**Abstract**

Avian influenza A (H5N6) virus is widespread in China Mainland, with frequent outbreaks in poultry, causing unnecessary damage and potentially infecting humans and posing a serious threat to public health. In this study, we analyzed the HA gene segment for better understanding the evolution and genetic variation of the virus in China Mainland. Using phytogeographic and spatial epidemiological methods to discern the genetic variation, predictors and effects of transmission of H5N6 virus. We found that HA gene selection pressure is relatively conserved, may produce key mutations that lead to spread in the population. H5N6 is predominantly concentrated in south-central China and tends to spread like around. Guangdong is the main exporting province. High risk of outbreak in South-west region, GLM analysis reveals several ecological factors, population, water resources and relative humidity associated with the spread of the virus and provides information for preventing and controlling the spread of the virus. These findings suggested a complex interaction between virus evolution, epidemiology, ecological factors and human behavior.

**Introduction**

Avian influenza A (H5N6) viruses is a subtype of the Avian Influenza A virus, Avian influenza A viruses generally do not cause disease in aquatic birds, the natural reservoir of these viruses (The evolution of H5N1 influenza viruses in ducks in southern China). Influenza A(H5N6) was first isolated from mallards by García et al. in 1975 (Evolution of H5 subtype avian influenza A viruses in North America). Influenza A(H5N6) outbreak caused the death of birds and ducks and found infection in a Domestic Cat (Fatal H5N6 Avian Influenza Virus Infection in a Domestic Cat and Wild Birds in China). In April 2014, the World Health Organisation (WHO) first reported a case of human infection with avian influenza A(H5N6) from China. Surveillance of AIVs in live poultry markets demonstrated that they are the source of H5N6 human infections with (Two Novel Reassortants of Avian Influenza A(H5N6) Virus in China). To 3 April 2020, a total of 24 laboratory-confirmed cases of human infection with influenza A(H5N6) virus, including seven deaths at time of IHR report, have been reported to WHO from China since 2014 ([www.who.int](http://www.who.int)). H5N6 cause tremendous economic losses to the poultry industry and represent a serious threat to public health.

There is already evidence that H5N6 is pathogenic and transmissible in mammals and H5N6 viruses acquired varying degrees of binding affinity for human-like receptors (Avian Influenza H5N6 Viruses Exhibit Differing Pathogenicities and Transmissibilities in Mammals). Fortunately, the weak affinity of these viruses for α-2,6-linked sialic acid receptors imited human transmissibility (The role of receptor binding specificity in interspecies transmission of influenza viruses). The haemagglutinin (HA) gene binding the HA glycoprotein binding sialic acid receptors on the surface of host cells (The role of receptor binding specificity in interspecies transmission of influenza viruses) (Predicting ‘airborne’ influenza viruses: (trans-) mission impossible?). Evolution and variation of HA gene of virus determines whether the virus can spread efficiently in the host population. We analyzed publicly available H5N6 HA gene sequence data covering 2013 to 2017 in China Mainland. Here we provide insight into the natural selection of HA gene that was tested using a variety of methods and spatially and temporally heterogeneous evolutionary dynamics and migration of H5N6 viruses. Different geographic division schemes, including one that subdivision by city and subdivision by economic region, were applicable to phylogeographic analysis of continuous and discrete space (Phylogeography takes a relaxed random walk in continuous space and time) (Bayesian phylogeography finds its roots). Phylogeny-trait associations for different location states were estimated by BaTS. A model-based approach to explicitly tests spatial epidemiological hypotheses by integrating empirical data on ecological factors with viral genetic data (Unifying Viral Genetics and Human Transportation Data to Predict the Global Transmission Dynamics of Human Influenza H3N2). This framework enables us to measure the relative contribution of different predictors and identify key drivers of viral spatial diffusion. Surface water resources, resident population, total number of sequences used in the analysis, rural population, sales of poultry per capita rural household, percentage of nature reserves in the region, sum of smoke and dust, sulphur dioxide, nitrogen oxides, average relative humidity of major cities and the average distance between centroid point of each locations were used in the model. Therefore, we focus on evolutionary analysis of HA gene and try to reveal the natural selection, population dynamics, migration direction and geographic origin of H5N6 in China Mainland.

