Virologica Sinica 39 (2024) 525–536

Contents lists available at ScienceDirect

Virologica Sinica

journal homepage: www.keaipublishing.com/en/journals/virologica-sinica
www.virosin.org

Research Article

Transmission restriction and genomic evolution co-shape the genetic diversity patterns of influenza A virus

Xiao Ding [a,b,1], Jingze Liu [a,b,1], Taijiao Jiang [a,c,d,*], Aiping Wu [a,b,*]

[a] State Key Laboratory of Common Mechanism Research for Major Diseases, Suzhou Institute of Systems Medicine, Chinese Academy of Medical Sciences & Peking Union Medical College, Suzhou, 215123, China
[b] Key Laboratory of Pathogen Infection Prevention and Control (Peking Union Medical College), Ministry of Education, Beijing, 100730, China
[c] Guangzhou National Laboratory, Guangzhou, 510006, China
[d] State Key Laboratory of Respiratory Disease, The Key Laboratory of Advanced Interdisciplinary Studies Center, The First Affiliated Hospital of Guangzhou Medical University, Guangzhou, 510030, China

ARTICLE INFO

Keywords:
Influenza A virus
Genetic diversity
Transmission
Genetic pattern
Genotype

ABSTRACT

Influenza A virus (IAV) shows an extensive host range and rapid genomic variations, leading to continuous emergence of novel viruses with significant antigenic variations and the potential for cross-species transmission. This causes global pandemics and seasonal flu outbreaks, posing sustained threats worldwide. Thus, studying all IAVs' evolutionary patterns and underlying mechanisms is crucial for effective prevention and control. We developed FluTyping to identify IAV genotypes, to explore overall genetic diversity patterns and their restriction factors. FluTyping groups isolates based on genetic distance and phylogenetic relationships using whole genomes, enabling identification of each isolate's genotype. Three distinct genetic diversity patterns were observed: one genotype domination pattern comprising only H1N1 and H3N2 seasonal influenza subtypes, multi-genotypes co-circulation pattern including majority avian influenza subtypes and swine influenza H1N2, and hybrid-circulation pattern involving H7N9 and three H5 subtypes of influenza viruses. Furthermore, the IAVs in multi-genotypes co-circulation pattern showed region-specific dominant genotypes, implying the restriction of virus transmission is a key factor contributing to distinct genetic diversity patterns, and the genomic evolution underlying different patterns was more influenced by host-specific factors. In summary, a comprehensive picture of the evolutionary patterns of overall IAVs is provided by the FluTyping's identified genotypes, offering important theoretical foundations for future prevention and control of these viruses.

1. Introduction

Influenza A virus (IAV) is a negative-sense single-stranded RNA virus with eight genomic segments, resulting in over 130 reported IAV subtypes (Pineo, 2021). Rapid genomic mutation and frequent reassortment constantly give rise to new viruses (Barrat-Charlaix et al., 2021; Du Toit, 2023; Müller et al., 2020; Yang et al., 2022), enabling them to infect various hosts (Ciminski et al., 2021; Ganti et al., 2022) and cause significant global impacts (Ali and Cowling, 2021; Krammer et al., 2018). For instance, the H1N1 pandemic in 2009 resulted from a triple reassortant virus, causing 1.4 billion infections and 151,700 to 575,400 deaths (Waters et al., 2021). Seasonal flu epidemics cause 3 to 5 million severe cases and 290,000 to 650,000 deaths worldwide annually, according to a report from the World Health Organization (WHO Working Group, 2023).

Understanding IAVs' genetic diversity is crucial for evolutionary analysis, risk assessment, source tracing, and vaccine recommendation (Borkenhagen et al., 2021; Carter et al., 2016; Gao et al., 2013; Naguib et al., 2019; Patrono et al., 2022; Ping et al., 2018; Shi et al., 2018; Zaraket et al., 2015). Previous studies have revealed significant scientific findings. For example, it has been determined that the H7N9 virus infecting humans in 2013 originated from genomic reassortments between the original H7N9 and H9N2 viruses (Gao et al., 2013). Research on the genotypes of H7N9 viruses in chickens and ferrets highlighted a narrow bottleneck limiting avian-to-mammalian transmission (Zaraket et al., 2015). Furthermore, broadly reactive hemagglutinin vaccines for

* Corresponding authors.
E-mail addresses: wap@ism.cams.cn (A. Wu), taijiao@ibms.pumc.edu.cn (T. Jiang).
[1] Xiao Ding and Jingze Liu contributed equally to this work.

https://doi.org/10.1016/j.virs.2024.02.005
Received 4 December 2023; Accepted 22 February 2024
Available online 27 February 2024

H5N1 and H1N1 were designed using the computationally optimized broadly reactive antigen strategy also based on genetic diversity analysis (Carter et al., 2016; Ping et al., 2018). However, these studies mainly focused on specific IAV subtypes, neglecting systematic analysis of overall IAV genetic diversity dynamics.

