Avian influenza viruses in wild birds: virus evolution in a multi-host ecosystem

Divya Venkatesh[1], Marjolein J. Poen[2,] Theo M. Bestebroer[2], Rachel D. Scheuer[2] , Oanh Vuong[2], Mzia Chkhaidze[3], Anna Machablishvili[3], Jimsher Mamuchadze[4], Levan Ninua[4], Nadia B. Fedorova[5], Rebecca A. Halpin[5], Xudong Lin[5], Amy Ransier[5], Timothy B Stockwell[5], David E. Wentworth[5*], Divya Kriti[6], Jayeeta Dutta[6], Harm van Bakel[6], Anita Puranik[7], Marek J Slomka[7], Steve Essen[7], Ian H. Brown[7], Ron A.M. Fouchier[2], **Nicola S. Lewis[1,7]#**

[1]Department of Zoology, University of Cambridge, Downing Street, Cambridge CB2 3EJ, United Kingdom

[2]Department of Viroscience, Erasmus MC, P.O. Box 2040, 3000CA Rotterdam, Netherlands

[3]National Centre for Disease Control, Tbilisi, Georgia

[4]Institute of Ecology, Ilia State University, 3/5 Cholokashvili, Tbilisi, Georgia.

[5]J. Craig Venter Institute, Rockville, Maryland, United States of America

[6]Icahn School of Medicine at Mount Sinai, New York, United States of America

[7]Animal and Plant Health Agency-Weybridge, United Kingdom

Running title: Evolution of avian influenza viruses in wild birds

#Address correspondence to **Nicola S. Lewis:** nsl25@cam.ac.uk

Abstract

Wild ducks and gulls are the major reservoirs for avian influenza A viruses (AIVs). The mechanisms that drive AIV evolution are complex at sites where various duck and gull species from multiple flyways breed, winter or stage. The Republic of Georgia is located at the intersection of three migratory flyways: Central Asian Flyway, East Asian/East African Flyway and Black Sea/Mediterranean Flyway. For six consecutive years (2010-2016), we collected AIV samples from various duck and gull species that breed, migrate and overwinter in Georgia. We found substantial subtype diversity of viruses that varied in prevalence from year to year. Low pathogenic (LP)AIV subtypes included H1N1, H2N3, H2N5, H2N7, H3N8, H4N2, H6N2, H7N3, H7N7, H9N1, H9N3, H10N4, H10N7, H11N1, H13N2, H13N6, H13N8, H16N3, plus two H5N5 and H5N8 highly pathogenic (HP)AIVs belonging to clade 2.3.4.4. Whole genome phylogenetic trees showed significant host species lineage restriction for nearly all gene segments and significant differences for LPAIVs among different host species in observed reassortment rates, as defined by quantification of phylogenetic incongruence, and in nucleotide diversity. Hemagglutinin clade 2.3.4.4 H5N8 viruses, circulated in Eurasia during 2014-2015 did not reassort, but analysis after its subsequent dissemination during 2016-2017 revealed reassortment in all gene segments except NP and NS. Some virus lineages appeared to be unrelated to AIVs in wild bird populations in other regions with maintenance of local AIV viruses in Georgia, whereas other lineages showed considerable genetic inter-relationship with viruses circulating in other parts of Eurasia and Africa, despite relative under-sampling in the area.

Importance

Waterbirds (e.g., gulls/ducks) are natural reservoirs of avian influenza viruses (AIVs) and have been shown to mediate dispersal of AIV at inter-continental scales during seasonal migration. The segmented genome of influenza viruses enables viral RNA from different lineages to mix or re-assort when two viruses infect the same host. Such reassortant viruses have been identified in most major human influenza pandemics and several poultry outbreaks. Despite their importance, we have only recently begun to understand AIV evolution and reassortment in their natural host reservoirs. This comprehensive study illustrates of AIV evolutionary dynamics within a multi-host ecosystem at a stop-over site where three major migratory flyways intersect. Our analysis of this ecosystem over a six-year period provides a snapshot of how these viruses are linked to global AIV populations. Understanding the evolution of AIVs in the natural host is imperative to both mitigating the risk of incursion into domestic poultry and potential risk to mammalian hosts including humans.

Introduction

Avian influenza viruses (AIVs) have been identified in a wide diversity of wild and domestic bird species but wild waterbirds of the Orders *Anseriformes and Charadriformes,* such as ducks, geese, swans and shorebirds (1, 2) form their natural reservoir. These birds maintain diverse group of low pathogenic avian influenza A viruses (LPAIVs), which cause limited morbidity in these host species in experimental settings (3). The effect of AIV infection in wild birds in non-experimental settings is more contradictory. Body mass was significantly lower in infected mallards (*Anas playrhynchos)* and the amount of virus shed by infected juveniles was negatively correlated with body mass. However, there was no general effect of infection on staging time (duration of stopover for migratory birds), except for juveniles in September and LPAIV infection did not affect speed or distance of subsequent migration (4). Conversely, a recent mallard study demonstrated no obvious detriment to the bird as movement patterns did not differ between LPAIV infected and uninfected birds. Hence, LPAIV infection probably does not affect mallard movements during stopover, consequently resulting in the potential for virus spread along the migration route (5). The precise role of migrants and resident birds in amplifying and dispersing AIVs however, remains unclear. In another study the migrant arrivals played a role in virus amplification rather than seeding a novel variant into a resident population (6). It has also been suggested that switching transmission dynamics might be a critical strategy for pathogens such as influenza A viruses associated with mobile hosts such as wild waterbirds, and that both intra and inter-species transmission are important to maintaining gene flow across seasons (7).

AIVs continue to cause both morbidity and mortality in poultry worldwide. Increased mortality is strongly related to infection with highly pathogenic influenza A viruses (HPAIVs), characterised by mortality in gallinaceous poultry (8). Periodically, human infections associated with HPAIV of both the H5 and H7 subtypes have been detected. In particular, parts of Asia and Africa have been significantly affected by the Eurasian (goose/Guangdong/1996) lineage H5 HPAIV epizootic for two decades, becoming enzootic in some areas and multiple waves of influenza with evolving viruses in others (9). More recently, H5Nx reassortants of the Eurasian lineage HPAIVs from clade 2.3.4.4 have been introduced into wild birds from poultry and spread to new geographic regions (10).

The Caucasus, at the border of Europe and Asia, is important for migration and over-wintering of wild waterbirds. Three flyways, the Central Asian, East Africa-West Asia, and Mediterranean/Black Sea flyways, converge in this region. Understanding the ecology and evolution of AIVs in wild birds is complex, particularly at sites where multiple species co-habit and in those ecosystems which support different annual life-cycle stages and where multiple migratory flyways intersect.

At a population level, Eurasian dabbling ducks were found to be more frequently infected than other ducks and Anseriformes (11) with most AIV subtypes detected in ducks, except H13 and H16 subtypes which were detected primarily in gulls (11, 12). Temporal and spatial variation in influenza virus prevalence in wild birds was observed, with AIV prevalence varying by sampling location. In this study site in the Republic of Georgia, we observed peak prevalence in large gulls during the autumn migration (5.3-9.8%), but peak prevalence in Black-headed Gulls (*Chroicocephalus ridibundus*) in spring (4.2-13%)(13). In ducks, we observed increased AIV prevalence

106 during the autumn post-moult aggregations and migration stop-over period (6.3%) but
107 at lower levels to those observed in other more northerly post-moult areas in Eurasia.
108
109 In North America, studies have primarily focused on Anseriformes species with
110 sampling during late summer and autumn southern migration (14-16), rather than
111 longitudinally throughout the annual lifecycle of the host or within an ecosystem. The
112 southwestern Lake Erie Basin is an important stopover site for waterfowl during
113 migration periods, and over the past 28 years, 8.72% of waterfowl sampled in this
114 geographic location have been positive for AIV recovery during summer and autumn
115 (June – December) (17). More recent studies which targeted overwintering and
116 returning migratory birds during February – April showed the presence of diverse AIV
117 subtypes in waterbirds at northern latitudes in the United States (17).
118
119 Previous genetic studies of the viruses isolated from wild birds have focused on gene
120 flow at an intra- or intercontinental level involving multiple hosts, rather than on virus
121 gene flow among species within an ecosystem (16, 18-20). Indeed, the conclusions of
122 such studies have been somewhat limited at times by statistical power owing to
123 insufficient sequence data from enough hosts relevant to virus dynamics across the
124 geographic study area. (21). In Eurasia, frequent reassortment and co-circulating
125 lineages were observed for all eight genomic RNA segments over time. Although,
126 there was no apparent species-specific effect on the diversity of the AIVs, there was
127 a spatial and temporal relationship between the Eurasian sequences and significant
128 viral migration of AIVs from West Eurasia towards Central Eurasia (22).
129

130 This study presents novel findings concerning the ecology and evolution of both
131 LPAIVs and HPAIVs circulating in wild birds in a key active surveillance site in Eurasia.
132 We investigated the diffusion of AIV gene segments within different wild bird hosts
133 occupying the same ecosystem. There was substantial diversity in surface
134 glycoprotein HA (heamagglutinin) and NA (neuraminidase) subtypes, which varied
135 year to year and with the host species. M, NS, NP, PB1, PB2 and PA (henceforth
136 referred to as “internal” gene segments) also showed host restriction to various
137 degrees. There were differences in genetic diversity, reassortment rates, and inter-
138 species transmission rates in the internal gene segments associated with different
139 host species and HA subtypes. We also examined how closely related the Georgian
140 AIV gene segments were to AIV globally. We found evidence for genetic inter-
141 relationship of Georgian AIV with AIV in mainly Africa and Eurasia but several lineages
142 appear to be maintained locally.
143
144

Methods

Active surveillance for influenza A viruses was carried out from 2010-2016 as described previously (13)

Dataset and genomic sequencing

Over a period of six years, 30,911 samples from 105 different bird species were analysed for the presence of AIVs. Positive isolates were obtained by standard approaches (23), and where possible, subtyped and sequence generated from extracted RNA as described below.