Result

1. **Maximum Likelihood Phylogenetic Analysis**

A maximum-likelihood (ML) phylogeny with bootstrap node support values was reconstructed from HA codon alignments. Branch lengths in the ML tree were rescaled into units of real calendar time using Least-Squares Dating. The ancestral location of each internal node in the rooted ML tree was then reconstructed using maximum parsimony to reveal how viral lineages have moved over time. The ML tree allows clearly distinguish the topological structure, the bootstrap values (>80) are displayed by a thick branch line, the ancestral location are colored and pie chart in each node represents the probability of each node being in each state (Fig. 1). The location annotated ML tree indicated that Guangdong is the main area of early virus transmission. A ML tree annotated by five regions (CT, NW, PRD, SW, YRD) was also generated (Fig. 1S). Although the other four regions have the same probability at the root except for NW, the PRD, the area where Guangdong is located, still occupies the main trunk of the previous period. Consistent with the conclusions obtained from the ML tree annotated with provincial administrative regions. Linear regression analysis of sequence sampling dates and root-to-tip genetic distances after obtaining root-to-tip genetic distances from the ML tree using TempEst (Fig. 2a). The result supports the use of a Molecular clock analysis in this study.

1. **Natural selection**

Selective pressure analysis was performed on the HA codon alignments using five methods: FEL (Fixed Effects Likelihood), FUBAR (Fast, Unconstrained Bayesian AppRoximation), MEME (Mixed Effects Model of Evolution) and SLAC (Single-Likelihood Ancestor Counting) in HYPHY and BEAST to infer nonsynoymous (dN) and synonymous (dS) substitution rates on a per-site basis. Positively selected sites detected by at least four methods were considered statistically significant. Six positive selection sites were considered potential positive selection sites (Fig. 2c), of which 154 and 230 were detected by all methods in in the HA1 gene (Table 3). Obviously, most of sites in the HA1 gene are under strong negative selection, indicating strongly conserved. For viruses from the pandemic period, residues under positive selection were mostly located on the HA stem region that is generally conserved.

1. **Evolutionary rate and divergence time estimation**

A relaxed molecular clock along with a Bayesian SkyGrid model was selected by marginal likelihood comparison using PS and SS estimation approaches as the best fit model combination (Tabel1, Fig. 2b). The mean rate of evolution was estimated to be 7.34 × 10^(-3) substitutions per site per year (subs/site/year) (95% highest posterior density (HPD) interval = 5.91 × 10^(-3) to 8.80 × 10^(-3)). Correspondingly, mean root height was estimated to be 6.865 (95% highest posterior density (HPD) interval = 4.717 to 11.165), indicating that the date of the most recent common ancestor (TMRCA) was 2010.66 (August 30, 2010) (95% highest posterior density (HPD) interval = 2006.369 to 2012.817). In root-to-tip regression analysis, the evolutionary rate was estimated to be 6.09 × 10 ^ (-3) and the TMRCA 2009.97 (December 21, 2009). The estimated TMRCA using the 'Least Squares Dating' method of lsd-0.3beta was 2009.359. The results obtained with these different methods are similar. Date-randomization tests were performed 20 times with TipDatingBeast to evaluate the mean rate under the uncorrelated lognormal relaxed molecular clock (ucld.mean parameter).

1. **Reconstruction demographic history**

To investigate the demographic dynamic of H5N6 in China Mainland, a Bayesian skygrid of influenza A H5N6 viruses over 5 years was reconstructed. Fig. 2d shows the demographic patterns of H5N6 in mainland China. The effective population size from the end of 2013 to the first half of 2014 was in a sharp rise and reached a peak at the end of 2014. There was a relatively slight decline in early 2015. The followed years, there is only a slight fluctuation.

