Quantifying within-host evolution of H5N1 influenza in humans and poultry in Cambodia

Louise H. Moncla*[1], Trevor Bedford[1,2], Philippe Dussart[3], Srey Viseth Horm[3], Sareth Rith[3], Philippe Buchy[4], Erik A Karlsson[3], Lifeng Li[5,6], Yongmei Liu[5,6], Huachen Zhu[5,6], Yi Guan[5,6], Thomas C. Friedrich[7,8], Paul F. Horwood*[3,9]

Author affiliations

1. Fred Hutchinson Cancer Research Center, Seattle, Washington, United States.

2. University of Washington, Seattle, Washington, United States.

3. Virology Unit, Institut Pasteur du Cambodge, Institut Pasteur International Network, Phnom Penh, Cambodia.

4. GlaxoSmithKline, Vaccines R&D, Singapore, Singapore.

5. Joint Influenza Research Centre (SUMC/HKU), Shantou University Medical College, Shantou, People's Republic of China.

6. State Key Laboratory of Emerging Infectious Diseases/Centre of Influenza Research, School of Public Health, The University of Hong Kong, Hong Kong, SAR, People's Republic of China.

7. Department of Pathobiological Sciences, University of Wisconsin School of Veterinary Medicine, Madison, WI, United States.

8. Wisconsin National Primate Research Center, Madison, WI, United States.

9. College of Public Health, Medical and Veterinary Sciences, James Cook University, Townsville, Australia.

* correspondence: lhmoncla@gmail.com and paul.horwood@jcu.edu.au

Abstract

Avian influenza viruses (AIVs) periodically cross species barriers and infect humans. The likelihood that an AIV will evolve mammalian transmissibility depends on acquiring and selecting mutations during spillover. We analyze deep sequencing data from infected humans and ducks in Cambodia to examine H5N1 evolution during spillover. Viral populations in both species are predominated by low-frequency (<10%) variation shaped by purifying selection and genetic drift. Viruses from humans contain some human-adapting mutations (PB2 E627K, HA A150V, and HA Q238L), but these mutations remain low-frequency. Within-host variants are not enriched along phylogenetic branches leading to human infections. Our data show that H5N1 viruses generate putative human-adapting mutations during natural spillover infection. However, short infections, randomness, and purifying selection limit the evolutionary capacity of H5N1 viruses within-host. Applying evolutionary methods to sequence data, we reveal a detailed view of H5N1 adaptive potential, and develop a foundation for studying host-adaptation in other zoonotic viruses.

Introduction

Influenza cross-species transmission poses a continual threat to human health. Since emerging in 1997, H5N1 avian influenza viruses (AIVs) have caused 860 confirmed infections and 454 deaths in humans[1]. H5N1 naturally circulates in aquatic birds, but some lineages have integrated into poultry populations. H5N1 is now endemic in domestic birds in some countries[2–4], and concern remains that continued human infection may one day facilitate human adaptation.

The likelihood that an AIV will adapt to replicate and transmit among humans depends generating and selecting human-adaptive mutations during spillover. Influenza viruses have high mutation rates[5–8], short generation times[9], and large populations, and rapidly generate diversity within-host. Laboratory studies using animal models[10–12] show that only 3-5 amino acid substitutions may be required to render H5N1 viruses mammalian-transmissible[10–12], and that viral variants present at frequencies as low as 5% may be transmitted by respiratory droplets[13]. Subsequent modeling studies suggest that within-host dynamics are conducive to generating human-transmissible viruses, but that these viruses may remain at frequencies too low for transmission[14,15]. Although these studies offer critical insight for H5N1 risk assessment, it is unclear whether they adequately describe how cross-species transmission proceeds in nature.

H5N1 outbreaks offer rare opportunities to study natural cross-species transmission, but data are limited. One study of H5N1-infected humans in Vietnam identified mutations affecting receptor binding, polymerase activity, and interferon antagonism; however, they remained at low frequencies throughout infection[16]. Recent characterization of H5N1-infected humans in Indonesia identified novel mutations within-host that enhance polymerase activity in human cells[17]. Unfortunately, neither of these studies include data from naturally infected poultry, which would provide a critical comparison for assessing whether infected humans exhibit signs of adaptive evolution. A small number of studies have examined within-host diversity in experimentally infected poultry[18–20], but these may not recapitulate the dynamics of natural infection.

As part of ongoing diagnostic and surveillance effort, the Institut Pasteur du Cambodge collects and confirms samples from AIV-infected poultry during routine market surveillance, and from human cases and poultry during AIV outbreaks. Since H5N1 was first detected in Cambodia in 2004, 56 human cases and 58 poultry outbreaks have been confirmed and many more have gone undetected. Here we analyze previously generated deep sequence data[21] from 8 infected humans and 5 infected domestic ducks collected in Cambodia between 2010 and 2014. We find that viral populations in both species are dominated by low-frequency variation, shaped by purifying selection, population expansion, and genetic drift. We identify a handful of mutations in humans linked to improved mammalian replication and transmissibility, two of which were detected in multiple samples, suggesting that adaptive mutations arise during natural spillover infection. Although most within-host mutations are not linked to human infections on the H5N1 phylogeny, three mutations identified within-host are enriched on phylogenetic branches leading to human infections. Our data suggest that H5N1 viruses exhibit clear potential for within-host adaptation, but that a short duration of infection, randomness, and purifying selection may together limit the evolutionary capacity of these viruses to evolve extensively during any individual spillover event.

Methods

Viral sample collection

The Institute Pasteur in Cambodia is a WHO H5 Reference Laboratory (H5RL) and has a mandate to assist the Cambodian Ministry of Health and the Ministry of Agriculture, Forestry, and Fisheries in conducting investigations into human cases and poultry outbreaks of H5N1, respectively. Surveillance for human cases of H5N1 infection is conducted through influenza-like-illness, severe acute respiratory illness and event-based surveillance in a network of hospitals throughout the country. Poultry outbreaks of H5N1 are detected through passive surveillance following reports from farmers and villagers of livestock illness or deaths. The H5RL conducts confirmation of H5N1 detection and further characterisation (genetic and antigenic) of H5N1 strains.

RNA isolation and RT-qPCR

RNA was extracted from swab samples using the QIAmp Viral RNA Mini Kit (Qiagen, Valencia, CA, USA), following manufacturer's guidelines. Extracts were tested for influenza A (M-gene) and subtypes H5 (primer sets H5a and H5b), N1, H7, and H9 by using quantitative RT-PCR (qRT-PCR) using assays sourced from the International Reagent Resource (https://www.internationalreagentresource.org/Home.aspx), as previously outlined[22]. Only samples with high viral load (≥10^3 copies/µl), as assessed by RT-qPCR, were selected for sequence analysis. All samples were sequenced directly from the original specimen, without passaging in cell culture or eggs. Information on the samples included in the present analyses are presented in Table 1.

cDNA generation and PCR

cDNA was generated using the Superscript IV Reverse Transcriptase (Invitrogen, Carlsbad, CA, USA) and custom influenza primers targeting the conserved ends. The following primers were pooled together in a 1.5 : 0.5 : 2.0 : 1.0 ratio: Uni-1.5: ACGCGTGATCAGCAAAAGCAGG, Uni-0.5: ACGCGTGATCAGCGAAAGCAGG, Uni-2.0: ACGCGTGATCAGTAGAAACAAGG, and Uni-1.0: AGCAAAAGCAGG. 1 µl of this primer pool were added to 1 µl of 10 mM dNTP mix (Invitrogen) and 11 µl of RNA. Contents were briefly mixed and heated for 5 minutes at 65°C, followed by immediate incubation on ice for at least 1 minute. Next, a second mastermix was made with 4 µl of 5X Superscript IV Buffer, 1 µl of 100 mM DTT, 1 µl of RNaseOut Recombinant RNase Inhibitor, and 1 µl of SuperScript IV Reverse Transcriptase (200 U/µl) (Invitrogen). 7 µl of mastermix was added to each sample, for a total volume of 20 µl. This mixture was briefly mixed, incubated at 55°C for 20 minutes, then inactivated by incubating at 80°C for 10 minutes. Amplicons were generated with PCR, with primers targeting the conserved 3' influenza UTRs.

Library preparation and sequencing

For each sample, amplicons were pooled in equimolar concentrations for input into the NEBNext Ultra DNA Library Prep Kit for Illumina (New England BioLabs, Ipswich, MA). Prepared libraries were pooled in equimolar concentrations to a final concentration of 1 nM, and run using an Illumina MiSeq Reagent Kit v2 (Illumina, San Diego, CA) for 500 cycles (2 x 250 bp). Demultiplexed files were output in FASTQ format.

Processing of raw sequence data, mapping, and variant calling

Human reads were removed from raw FASTQ files by mapping to the human reference genome GRCH38 with bowtie2[24] version 2.3.2 (http://bowtie-bio.sourceforge.net/bowtie2/index.shtml). Reads that did not map to human genome were output to separate FASTQ files and used for all subsequent analyses. Illumina data was analyzed using the pipeline described in detail at https://github.com/lmoncla/illumina_pipeline. Briefly, raw FASTQ files were trimmed using Trimmomatic[23] (http://www.usadellab.org/cms/?page=trimmomatic), trimming in sliding windows of 5 base pairs and requiring a minimum Q-score of 30. Reads that were trimmed to a length of <100 base pairs were discarded. Trimming was performed with the following command: java -jar Trimmomatic-0.36/trimmomatic-0.36.jar SE input.fastq output.fastq SLIDINGWINDOW:5:30 MINLEN:100. Trimmed reads were mapped to consensus sequences previously derived[21] using bowtie2[24] version 2.3.2 (http://bowtie-bio.sourceforge.net/bowtie2/index.shtml), using the following command: bowtie2 -x reference_sequence.fasta -U read1.trimmed.fastq,read2.trimmed.fastq -S output.sam --local. Duplicate reads were removed with Picard (http://broadinstitute.github.io/picard/) with: java -jar picard.jar MarkDuplicates I=input.sam O=output.sam REMOVE_DUPLICATES=true. Mapped reads were imported into Geneious (https://www.geneious.com/) for visual inspection and consensus calling, with nucleotide sites with <100x coverage called as Ns. To avoid issues with mapping to an improper reference sequence, we then remapped each sample's trimmed FASTQ files to its own consensus sequence. These bam files were again manually inspected in Geneious, and a final consensus sequence was called. We were able to generate full-genome data for all samples except for A/Cambodia/X0128304/2013, for which we were lacked data for PB1. These BAM files were then exported and converted to mpileup files with samtools[25] (http://samtools.sourceforge.net/), and within-host variants were called using VarScan[26,27] (http://varscan.sourceforge.net/). For a variant to be reported, we required the variant site to be sequenced to a depth of at least 100x with a mean quality of Q30, and for the variant to be detected in both forward and reverse reads at a frequency of at least 1%. We called variants using the following command: java -jar VarScan.v2.3.9.jar mpileup2snp input.pileup --min-coverage 100 --min-avg-qual 30 --min-var-freq 0.01 --strand-filter 1 --output-vcf 1 > output.vcf. VCF files were parsed and annotated with coding region changes using custom software available here (https://github.com/blab/h5n1-cambodia/blob/master/scripts/H5N1_vcf_parser.py). All amino acid changes for HA are reported and plotted using native H5 numbering, including the signal peptide, which is 16 amino acids in length. For ease of comparison, some amino acid changes are also reported with mature H5 peptide numbering in the manuscript when indicated, and in **Table 2**.