For virus samples from 2010-2012, codon complete genomes of IAV were sequenced as part of the Influenza Genome Project (http://gcid.jcvi.org/projects/gsc/influenza/index.php), an initiative by the National Institute of Allergies and Infectious Diseases (NIAID). IAV viral RNA (vRNA) was isolated from the samples/specimens, and the entire genome was amplified from 3 ul of RNA template using a multi-segment RT-PCR strategy (M-RTPCR) (24, 25). The amplicons were sequenced using the Ion Torrent PGM (Thermo Fisher Scientific, Waltham, Massachusetts, USA) and/or the Illumina MiSeq v2 (Illumina, Inc., San Diego, California, USA) instruments. When sequencing data from both platforms was available, the data were merged and assembled together; the resulting consensus sequences were supported by reads from both technologies. Sequence data for Georgia was downloaded from the NIAID Influenza Research Database (IRD) (Squires et al. 2012) through the web site at http://www.fludb.org on 11/5/2016. To this dataset, we added sequence data for isolates from 2013 and 2016 which were sequenced at either Erasmus MC, Animal and Plant Health Agency (APHA) or the Icahn School of Medicine at Mount Sinai (ISMMS). At Erasmus MC sequencing was performed as described previously by V. J. Munster et al. (26), with modifications.

Primer sequences are available upon request.

At APHA, viral RNA was extracted using the QIAquick Viral RNA extraction kit (Qiagen, UK) without the addition of carrier. Double stranded cDNA (cDNA synthesis system, Roche, UK) was generated from RNA according to the manufacturer's instructions. This was quantified using the fluorescent PicoGreen reagent and 1ng was used as a template for the preparation of the sequencing library (NexteraXT, Illumina, Cambridge, UK). Sequencing libraries were run on a MiSeq instrument (Illumina, Cambridge, UK) with 2x75 base paired end reads. Data handling of raw sequence reads and extraction of consensus sequences were performed at APHA.

For the Icahn School of medicine at Mount Sinai, RNA was extracted using the QIAamp Viral RNA Mini Kit (52904, Qiagen, UK). MS-RTPCR amplification was performed with the Superscript III high-fidelity RT-PCR kit (12574-023, Invitrogen) according to manufacturer's instructions using the Opti1 primer set: Opti1-F1 5' GTTACGCGCCAGCAAAAGCAGG, Opti1-F2 5'GTTACGCGCCAGC**G**AAAGCAGG and Opti1-R15'GTTACGCGCCAGTAGAAACAAGG. DNA amplicons were purified using Agencourt AMPure XP 5ml Kit (A63880, Beckman Coulter). At the Icahn School of Medicine, sequencing libraries were prepared and sequencing was performed on a MiSeq instrument (Illumina, Cambridge, UK) with 2x150 base paired end reads. Data handling of raw sequence reads and extraction of consensus sequences were performed at ISMMS, as described previously (27).

Genetic analyses

Sequence alignment preparation

Whole genome sequences from 81 Georgian strains isolated between 2010 and 2016 are used in this analysis. We aligned sequences from each gene segment separately using MAFFT v7.305b (28) and trimmed to starting ATG and STOP codon

in Aliview v1.18. Hemagglutinin (HA) sequences were further trimmed to exclude the initial signal sequence (29, 30). Sequences were then aligned using “muscle-codon” option with default settings in MEGA7 (31).

The NS gene has two alleles A and B, with significant difference in sequence composition, which could skew analyses of sequence diversity. The NS gene sequences were therefore considered both as a complete dataset (NS) and subdivided into NS-A and NS-B datasets where required. As only six out of 81 sequenced strains had the NS-A allele, only NS and NS-B datasets were used in the analyses.

We then subdivided the complete datasets of each gene according to viral traits, namely:

- host group (gull and duck)
- host type
    - **BMG**: Black-headed Gulls (*Chroicocephalus ridibundus*) and Mediterranean Gulls (*Ichthyaetus melanocephalus*).
    - **YAG**: Yellow-legged Gulls (*Larus michahellis*) and Armenian Gulls (*Larus armenicus*).
    - **MD**: Mallards (*Anas platyrhynchos*).
    - **OD**: Other ducks. This includes the common teal (*Anas crecca*), domestic duck (*Anas platyrhynchos domesticus*), garganey (*Anas querquedula*), northern shoveler (*Anas clypeata*), common coot (*Fulica atra*), and tufted duck (*Aythya fuligula*).

- HA subtype. Dataset was reduced to include subtypes H1, 2, 3,4, 5, 6, 7, 9,10, 11, 13 where greater than three sequences were available for statistical analyses.

Visualisation of phylogenetic incongruence

We inferred Maximum Likelihood (ML) phylogenetic trees for each gene segment using IQ-TREE, 1.5.5 (32) and ModelFinder (33) and obtained branch supports with SH-like approximate Likelihood Ratio Test (aLRT) and standard non-parametric bootstrap. All trees were rooted using the "best-fitting-root" function in Tempest v1.5 (34) and visualised in FigTree v1.4.2, with increasing node-order. To visualise incongruence, we traced the phylogenetic position of each sequence, coloured according to host, across unrooted ML trees for all internal gene segments. Figures were generated by modifying scripts from a similar analysis (35).

Quantification of nucleotide diversity

Complete alignments of each internal gene, as well as alignment subsets by host group, host type and HA subtype were used in "PopGenome" package in R v3.2 (36) to estimate nucleotide diversity. Per-site diversity was calculated by dividing the nucleotide diversity output by number of sites present in each alignment. As each subset contained different numbers of sequences, this value was normalised by dividing by the number of sequences in each respective dataset. Heat maps from this data were plotted in R v3.2.

Correlating traits with phylogeny (BaTS)

Null hypothesis of no association between phylogenetic ancestry and traits (host group, host type and HA subtype) was tested using Bayesian Tip-association Significance Testing (BaTS) beta build 2 (37) for all internal gene segments. Bayesian

241 posterior sets of trees were inferred using MrBayes v3.2.6 (38) using the same
242 segment-wise alignments generated for ML tree estimation. A set of scripts and
243 commands used to generate the input file for BaTS are provided in [Supplementary
244 materials]. Ratio of clustering by each trait on the gene segment trees that is expected
245 by chance alone (Null mean), with the association that is observed in the data
246 (Observed mean) was calculated. These expected/observed ratios were summarized
247 in a heat-map with the y-axis ordered by the amount of reassortment observed. Data
248 manipulation and figure preparation was done in R v3.2.

249 ***Quantification of diversity and between host transmission***
250 Alignments generated for ML trees were also used in Bayesian phylodynamic
251 analyses using BEAST v1.8.4 (39). We employed a strict molecular clock, a
252 coalescent constant tree prior and the SRD06 site model with two partitions for codon
253 positions (1^{st}+2^{nd} positions, 3^{rd} position), with base frequencies unlinked across all
254 codon positions. The MCMC chain was run twice for 100 million iterations, with sub-
255 sampling every 10,000 iterations. All parameters reached convergence, as assessed
256 visually using Tracer (v.1.6.0). Log combiner (v1.8.4) was used to remove initial 10%
257 of the chain as burn-in, and merge log and trees files output from the two MCMC runs.
258 Maximum clade credibility (MCC) trees were summarized using TreeAnnotator
259 (v.1.8.4). After removal of burn-in, the trees were analysed using PACT (Posterior
260 analysis of coalescent trees) (https://github.com/trvrb/PACT.git) to determine
261 measures of diversity, and migration rates between hosts over time.

262 ***Geographical context for 'Georgian origin' internal protein coding gene***
263 ***segments***
264 Internal gene sequences from, avian hosts, sampled across the world between 2005
265 and 2017 were obtained from gisaid.org (downloaded November 2017). Sequences

(each segment separately) were divided into regions namely Asia (including Oceania), Europe, Africa, North America and South America. The program cd-hit-est (40, 41) was used to down-sample each regional dataset to 0.9 similarity cut-off level. These down-sampled sequences were then merged with the Georgian dataset. Discrete trait ancestral reconstruction with symmetric and asymmetric models were implemented in BEAST v1.8.4 (39) together with marginal likelihood estimation using path-sampling/stepping-stone analysis. The symmetric model was chosen over the asymmetric (log Bayes factor =14). The MCMC chain was run twice for 100 million iterations, with sub-sampling every 10,000 iterations. All parameters reached convergence, as assessed visually using Tracer (v.1.6.0). Log combiner (v1.8.4) was used to remove initial 10% of the chain as burn-in, and merge log and trees files output from the two MCMC runs. Maximum clade credibility (MCC) trees were summarized using TreeAnnotator (v.1.8.4). PACT was used to extract overall migration rates between trait locations.

Results

HA-NA subtype diversity and host-specificity

Over the six-year period between 2010 and 2016, 24 HA/NA subtypes of influenza A virus, including 12 different HA subtypes (H1, 2, 3, 4, 5, 6, 7, 9, 10, 11, 13, and 16) were isolated (Figure 1A). The diversity of subtypes varied from year to year, and associated with the level of prevalence in duck versus gull hosts. Within our sampling in Georgia, H9, H13 subtypes are found exclusively in gulls, while H1, H5, and H7 were detected exclusively in mallards. H3, H4, H6, and H10 were found in mallards and various other ducks. Positive evidence for multiple-species infection (ducks and

gulls) was found only for H2 and H11 viruses in this dataset even though globally, many other subtypes are found in multiple hosts.

In the first three years between 2010-12, up to seven different HA subtypes were found every year. These included H1, 2, 3, 4, 6, 10, 11, 13, and 16. H13, which was found in the greatest proportion of sequenced samples in 2011 and 2012 and was the sole type sequenced in 2013. In 2014, again only a single subtype was found (H10). The absence of more subtypes in these years could be explained by the comparatively low prevalence of IAV in these years, in both gulls and ducks in 2014 and especially ducks in 2013 (Figure 1B). In 2015, where prevalence was nearly zero in gulls, we saw HPAI H5 type viruses detected along with an H6. H4, which was previously isolated only in 2011, was the predominant type in 2016, followed by H5 and H7.

Genetic structure of AIV detected in Georgia in 2010-16

For all gene segments except PA, there were two major subdivisions in tree topology – one clade containing sequences predominantly from ducks and one clade entirely derived from gull sequences (Figure 2, S2). The internal protein coding gene segments from certain subtypes formed sub-clades that were defined by year of circulation suggesting single-variant epidemic-like transmission within the population. This was seen in H13N8 in gulls and H4N6 and H5N8 in ducks. There were several examples of gull-derived viruses, which had several internal gene segments (other than NP) located in the 'duck' clade, mostly derived from Black-headed and Mediterranean Gulls (BMG). Only the PA gene phylogeny had an occurrence of a small sub-clade of Yellow-legged and Armenian Gull-derived (YAG) viruses clustered within the duck-derived viruses. For M gene segment, there were two major clades entirely defined by host species (except for 2 BMG viruses), and an outlier sub-clade consisting of H2 and H9 gull lineage viruses from BMGs. In PB1, PB2 and PA, these outlier- sub-clade

viruses were found in various configurations in the tree. For NS, the tree topology divided into two alleles as reported previously (42). However, there were only six viruses from Allele *A* isolated from four mallards *(*MD), a garganey (OD) and a common teal (OD). Allele *B* splits into two sub-clades again defined by whether the viruses were isolated from gulls or ducks. The 'duck' sub-clade includes the outlier BMG viruses identified above for M. The long branch length to the gull sub-clade from the duck sub-clade in Allele *B* would suggest that there might be host-specificity in NS evolution, perhaps in response to differences between avian host innate immune responses.