1. **Discrete Phylogeographic inference**

To estimate the spatial diffusion processes underlying H5N6 spread across China, a discrete Bayesian phylogeographic method was implemented. The MCC (Maximum clade credibility) phylogenetic tree annotated with 4 geographic regions is plotted in Fig. 3a and the MCC tree annotated with provincial administrative regions is plotted in Fig. 2S. The MCC trees indicate that the most recent common ancestor of the epidemic existed between November 2009 and December 2011 (mean, 23 Dec 2010; 95% credible interval (CI), 8 Nov 2009 - 29 Dec 2011) and phylogeographic estimation assigns this ancestor to the PRD region (Guangdong, Fujian) and Guangdong province, with high credibility (Fig. 3a). For PRD region and Guangdong province, the mean duration of persistence are longer than other locations and (Fig. 3a, Fig. 2S) and have the highest posterior probability. Null hypothesis of no association between phylogenetic tree and locations (4 geographic regions (CT, PRD, SW, YRD) and provincial administrative regions) was tested using Bayesian Tip-association Significance Testing (BaTS). The analysis shows a very strong geographic clustering of strains by 4 geographic regions (P = 0 for both association index (AI) and parsimony score (PS) statistics) (Table 3). The monophyletic clade (MC) statistic was used to test the extent of phylogenetic clustering of individual regions, the result showed that population subdivision was significant for most of the localities (P < 0.05). In contrast, strains by province have no such strong of geographical association (Table S2). We obtain the Bayes factor support for each pairwise rate of diffusion between 4 geographic regions. The phylogeographic analysis indicate three significant migration links with mean rates from 1.18 to 1.92 (decisive support with BF > 1000), with all originating from PRD to SW, CT, YRD (Fig. 3b, Table S3).Transmission routes SW to PRD also have strongly supported diffusion rates (BF >10). The phylogeographic link between SW and CT is considered to be statistically significant. In summary, the inferred spatial dynamics suggested PRD played a very important role in the H5N6 virus epidemic in China Mainland. The count of observed state changes (the number of geo graphical state transition / year) further demonstrates that migration from PRD is significantly greater than in other regions. Although the geographical association of the strains by province is not so strong, we conducted the BSSVS analysis and the results were plot in Fig. S3. The results show that there are a large number of migrations from Guangdong, provinces belonging to the PRD region. Supporting our previous analysis. A detailed dynamic process of the phylogeographic dispersal can be assessed in supplementary video (Supplementary Video 1,2). The proportion of the trunk belonging each geographic region was caculate by a structured coalescent approach. A higher trunk proportion indicates that the corresponding region take up more virus variation and evlution and the greater probability of this geographic region is the origin of H5N6 virus in China Mainland. The mean and 95% credible interval of the trunk proportion of each region were summarized in Table 3. Consistent with previous analysis, PRD occupies more than half of the trunk of the genealogy (52%). SW and CT account for 29% and 13% respectively. YRD only occupies 6%. The results show that PRD is the most likely origin, and SW and CT also have a significant contribution to transmission. A further analysis of reward times for the proportional trunk location during 2011-2017, revealing that the contribution of each geographic region to the variation and evolution of the virus changes over time. The PRD was identified as the primary trunk location before 2015, while SW began to act as dominant source for novel strains after 2015. Since mid-2016, H5N6 viruses on the trunk have settled all in SW.

1. **Continuous Phylogeographic inference**

A more detailed spatial dispersal process over time was reconstructed by a continuous Bayesian phylogeographic method. The marginal likelihood comparison using PS and SS estimation approaches rejects a Brownian random walk model in favor of Cauchy RRW model. The MCC tree estimated under the Cauchy RRW model of continuous space phylogeographic diffusion was mapped onto the study area by SpreaD3 (Fig. 5). Figure 5 presents that H5N6 spread remained largely restricted to Hunan (CT region), Hubei (CT) up 2012, but migration events to Guangdong (PRD), Sichuan (SW) and Jiangsu (YRD) are starting to emerge from this time. These migrations became more pronounced by 2013, include Jiangxi (CT), Jiangsu (YRD) and start a new migration events to Chongqing (SW) and Xinjiang (NW). Seeding of Fujian (PRD) and Guangxi (SW) is about 2014 and Yunnan (SW) are starting to emerge. Transmission of the virus has become more frequent in these areas by 2015 to 2017 and include the seeding of Anhui. The dynamics are mainly in the CT region and have since spread to the periphery. The spatial expansion in continuous space is characterized by an average diffusion rate of about 672.5 km/year.