General availability of analysis software and data

All code used to analyze data and generate figures for this manuscript are publicly available at https://github.com/blab/h5n1-cambodia. Raw FASTQ files with human reads removed are available under SRA accession number PRJNA547644, and accessions SRX5984186-SRX5984198. All reported variant calls and phylogenetic trees are available at https://github.com/blab/h5n1-cambodia/tree/master/data.

Phylogenetic reconstruction

We downloaded all currently available H5N1 genomes from the EpiFlu Database of the Global Initiative for Sharing All Influenza Data[28,29] (GISAID, https://www.gisaid.org/) and all currently available full H5N1 genomes from the Influenza Research Database (IRD, http://www.fludb.org)[30] and added consensus genomes from our 5 duck samples and 8 human samples. Sequences and metadata were cleaned and organized using fauna (https://github.com/nextstrain/fauna), a database system part of the Nextstrain platform. Sequences were then processed using Nextstrain's augur software[31] (https://github.com/nextstrain/augur). Sequences were filtered by length to remove short sequences using the following length filters: PB2: 2100 bp, PB1: 2100 bp, PA: 2000 bp, HA: 1600 bp, NP: 1400 bp, NA: 1270 bp, MP: 900 bp, and NS: 800 bp. We excluded sequences with sample collection dates prior to 1996, and those for which the host was annotated as laboratory derived, ferret, or unknown. We also excluded sequences for which the country or geographic region was unknown. Sequences for each gene were aligned using MAFFT[32], and then trimmed to the reference sequence. We chose the A/Goose/Guangdong/1/96(H5N1) genome (GenBank accession numbers: AF144300-AF144307) as the reference genome. IQTREE[33,34] was then used to infer a maximum likelihood phylogeny, and TreeTime[35] was used to infer a molecular clock and temporally-resolved phylogeny. Tips which fell outside of 4 standard deviations away from the inferred molecular clock were removed. Finally, TreeTime[35] was used to infer ancestral sequence states at internal nodes and the geographic migration history across the phylogeny. We inferred migration among 9 defined geographic regions, China, Southeast Asia, South Asia, Japan and Korea, West Asia, Africa, Europe, South America, and North America. Our final trees are available at https://github.com/blab/h5n1-cambodia/tree/master/data/tree-jsons, and include the following number of sequences: PB2: 4063, PB1: 3867, PA: 4082, HA: 6431, NP: 4070, NA: 5357, MP: 3940, NS: 3678. Plotting was performed using baltic (https://github.com/evogytis/baltic).

Diversity analyses

Within-host variants were called as described above, requiring a minimum coverage of 100x, a minimum frequency of 1%, a minimal base quality score of Q30, and detection on both forward and reverse reads. Variants were annotated as nonsynonymous or synonymous. For each sample, we computed the number of synonymous and nonsynonymous sites for each coding region with SNPGenie[36,37] (https://github.com/chasewnelson/SNPGenie). For each sample and coding region, we then computed π_N as the number of nonsynonymous mutations per nonsynonymous site, and π_S as the number of synonymous mutations per synonymous site. Bars shown in **Fig. 1c** and values in **Supplementary Table 1** depict mean π_N (dark colors) or π_S (light colors) when values were combined across all humans (red bars) or ducks (blue bars). Error bars represent the standard deviations.

Comparison to functional sites

We used the Sequence Feature Variant Types tool from the Influenza Research Database[30] to download all currently available annotations for H5 hemagglutinins, N1 neuraminidases, and all subtypes for the remaining gene segments. We then annotated each within-host SNV identified in our dataset that fell within an annotated region or site. The complete results of this annotation are available in **Supplementary Table 2**. We next filtered our annotated SNVs to include only those located in sites involved in “host-specific” functions or interactions, i.e., those that are distinct between human and avian hosts. We defined host-specific functions/interactions as receptor binding, interaction with host cellular machinery, nuclear import and export, immune antagonism, 5' cap binding, temperature sensitivity, and glycosylation. We also included sites that have been phenotypically identified as determinants of transmissibility and virulence. Sites that participate in binding interactions with other viral subunits or vRNP, conserved active site domains, drug resistance mutations, and epitope sites were not categorized as host-specific for this analysis. We annotated both synonymous and nonsynonymous mutations in our dataset, but only highlight nonsynonymous changes in **Fig. 2** and **Table 2**.

Shared sites permutation test

To test whether humans or duck samples shared more polymorphisms than expected by chance, we performed a permutation test. We first counted the number of sites, *n*, in which an SNV altered amino acid used, across coding regions and samples. For example, if two SNVs occurred in the same codon site, we counted this as 1 variable amino acid site. Next, for each gene and sample, we calculated the number of amino acid sites that were covered with with sufficient sequencing depth that a mutation could have been called using our SNV calling criteria. To do this, we calculated the length in amino acids of each coding region, *L*, that was covered by at least 100 reads. Non-coding regions were not included. For each coding region and sample, we then simulated the effect of having *n* variable amino acid sites placed randomly along the coding region between sites 1 to *L*, and recorded the site where the polymorphism was placed. After simulating this for each gene and sample, we counted the number of sites that were shared between at least 2 human or at least 2 duck samples. This process was repeated 100,000 times. The number of shared polymorphisms at each iteration was used to generate a null distribution, as shown in **Fig. 3b**. We calculated p-values as the number of iterations for which there were at least as many shared sites as observed in our actual data, divided by

100,000. For the simulations displayed in **Fig. 3c** and **Fig. 3d**, we wanted to simulate the effect of genomic constraint, meaning that only some fraction of the genome could tolerate mutation. For these analyses, simulations were done exactly the same, except that the number of sites at which a mutation could occur was reduced to 70% (**Fig. 3c**) or 50% (**Fig. 3d**). Code for performing the shared sites permutation test is freely available at https://github.com/blab/h5n1-cambodia/blob/master/figures/figure-5b-shared-sites-permutation-test.ipynb.

Reconstruction of host transitions along the phylogeny

We used the phylogenetic trees in **Supplementary Figure 2** to infer host transitions along each gene's phylogeny. As described above, we used TreeTime[35] to reconstruct ancestral nucleotide states at each internal node and infer amino acid mutations along each branch along these phylogenetic trees. We then classified host transition mutations along branches that lead to human or avian tips as follows (**Fig. 4a**). For each branch in the phylogeny, we enumerated all tips descending from that branch. If all descendent tips were human, we considered this a monophyletic human clade. If the current branch's ancestral node also led to only human descendants, we labelled the current branch a "human-to-human" branch. If a branch leading to a monophyletic human clade had an ancestral node that included avian and human descendants, then we considered the current branch a "avian-to-human" branch. All other branches were considered "avian-to-avian" branches. We did not explicitly allow for human-to-avian branches in this analysis. Because avian sampling is poor relative to human sampling, and because H5N1 circulation is thought to be maintained by transmission in birds, we chose to only label branches explicitly leading to human infections as human branches. We also reasoned that for instances in which a human tip appears to be ancestral to an avian clade, this more likely results from poor avian sampling than from true human-to-avian transmission. Using these criteria, we then gathered the inferred amino acid mutations that occurred along each branch in the phylogeny, and counted the number of times they were associated with each type of host transition. We then queried each SNV detected within-host in our dataset, in both human and duck samples, to determine the number of host transitions that they occurred on in the phylogeny, as shown in **Fig. 4b**. For ease of plotting and viewing, we combined counts for avian to human and human-to-human transitions. To test whether individual mutations were enriched along branches leading to human infections, we performed Fisher's exact tests comparing the number of avian-to-avian and avian/human-to-human transitions along which the mutation was detected vs. the overall number of avian-to-avian and avian/human-to-avian transitions that were observed along the tree. Mutations that showed statistically significant enrichment are annotated in **Fig. 4b**.

Results

Sample selection and dataset information

We analyzed full-genome sequence data from primary, influenza-confirmed samples from infected humans and domestic ducks from Cambodia (**Table 1**). Four domestic duck samples (pooled organs) were collected as part of poultry outbreak investigations, while one was collected during live bird market surveillance (pooled throat and cloacal swab). All human samples (throat swabs) were collected via event-based surveillance upon admittance to various hospitals throughout Cambodia[21]. Because of limited sample availability and long storage times, generating duplicate sequence data for each sample was not possible. We therefore focused on samples whose viral RNA copy numbers were ≥10^3 copies/µl as assessed by RT-qPCR (**Table 1**), and whose mean coverage depth exceeded 100x (**Supplementary Figure 1**).

We analyzed full genome data for 7 human and 5 duck samples, and near complete genome data for A/Cambodia/X0128304/2013, for which we lack data from the PB1 gene.

H5 viruses circulating in Cambodia were exclusively clade 1.1.2[4] until 2013, when a novel reassortant virus emerged[38]. This reassortant virus expressed a hemagglutinin (HA) and neuraminidase (NA) from clade 1.1.2, with internal genes from clade 2.3.2.1a[21]. All 2013/2014 samples in our dataset come from this outbreak, while samples collected prior to 2013 are clade 1.1.2 (**Table 1** and **Supplementary Figure 2**). The 2013 reassortant viruses share 4 amino acid substitutions in HA, S123P, S133A, S155N, and K266R[21] (H5, mature peptide numbering). S133A and S155N have been linked to improved α-2,6 linked sialic acid binding, independently and in combination with S123P[39–41]. All samples encode a polybasic cleavage site in HA (XRRKRR) between amino acids 325-330 (H5, mature peptide numbering), a virulence determinant for H5N1 AIVs[42,43], and a 20 amino acid deletion in NA. This NA deletion is a well-documented host range determinant[44–47].