Variation in nucleotide diversity

We used the PopGenome package in R to calculate the per-site nucleotide diversity for all internal gene segments (Figure 3A-C). Nucleotide diversity of the internal gene segments in one surveillance site may be an indication of the breadth of sources where the viruses have been derived from. We found greater diversity in both gulls and ducks in gene segment NS (possibly because of the presence of both A and B alleles of this gene in the dataset) and PB2 (Figure 3A). When further sub-divided into "host types" as described in the methods, we found that the group of Black-headed and Mediterranean Gulls (BMG) had the highest per-site diversity. In comparison, the mallards (MD), the Yellow-legged and Armenian Gulls (YAG) and other ducks (OD) had relatively lower values across all internal gene segments, despite the OD comprising of a variety of ducks. Only the PA gene had greater diversity in Yellow-legged and Armenian Gulls than in Black-headed and Mediterranean Gulls (Figure 3B). When subset by HA subtype (Figure 3C), the internal gene segments associated with H4 and H13, the most abundant types found in our dataset, had the lowest diversity – possibly because several of the isolates were detected at the same time.

Those less commonly isolated, such as H11 was detected in different years (2011, 2014) which may explain the high diversity of its NS, M, NP, PA, PB1, and PB2 gene segments. However, H3, which also has relatively high diversity were both detected at the same time (September 2011). Both NS and NS-B datasets were used in the analysis and as expected, the exclusion of sequences of NS-A (found exclusively in viruses from duck hosts), lowers the overall diversity within the ducks even when the values are normalised for the number of sequences found in each subset.

We tested the root-to-tip regression for ML trees for each of the six internal protein coding gene segments using Tempest v1.5 (34) to look for temporal signatures. All except NS gene showed positive correlation of distance with time, despite the short window of six years (Figure S1A). NS root to tip regression shows a negative slope, and it is likely confounded by the presence of two alleles A and B. Therefore, only NS-B allele, which forms a dominant portion of the NS gene segments in the data-set (75 out of 81), and shows clock-likeness (Figure S1B) were used for further analysis using BEAST v1.8.4. PACT analysis showed that the overall and yearly host-related diversity measures (Figure 4 A and B) show similar trends as seen in Figure 3.

Correlation of traits with phylogeny

We tested the null hypothesis that there is no association between phylogenetic ancestry and traits (host group, host type and HA subtype) using Bayesian Tip-association Significance Testing (BaTS). Ratio of clustering by each trait on the gene segment trees that is expected by chance alone (Null mean), with the association that is observed in the data (Observed mean) are presented in Figure 5 (A-C). The higher the value of null/observed, the lower is the support for phylogenetic clustering of the given trait. Therefore, a higher value indicates a different ancestry. Hence, when we consider the HA subtype trait as “lineage”, it provides a measure of reassortment as

364 described (43). Again, NS-B dataset was considered along with the complete NS
365 dataset but no significant differences in trends were found. Panel A shows that gull
366 viruses are more likely to cluster together in a phylogenetic tree than duck viruses in
367 general. When viruses of gulls and ducks were further subdivided, panel B shows that
368 OD viruses are less likely to cluster together in the tree, which is expected given that
369 we have grouped together several duck species under this category. Among the rest,
370 again it is the duck species (MD) that exhibit dynamic phylogenetic placing compared
371 to both the gull types. The only exception is with the PB2 gene segment, for which the
372 BMG show a lower level of phylogenetic clustering by species indicating putative
373 reassortment events. When we consider the HA subtype (lineage) of the viruses, we
374 find that H4 and H13, which showed the lowest nucleotide diversity, also show very
375 low levels of reassortment, as does H5. There was not enough statistical power to
376 interpret events in H1, 3, 6, 7, 9 or 11 viruses. Where statistically significant values
377 were found, lower levels of clustering were observed.

378 ***Directionality of viral gene segment transfer***

379 Figure S3 shows ancestral reconstruction of the host state along time-scaled
380 phylogenies for five of six internal gene segments. The results are summarised in
381 Figure 6A showing the mean number of host jump events from duck to gull and vice-
382 versa. For all gene segments, most of the host spillover events are in the direction
383 from ducks to gulls. In figure 6B we see that at a finer level, most of the host jump
384 events happen within the duck (mallards (MD) to other ducks (OD)) and gull (Black-
385 headed and Mediterranean Gulls (BMG) to Yellow-legged and Armenian Gulls (YAG)
386 and *vice versa*) species. In transmissions from ducks to gulls it is largely noticeable
387 only from MD to BMG. This likely explains the higher levels of nucleotide diversity and
388 reassortment rates in the BMG viruses relative to YAG seen above.

Geographical context for GE NS, M, NP, PA, PB1, PB2 segments

To determine the origin and destination of the internal protein coding gene segments found in viruses isolated in Georgia, we analysed our sequence dataset together with avian influenza sequences from a broader timeframe (2005-2016) and regional sampling. Figure 7A shows the genealogy for the NP gene for whose tips we know the location of sampling and whose internal nodes are estimated using discrete-state ancestral reconstruction in BEAST. Clades in which Georgian sequences occur are highlighted. Figure 7B summarises the genealogy in a circularised graph in which the arrowheads indicate the direction of transfer and the width of the arrow indicate the rate of transfer to different locations. The analyses reveal viruses from the Atlantic and Afro-Eurasian locations form largely separate clades, which is consistent with previous studies (44, 45). However, we do find instances of transmission across this divide, most notably to and from Asia and Europe. Many NP genes from Georgia cluster with other Georgian NP genes, in some cases forming the terminal branches spanning years indicating restriction to local spread. However, our dataset contains the latest Georgian sequences, and sequences from this timeframe were not available from the rest of Eurasia. Hence, we can expect to have missed identifying onward transmission. From the transmission we do identify, it appears that there is considerable migration into Africa and Europe and to a lesser extent to Southern/Eastern Asia. Most of the sequences transmitted into Georgia come from Asia and Europe, along with a single identified instance of direct transfer from North America.

Discussion

Wild birds have been shown to harbor substantial genetic diversity of avian influenza viruses. This study showed the diversity not only varied by year but was associated with the level of overall prevalence in different wild bird host species, perhaps

414 influencing the observed rates and diversity if prevalence were low. From these
415 results, there is little evidence that one species group maintains all influenza A virus
416 diversity, there appears to be relative host-restriction in many subtypes (except for H2
417 and H11 viruses) and there are differences in prevalence dynamics depending on
418 host. Therefore, one host is not representative of influenza A virus prevalence,
419 dynamics and diversity across the wild bird reservoir. Within both ducks and gulls
420 however, peak prevalence was consistently observed in hatch-year birds and with a
421 more restricted subtype diversity, suggesting that there is an initial influenza A virus
422 epidemic wave as naïve birds aggregate in their first year. Subsequently in the over-
423 wintering period, a wider subtype diversity was observed in both host groups and
424 adults were more frequently infected. This suggests that disease dynamics are
425 complex and influenced by multiple host factors including age and annual life cycle
426 stage.

427 It has previously been observed that some subtypes are routinely and nearly
428 exclusively isolated from certain host families/genus, the most notable example being
429 H13 and H16 viruses from gulls. However, mixed infections are relatively common but
430 might be masked if subtype characterization requires virus isolation, therefore putting
431 the clinical specimen through a culture bottleneck. Advances in sequencing direct from
432 clinical material would more accurately (remove possible culture selection bias)
433 establish the prevalence, subtype diversity and genetic diversity within wild birds.

434 In general, for all gene segments except PA, we identify strong patterns of clade
435 topology defined by host. This suggests that there is segregated gene flow through
436 these host populations with little inter-host reassortment. Additionally, within our study
437 period there were large scale perturbations in ecology which might also influence our

prevalence and subtype diversity estimates. For example, in 2014 and 2015 there was widespread reproductive failure in two gull host species due to nest flooding (Yellow-legged Gulls) and few returning adults to the colony (Armenian Gulls), and therefore few juveniles from which to detect the annual epidemic wave. The occurrence and significance of such ecological fluctuations on disease dynamics are unclear. We also increased the ability to sample migrant ducks in late summer and early autumn from August 2015 by constructing a duck trap in the newly created National Park. Again, this addition to sampling strategy likely increased the detection of influenza in these anseriform hosts as they were previously under-sampled.

We tested whether certain hosts maintained higher levels of nucleotide diversity in the non-immune related internal genes. PB2 and NS were the most genetically diverse in both gulls and ducks. Within host-group, Black-headed and Mediterranean Gull-derived viruses showed highest per-site diversity, Yellow-legged and Armenian Gulls lower diversity, likely because some of the viruses of the former were associated with reassortants probably derived from ducks (or another unsampled host group). While despite high rates of reassortment and spillover between duck subgroups mallards (MD) and other ducks (OD), the absence of any gull derived viruses in these ducks keeps their diversity levels lower compared to gulls/BMG.

Where gene flow does occur between host groups, for all gene segments, host-spillover events were in the direction of ducks to gulls and from other ducks to Black-headed and Mediterranean Gulls, likely explaining the higher levels of nucleotide diversity in these gulls observed above. Where HA and NA gene segments were acquired by gulls from ducks, there was a pre-requisite for a gull-clade internal gene cassette suggesting a host-restrictive effect for onward maintenance within the gull

462 population (11, 46). Interestingly, Black-headed and Mediterranean Gulls only occur
463 on the study site in the over-wintering period where there are also high densities of
464 over-wintering ducks from other geographic areas. Although there is a duck-gull
465 interface on the breeding grounds in summer, the duck densities are very much lower,
466 perhaps suggesting that there is a threshold level of bird density that allows gene flow
467 among hosts.

468 If we look at diversity by HA subtype, H4 and H13 were the least diverse and
469 showed the lowest rates of reassortment and were also associated with hatch-year
470 bird infections, suggesting a clonal expansion and epidemic gene flow through these
471 birds. The 2014-2015 HPAI H5 epizootic also showed no reassortment unlike the
472 2016-2017 HPAI H5 viruses, perhaps indicating that the first wave of 2.3.4.4 viruses
473 diffused through the wild bird population similarly to a 'naïve' infection, and subsequent
474 epizootics have resulted in altered pathogen evolution strategies to maintain gene
475 flow, similar to those previously observed in North America when considering the effect
476 of latitude on gene flow (7).