1. **GLM analysis**

A Bayesian phylogeographic generalized linear model (GLM) was used to determine the factors that influenced the spread of H5N6 among provincial administrative regions. Of the 17 factors assessed predictors were included in the GLM model (see Table S1 for a full list and description), 5 were included in the model with Bayes factor > 3 (Table S2). In summary, H5N6 tends to disperse in rural areas and the place with a large population (The origin percentage of the rural population, Bayes factor (BF) support for inclusion: BF = 19.13; The origin resident population, BF = 9.24). The humidity (BF = 7.60) of the origin is also positively coorrelated with viral dissemination. The positive effect of population combined with the positive effect of the humidity implies that the epidemic originates from the place with many people and a humid climate. Although a humid climate has a positive impact on the spread of the virus, the surface water resources (BF = 3) have a inverse effect, suggesting that surface water resources curbed the geographic spread of H5N6. Sales of poultry per capita rural household (BF = 5.43) of origin suggested that the poultry trade in the rural areas of origin is also one of the factors contributing to the spread. We found that each predictor has significant impacts on the origin, indicating that the factors in the origin region play a key role in the h5n6 mode of transmission. There are many other possible predictors for H5N6 transmission, such as nature reserves, pollution and great circle distance, were not significantly associated with virus spread.

**Discussion**

Avian influenza A (H5N6) viruses evolve fast in China Mainland and are likely to become candidates to cause a new influenza pandemic in humans. Results of selection stress analysis show that positive selection mutations in the HA gene occur mainly in the stalk of the HA. Interestingly, mutations in HA stalks are associated with increased stability of HA protein and play a critical role in airborne transmission of avian viruses in mammals (Identification, characterization, and natural selection of mutations driving airborne transmission of A/H5N1 virus). It’s possible that the virus will pre-adapted to humans during circulation in avian. Therefore, these amino acid residues deserve further study as potential targets for therapeutic intervention. Frequent migration may bring increased risk of infection, and analysis of natural selection pressures shows no signs of stopping the spread of the viruses. In summary, the CT region and Guangdong of PRD region, a reservoir for the spread of the virus. Of the nine ecological factors, Surface Water Resources (100 million cu.m), Resident Population (10, 000 persons), Rural Population (%), Sales of Poultry Per Capita Rural Household (kg) and Average Relative Humidity of Major Cities (%) are the main factors affecting the virus and depend largely on the source of origin of these factors. It can be inferred that certain conditions need to be in place for the source of infection to spread, such as the five predictors above. There may be other potential factors that are not accounted for in the analysis, which are limited by the ability to obtain and process data. From these factors, it may be possible to control the outbreak and spread of the disease with greater precision. Despite this, continuous surveillance of the virus can facilitate the early detection of novel emerging variants, and acts as a fundamental part of integrated preventive strategy. Our analysis used a relatively comprehensive approach to study the evolutionary dynamics of Avian influenza A (H5N6) viruses. Similar results from different methods increase the robustness of the estimates. The maximum likelihood phylogenetic analysis and Bayesian methods obtained extremely similar results for tMRCA (2010.66 and 210.97) and evolutionary rates (6.09 × 10^(-3) and 7.34 × 10^(-3)). The ancestral location of the root reconstructed using maximum parsimony is also consistent with that constructed by the Bayesian method. Mathematical models provide a powerful tool for epidemiological analysis to estimate an underlying infectious disease transmission process and are widely used, but subject to the assumptions on which the model is based and the data provided. Simplicity and impractical are often questioned as drawing conclusions that are far from reality. Obviously, reality is always more complicated, so epidemiology or H5N6 transmission will always involve some approximation, and parameter estimates will always involve a certain degree of confidence specified. If people can understand the nature of the problem, a certain degree of confidence specified. This study is limited to a sample size and does not cover all instances where it occurred, while it's possible that the populations not sampled are carrying key information on spread and persistence of H5N6. Sample size is closely related to strength of migration in our analysis. A migration signal can be weak because of inadequate of samples. In contrast, A small sample size may be a result of the low incidence of H5N6 and correspondingly weak migratory links with other regions. To overcome this bias, subsampling is usually performed to ensure the similar number of sequences in each region and locations with only a small number of sequences were removed completely from the analysis. For the phylogeographic inference in discrete space, we only used 125 sequences from four region (CT, PRD, SW, YRD) because NW was removed as there were only 4 sequences, which is far less than other areas. For the phylogeographic inference in continuous space, we only used 124 sequences because there were only two sequences in Anhui and Zhejiang and one in Sichuan, and they were all removed. This way both methods of analysis get the most suitable data set and make the most of the data. Comparing continuous and discrete phylogeographic inference, Guangdong Province and PRD region are the main output areas, and the results of both analyses are consistent. The sampling times of the viral sequences are used to calibrate molecular clock methods, while the distribution of these sampling times is not always enough to allow accurate estimates of substitution rates. The date-randomization test was used to prevent data without a time signal.