Using this subset of 8 human and 5 duck samples, we aimed to determine whether positive selection would promote adaptation in humans. Positive selection increases the frequency of beneficial variants, and is often identified by tracking mutations' frequencies over time. While multiple time points were not available in our dataset, all human samples were collected 5-12 days after reported symptom onset[21]. Animal infection studies have observed drastic changes in within-host variant frequencies in 3-7 days[11,13], suggesting that 5-12 days post symptom onset may provide sufficient time to observe within-host evolution. We reasoned that strong within-host positive selection should result in the following patterns: (1) Positive selection should increase the frequencies of human-adaptive mutations during human infection. Therefore, viral populations in humans should exhibit more high-frequency polymorphisms, and a higher mean variant frequency, than viral populations in ducks. (2) Viruses in humans should harbor mutations phenotypically linked to mammalian adaptation. (3) Viruses in humans should exhibit evidence for convergent evolution, i.e., the same mutation arising across multiple samples. (4) Variants arising within humans should be enriched among viruses leading to human infections on the H5N1 phylogeny.

Purifying selection predominates in humans and ducks

We called within-host variants across the genome that were present in ≥1% of sequencing reads and occurred at a site with a minimum read depth of 100x and a minimum quality score of Q30 (see Methods for details). All coding region changes are reported using native H5 numbering, including the signal peptide, unless otherwise noted. Most single nucleotide variants (SNVs) were present at low frequencies (**Fig. 1a**). We identified a total of 198 SNVs in humans (111 nonsynonymous, 91 synonymous, 4 missense) and 40 in ducks (16 nonsynonymous, 23 synonymous, 1 missense). Human samples had more SNVs than duck samples on average (mean SNVs per sample: humans = 26 ± 19, ducks = 8 ± 3, $p = 2.79 \times 10^{-17}$, Fisher's exact test), although the number of SNVs per sample was variable among samples in both species (**Supplementary Figure 3**). To determine whether humans had more high-frequency variants than ducks, we generated a site frequency spectrum (**Fig. 1b**). Purifying selection removes new variants from the population, generating an excess of low-frequency variants, while positive selection promotes accumulation of high-frequency polymorphisms. Exponential population expansion also causes excess low-frequency variation; however, while selection disproportionately affects nonsynonymous variants, demographic factors affect synonymous and nonsynonymous variants equally. In both humans and ducks, over 80% of variants (both synonymous and nonsynonymous) were present in <10% of the population, and the distribution of SNV frequencies were strikingly similar (**Fig. 1b**). The mean SNV frequency in human (5.8%) and duck samples (6.6%) were not statistically different (p=0.11, Mann Whitney U test).

Comparing nonsynonymous (π_N) and synonymous (π_S) polymorphism in a population is another common measure for selection. An excess of synonymous polymorphism ($\pi_N/\pi_S < 1$) indicates purifying selection, an excess of nonsynonymous variation ($\pi_N/\pi_S > 1$) suggests positive selection, and approximately equal rates ($\pi_N/\pi_S \sim 1$) suggest that genetic drift is the predominant force shaping diversity. We calculated the number of synonymous and nonsynonymous variants for each gene in each sample, and normalized these counts to the number of synonymous and nonsynonymous sites. In both species, most genes exhibited $\pi_N < \pi_S$, although there was substantial variation among samples (**Supplementary Table 1** and **Fig. 1c**). The difference between π_s and π_N was generally not statistically significant (**Supplementary Table 1**). The exception was human M2 (π_N = 0.0028, π_S = 0, p = 0.049, paired t-test) and NA (πN/πS = 0.21, p = 0.033, paired t-test), which exhibited evidence of purifying selection. When diversity estimates across all genes were combined, both species exhibited $\pi_N/\pi_S < 1$ (**Fig. 1c**) (human π_N/π_S= 0.41, p = 0.00028, unpaired t-test; duck π_N/π_S= 0.29, p = 0.022, unpaired t-test). Taken together, our data suggest that H5N1 within-host populations in both humans and ducks are broadly shaped by a combination of purifying selection, population growth, and genetic drift. We do not find evidence for widespread positive selection in any individual coding region.

SNVs are identified in humans at functionally relevant sites

Influenza phenotypes can be drastically altered by single amino acid changes. We took advantage of the Influenza Research Database[29] Sequence Feature Variant Types tool, a catalogue of amino acids critical to protein structure and function and experimentally linked to functional alteration. We downloaded all available annotations for H5 HAs, N1 NAs, and all subtypes for the remaining proteins, and annotated each mutation in our dataset that fell within an annotated region (**Supplementary Table 2**). We then filtered these annotated amino acids to include only those located in sites involved in host-specific functions (see Methods for details).

Of the 218 unique, polymorphic amino acid sites in our dataset, we identified 34 nonsynonymous mutations at sites involved in viral replication, receptor binding, virulence, and interaction with host cell machinery (**Fig. 2**). Some sites are explicitly linked to H5N1 mammalian adaptation (**Table 2**). PB2 E627K was detected as a minor variant in A/Cambodia/W0112303/2012, and in A/Cambodia/V0417301/2011 at consensus. A lysine at position 627 is a conserved marker of human adaptation[47,48] that enhances H5N1 replication in mammals[11,12,47,49]. A/Cambodia/W0112303/2012 also encoded PB2 D701N at consensus. Curiously, this patient also harbored the reversion mutation, N701D, at low-frequency within-host. An asparagine (N) at PB2 701 enhances viral replication and transmission in mammals[50,51], while an aspartate (D) is commonly identified in birds. We cannot distinguish whether the founding virus harbord an asparagine or aspartate, so our data are consistent with two possibilities: transmission of a virus harboring asparagine and within-host generation of aspartate; or, transmission of a virus with asparate followed by within-host selection but incomplete fixation of asparagine. All other human and avian samples in our dataset encoded the “avian-like” amino acids, glutamate at PB2 627, and aspartate at PB2 701. None of the adaptive polymerase mutations that recently identified by Welkers et al.[17] in H5N1-infected humans in Indonesia were present in our samples, nor were any of the human-adaptive mutations identified in a recent deep mutational scan of PB2[52].

We also identified HA mutations linked to human receptor binding. Two human samples encoded an HA A150V mutation (134 in mature, H5 peptide numbering, **Fig. 2**). A valine at HA 150 improves α-2,6 linked sialic acid binding in H5N1 viruses[53,54], and was also identified in H5N1-infected humans in Vietnam[16]. Finally, HA Q238L was detected in A/Cambodia/V0417301/2011 and A/Cambodia/V0401301/2011. HA 238L (222 in mature, H5 peptide numbering) was shown in H5N1 transmission studies to confer a switch from α-2,3 to α-2,6 linked sialic acid binding[11] and mediate transmission[11,12]. An HA Q238R mutation was identified in A/Cambodia/X0125302/2013, although nothing is known regarding an arganine (R) at this site.

Mutations annotated as host-specific were not detected at higher frequencies than non-host-specific mutations (mean frequency for host-specific mutations = 6.8% ± 7.5%, mean frequency for non-host-specific mutations = 5.5% ± 5.4%, p-value = 0.129, unpaired t-test). All 8 human samples harbored at least 1 mutant in a host-specific site. Critically though, the functional impacts of influenza mutations strongly depend on sequence context[55], and we did not phenotypically test these mutations. We caution that confirming functional impacts for these mutations would require further study. Still, our data show that putative human-adapting mutations are generated during natural spillover. Our results also highlight that even mutations that have been predicted to be strongly beneficial (e.g., PB2 627K and HA 238L) may remain at low frequencies in vivo.

Shared diversity is limited

Each human H5N1 infection is thought to represent a unique avian spillover event. If selection is strong at a given site in the genome, then mutations may arise at that site independently across multiple patients. We identified 13 amino acid sites in our dataset that were polymorphic in at least 2 samples, 4 of which were detected in both species (PB1 371, PA 307, HA 265 and NP 201). Of the 34 unique polymorphic amino acid sites in ducks, 3 sites were shared by at least 2 duck samples; of the 188 unique polymorphic amino acid sites in humans, 9 were shared by at least 2 human samples (**Fig. 3a**). Two of these shared sites, HA 150 and HA 238, are linked to human-adapting phenotypes **(Table 2)**. To determine whether the number of shared sites we observe is more or less than expected by chance, we performed a permutation test. For each species, we simulated datasets with the same number of sequences and amino acid polymorphisms as our actual dataset, but assigned each polymorphism to a random amino acid site. For each iteration, we then counted the number of polymorphic sites that were shared by ≥2 samples. We ran this simulation for 100,000 iterations for each species, and used the number of shared sites per iteration to generate a null distribution (**Fig. 3b**, colored bars). Comparison to the observed number of shared sites (3 and 9, dashed lines in **Fig. 3b**), confirmed that humans share slightly more polymorphisms than expected by chance ($p = 0.046$), while ducks share significantly more ($p = 0.00006$).

Viral genomes are highly constrained[56], which could account for the convergence we observe. To test this, we repeated our simulations to restrict the number of amino acid sites that could tolerate a mutation to 70% or 50%. When 70% of the coding region was permitted to mutate, ~23% of simulations resulted in ≥9 shared sites in humans ($p = 0.23$), and when 50% of the genome was permitted to mutate, ~61% of simulations resulted in ≥9 shared sites ($p = 0.608$). In contrast, the probability of observing 3 shared sites among duck samples remained low regardless of genome constraint (70% of genome tolerates mutation: p = 0.00014; 50% of genome tolerates mutation: p = 0.00051), suggesting a significant, although low, level of convergence. Our results suggest that the shared sites we observe in humans could be explained by genome constraint. However, given the presence of functionally relevant shared polymorphisms in humans, we speculate that the shared diversity we observe reflects a combination of host-specific positive selection at isolated sites, amongst a background of genomic constraint.

Within-host SNVs are not enriched on spillover branches

If within-host mutations are human-adapting, then those mutations should be enriched among H5N1 viruses that have caused human infections in the past. To test this hypothesis, we inferred full genome phylogenies using all available full-genome H5N1 viruses from the EpiFlu[28,29] and IRD[30] databases (**Supplementary Figure 2)**, reconstructed ancestral nucleotide states at each internal node, and inferred amino acid mutations along each branch. We then classified host transition mutations along branches that led to human or avian tips (**Fig. 4a)**. If a branch fell within a clade that included only human tips, that branch was labelled as a human-to-human transition. If a branch led to a human-only clade but its ancestral branch included avian descendants, this was labelled as an avian-to-human transition. All other transitions were labelled avian-to-avian (**Fig. 4a,** see Methods for more details). We then curated the mutations that occurred on each type of host transition, and compared these counts to the mutations identified within-host in our dataset.