477 When we examine the internal gene segments of the Georgian AIV in a broader
478 geographical context, we find significant gene flow to and from Georgia with Europe
479 and the rest of Asia, although data for Africa is very limited. Crossover into the Atlantic
480 flyway appears to be mediated largely by gulls with some exceptions, notably the
481 H5N1-NP gene that was transmitted between ducks.

482 From this study, the diffusion of avian influenza viruses within a multi-host
483 ecosystem is heterogeneous. One host group cannot therefore be used as a surrogate
484 for others. It is likely that virus evolution in these natural eco-systems is a complex mix
485 of host-pathogen interface and ecological factors. Understanding such drivers is key

to investigating these emerging pathogens, interpreting the data from different sites around the world and ultimately informing risk of incursion of emerging variants from one geographic region to another.

Acknowledgements:

This study including field work and sequencing was funded by National Institute of Allergy and Infectious Diseases, National Institutes of Health, Department of Health and Human Services contract No.HHSN272200090007C and HHSN266200700010C “NIAID Centres of Excellence for Influenza Research and Surveillance” http://www.niaid.nih.gov/LabsAndResources/resources/ceirs/Pages/crip.aspx, and a DTRA FRCWMD Broad Agency Announcement (HDTRA1-09-14-FRCWMD GRANT11177182). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript. The sequencing data for this manuscript was generated while D. E. Wentworth was employed at the J. Craig Venter Institute. The opinions expressed in this article are the author's own and do not reflect the view of the Centers for Disease Control and Prevention, the Department of Health and Human Services, or the United States government.

References

1. Olsen B, Munster VJ, Wallensten A, Waldenstrom J, Osterhaus AD, Fouchier RA. 2006. Global patterns of influenza a virus in wild birds. Science 312:384-8.
2. Webster RG, Bean WJ, Gorman OT, Chambers TM, Kawaoka Y. 1992. Evolution and ecology of influenza A viruses. Microbiol Rev 56:152-79.
3. Gunnarsson G, Jourdain E, Waldenstrom J, Helander B, Lindberg P, Elmberg J, Latorre-Margalef N, Olsen B. 2010. Zero prevalence of influenza A virus in two raptor species by standard screening. Vector Borne Zoonotic Dis 10:387-90.
4. Latorre-Margalef N, Gunnarsson G, Munster VJ, Fouchier RA, Osterhaus AD, Elmberg J, Olsen B, Wallensten A, Haemig PD, Fransson T, Brudin L, Waldenstrom J. 2009. Effects of influenza A virus infection on migrating mallard ducks. Proc Biol Sci 276:1029-36.
5. Bengtsson D, Safi K, Avril A, Fiedler W, Wikelski M, Gunnarsson G, Elmberg J, Tolf C, Olsen B, Waldenstrom J. 2016. Does influenza A virus infection affect movement behaviour during stopover in its wild reservoir host? R Soc Open Sci 3:150633.
6. Verhagen JH, van Dijk JG, Vuong O, Bestebroer T, Lexmond P, Klaassen M, Fouchier RA. 2014. Migratory birds reinforce local circulation of avian influenza viruses. PLoS One 9:e112366.

7. Hill NJ, Ma EJ, Meixell BW, Lindberg MS, Boyce WM, Runstadler JA. 2016. Transmission of influenza reflects seasonality of wild birds across the annual cycle. Ecol Lett doi:10.1111/ele.12629.
8. Alexander DJ. 2007. An overview of the epidemiology of avian influenza. Vaccine 25:5637-44.
9. Lee DH, Bertran K, Kwon JH, Swayne DE. 2017. Evolution, global spread, and pathogenicity of highly pathogenic avian influenza H5Nx clade 2.3.4.4. J Vet Sci 18:269-280.
10. Verhagen JH, Herfst S, Fouchier RA. 2015. Infectious disease. How a virus travels the world. Science 347:616-7.
11. Munster VJ, Baas C, Lexmond P, Waldenstrom J, Wallensten A, Fransson T, Rimmelzwaan GF, Beyer WE, Schutten M, Olsen B, Osterhaus AD, Fouchier RA. 2007. Spatial, temporal, and species variation in prevalence of influenza A viruses in wild migratory birds. PLoS Pathog 3:e61.
12. Latorre-Margalef N, Tolf C, Grosbois V, Avril A, Bengtsson D, Wille M, Osterhaus AD, Fouchier RA, Olsen B, Waldenstrom J. 2014. Long-term variation in influenza A virus prevalence and subtype diversity in migratory mallards in northern Europe. Proc Biol Sci 281:20140098.
13. Lewis NS, Javakhishvili Z, Russell CA, Machablishvili A, Lexmond P, Verhagen JH, Vuong O, Onashvili T, Donduashvili M, Smith DJ, Fouchier RA. 2013. Avian influenza virus surveillance in wild birds in Georgia: 2009-2011. PLoS One 8:e58534.
14. Dusek RJ, Hallgrimsson GT, Ip HS, Jonsson JE, Sreevatsan S, Nashold SW, TeSlaa JL, Enomoto S, Halpin RA, Lin X, Fedorova N, Stockwell TB, Dugan VG, Wentworth DE, Hall JS. 2014. North Atlantic migratory bird flyways provide routes for intercontinental movement of avian influenza viruses. PLoS One 9:e92075.
15. Lindsay LL, Kelly TR, Plancarte M, Schobel S, Lin X, Dugan VG, Wentworth DE, Boyce WM. 2013. Avian influenza: mixed infections and missing viruses. Viruses 5:1964-77.
16. Fries AC, Nolting JM, Bowman AS, Lin X, Halpin RA, Wester E, Fedorova N, Stockwell TB, Das SR, Dugan VG, Wentworth DE, Gibbs HL, Slemons RD. 2015. Spread and persistence of influenza A viruses in waterfowl hosts in the North American Mississippi migratory flyway. J Virol 89:5371-81.
17. Nolting JM, Fries AC, Gates RJ, Bowman AS, Slemons RD. 2016. Influenza A Viruses from Overwintering and Spring-Migrating Waterfowl in the Lake Erie Basin, United States. Avian Dis 60:241-4.
18. Bahl J, Vijaykrishna D, Holmes EC, Smith GJ, Guan Y. 2009. Gene flow and competitive exclusion of avian influenza A virus in natural reservoir hosts. Virology 390:289-97.
19. Fourment M, Darling AE, Holmes EC. 2017. The impact of migratory flyways on the spread of avian influenza virus in North America. BMC Evol Biol 17:118.
20. Chen R, Holmes EC. 2009. Frequent inter-species transmission and geographic subdivision in avian influenza viruses from wild birds. Virology 383:156-61.
21. Anderson TK, Campbell BA, Nelson MI, Lewis NS, Janas-Martindale A, Killian ML, Vincent AL. 2015. Characterization of co-circulating swine influenza A viruses in North America and the identification of a novel H1 genetic clade with antigenic significance. Virus Res 201:24-31.
22. Lewis NS, Verhagen JH, Javakhishvili Z, Russell CA, Lexmond P, Westgeest KB, Bestebroer TM, Halpin RA, Lin X, Ransier A, Fedorova NB, Stockwell TB, Latorre-Margalef N, Olsen B, Smith G, Bahl J, Wentworth DE, Waldenstrom J, Fouchier RA, de Graaf M. 2015. Influenza A virus evolution and spatio-temporal dynamics in Eurasian wild birds: a phylogenetic and phylogeographical study of whole-genome sequence data. J Gen Virol 96:2050-60.
23. OIE. 2015. Avian influenza (Infection with avian influenza viruses), Avian influenza (infection with avian influenza viruses): Manual of Diagnostic Tests and Vaccines for Terrestrial Animals. World Organisation for Animal Health (OIE), Paris, France.