**Conclusion**

Our study provides novel and significant insights into the temporal and spatial dynamics of the H5N6. The estimated TMRCA suggested that H5N6 emerged in 2009. Purifying selection is a dominant force acting on HA gene to drive the evolution of the virus. Our study suggested that Central and Southern China region is the origin of H5N6, and Guangdong is the most exported province. Positive selection sites for the HA gene are mainly in the stalk. There are several environmental factors where migration originates that are highly relevant to transmission. The research into virus intercontinental transmission is necessary to understand molecular diversity and to manage public health.

**Materials and Methods**

**Sequence datasets**

All complete Haemagglutinin (HA) proteins and the corresponding DNA sequences of influenza A H5N6 viruses (covering 2013 to 2017 in China Mainland) were collected from the GISAID EpiFlu database (<http://www.gisaid.org>)GISAID: Global initiative on sharing all influenza data–from vision to reality. After aligning with MAFFT version 7.407MAFFT multiple sequence alignment software version 7: improvements in performance and usability, we obtained the corresponding codon alignments by PAL2NAL version 14PAL2NAL: robust conversion of protein sequence alignments into the corresponding codon alignments. Duplicate sequences and sequences overly divergent based on root-to-tip distances were removed. A recombination may impact evolutionary estimatesConsequences of recombination on traditional phylogenetic analysis. We employed RDP, Chimaera, BootStan, GENECONV, MaxChi and SiScan methods available in RDP version 4RDP4: Detection and analysis of recombination patterns in virus genomes to avoid inferential biases. The recombination sequences were removed. All sequences were annotated with available collection dates and location and grouped into 5 geographic regions: Central (CT), North-west (NW), Pan-Pearl River Delta (PRD), South-west (SW), Yangtze-River Delta (YRD). And accession numbers are given in supplementary file.

**Maximum likelihood estimation**

To analyze the temporal molecular evolutionary signal of the sequence dataset, a maximum likelihood (ML) tree generated using a multiple sequence alignment. We ran IQ-TREE multicore version 1.6.1IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies with the (K3Pu+F+G4), which was identified as the best fitting model for ML inference by ModelFinder and chosen by Bayesian Information Criterion (BIC). Ultrafast bootstrap with 1000 replicates were performed for branch support inference. The TempEst version 1.5.1Exploring the temporal structure of heterochronous sequences using TempEst (formerly Path-O-Gen) was implemented to estimate root-to-tip genetic distances after selecting the best-fitting root with a residual mean squared function. Dates of sampling were regressed against root-to-tip genetic distances and regression were performed in R with the lm functionR: A language and environment for statistical computing of distances on dates. The ML trees were then dated using lsd-0.3betaFast dating using least-squares criteria and algorithms and reconstructed the ancestral location of each internal node in the ML trees using the asr\_max\_parsimony function in castorEfficient comparative phylogenetics on large trees.