Of the 120 nonsynonymous within-host SNVs we identified in our dataset, 60 (50%) were not detected in the phylogeny at all. This suggests that many of the mutations generated within-host are likely deleterious, and are purged from the H5N1 population over time. Additionally, because humans are generally dead-end hosts for H5N1, even human-adapting variants arising within-host are likely to be lost due to terminal human transmission chains. Of the within-host mutations that were detected on the phylogeny, most occurred on branches leading to avian infections (**Fig 4b,** blue bars). However, there were a few exceptions (**Fig 4b,** red bars). Across the phylogeny, we enumerated a total of 31,939 avian-to-avian transitions, and 2,787 human/avian-to-human transitions, so that we expect a 11.46:1 ratio of avian-to-avian transitions relative to human/avian-to-human transitions. In contrast, PB2 E627K was heavily enriched among human infections, detected on 15 avian-to-avian transitions and 36 human/avian-to-human transitions ($p = 4.21 \times 10^{-28}$, Fisher's exact test). HA A150V was detected in only one avian-to-avian transition, but in 8 human/avian-to-human transitions ($p = 1.46 \times 10^{-8}$, Fisher's exact test), and HA N198S was detected on 4 avian-to-avian transitions and 3 avian-to-human transitions ($p = 0.014$, Fisher's exact test). Although nothing is known regarding a serine at HA 198, a lysine at that site can confer α-2,6-linked sialic acid binding[39,57]. Taken together, these data suggest that the majority of mutations detected within-host are not associated with human spillover. However, they agree with selection for human-adapting phenotypes at a small subset of sites (PB2 E627K, HA A150V, HA N198S).

Discussion

Our study utilizes a unique dataset of to quantify how viruses H5N1 evolve during natural spillover infection. We find that purifying selection, population growth, and genetic drift broadly shape viral diversity in both hosts. Half of the within-host variants identified within-host are never detected in the H5N1 phylogeny and are likely deleterious. We detect putative human-adapting mutations (PB2 E627K, HA A150V, and HA Q238L) during human infection, two of which arose multiple times. PB2 E627K and HA A150V are enriched along phylogenetic branches leading to human infections, supporting their potential role in human adaptation. Our data show that during spillover, H5N1 viruses have the capacity to generate well-known markers of mammalian adaptation in multiple, independent hosts. However, they also highlight that within-host diversity is shaped heavily by purifying selection and randomness as these markers do not reach high-frequency during a single spill-over human infection. We speculate that during spillover, short infection times, randomness, and purifying selection may together limit the capacity of H5N1 viruses to evolve within-host.

Although data from spillovers are limited, our results align with data from Vietnam[16] and Indonesia[17]. Welkers et al.[17] identified markers of mammalian replication (PB2 627K) and transmission (HA 220K) in humans, but found that adaptive markers were not widespread. Welkers et al. also characterized new mutations that improved human replication, suggesting that there are yet undiscovered pathways for adaptation. Imai et al.[16] characterized SNVs in H5N1-infected humans that altered viral replication, receptor binding, and interferon antagonism, but these mutations stayed at low frequencies. Imai et al. also showed that most within-host variants elicited neutral or deleterious effects on protein function in humans, aligning with the widespread purifying selection we detect within-host, and the absence of ~50% of within-host variants in the phylogeny. These findings also agree with predictions by Russell et al.[14], who hypothesized that H5N1 viruses would generate human-adapting mutations during infection, but that these mutations would remain at low frequencies and fail to be transmitted.

One unexpected result is that mutations that hypothesized to be strongly beneficial, like PB2 627K and HA 238L, remained low-frequency during infection. These mutations could have arisen late in infection or been linked to deleterious mutations. Additionally, epistasis is crucial to influenza evolution, and mutations that promote human adaptation in one background may not be well-tolerated in others. PB2 E627K is widespread among clade 2.2.1 H5N1 viruses, but only sparsely detected in other H5N1 clades. Soh et al.[52] recently uncovered strongly human-adapting PB2 mutations that are rare in nature, likely because they are inaccessible via single site mutations. Genetic background plays a vital role in determining how AIVs evolve, and may at least partially explain our findings. Importantly, our study involves a small number of samples from a single geographic location, and two H5N1 clades. Continued characterization of H5N1 spillover in other clades is necessary to define whether our observations are generalizable across H5N1 outbreaks.

Assessing zoonotic risk is critical but challenging. By quantifying within-host selection, identifying mutations at adaptive sites, measuring convergent evolution, and comparing within-host diversity to long-term evolution, we can assemble a nuanced understanding of AIV evolution. These methods provide a foundation for understanding cross-species transmission that can readily be applied to other avian influenza datasets, as well as newly emerging zoonotic viruses.

References

1. Organization, W. H. *Cumulative number of confirmed human cases for avian influenza A(H5N1) reported to WHO, 2003-2018*.
2. Chen, H. *et al.* Establishment of multiple sublineages of H5N1 influenza virus in Asia: Implications for pandemic control. *Proceedings of the National Academy of Sciences* (2006). doi:10.1073/pnas.0511120103
3. Nguyen, D. T. *et al.* Shifting Clade Distribution, Reassortment, and Emergence of New

Subtypes of Highly Pathogenic Avian Influenza A(H5) Viruses Collected from Vietnamese Poultry from 2012 to 2015. *J. Virol.* (2017). doi:10.1128/JVI.01708-16

4. Horm, S. V. *et al.* Intense circulation of A/H5N1 and other avian influenza viruses in Cambodian live-bird markets with serological evidence of sub-clinical human infections. *Emerg. Microbes Infect.* (2016). doi:10.1038/emi.2016.69
5. Nobusawa, E. & Sato, K. Comparison of the Mutation Rates of Human Influenza A and B Viruses. *J. Virol.* **80**, 3675–3678 (2006).
6. Parvin, J. D., Moscona, A., Pan, W. T., Leider, J. M. & Palese, P. Measurement of the mutation rates of animal viruses: influenza A virus and poliovirus type 1. *J. Virol.* **59**, 377–383 (1986).
7. Pauly, M. D., Procario, M. C. & Lauring, A. S. A novel twelve class fluctuation test reveals higher than expected mutation rates for influenza A viruses. *Elife* **6**, (2017).
8. Suárez, P., Valcárcel, J. & Ortín, J. Heterogeneity of the mutation rates of influenza A viruses: isolation of mutator mutants. *J. Virol.* **66**, 2491–2494 (1992).
9. Baccam, P., Beauchemin, C., Macken, C. A., Hayden, F. G. & Perelson, A. S. Kinetics of influenza A virus infection in humans. *J. Virol.* **80**, 7590–7599 (2006).
10. Imai, M. *et al.* Experimental adaptation of an influenza H5 HA confers respiratory droplet transmission to a reassortant H5 HA/H1N1 virus in ferrets. *Nature* **486**, 420–428 (2012).
11. Linster, M. *et al.* Identification, Characterization, and Natural Selection of Mutations Driving Airborne Transmission of A/H5N1 Virus. *Cell* **157**, 329–339 (2014).
12. Herfst, S. *et al.* Airborne transmission of influenza A/H5N1 virus between ferrets. *Science* **336**, 1534–1541 (2012).
13. Wilker, P. R. *et al.* Selection on haemagglutinin imposes a bottleneck during mammalian transmission of reassortant H5N1 influenza viruses. *Nat. Commun.* **4**, 2636 (2013).
14. Russell, C. A. *et al.* The Potential for Respiratory Droplet–Transmissible A/H5N1 Influenza

Virus to Evolve in a Mammalian Host. *Science* **336**, 1541–1547 (2012).

15. Sigal, D., Reid, J. N. S. & Wahl, L. M. Effects of Transmission Bottlenecks on the Diversity of Influenza A Virus. *Genetics* **210**, 1075–1088 (2018).
16. Imai, H. *et al.* Diversity of Influenza A(H5N1) Viruses in Infected Humans, Northern Vietnam, 2004–2010. *Emerg. Infect. Dis.* **24**, 1128–1238 (2018).
17. Welkers, M. R. A. *et al.* Genetic diversity and host adaptation of avian H5N1 influenza viruses during human infection. *Emerg. Microbes Infect.* **8**, 262–271 (2019).
18. Milani, A. *et al.* Viral population diversity in vaccinated poultry host infected with H5N1 highly pathogenic avian influenza virus. *Int. J. Infect. Dis.* **53**, 104 (2016).
19. Iqbal, M. *et al.* Within-host variation of avian influenza viruses. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **364**, 2739–2747 (2009).
20. Gutiérrez, R. A., Viari, A., Godelle, B. & Buchy, P. Biased mutational pattern and quasispecies hypothesis in H5N1 virus. *Infect. Genet. Evol.* **15**, 69–76 (2013).
21. Rith, S. *et al.* Identification of Molecular Markers Associated with Alteration of Receptor-Binding Specificity in a Novel Genotype of Highly Pathogenic Avian Influenza A(H5N1) Viruses Detected in Cambodia in 2013. *J. Virol.* **88**, 13897–13909 (2014).
22. Horwood, P. F. *et al.* Co-circulation of Influenza A H5, H7, and H9 Viruses and Co-infected Poultry in Live Bird Markets, Cambodia. *Emerg. Infect. Dis.* **24**, 352–355 (2018).
23. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114–2120 (2014).
24. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**, 357–359 (2012).
25. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
26. Koboldt, D. C. *et al.* VarScan: variant detection in massively parallel sequencing of

individual and pooled samples. *Bioinformatics* **25**, 2283–2285 (2009).

27. Koboldt, D. C. *et al.* VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome Res.* **22**, 568–576 (2012).
28. Elbe, S. & Buckland-Merrett, G. Data, disease and diplomacy: GISAID's innovative contribution to global health. *Global Challenges* **1**, 33–46 (2017).
29. Shu, Y. & McCauley, J. GISAID: Global initiative on sharing all influenza data – from vision to reality. *Eurosurveillance* **22**, 30494 (2017).
30. Zhang, Y. *et al.* Influenza Research Database: An integrated bioinformatics resource for influenza virus research. *Nucleic Acids Res.* **45**, D466–D474 (2017).
31. Hadfield, J. *et al.* Nextstrain: real-time tracking of pathogen evolution. *Bioinformatics* **34**, 4121–4123 (2018).
32. Katoh, K., Misawa, K., Kuma, K. K.-I. & Miyata, T. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Res.* **30**, 3059–3066 (2002).
33. Nguyen, L.-T., Schmidt, H. A., von Haeseler, A. & Minh, B. Q. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol. Biol. Evol.* **32**, 268–274 (2015).
34. Chernomor, O., von Haeseler, A. & Minh, B. Q. Terrace Aware Data Structure for Phylogenomic Inference from Supermatrices. *Syst. Biol.* **65**, 997–1008 (2016).
35. Sagulenko, P., Puller, V. & Neher, R. A. TreeTime: Maximum-likelihood phylodynamic analysis. *Virus Evolution* **4**, (2018).
36. Nelson, C. W., Moncla, L. H. & Hughes, A. L. SNPGenie: estimating evolutionary parameters to detect natural selection using pooled next-generation sequencing data. *Bioinformatics* **31**, 3709–3711 (2015).
37. Nei, M. & Gojobori, T. Simple methods for estimating the numbers of synonymous and

nonsynonymous nucleotide substitutions. *Mol. Biol. Evol.* **3**, 418–426 (1986).