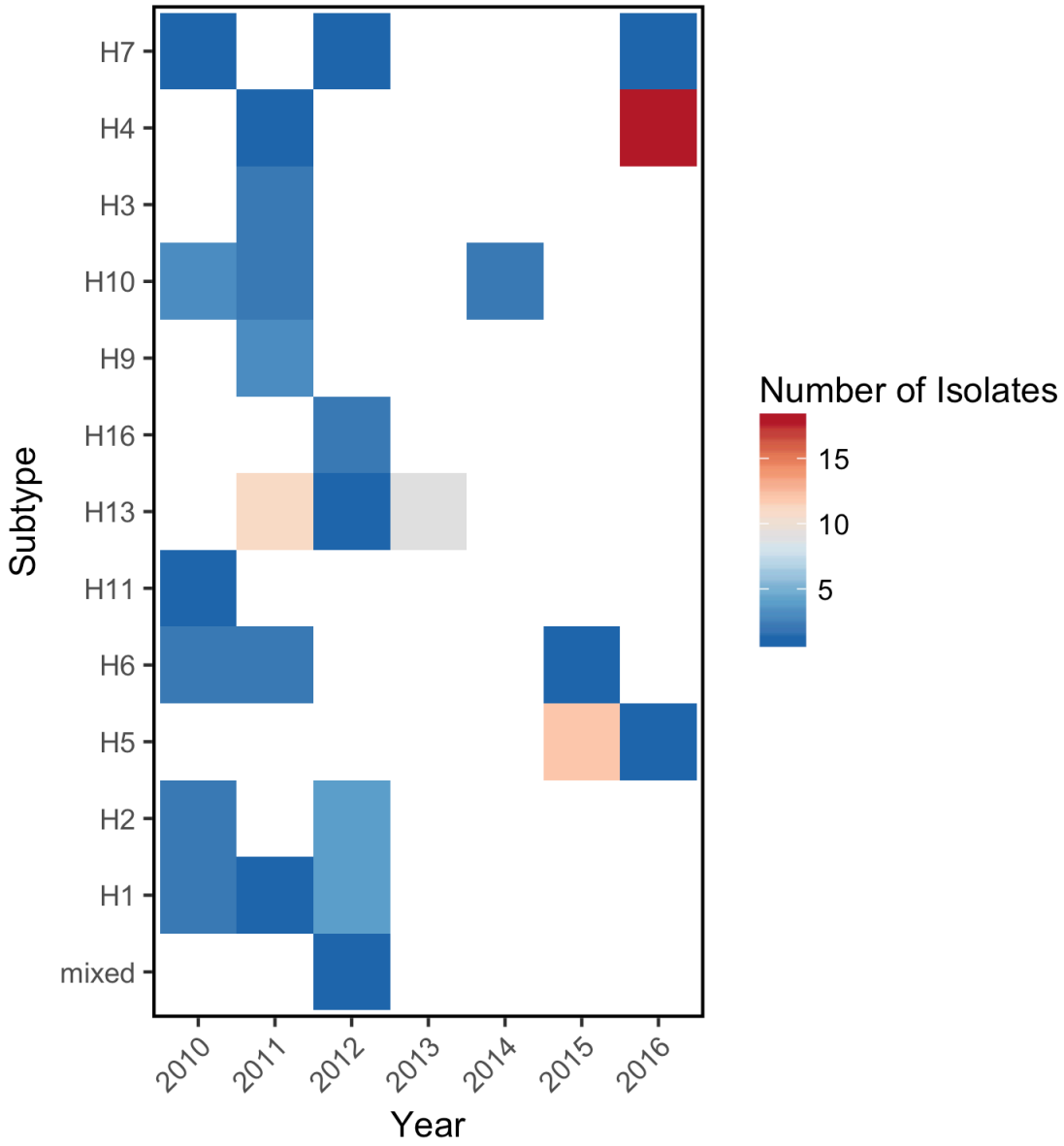
24. Zhou B, Donnelly ME, Scholes DT, St George K, Hatta M, Kawaoka Y, Wentworth DE. 2009. Single-reaction genomic amplification accelerates sequencing and vaccine production for classical and Swine origin human influenza a viruses. J Virol 83:10309-13.
25. Zhou B, Wentworth DE. 2012. Influenza A virus molecular virology techniques. Methods Mol Biol 865:175-92.
26. Munster VJ, Baas C, Lexmond P, Bestebroer TM, Guldemeester J, Beyer WEP, de Wit E, Schutten M, Rimmelzwaan GF, Osterhaus ADME, Fouchier RAM. 2009. Practical considerations for high-throughput influenza A virus surveillance studies of wild birds by use of molecular diagnostic tests. Journal of Clinical Microbiology 47:666-673.
27. Mena I, Nelson MI, Quezada-Monroy F, Dutta J, Cortes-Fernandez R, Lara-Puente JH, Castro-Peralta F, Cunha LF, Trovao NS, Lozano-Dubernard B, Rambaut A, van Bakel H, Garcia-Sastre A. 2016. Origins of the 2009 H1N1 influenza pandemic in swine in Mexico. Elife 5.
28. Katoh K, Standley DM. 2013. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol Biol Evol 30:772-80.
29. McCauley J, Bye J, Elder K, Gething MJ, Skehel JJ, Smith A, Waterfield MD. 1979. Influenza virus haemagglutinin signal sequence. FEBS Lett 108:422-6.
30. Burke DF, Smith DJ. 2014. A recommended numbering scheme for influenza A HA subtypes. PLoS One 9:e112302.
31. Kumar S, Stecher G, Tamura K. 2016. MEGA7: Molecular Evolutionary Genetics Analysis Version 7.0 for Bigger Datasets. Mol Biol Evol 33:1870-4.
32. Nguyen LT, Schmidt HA, von Haeseler A, Minh BQ. 2015. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. Mol Biol Evol 32:268-74.
33. Kalyaanamoorthy S, Minh BQ, Wong TKF, von Haeseler A, Jermiin LS. 2017. ModelFinder: fast model selection for accurate phylogenetic estimates. Nat Methods 14:587-589.
34. Rambaut A, Lam TT, Max Carvalho L, Pybus OG. 2016. Exploring the temporal structure of heterochronous sequences using TempEst (formerly Path-O-Gen). Virus Evol 2:vew007.
35. Bell SM, Bedford T. 2017. Modern-day SIV viral diversity generated by extensive recombination and cross-species transmission. PLoS Pathog 13:e1006466.
36. Pfeifer B, Wittelsburger U, Ramos-Onsins SE, Lercher MJ. 2014. PopGenome: an efficient Swiss army knife for population genomic analyses in R. Mol Biol Evol 31:1929-36.
37. Parker J, Rambaut A, Pybus OG. 2008. Correlating viral phenotypes with phylogeny: accounting for phylogenetic uncertainty. Infect Genet Evol 8:239-46.
38. Ronquist F, Teslenko M, van der Mark P, Ayres DL, Darling A, Hohna S, Larget B, Liu L, Suchard MA, Huelsenbeck JP. 2012. MrBayes 3.2: efficient Bayesian phylogenetic inference and model choice across a large model space. Syst Biol 61:539-42.
39. Drummond AJ, Suchard MA, Xie D, Rambaut A. 2012. Bayesian phylogenetics with BEAUti and the BEAST 1.7. Mol Biol Evol 29:1969-73.
40. Fu L, Niu B, Zhu Z, Wu S, Li W. 2012. CD-HIT: accelerated for clustering the next-generation sequencing data. Bioinformatics 28:3150-2.
41. Li W, Godzik A. 2006. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics 22:1658-9.
42. Kawaoka Y, Gorman OT, Ito T, Wells K, Donis RO, Castrucci MR, Donatelli I, Webster RG. 1998. Influence of host species on the evolution of the nonstructural (NS) gene of influenza A viruses. Virus Res 55:143-56.
43. Nelson MI, Detmer SE, Wentworth DE, Tan Y, Schwartzbard A, Halpin RA, Stockwell TB, Lin X, Vincent AL, Gramer MR, Holmes EC. 2012. Genomic reassortment of influenza A virus in North American swine, 1998-2011. J Gen Virol 93:2584-9.

629 44. Lu L, Lycett SJ, Leigh Brown AJ. 2014. Reassortment patterns of avian influenza virus
630 internal segments among different subtypes. BMC Evol Biol 14:16.
631 45. Ma HC, Chen JM, Chen JW, Sun YX, Li JM, Wang ZL. 2007. The panorama of the
632 diversity of H5 subtype influenza viruses. Virus Genes 34:283-7.
633 46. Tonnessen R, Hauge AG, Hansen EF, Rimstad E, Jonassen CM. 2013. Host
634 restrictions of avian influenza viruses: in silico analysis of H13 and H16 specific
635 signatures in the internal proteins. PLoS One 8:e63270.

636

Figure 1. HA subtype-wise breakdown (A) and overall (B) yearly prevalence of viruses in Georgia during 2010-16. X-axis marks year of isolation. In bar chart 1A the Y-axis marks the proportion virus-positive samples +/- standard deviation and bars are colored according to host from which virus was isolated (duck in pink and gull in green). In heat map 1B, the Y-axis shows the HA subtypes of viruses isolated and squares are colored according to the number of isolates of each type identified.

A



B

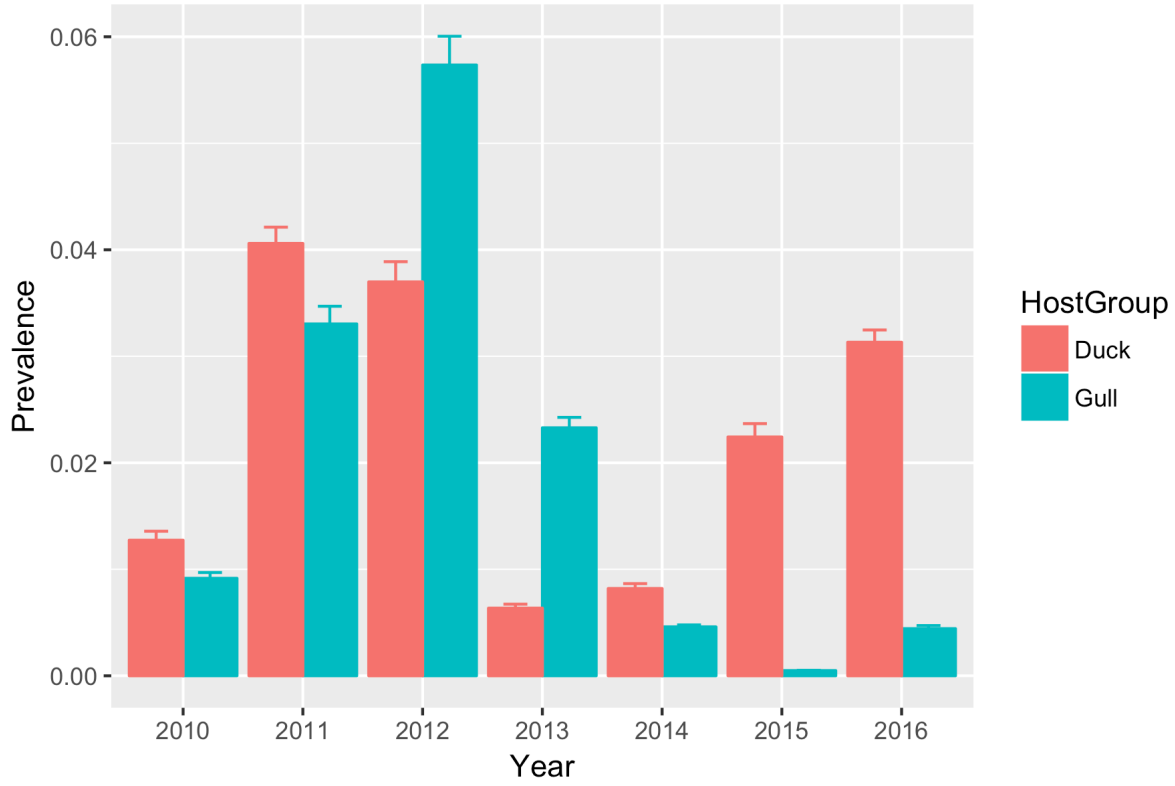


Figure 2. Maximum-likelihood trees for all internal genes – PB2, PB1, MP, NS, NP and PA, from equivalent strains connected across the trees. Tips and connecting lines are coloured according to host type –mallards (MD) bright red and other ducks (OD) orange, yellow-legged and armenian gulls (YAG) in blue and black-headed and mediterrenean gulls (BMG) in light blue).