**Natural selection**

Site-specific positive selection was identified by four different methods in hyphy: FELNot So Different After All: A Comparison of Methods for Detecting Amino Acid Sites Under Selection (Fixed Effects Likelihood), FUBARFUBAR: A Fast, Unconstrained Bayesian AppRoximation for Inferring Selection (Fast, Unconstrained Bayesian AppRoximation), MEMEDetecting Individual Sites Subject to Episodic Diversifying Selection (Mixed Effects Model of Evolution) and SLACNot So Different After All: A Comparison of Methods for Detecting Amino Acid Sites Under Selection (Single-Likelihood Ancestor Counting) implemented in HYPHY version 2.3 for linuxHyPhy: hypothesis testing using phylogenies and also available through the web-based interface Datamonkey (http://www.datamonkey.org/). To select sites under positive selective pressure with a statistically significant result, the critical values was provided as a P value of <0.1 or a posterior probability of >0.9. We also performed a ‘Renaissance Counting’A counting renaissance: combining stochastic mapping and empirical Bayes to quickly detect amino acid sites under positive selection procedure to estimate the ratio of nonsynonymous to synonymous substitutions per site (dN/dS) in BEAST version 1.10.4Bayesian phylogenetic and phylodynamic data integration using BEAST 1.10. In order to visualize the locations of the positively selected sites identified by at least four methods, each positively selected site was mapped onto a three-dimensional structure of HA protein (Protein Data Bank code: 5HU8) and visualization was generated with pymol open source program (https://github.com/schrodinger/pymol-open-source).

**Molecular clock phylogenetics and Demographic inference**

Time-resolved phylogenetic trees were estimated using BEAST version 1.10.4, which employs a Bayesian Markov Chain Monte Carlo (MCMC) algorithms to generate population genetic phylogenies from molecular sequences. We used the SRD06 model on CDS, which uses HKY substitution model and partitioned codons into two partitions (positions (1+ 2) and 3)Choosing appropriate substitution models for the phylogenetic analysis of protein-coding sequences. To select the suitable model, we performed marginal likelihood estimation (MLE) using path sampling (PS) / stepping-stone sampling (SS) in BEAST version 1.10.4Improving the accuracy of demographic and molecular clock model comparison while accommodating phylogenetic uncertaintyAccurate model selection of relaxed molecular clocks in bayesian phylogenetics. We set 100 path steps for a chain length of 1 million, with power posteriors are determined according to evenly spaced quantiles of a Beta (0.3, 1.0) distribution. Four coalescent tree priors: a constant-size population, an exponential growth population, a Bayesian Skyline tree prior (ten groups, piecewise-constant model)Bayesian coalescent inference of past population dynamics from molecular sequences and a Bayesian Sky Grid (a grid of 50 intervals over 7 years)Improving Bayesian population dynamics inference: a coalescent-based model for multiple loci were tested. For each tree prior, two clock models: a strict clock and an uncorrelated relaxed clock with log-normal distribution (UCLN)Relaxed phylogenetics and dating with confidence was tested. BEAST was run for 200 million Markov Chain Monte Carlo (MCMC) steps, sampling parameters and trees were sampled every 10,000 generations after allowing a burn-in of 20 million generations. The BEAGLE library was used to accelerate BEAST’s computationBEAGLE: An Application Programming Interface and High-Performance Computing Library for Statistical Phylogenetics. Convergence of MCMC chains was checked with Tracer version 1.7.1 (available from http://beast.bio.ed.ac.uk/Tracer) by inspecting the Effective Sample Sizes (ESS > 200). The Sky Grid analysis run on Trace version 1.7.1 to reconstruct a Bayesian skyline plot. Clock rate and tMRCA estimates, and their distributions extracted by R and plotted with ggplot2ggplot2.