38. Sorn, S. *et al.* Dynamic of H5N1 virus in Cambodia and emergence of a novel endemic sub-clade. *Infect. Genet. Evol.* **15**, 87–94 (2013).

39. Yamada, S. *et al.* Haemagglutinin mutations responsible for the binding of H5N1 influenza A viruses to human-type receptors. *Nature* **444**, 378–382 (2006).

40. Yang, Z.-Y. *et al.* Immunization by avian H5 influenza hemagglutinin mutants with altered receptor binding specificity. *Science* **317**, 825–828 (2007).

41. Wang, M. *et al.* Residue Y161 of influenza virus hemagglutinin is involved in viral recognition of sialylated complexes from different hosts. *J. Virol.* **86**, 4455–4462 (2012).

42. Suguitan, A. L. *et al.* The multibasic cleavage site of the hemagglutinin of highly pathogenic A/Vietnam/1203/2004 (H5N1) avian influenza virus acts as a virulence factor in a host-specific manner in mammals. *J. Virol.* **86**, 2706–2714 (2012).

43. Schrauwen, E. J. A. *et al.* The multibasic cleavage site in H5N1 virus is critical for systemic spread along the olfactory and hematogenous routes in ferrets. *J. Virol.* **86**, 3975–3984 (2012).

44. Zhou, H. *et al.* The Special Neuraminidase Stalk-Motif Responsible for Increased Virulence and Pathogenesis of H5N1 Influenza A Virus. *PLoS One* **4**, e6277 (2009).

45. Zhou, H., Jin, M., Chen, H., Huag, Q. & Yu, Z. Genome-sequenee Analysis of the Pathogenic H5N1 Avian Influenza A Virus Isolated in China in 2004. *Virus Genes* **32**, 85–95 (2006).

46. Matsuoka, Y. *et al.* Neuraminidase Stalk Length and Additional Glycosylation of the Hemagglutinin Influence the Virulence of Influenza H5N1 Viruses for Mice. *J. Virol.* **83**, 4704–4708 (2009).

47. Hatta, M., Gao, P., Halfmann, P. & Kawaoka, Y. Molecular basis for high virulence of Hong Kong H5N1 influenza A viruses. *Science* **293**, 1840–1842 (2001).

48. Subbarao, E. K., Kawaoka, Y. & Murphy, B. R. Rescue of an influenza A virus wild-type PB2 gene and a mutant derivative bearing a site-specific temperature-sensitive and attenuating mutation. *J. Virol.* **67**, 7223–7228 (1993).

49. Le, Q. M., Sakai-Tagawa, Y., Ozawa, M., Ito, M. & Kawaoka, Y. Selection of H5N1 influenza virus PB2 during replication in humans. *J. Virol.* **83**, 5278–5281 (2009).

50. Gabriel, G. *et al.* The viral polymerase mediates adaptation of an avian influenza virus to a mammalian host. *Proc. Natl. Acad. Sci. U. S. A.* **102**, 18590–18595 (2005).

51. Steel, J., Lowen, A. C., Mubareka, S. & Palese, P. Transmission of Influenza Virus in a Mammalian Host Is Increased by PB2 Amino Acids 627K or 627E/701N. *PLoS Pathog.* **5**, e1000252 (2009).

52. Soh, Y. Q. S., Moncla, L. H., Eguia, R., Bedford, T. & Bloom, J. D. Comprehensive mapping of adaptation of the avian influenza polymerase protein PB2 to humans. *Elife* **8**, (2019).

53. Auewarakul, P. *et al.* An avian influenza H5N1 virus that binds to a human-type receptor. *J. Virol.* **81**, 9950–9955 (2007).

54. Naughtin, M. *et al.* Neuraminidase inhibitor sensitivity and receptor-binding specificity of Cambodian clade 1 highly pathogenic H5N1 influenza virus. *Antimicrob. Agents Chemother.* **55**, 2004–2010 (2011).

55. Lyons, D. M. & Lauring, A. S. *Mutation and epistasis in influenza virus evolution*. *Viruses* **10**, (2018).

56. Visher, E., Whitefield, S. E., McCrone, J. T., Fitzsimmons, W. & Lauring, A. S. The Mutational Robustness of Influenza A Virus. *PLoS Pathog.* **12**, e1005856 (2016).

57. Watanabe, Y. *et al.* Acquisition of Human-Type Receptor Binding Specificity by New H5N1 Influenza Virus Sublineages during Their Emergence in Birds in Egypt. *PLoS Pathog.* **7**, e1002068 (2011).

58. Treanor, J., Perkins, M., Battaglia, R. & Murphy, B. R. Evaluation of the genetic stability of

the temperature-sensitive PB2 gene mutation of the influenza A/Ann Arbor/6/60 cold-adapted vaccine virus. *J. Virol.* **68**, 7684–7688 (1994).

59. Guilligay, D. *et al.* The structural basis for cap binding by influenza virus polymerase subunit PB2. *Nat. Struct. Mol. Biol.* **15**, 500–506 (2008).

60. Nerome, R. *et al.* Evolutionary characterization of the six internal genes of H5N1 human influenza A virus. *J. Gen. Virol.* **81**, 1293–1303 (2000).

61. Xu, L. *et al.* Genomic Polymorphism of the Pandemic A (H1N1) Influenza Viruses Correlates with Viral Replication, Virulence, and Pathogenicity In Vitro and In Vivo. *PLoS One* **6**, e20698 (2011).

62. Bussey, K. A. *et al.* PA Residues in the 2009 H1N1 Pandemic Influenza Virus Enhance Avian Influenza Virus Polymerase Activity in Mammalian Cells. *J. Virol.* **85**, 7020–7028 (2011).

63. Hiromoto, Y., Saito, T., Lindstrom, S. & Nerome, K. Characterization of Low Virulent Strains of Highly Pathogenic A/Hong Kong/156/97 (H5N1) Virus in Mice after Passage in Embryonated Hens' Eggs. *Virology* **272**, 429–437 (2000).

64. Webster, R. G. *et al.* Structure of antigenic sites on the haemagglutinin molecule of H5 avian influenza virus and phenotypic variation of escape mutants. *J. Gen. Virol.* **83**, 2497–2505 (2002).

65. Yen, H.-L. *et al.* Changes in H5N1 influenza virus hemagglutinin receptor binding domain affect systemic spread. *Proc. Natl. Acad. Sci. U. S. A.* **106**, 286–291 (2009).

66. Stevens, J. *et al.* Recent Avian H5N1 Viruses Exhibit Increased Propensity for Acquiring Human Receptor Specificity. *J. Mol. Biol.* **381**, 1382–1394 (2008).

67. Wang, W. *et al.* Glycosylation at 158N of the hemagglutinin protein and receptor binding specificity synergistically affect the antigenicity and immunogenicity of a live attenuated H5N1 A/Vietnam/1203/2004 vaccine virus in ferrets. *J. Virol.* **84**, 6570–6577 (2010).

68. Chutinimitkul, S. *et al.* In vitro assessment of attachment pattern and replication efficiency of H5N1 influenza A viruses with altered receptor specificity. *J. Virol.* **84**, 6825–6833 (2010).

69. Stevens, J. *et al.* Structure and receptor specificity of the hemagglutinin from an H5N1 influenza virus. *Science* **312**, 404–410 (2006).

70. Maines, T. R. *et al.* Effect of receptor binding domain mutations on receptor binding and transmissibility of avian influenza H5N1 viruses. *Virology* **413**, 139–147 (2011).

71. Chen, L.-M. *et al.* In vitro evolution of H5N1 avian influenza virus toward human-type receptor specificity. *Virology* **422**, 105–113 (2012).

72. Weber, F., Kochs, G., Gruber, S. & Haller, O. A Classical Bipartite Nuclear Localization Signal on Thogoto and Influenza A Virus Nucleoproteins. *Virology* **250**, 9–18 (1998).

73. Grantham, M. L. *et al.* Palmitoylation of the influenza A virus M2 protein is not required for virus replication in vitro but contributes to virus virulence. *J. Virol.* **83**, 8655–8661 (2009).

74. Holsinger, L. J., Shaughnessy, M. A., Micko, A., Pinto, L. H. & Lamb, R. A. Analysis of the posttranslational modifications of the influenza virus M2 protein. *J. Virol.* **69**, 1219–1225 (1995).

75. Li, Y., Yamakita, Y. & Krug, R. M. Regulation of a nuclear export signal by an adjacent inhibitory sequence: The effector domain of the influenza virus NS1 protein. *Proceedings of the National Academy of Sciences* **95**, 4864–4869 (1998).

76. Hale, B. G., Barclay, W. S., Randall, R. E. & Russell, R. J. Structure of an avian influenza A virus NS1 protein effector domain. *Virology* **378**, 1–5 (2008).

77. Imai, H. *et al.* The HA and NS Genes of Human H5N1 Influenza A Virus Contribute to High Virulence in Ferrets. *PLoS Pathog.* **6**, e1001106 (2010).

Acknowledgments

We would like to thank Katherine Xue for her careful reading and comments on the manuscript.

Author contributions

LHM, TB, PD, PB, TCF, and PFH contributed the conception and design of the experiments. PD, SVH, SR, PB, EAK, LL, YL, HZ, YG, and PFH acquired samples and generated data. LHM, TB, TCF, and PFH analyzed and interpreted data. LHM. TB, EAK, TCF, and PFH wrote the manuscript.

Competing interests

Dr. Philippe Buchy is a former Head of Virology at Institut Pasteur du Cambodge and is currently an employee of GSK Vaccines, Singapore. The other authors declare no conflict of interest.

Funding statement

The study was funded by the US Agency for International Development (grant No. AID-442-G-14-00005).