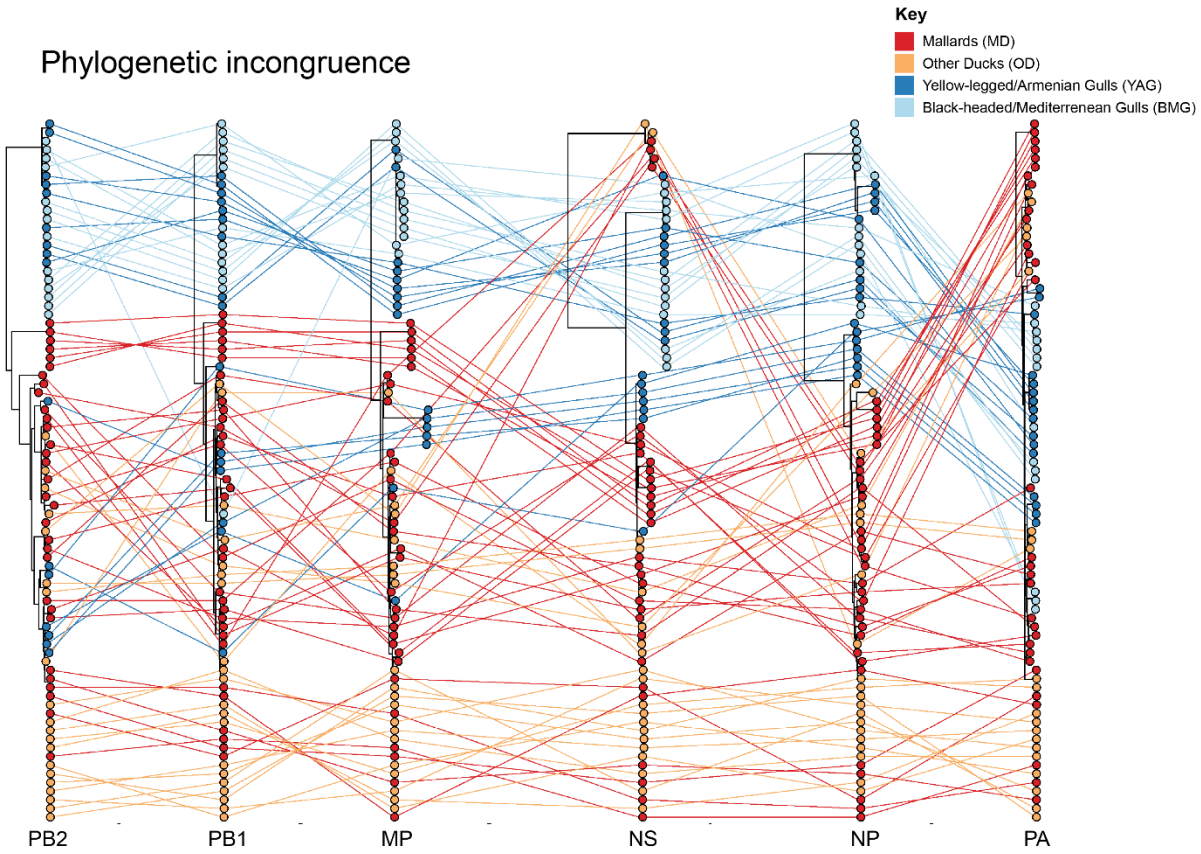
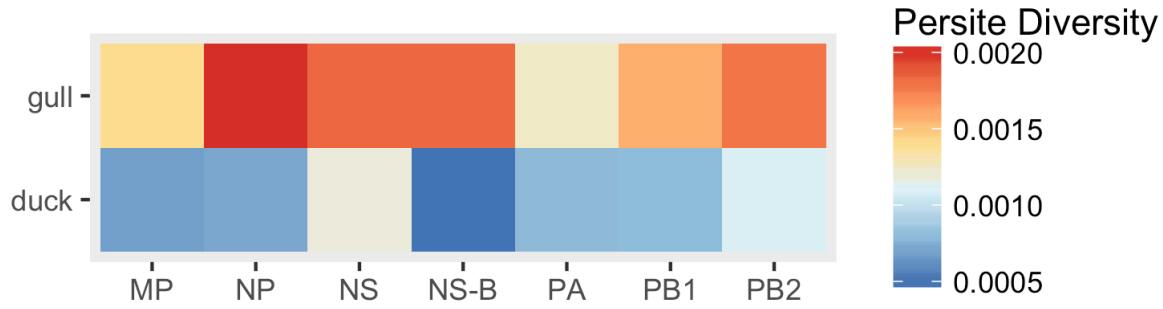
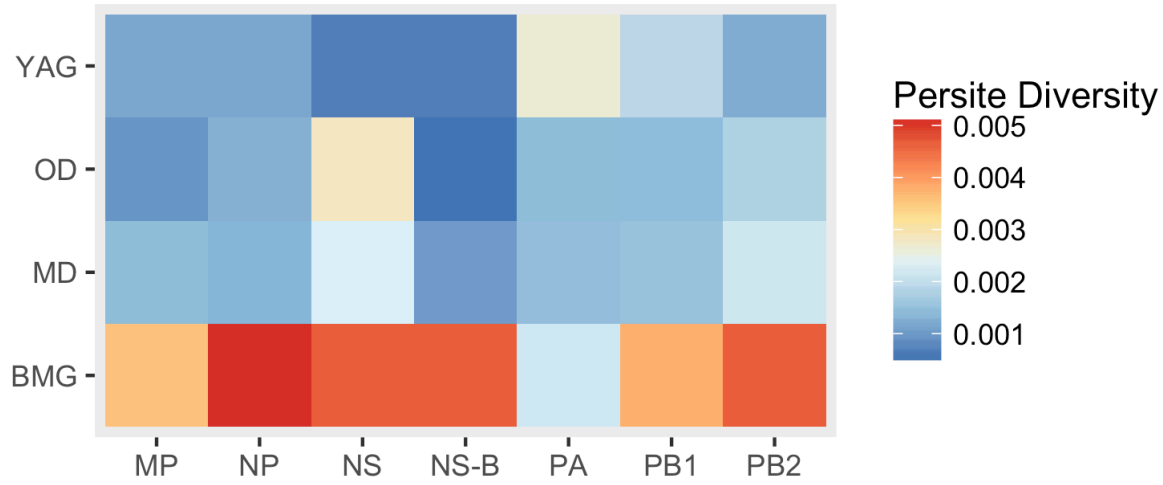


Figure 3. Overall per-site nucleotide diversity defined as average number of nucleotide differences per site between two sequences in all possible pairs in the sample population, normalised to the number of sequences in each population. Comparison between (A) gulls and ducks. (B) host-types (MD, OD, YAG, BMG) and (C) HA type are shown.

A



B



C

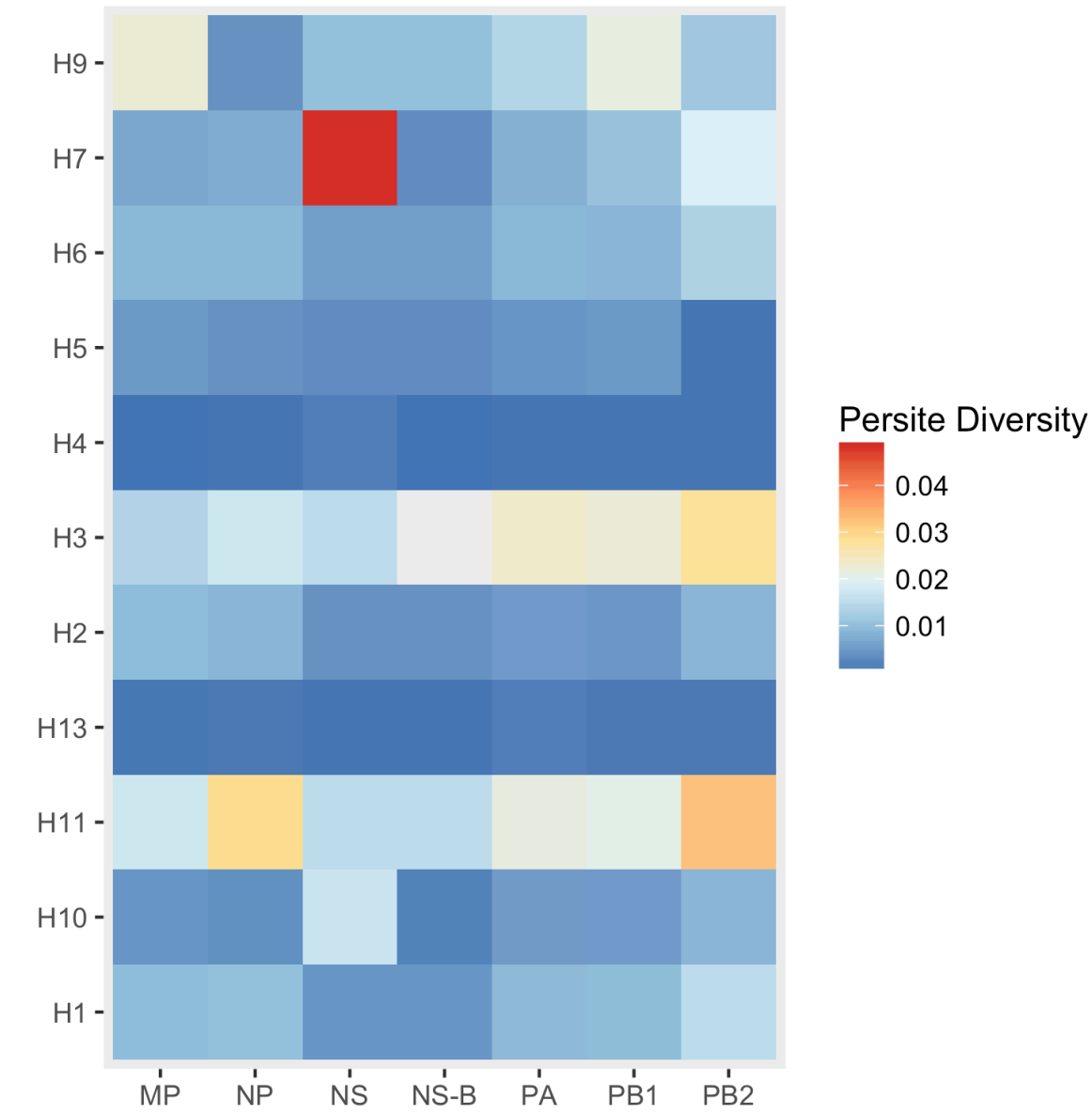
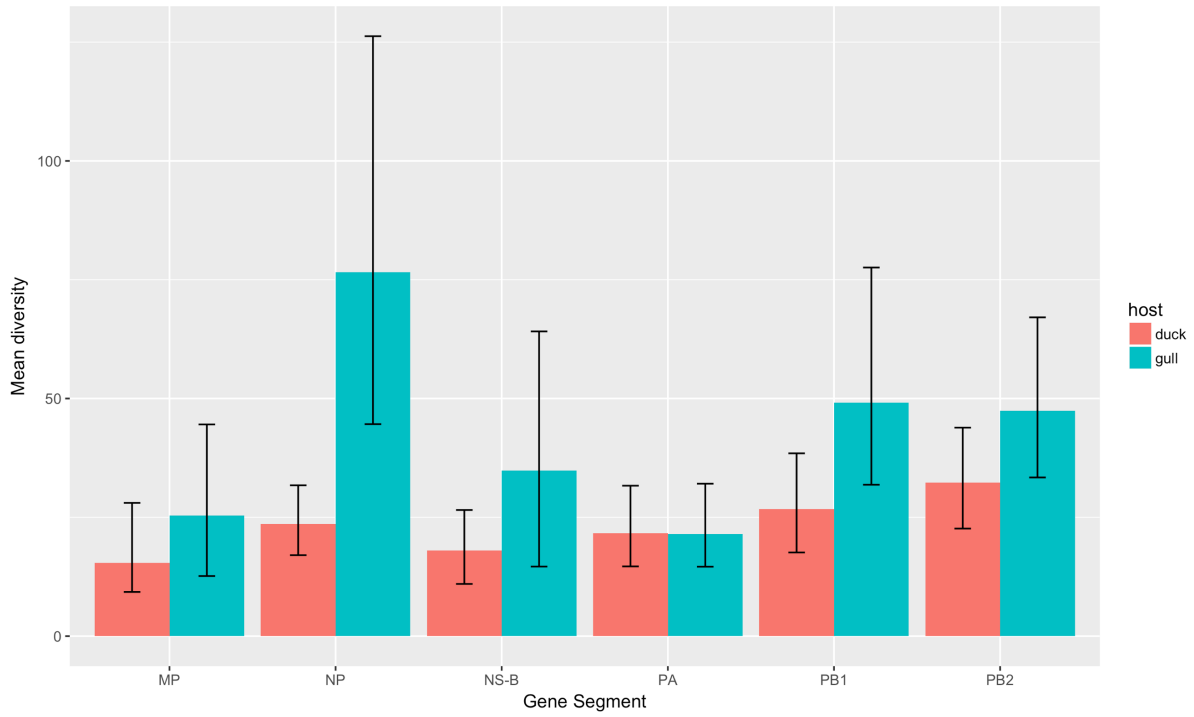
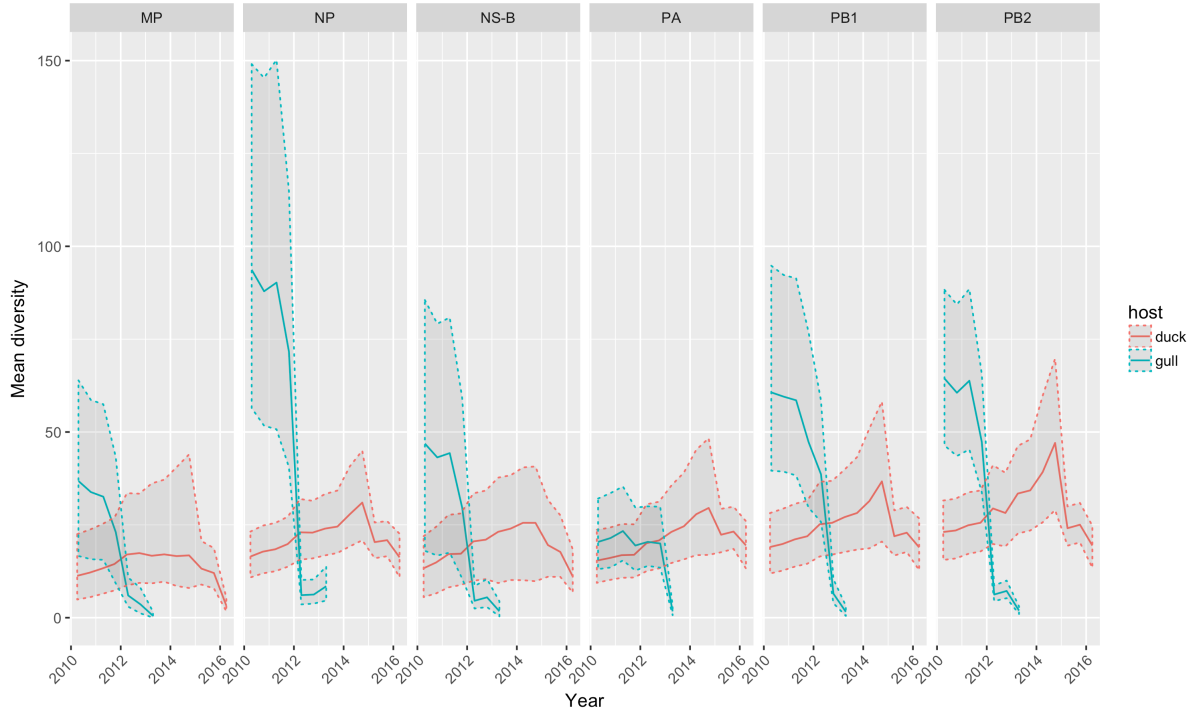


