 1 and wave 2 from alanine (87.4%) to valine (64.4%)

and asparagine (84.6%) to aspartic acid (65.2%), respectively. PA A100V might reflect an increased ability to infect and replicate in human cells[39].

Most of these changes in amino acid frequency occurred when wave 2 viruses reassorted with and obtained internal genes from co-circulating H9N2 viruses of clades 2 or 3, which had molecular markers different from those of clade 1. While most amino acid differences between internal genes of clades 1 to 3 or their sub-clades are due to single substitution events in ancestral viruses (Extended Data Fig. 1), some arose through multiple independent substitutions – PB2-M570I occurred three times; prior to the divergence of clade 1.3, early in clade 2 before clades 2.1 and 2.2 diverged, and prior to the divergence of clade3.1 (Extended Data Fig. 1a). PB2 559 and 570 are structurally close on the surface of the 627-domain. However, how these substitutions affect protein function and viral fitness is currently unclear.

In NP, the Q357R substitution occurred in one wave 2 human isolate from Shenzhen (A/SZ/SP38/2014), which also possess 627K in PB2 (Extended Data Table 4). This combination of substitutions has been shown to enhance virulence in mice[40]. All wave 1 and wave 2 viruses are resistant to adamantanes, as they have the 31N in their M2 proteins[41].

In the NS1 protein, 212S and 216T were predominant in wave 1 viruses (93.3% & 96%) whereas 212P and 216P became the most common in wave 2 viruses (60.1%

& 60.3%). These changes were obtained through reassortment with co-circulating H9N2 viruses of clade 3 (Extended Data Fig. 1f). The SH3 binding motif (PPΦXPK/R, Φ hydrophobic) had been lost in the clade 1.1 viruses but was retained in the clade 2 and clade 3 H9N2 viruses and may modulate binding of Crk/CrkL to increase phosphorylation of Akt that favours virus replication[42,43].

1.3 Genomic similarity between human and avian H7N9 isolates

Amino acid substitutions between human H7N9 viruses and their most closely related avian viruses were summarized from the phylogenies (Extended Data Table 5). In the PB2 protein, E627K substitutions are found in 71% of wave 1 human viruses and 57% of wave 2 viruses (overall frequency = 55%). D701N was found in 9% of wave 1 and 17% of wave 2 human isolates (overall frequency = 12%). No H7N9 isolates had both 627K and 701N substitutions. These substitutions are the main markers of enhanced polymerase activity in mammalian hosts[16-19].

1.4 Phylogeographic inference of H7N9 virus dissemination in China

We employed Bayesian phylogeographic methods to estimate the ancestral geographic states, together with the evolutionary time scale, of the H7N9 and related

viruses. The maximum clade credibility (MCC) trees summarised from the BMCMC samples are available from http://dx.doi.org/10.5061/dryad.5q7kf.

Based on the maximum likelihood H7 and N9 phylogenies, the W2-B clade established in Guangdong originated from one of the several virus strains introduced from eastern China during the first wave. The phylogeographic analysis of the H7 and N9 genes confirmed this finding (the H7-N9 MCC tree is available at http://dx.doi.org/10.5061/dryad.5q7kf). It shows that the W2-B clade clusters with wave1 Guangdong viruses, which originated from Zhejiang, with the best geographic state estimate (Probabilities: Zhejiang=0.92, Shanghai=0.07). The phylogeographic analysis also demonstrated repeated disseminations of the virus from eastern China to other regions (e.g. Jiangxi) in the W2-A clade. The analysis could not estimate the geographic state for the immediate ancestral node of the W2-C clade, highlighting the need for more data and surveillance to better understand the source of this clade.

The Bayesian MCC trees of the internal genes consistently show frequent interactions (i.e. reassortment) between H7N9 and H9N2 viruses, as shown in the maximum likelihood trees (available from http://dx.doi.org/10.5061/dryad.5q7kf). The internal genes of the H7N9 clade established in Guangdong were found to originate from the local (Guangdong) H9N2 viruses (highlighted with green boxes in the MCC trees available from http://dx.doi.org/10.5061/dryad.5q7kf).

39 Yamayoshi, S. *et al.* Virulence-affecting amino acid changes in the PA protein of H7N9 influenza A viruses. *J Virol* **88**, 3127-3134 (2014).

40 Kim, J. H. *et al.* Role of host-specific amino acids in the pathogenicity of avian H5N1 influenza viruses in mice. *J Gen Virol* **91**, 1284-1289 (2010).

41 Lan, Y. *et al.* A comprehensive surveillance of adamantane resistance among human influenza A virus isolated from mainland China between 1956 and 2009. *Antivir Ther* **15**, 853-859 (2010).

42 Heikkinen, L. S. *et al.* Avian and 1918 Spanish influenza a virus NS1 proteins bind to Crk/CrkL Src homology 3 domains to activate host cell signaling. *J Biol Chem* **283**, 5719-5727 (2008).

43 Hrincius, E. R. *et al.* CRK adaptor protein expression is required for efficient replication of avian influenza A viruses and controls JNK-mediated apoptotic responses. *Cell Microbiol* **12**, 831-843 (2010).

2. Supplementary Data

Supplementary Data | Haemagglutinin inhibition (HI) assays of H7N9 viruses.

Homologous titres are shown in orange cells. Samples with low reactions to all antisera are in red. Ferret antisera were raised against A/duck/Wenzhou/47/13 (WZ47), A/chicken/Wenzhou/610/13 (WZ610), A/Anhui/1/13 (AH1, WHO vaccine candidate), A/Shanghai/2/13 (SH2, WHO vaccine candidate), A/Shanghai/5/13 (SH5) and A/Hong Kong/5942/13 (HK5942). Antisera strains (n=6) and test strains (n=92) are indicated in the complete HA phylogeny available from http://dx.doi.org/10.5061/dryad.5q7kf .