Figure legends

Figure 1: Purifying selection, population growth, and randomness shape within-host diversity in humans and ducks

(a) Within-host polymorphisms present in at least 1% of sequencing reads were called in all human (red) and duck (blue) samples. Each dot represents one unique single nucleotide variant (SNV), the x-axis represents the nucleotide site of the SNV, and the y-axis represents its frequency within-host. (b) or each sample in our dataset, we calculated the proportion of its synonymous (light blue and light red) and nonsynonymous (dark blue and dark red) within-host variants present at frequencies of 1-10%, 10-20%, 20-30%, 30-40%, and 40-50%. We then took the mean across all human (red) or duck (blue) samples. Bars represent the mean proportion of variants present in a particular frequency bin and error bars represent standard deviations. (c) For each sample and gene, we computed the number of nonsynonymous SNPs per nonsynonymous site, and the number of synonymous SNPs per synonymous site. We then calculated the mean for each gene and species. Each bar represents the mean and error bars represent the standard deviation. Human values are shown in red and duck values are shown in blue.

Figure 2: Mutations are present at functionally relevant sites.

We queried each amino acid changing mutation identified in our dataset against all known annotations present in the Influenza Research Database Sequence Feature Variant Types tool. Each mutation is colored according to its function. Shape represents whether the mutation was identified in a human (circle) or duck (square) sample. Mutations shown here were detected in at least 1 human or duck sample. Filled in shapes represent nonsynonymous changes and open shapes represent synonymous mutations. Grey, transparent dots represent mutations for which no host-related function was known. Each nonsynonymous colored mutation, its frequency, and its phenotypic effect is shown in Table 2, and a full list of all mutations and their annotations are available in **Supplementary Table 2**.

Figure 3: Humans and ducks share more polymorphisms than expected by chance

(a) All amino acid sites that were polymorphic in at least 2 samples are shown. This includes sites at which each sample had a polymorphism at the same site, but encoded different variant amino acids. There are 3 amino acid sites that are shared by at least 2 duck samples, and 9 polymorphic sites shared by at least 2 human samples. 3 synonymous changes are detected in both human and duck samples (PB1 371, PA 397, and NP 201). Frequency is shown on the y-axis. (b) To test whether the level of sharing we observed was more or less than expected by chance, we performed a permutation test. The x-axis represents the number of sites shared by at least 2 ducks (blue) or at least 2 humans (red), and the bar height represents the number of simulations in which that number of shared sites occurred. Actual observed number of shared sites (3 and 9) are shown with a dashed line. (c) The same permutation test as shown in (b), except that only 70% of available amino acid sites were permitted to mutate. (d) The same permutation test as shown in (b), except that only 50% of available amino acid sites were permitted to mutate.

Figure 4: A small subset of within-host variants are enriched on spillover branches

(a) A schematic for how we classified host transitions along the phylogeny. Branches within monophyletic human clades were labelled “human to human” (red branches). Branches leading to a monophyletic human clade, whose parent node had avian children were labelled as “avian to human” (half red, half blue branches labelled “A -> H”), and all other branches were labelled “avian to avian” (blue branches). (b) Each amino acid-changing SNV we detected within-host in either ducks (left) or humans (right) that was present in the H5N1 phylogeny is displayed. Each bar represents an amino acid mutation, and its height represents the number of avian to avian (blue) or avian/human to human (red) transitions in which this mutation was present along the H5N1 phylogeny. Avian/human to human transitions includes both avian-to-human and human-to-human transitions summed together. Significance was assessed with a Fisher’s exact test. * indicates $p < 0.05$, **** indicates $p < 0.0001$.

Supplementary Figure 1: Genome coverage

The mean coverage depth at each nucleotide site (x-axis) for each gene across our 8 human and 5 duck samples is shown. Solid black lines represent the mean coverage across samples, and the grey shaded area represents the standard deviation of coverage depth across samples.

Supplementary Figure 2: Phylogenetic placement of H5N1 samples from Cambodia

All currently available H5N1 sequences were downloaded from the Influenza Research Database and the Global Initiative on Sharing All Influenza Data and used to generate full genome phylogenies using Nextstrain's augur pipeline. Colors represent the geographic region in which the sample was collected and x-axis position indicates the date of sample collection (for tips) or the inferred time to the most recent common ancestor (for internal nodes). H5N1 viruses from Cambodia selected for within-host analysis are indicated by green circles with black outlines. All HA and NA sequences in this dataset, besides A/duck/Cambodia/Y0224304/2014, belong to clade 1.1.2. Internal genes from samples collected prior to 2013 belong to clade 1.1.2, while internal genes from samples collected in 2013 or later belong to clade 2.3.2.1a.

Supplementary Figure 3: All within-host variants detected in our dataset

All within-host variants detected in our study are shown. Each row represents one sample and each column represents one gene. The x-axis shows the nucleotide site and the y-axis shows the frequency that the variant was detected within-host. Filled circles represent nonsynonymous changes, while open circles represent synonymous changes. Green dots represent variants identified within duck samples, while maroon dots represent variants identified in human samples. Blank plots indicate that no variants were identified in that sample and gene.

Tables

Table 1: Sample information

Sample ID	Host	Sample type	Collection	Date	Days post-symptom onset	vRNA copies/µl	Clade
A/duck/Cambodia/PV027D1/2010	Domestic duck	Pooled organs	Poultry outbreak investigation	April 2010	NA	5.45 x 10^{6}	1.1.2
A/duck/Cambodia/083D1/2011	Domestic duck	Pooled organs	Poultry outbreak investigation	September 2011	NA	3.74 x 10^{7}	1.1.2
A/duck/Cambodia/381W11M4/2013	Domestic duck	Pooled throat and cloacal swab	Live bird market surveillance	March 2013	NA	7.37 x 10^{5}	1.1.2/2.3.2.1a reassortant
A/duck/Cambodia/Y0224301/2014	Domestic duck	Pooled organs	Poultry outbreak investigation	February 2014	NA	2.0 x 10^{5}	1.1.2/2.3.2.1a reassortant
A/duck/Cambodia/Y0224304/2014	Domestic duck	Pooled organs	Poultry outbreak investigation	February 2014	NA	5.0 x 10^{6}	1.1.2/2.3.2.1a reassortant
A/Cambodia/V0401301/2011	Human (10F, died)	Throat swab	Event-based surveillance	April 2011	9	5.02 x 10^{3}	1.1.2

A/Cambodia/V0417301/2011	Human (5F, died)	Throat swab	Event-based surveillance	April 2011	5	8.98 x 10^4	1.1.2
A/Cambodia/W0112303/2012	Human (2M, died)	Throat swab	Event-based surveillance	January 2012	7	2.05 x 10^3	1.1.2
A/Cambodia/X0125302/2013	Human (1F, died)	Throat swab	Event-based surveillance	January 2013	12	6.84 x 10^4	1.1.2/2.3.2.1a reassortant
A/Cambodia/X0128304/2013	Human (9F, died)	Throat swab	Event-based surveillance	January 2013	8	5.09 x 10^3	1.1.2/2.3.2.1a reassortant
A/Cambodia/X0207301/2013	Human (5F, died)	Throat swab	Event-based surveillance	February 2013	12	1.73 x 10^5	1.1.2/2.3.2.1a reassortant
A/Cambodia/X0219301/2013	Human (2M, died)	Throat swab	Event-based surveillance	February 2013	12	1.66 x 10^3	1.1.2/2.3.2.1a reassortant
A/Cambodia/X1030304/2013	Human (2F, died)	Throat swab	Event-based surveillance	October 2013	8	1.08 x 10^4	1.1.2/2.3.2.1a reassortant

Table 2: Mutations identified at functionally relevant sites

Sample	Gene	Nt site	Ref base	Variant base	Coding region change	Frequency	Description	Type
A/Cambodia/X0125302/2013	PB2	816	A	C	N265H	2.82%	Determinant of temperature sensitivity in an H3N2 virus[58].	replication
A/Cambodia/X0128304/2013	PB2	1069	A	T	N348Y	5.88%	Putative m7GTP cap binding site[59].	replication
A/Cambodia/V0401301/2011	PB2	1115	C	T	P363P	10%	Putative m7GTP cap binding site[59].	replication
A/Cambodia/V0401301/2011	PB2	1202	A	C	N392H	3.61%	Putative m7GTP cap binding site[59].	replication
A/Cambodia/W0112303/2012	PB2	1891	G	A	E627K	7.20%	A Lys at 627 enhances mammalian replication[47,48].	replication
A/Cambodia/X0125302/2013	PB2	2022	G	A	V667I	2.95%	An Ile at 667 was associated with human-infecting H5N1 strains[60].	replication
A/Cambodia/W0112303/2012	PB2	2113	A	G	N701D	16.26%	An Asn at 701 enhances mammalian replication[50,51].	replication
A/Cambodia/X0125302/2013	PB2	2163	A	G	S714G	8.31%	An Arg at 714 enhances mammalian replication[50].	replication
A/Cambodia/X1030304/2013	PB1	631	A	G	R211G	1.89%	Nuclear localization motif.	interaction with host machinery

A/Cambodia/X1030304/2013	PB1	643	A	G	R215G	1.91%	Nuclear localization motif.	interaction with host machinery
A/Cambodia/X0125302/2013	PB1	1078	A	G	K353R	2.58%	An Arg at 353 is associated with higher replication and pathogenicity of an H1N1 pandemic strain[61].	replication
A/Cambodia/X0125302/2013	PB1	1716	A	T	T566S	5.38%	An Ala at 566 is associated with higher replication and pathogenicity of an H1N1 pandemic virus[61].	replication
A/Cambodia/X0219301/2013	PA	265	A	G	T85A	2.36%	An Ile at 85 enhances polymerase activity of pandemic H1N1 in mammalian cells[62].	replication
A/Cambodia/X0207301/2013	PA	1903	A	G	S631G	1.90%	A Ser at 631 enhances virulence of H5N1 in mice[63]..	virulence
A/Cambodia/X0128304/2013	HA	299	A	G	E91G	7.22%	A Lys at 91 enhances α-2,6 binding[39]. (H5 mature: 75)	receptor binding
A/Cambodia/V0417301/2011	HA	425	A	G	E142G	2.51%	Putative glycosylation site[64]. (H5 mature: 126)	virulence
A/Cambodia/X1030304/2013	HA	448	G	A	A150T	1.65%	A Val at 150 confers enhanced α-2,6 sialic acid binding in H5N1 viruses[53,54]. (H5 mature: 134)	receptor binding