Figure 4. Overall/summary (A) and over-time/skyline (B) mean diversity for each segment from gulls (green) and ducks (pink) as determined by posterior analysis of coalescent trees (PACT). Here, diversity is defined as the average time to coalescence for pairs of lineages belonging to each host. Panel (C) shows overall/summary mean diversity values for ducks divided in to MD, OD (light and dark blue), and gulls divided into BMG and YAG (light and dark green).

A



B



C

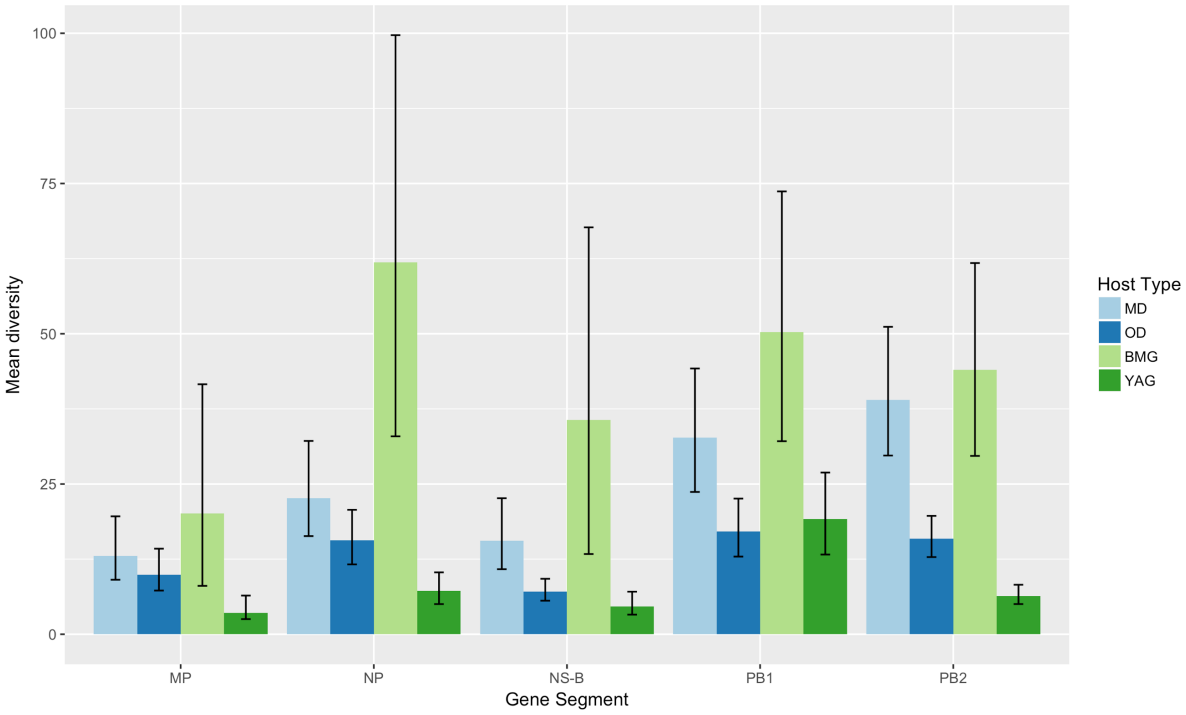
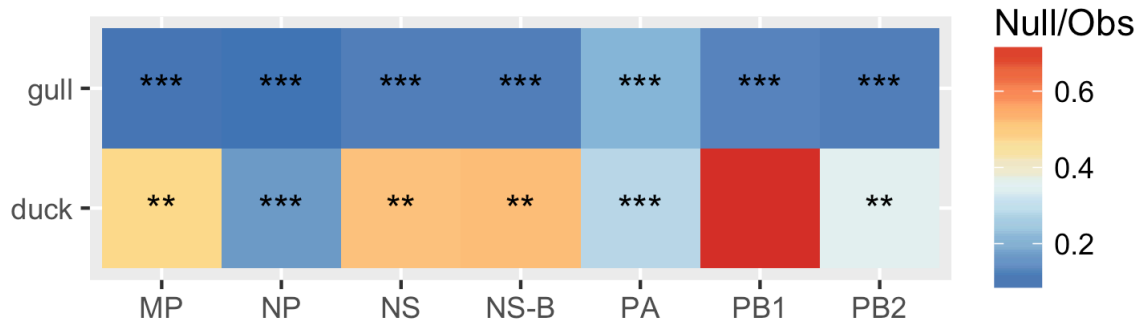
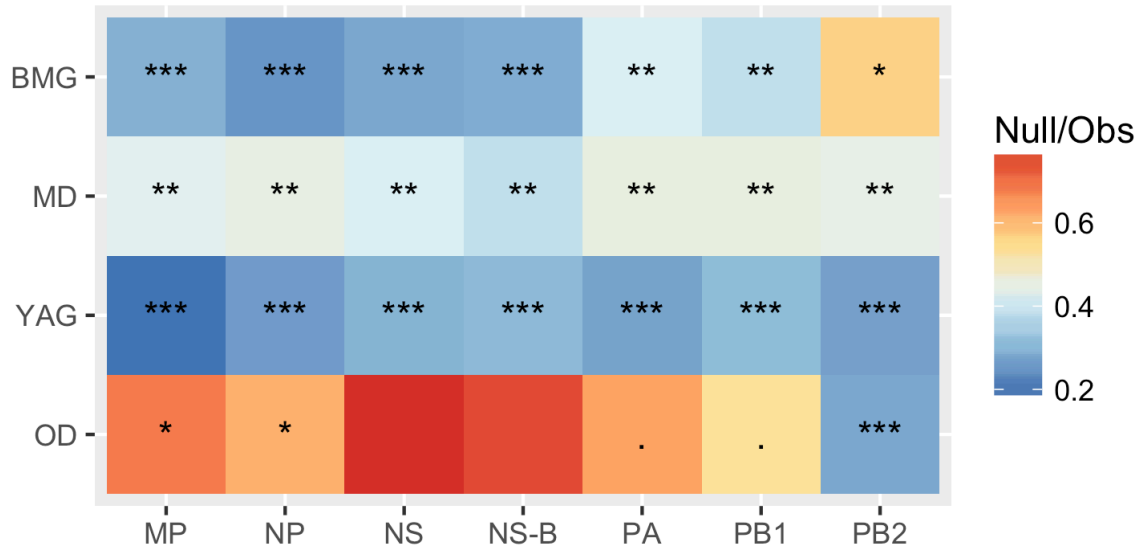


Figure 5. Summaries of expected/observed ratios from Bayesian Tip-association Significance testing (BaTS) for all internal genes. Higher values indicate less phylogenetic clustering by trait and hence higher rates of mixed ancestry. Comparison between (A) gulls and ducks. (B) host-types (MD, OD, YAG, BMG) and (C) HA type are shown. Asterisks indicate p-values (*** < 0.001, ** < 0.01, * < 0.05 and no asterisk > 0.05).