**Phylogeographic inference in discrete space**

For reconstructing the temporal and spatial dynamics of H5N6 in China Mainland, a Bayesian discrete phylogeographic approach was employed to estimate the ancestral locations of H5N6 and infer the migration history. Phylogeographic reconstruction was estimated using a discrete state continuous time Markov chain (CTMC) modelBayesian phylogeography finds its roots, in which transition rates were estimated among discrete location states. The asymmetric discrete trait substitution model with the Bayesian stochastic search variable selection (BSSVS) implemented in BEAST v1.10.4. To observe the transitions (Markov jump) between location states along phylogenetic branches and the time (Markov rewards) spent in the location states between two transitions included in the realization of model parameter estimates yielded by fitting a CTMC model to location data results, posterior inference of the complete Markov jump history through time was usedUnifying viral genetics and human transportation data to predict the global transmission dynamics of human influenza H3N2Fast, accurate and simulation-free stochastic mapping. The BSSVS enables us to obtain a Bayes factor (BF) test that provide an identification of the most parsimonious description of the phylogeographic diffusion process. BF tests for significance of individual rates be used to infer the significant epidemiological links. If 3 ≤ BF < 10, the phylogeographic link between two locations is considered to be statistically significant, 10 ≤ BF < 100 is regarded as a very strong support, and 100 ≤BF < 1000 indicates decisive of the degree of rates. All demographic parameter settings were the same as the best fit model combinations in the model select analysis described earlier. The analyses were applied to a geographic assignment of sequences to ten provincial administrative regions, and a geographic assignment of sequences to four economic regions. Then, we identified the maximum clade credibility (MCC) tree using TreeAnnotator version 1.10.4 after 10% burn-in and visualized the tree in ggtreeggtree: an R package for visualization and annotation of phylogenetic trees with their covariates and other associated data. SpreaD3 version 0.9SpreaD3: interactive visualization of spatiotemporal history and trait evolutionary processes was used to visualize the dispersal process through time and obtain the BF for the diffusion between discrete locations. We analyzed the posterior trees using the program PACT v0.9.5 (https://github.com/trvrb/PACT) to perform trunk extraction and processing, which is able to estimate the proportion of the trunk for each geographic region, the trunk location through time, the persistence and the migration rate. To estimate phylogeny-trait associations for different location states, we use the AI, PS and MC statistics, and provide 95% confidence intervals and significance estimation for theseCorrelating viral phenotypes with phylogeny: accounting for phylogenetic uncertainty.

**Phylogeographic inference in continuous space**

We also reconstructed the temporal and spatial dynamics of the H5N6 in continuous space as a complementary analysis. All demographic parameter settings were the same as in the BEAST analysis described earlier. To select the suitable continuous traits substitution model, we performed marginal likelihood estimation (MLE), which was the same as analysis described earlier. In continuous traits substitution model, the Cauchy RRW model and the Brownian random walk modelPhylogeography takes a relaxed random walk in continuous space and time were tested. Bivariate trait represents latitude and longitude. We used the TreeAnnotator version 1.10.4 to summarize the information contained within sampled trees and used the SpreaD3 to visualize the diffusion in continuous space.

**Identifying predictors for the diffusion process**

To identify the contribution of potential predictors for the CTMC transition rate matrix between locations, we use the methods that extends the Bayesian method of phylogeographic inference into a generalized linear model (GLM), by parameterizing the CTMC matrix as a generalized linear model (GLM), in which log CTMC rates are a log linear function of various potential predictors. We select nine predictors at last (Table S1). These potential predictors of previous related studies were considered. We used surface water resources, resident population, total number of sequences used in the analysis, rural population, sales of poultry per capita rural household, percentage of nature reserves in the region, sum of smoke and dust, sulphur dioxide, nitrogen oxides, average relative humidity of major cities and the average distance between centroid point of each locations, calculated using latitude and longitude coordinates. The average distance between regions was calculated by the distm function in geosphere (https://CRAN.R-project.org/package=geosphere). The raw data set was download from National Bureau of Statistics of China (www.stats.gov.cn). The potential predictors were grouped to origin and destination predictors, except for distance. All values were log-transformed and standardized. We run the GLM analysis in BEAST version 1.10.4 with the BEAGLE library. We calculated the inclusion probability for each predictor, BF support value, and coefficient. The result was visualized by ggplot2.

**Main Figure and Table Legends**

**Figure 1.** ML phylogenetic tree reconstructed from H5N6 sequences. Branches and tips are colored according to their ancestral location reconstructed in 13 provincial administrative regions. Pie chart in each node represents the probability of each node being in each state. The shape of tips represents different regions grouped by economic division. Central (CT): Anhui, Hubei, Hunan, Jiangxi; North-west (NW): Xinjiang; Pan-Pearl River Delta (PRD): Guangdong, Fujian; South-west (SW): Chongqing, Guangxi, Sichuan, Yunnan; Yangtze-River Delta (YRD): Zhejiang, Jiangsu. A thick branch line represents bootstrap value that is greater than or equal to 80. An identical ML tree reconstructs the ancestral location of each internal node in 5 economic regions shown in Figure S1.