A/Cambodia/V0401301/2011	HA	449	C	T	A150V	20.24%	A Val at 150 confers enhanced α-2,6 sialic acid binding in H5N1 viruses[53,54]. (H5 mature: 134)	receptor binding
A/Cambodia/X0125302/2013	HA	449	C	T	A150V	15.17%	A Val at 150 confers enhanced α-2,6 sialic acid binding in H5N1 viruses[53,54]. (H5 mature: 134)	receptor binding
A/Cambodia/X0128304/2013	HA	542	A	C	K172T	11.11%	Part of putative glycosylation motif that improves α-2,6 binding[65–67]. (H5 mature: 156)	receptor binding
A/Cambodia/V0401301/2011	HA	517	T	C	Y173H	5.04%	Residue involved in sialic acid recognition[41]. (H5 mature: 157)	receptor binding
A/Cambodia/V0401301/2011	HA	593	A	G	N198S	3.32%	A Lys at 198 confers α-2,6 sialic acid binding [39,68](H5 mature: 182)	receptor binding
A/Cambodia/X0128304/2013	HA	703	A	G	T226A	28.07%	An Ile at 226 enhanced α-2,6 sialic acid binding[57]. (H5 mature: 210)	receptor binding
A/Cambodia/V0401301/2011	HA	713	A	T	N238L	2.80%	A Leu at 238 confers a switch from α-2,3 to α-2,6 sialic acid binding and is a determinant of mammalian transmission[11,12,68–71]. (H5 mature: 222)	receptor binding
A/Cambodia/V0417301/2011	HA	713	A	T	N238L	8.05%	A Leu at 238 confers a switch from α-2,3 to α-2,6 sialic acid binding and is a determinant of mammalian transmission[11,12,68–71]. (H5 mature: 222)	receptor binding

A/Cambodia/X0125302/2013	HA	713	A	G	N238R	37.29%	A Leu at 238 confers a switch from α-2,3 to α-2,6 sialic acid binding and is a determinant of mammalian transmission[11,12,68–71]. (H5 mature: 222)	receptor binding
A/duck/Cambodia/Y0224304/2014	NP	674	C	T	T215I	3.69%	Nuclear targeting motif[72].	interaction with host machinery
A/Cambodia/X1030304/2013	M2	861	G	A	C50Y	1.88%	A Cys at position 50 is a palmitoylation site that enhances virulence[73,74].	virulence
A/Cambodia/X0128304/2013	NS1	502	C	T	P159L	2.98%	Part of the NS1 nuclear export signal mask[75].	interaction with host machinery
A/duck/Cambodia/Y0224301/2014	NS1	646	T	C	L207P	2.22%	NS1 flexible tail, which interacts with host machinery[76].	interaction with host machinery
A/duck/Cambodia/Y0224301/2014	NS1	654	C	T	P210S	2.55%	NS1 flexible tail, which interacts with host machinery[76].	interaction with host machinery
A/Cambodia/X0207301/2013	NEP	609	A	G	E47G	4.53%	This site was implicated in enhanced virulence of H5N1 in ferrets[77].	virulence

All nonsynonymous mutations that were identified in sites with putative links to host-specific phenotypes are shown. We identify a handful of amino acid mutations that have been explicitly linked to mammalian adaptation of avian influenza viruses. For HA mutations, all mutations use native H5 numbering, including the signal peptide. For ease of comparison, the corresponding amino acid number in mature, H5 peptide numbering is also provided in parentheses in the description column. Full annotations for all mutations in our data are shown in **Supplementary Table 2**.

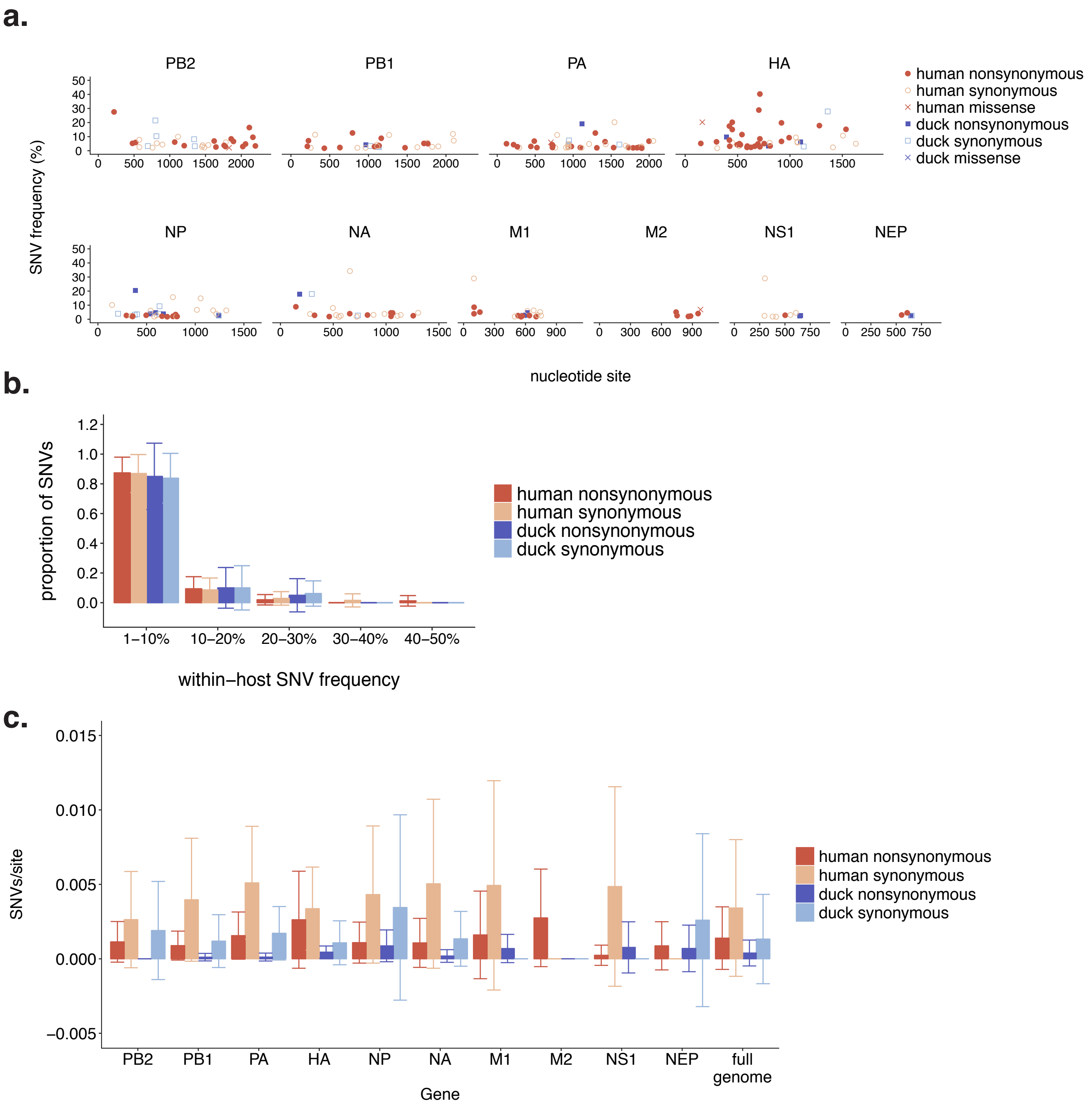

Figure 1: Purifying selection, population growth, and randomness shape within-host diversity in humans and ducks

(a) Within-host polymorphisms present in at least 1% of sequencing reads were called in all human (red) and duck (blue) samples. Each dot represents one unique single nucleotide variant (SNV), the x-axis represents the nucleotide site of the SNV, and the y-axis represents its frequency within-host. **(b)** or each sample in our dataset, we calculated the proportion of its synonymous (light blue and light red) and nonsynonymous (dark blue and dark red) within-host variants present at frequencies of 1-10%, 10-20%, 20-30%, 30-40%, and 40-50%. We then took the mean across all human (red) or duck (blue) samples. Bars represent the mean proportion of variants present in a particular frequency bin and error bars represent standard deviations. **(c)** For each sample and gene, we computed the number of nonsynonymous SNPs per nonsynonymous site, and the number of synonyous SNPs per synonymous site. We then calculated the mean for each gene and species. Each bar represents the mean and error bars represent the standard deviation. Human values are shown in red and duck values are shown in blue.

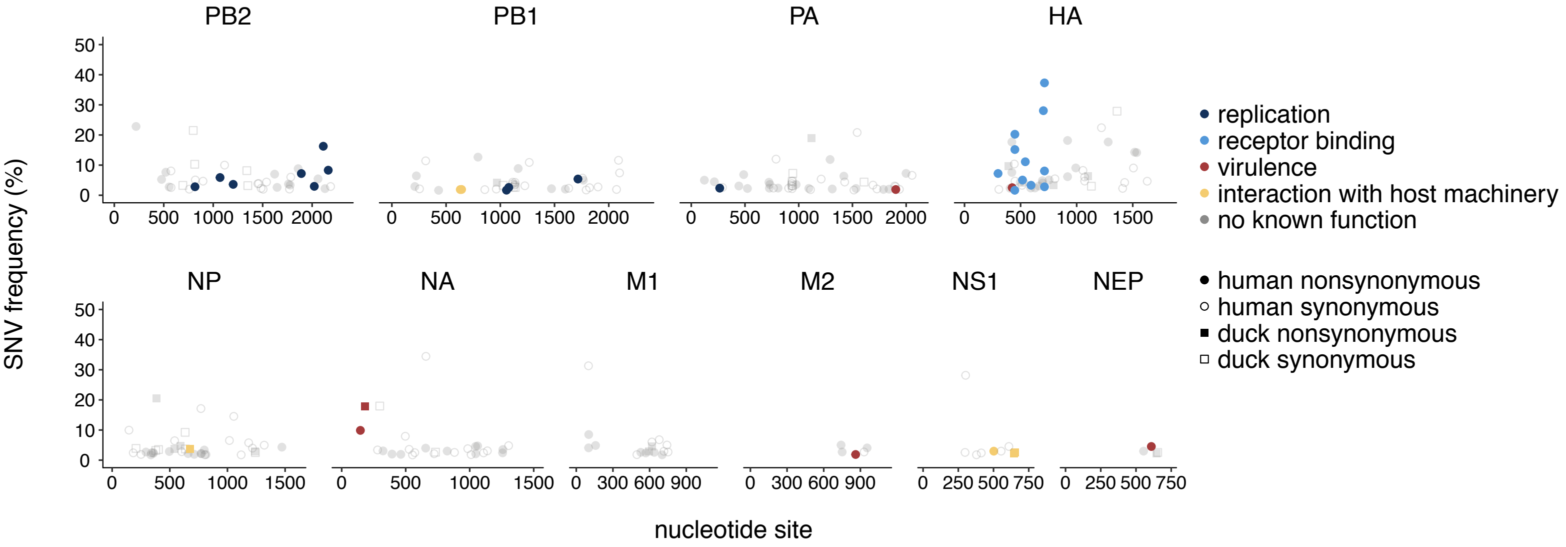

Figure 2: Mutations are present at functionally relevant sites.