A



B



C

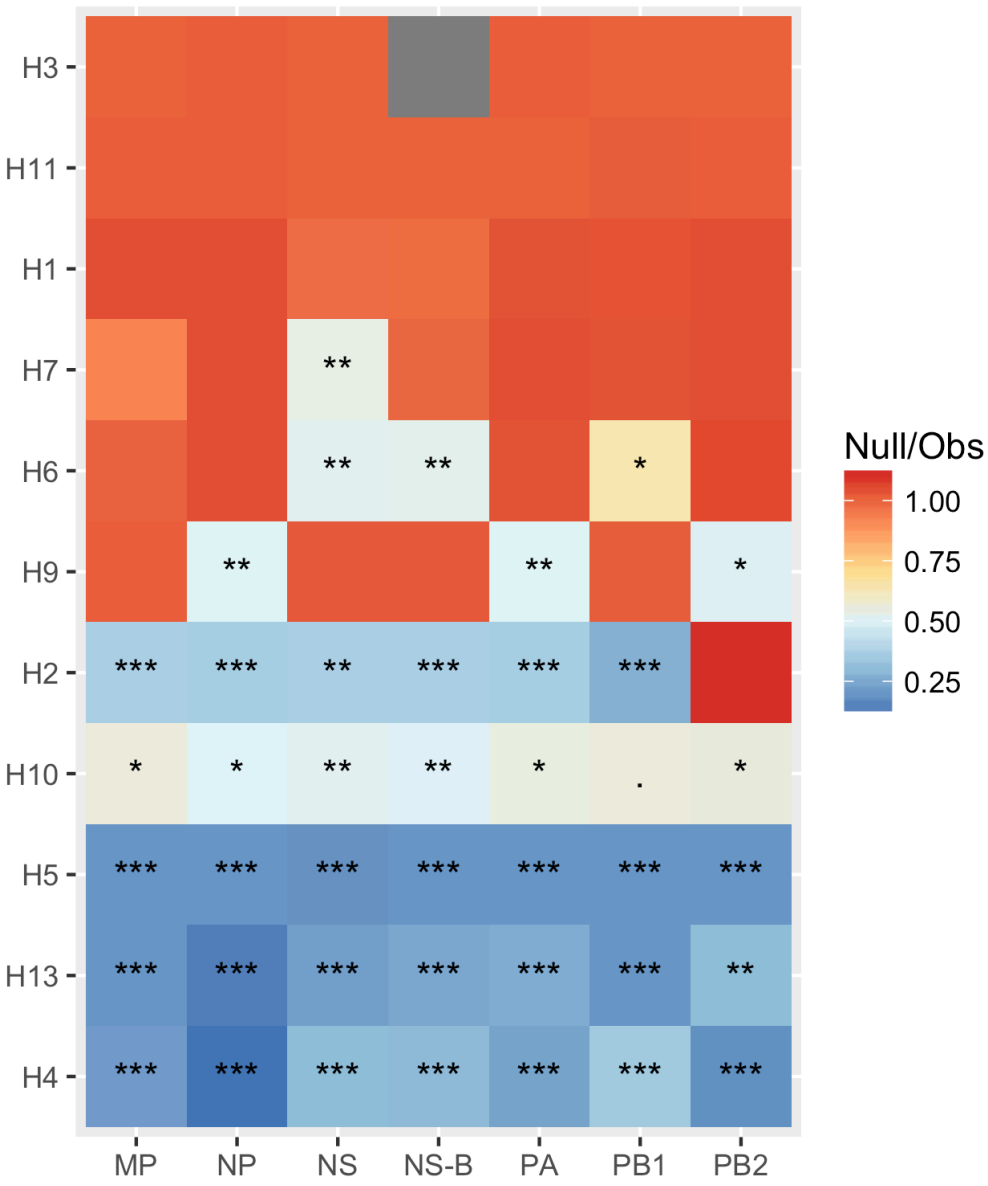
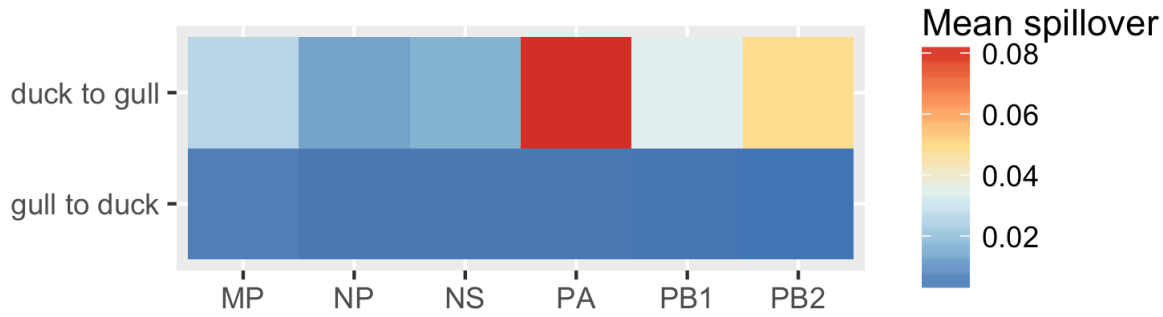


Figure 6A. Summary of mean migration events between hosts in the direction from (A) duck to gull and gull to duck, and (B) between different host types - Mallards (MD), other ducks (OD), Black-headed and Mediterranean gulls (BMG) and Yellow-legged and Armenian gulls (YAG) derived from the genealogy.

A



B

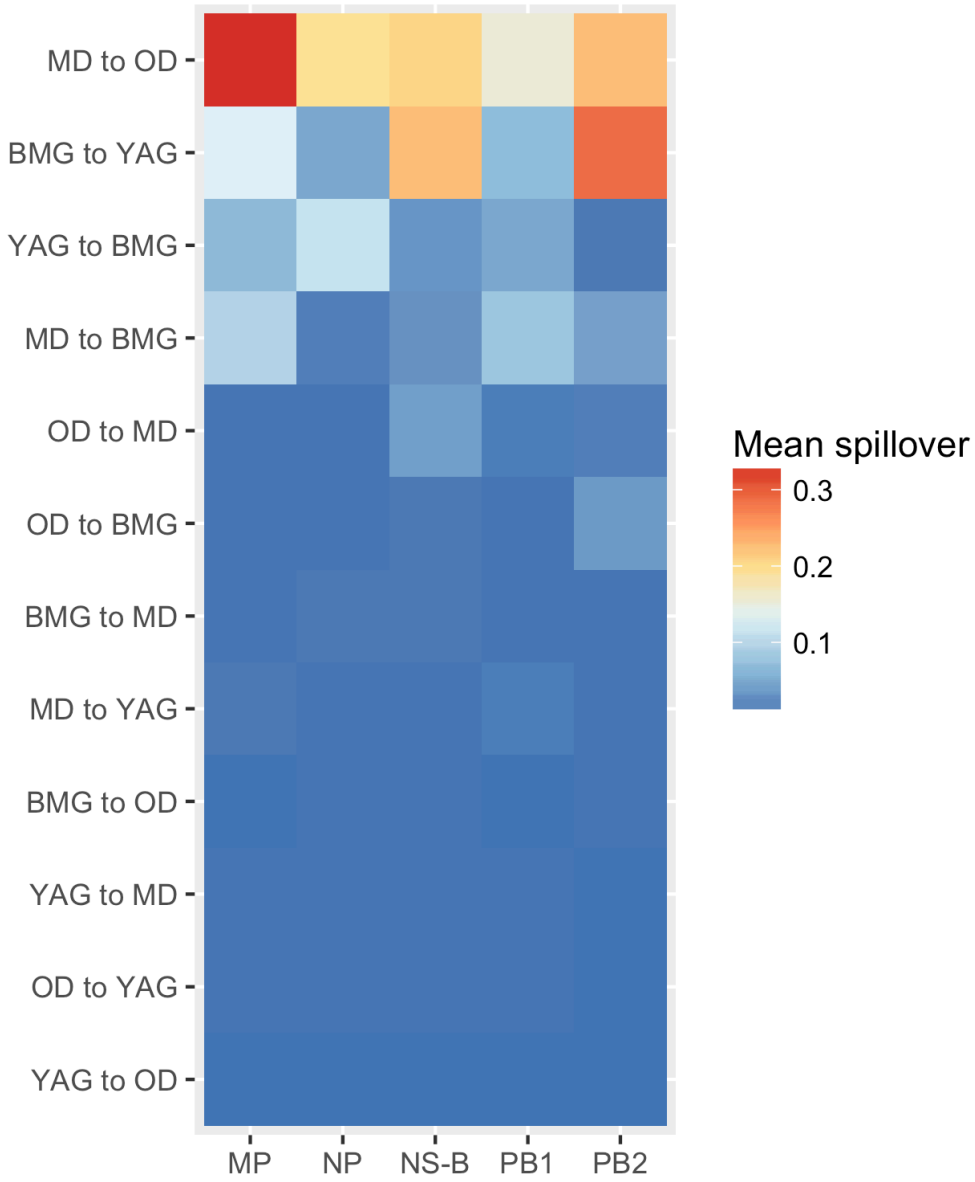


Figure 7A. BEAST MCC (median-clade credibility) trees from viral sequences NP gene sequences isolated world-wide from avian hosts between 2005 and 2016. Branches are coloured according to location, observed at the tips and estimated at internal nodes by ancestral reconstruction of discrete trait. Africa, Asia, Europe in very dark, dark and light purple, Georgian sequences from this study in light green, North and South America in dark and very dark green.

Figure 7B. Circularised graph shows overall rates of migration, defined as the rate at which labels (locations) change over the course of the genealogy, between Georgia and other locations. Arrow heads indicate direction of migration; rates are measured as migration events per lineage per year (indicated by the width of the arrow). Asia in blood orange, Africa in orange, Georgia in yellow, Europe in green, South America in teal and North America in blue.