**Figure 2.** Evolution of influenza A H5N6 viruses and population dynamics. (a) Linear regression analysis of the root to tip distance against sampling dates for the HA gene sequences. (b) Probability distribution for the tMRCA of each tested model. (c) Mapping of positively selected sites identified by at least four methods. (d) Bayesian skygrid of influenza A H5N6 viruses over 5 years shows the changes in effective population size of H5N6 over time. The thick solid line indicates the median estimates, and the blue area displays the 95% HPD.

**Figure 3.** Maximum clade credibility tree and Spatial diffusion of H5N6 viruses. (a, b) MCC tree created with HA gene segments of avian influenza A H5N6. The internal branches are colored according to the most probable ancestor location of their descendent nodes. Posterior probabilities

>0.5 are shown. 95% posterior densities of most recent common ancestor estimates for all lineages is shown at the bottom. To the upper left of phylogenies are their root location state posterior probability distributions. The inset to the top left of tree shows duration of region-specific persistence measured as the waiting time in years to leave its geographic region of origin. Circles represent mean persistence across sampled viruses. (c) Discrete spatial diffusion pathways, significant epidemiological unidirectional pathways from one location to another are indicated on the maps. Yellow arrows, decisive rates with BF > 10,000; deep blue arrows, very strongly supported rates with 10,000 > BF > 1,000; blue arrows, strongly supported rates with 1,000 > BF > 100; purple arrows, supported rates with 100 > BF > 10; and grey arrows, supported rates with 10 > BF >3. (d) Histograms of total number of state transitions.

**Figure 4.** Proportional ancestral region states on the phylogenetic tree trunk estimated for each region over time and migration rates were measured as migration events per lineage per year. (a) The waiting time between geographic location transitions was inferred using the continuous-time Markov chain model. The trunk reward proportion for each geographic location from 2004 through 2017 was determined from an analysis of HA sequences of H5N6 viruses. Colored areas represent the trunk proportions over time for the four geographical locations included in the analysis. (b) Distance to the trunk measured in terms of years. Scatter plot represents the distance to trunk of the tree. The height of each point on the y-axis shows the mean distance to the trunk. Dots are colored according to four sampled regions. (c, d) Demonstrated the rate at which labels change over the course of the genealogy. Each pair of labels is considered. Rates are measured as migration events per lineage per year. With the observed rate of transition between two locations (migration events per lineage per year), the dot size became bigger.

**Figure 5.** Spatiotemporal dispersal of avian influenza A H5N6 in China Mainland reconstructed using continuous phylogeographic analysis of HA. Dispersal patterns are shown up to four different time intervals: 2011-2012, 2013, 2014, and 2015-2017. The black lines project the part of the MCC tree up to each of those times, whereas the contours represent statistical uncertainty of the estimated locations at the internal nodes (95% credible contours based on kernel density estimates).

**Figure 6.** Predictors of avian influenza A H5N6 diffusion among ten cites. Seventeen predictors were considered: surface water resources, resident population, total number of sequences used in the analysis, rural population, sales of poultry per capita rural household, percentage of nature reserves in the region, sum of smoke and dust, sulphur dioxide, nitrogen oxides, average relative humidity of major cities and the average distance. Blue refers to the predictors of origin, and green refers to the predictors of destination. Support for each predictor is represented by an inclusion probability which is defined by the indicator expectations. The mean and credible intervals of the GLM coefficient (β|δ=1) on a log scale for each predictor were conditioned on the inclusion of that term in the model (*E*[δ]).

**Table 1.** Positively selected sites identified by FEL, MEME, SLAC, FUBAR, BEAST.

**Table 2.** Log-marginal likelihood estimates using the path-sampling (PS) and stepping-stone (SS) model selection approaches. The best-fitting combination is underscored. Two molecular clock models were tested here. SC, Strict clock model; UCLN, uncorrelated relaxed clock with log-normal distribution.

**Table 3.** Mean and 95% credible intervals over sampled genealogies for the location of the genealogy trunk between the years of 2014 and 2017.

**Table 4.** Phylogeny-trait association tests of the phylogeographic structure of avian influenza A H5N6 using BaTS.