We queried each amino acid changing mutation identified in our dataset against all known annotations present in the Influenza Research Database Sequence Feature Variant Types tool. Each mutation is colored according to its function. Shape represents whether the mutation was identified in a human (circle) or duck (square) sample. Mutations shown here were detected in at least 1 human or duck sample. Filled in shapes represent nonsynonymous changes and open shapes represent synonymous mutations. Grey, transparent dots represent mutations for which no host-related function was known. Each nonsynonymous colored mutation, its frequency, and its phenotypic effect is shown in Table 2, and a full list of all mutations and their annotations are available in Supplementary Table 2.

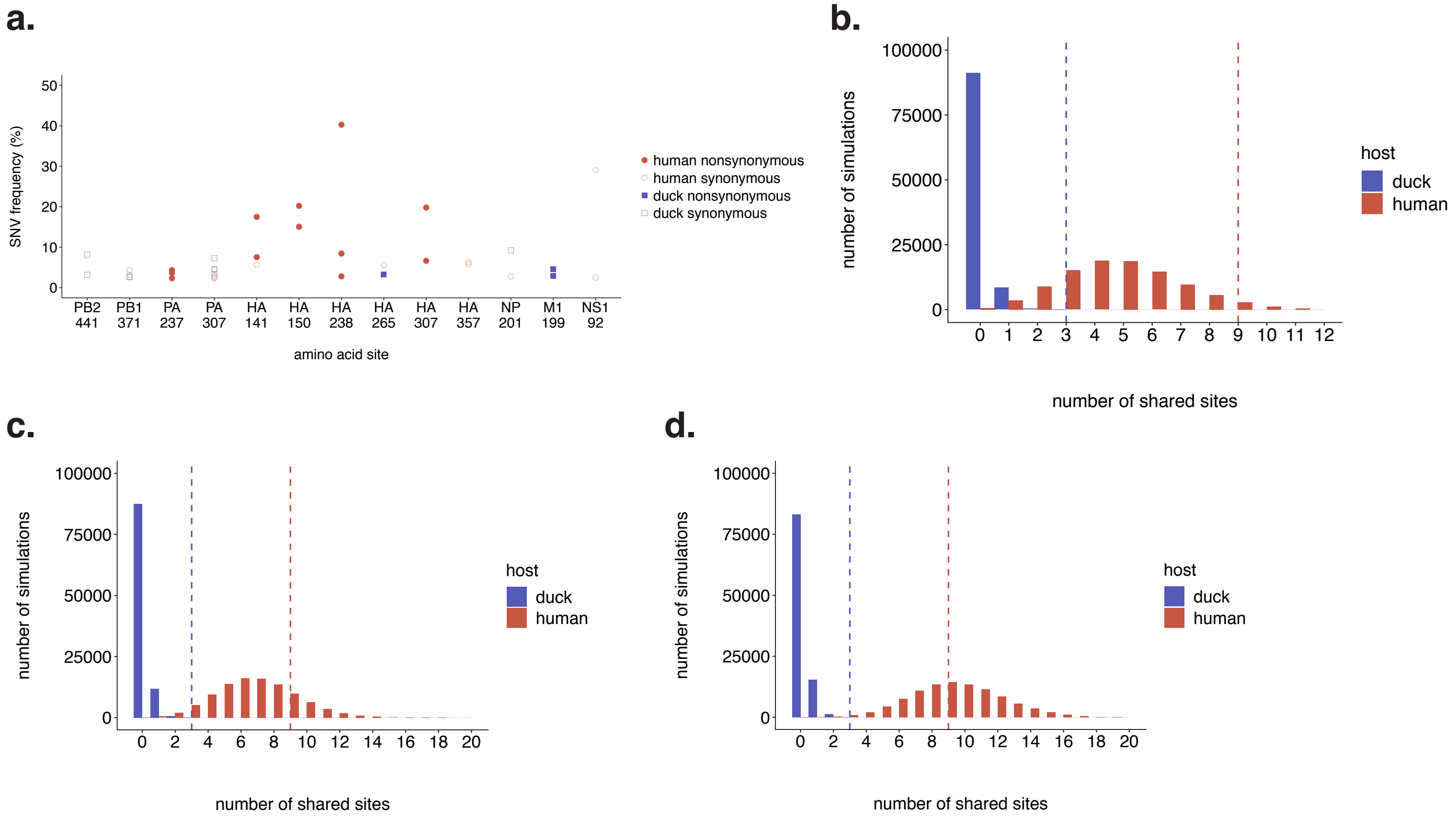

Figure 3: Humans and ducks share more polymorphisms than expected by chance

(a) All amino acid sites that were polymorphic in at least 2 samples are shown. This includes sites at which each sample had a polymorphism at the same site, but encoded different variant amino acids. There are 3 amino acid sites that are shared by at least 2 duck samples, and 9 polymorphic sites shared by at least 2 human samples. 3 synonymous changes are detected in both human and duck samples (PB1 371, PA 397, and NP 201). Frequency is shown on the y-axis. **(b)** To test whether the level of sharing we observed was more or less than expected by chance, we performed a permutation test. The x-axis represents the number of sites shared by at least 2 ducks (blue) or at least 2 humans (red), and the bar height represents the number of simulations in which that number of shared sites occurred. Actual observed number of shared sites (3 and 9) are shown with a dashed line. **(c)** The same permutation test as shown in **(b)**, except that only 70% of available amino acid sites were permitted to mutate. **(d)** The same permutation test as shown in **(b)**, except that only 50% of available amino acid sites were permitted to mutate.

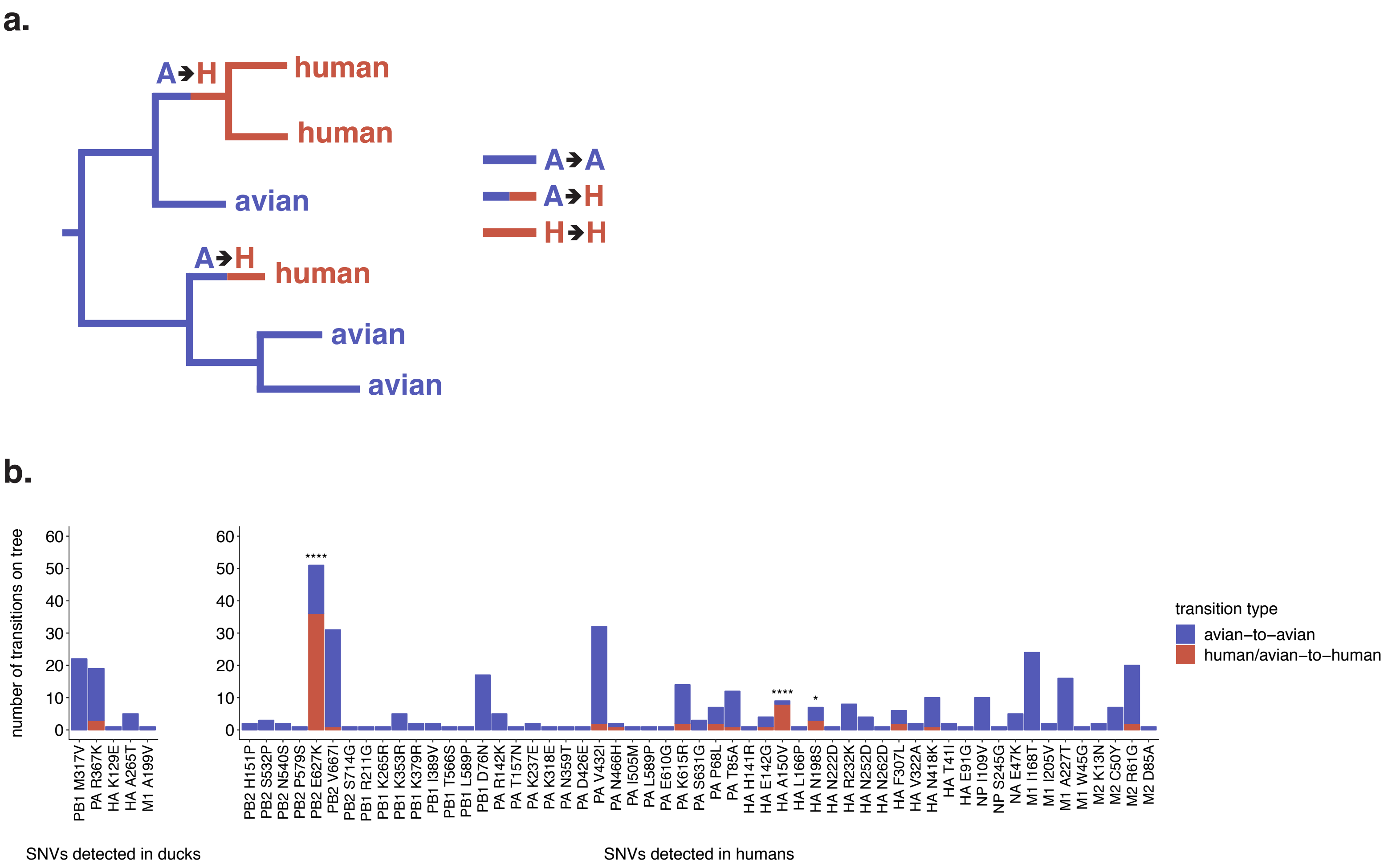

Figure 4: A small subset of within-host variants are enriched on spillover branches
(a) A schematic for how we classified host transitions along the phylogeny. Branches within monophyletic human clades were labelled “human to human” (red branches). Branches leading to a monophyletic human clade, whose parent node had avian children were labelled as “avian to human” (half red, half blue branches labelled “A -> H”), and all other branches were labelled “avian to avian” (blue branches). **(b)** Each amino acid-changing SNV we detected within-host in either ducks (left) or humans (right) that was present in the H5N1 phylogeny is displayed. Each bar represents an amino acid mutation, and its height represents the number of avian to avian (blue) or avian/human to human (red) transitions in which this mutation was present along the H5N1 phylogeny. Avian/human to human transitions includes both avian-to-human and human-to-human transitions summed together. Significance was assessed with a Fisher’s exact test. * indicates $p < 0.05$, **** indicates $p < 0.0001$.