The intricate evolution of IAVs is influenced by factors like diverse subtypes, hosts, and rapid genomic variations. Avian influenza virus (AIV) can jump to new species and occasionally acquire human-to-human transmission (Long et al., 2019). Different subtypes can reassort and generate novel viruses within the same host (Ganti et al., 2022). Moreover, IAVs of the same subtype undergo fast antigenic mutation, allowing immune evasion in hosts (Medina and García-Sastre, 2011). Current evolutionary studies tend to focus on specific viruses, resulting in a limited and potentially biased understanding of the overall picture. Thus, systematic research is necessary to explore the genetic patterns of overall IAVs.

Previous studies on IAV genetic diversity often categorized viruses into genotypes based on genetic differences from large-scale data (Yamayoshi and Kawaoka, 2019). Genetic clustering methods include phylogenetic-based and genetic distance-based approaches (Dong et al., 2020; Poon, 2016). Phylogenetic methods analyze phylogenetic relationships in trees or networks, considering additional data such as epidemiological data (Han et al., 2019a, 2019b; Ji et al., 2022; Prosperi et al., 2011; Ragonnet-Cronin et al., 2013; Tan et al., 2019; Wu et al., 2013). However, the validity and complexity involved in constructing phylogenetic relationships pose limitations on genomic classifications. Genetic distance methods quickly partition sequences without trees or networks, requiring fewer calculations (Han et al., 2019a, 2019b; Prosperi et al., 2011; Ragonnet-Cronin et al., 2013; Tan et al., 2019). Yet, they lack phylogenetic information, affecting identifications for complex events like reassortments. A valid genomic clustering strategy for studying virus evolution based on genotypes remains missing.

To address these challenges, we developed FluTyping, which classifies each isolate of all IAVs based on distinct genotypes identified with different genomic segment clusters (Fig. 1A). Both genetic distances and phylogenetic relationships (Fig. 1B) define these classifications together, providing a macro perspective of IAVs' genetic diversity. Our study reveals three distinct diversity patterns, each with different influenza subtypes. We explore factors influencing these patterns, considering epidemiological and genomic evolution aspects. In summary, this study unveils valid genotypes, enabling quick tracing of emerging viruses and early warning for potential influenza outbreaks.

2. Methods and materials

2.1. Data collection and processing

The genomic sequences of 340,862 influenza A strains, collected before January 1, 2023, were obtained from the Global Initiative on Sharing All Influenza Data (GISAID) database (Elbe and Buckland-Merrett, 2017). To ensure data quality, we performed four steps of sequence preprocessing.

Firstly, we removed genomic sequences with a length that deviated from the standard length of the corresponding gene by more than 10%. Additionally, sequences containing ambiguous bases (i.e., bases that were not "a", "t", "c", or "g") greater than 1% of the sequence length were also excluded. Secondly, in cases where there were duplicate strain IDs, we retained only the sequences with the least number of ambiguous bases and the minimal difference in length from the standard length of the corresponding gene. Thirdly, we kept only the sequences from strains with explicit subtype, collection year, country, and host information. Lastly, we focused on the whole genomic sequences containing *PB2*, *PB1*, *PA*, *HA*, *NP*, *NA*, *M*, and *NS* genes for our analysis.

After these preprocessing steps, a total of 133,249 whole genomes of influenza A virus were selected for further analysis. This dataset constitutes the basis for our study on the genetic diversity patterns of influenza A viruses using our FluTyping method.

2.2. Framework for FluTyping

The FluTyping pipeline consists of three main steps: clustering, phylogenetic calibration, and genotyping.

In the clustering step, the genomic sequences of all IAVs are semi-automatically grouped into distinct phylogenetic classes based on each genomic segment. For all influenza virus genomic sequences, representative sequences for each genomic segment were acquired based on their collection years, countries, and pairwise sequence similarity primarily. Subsequently, specific clusters of representative sequences were organized using average intra-MCU sequence similarity, entropy change post combining MCUs, and the overlap of unit-specific genomic loci in the MCU-based combination. The resulting clusters underwent further grouping through hierarchical clustering, employing the estimated optimal cluster number. These steps involve analyzing the genetic distance and phylogenetic relationship between isolates to identify clusters of related strains. Next, in the phylogenetic calibration step, the obtained clusters are further optimized to improve the accuracy of the phylogenetic relationships. Mixed clusters, which contain strains from different phylogenetic clades, and outliers, which do not fit well into any cluster, are manually combined to better reflect the true evolutionary relationships between isolates. The final classification was determined by sorting the quantity of isolates included in each cluster. For instance, the category containing the most isolates is designated as 1. Finally, in the genotyping step, the genotype of each isolate was ascertained by consolidating the relevant clusters in the sequence of *PB2*, *PB1*, *PA*, *HA*, *NP*, *NA*, *M*, and *NS* genes. For example, the genotype assigned to EPI_ISL_69853, represented as 5|2|6|H7.1|3|N7.2|3|1 (Supplementary Table S1), signifies that the identified clusters of PB2, PB1, PA, HA, NP, NA, M, and NS genomic segments for EPI_ISL_69853 are 5, 2, 6, H7.1, 3, N7.2, 3, and 1, respectively. This genotype represents the overall genetic characteristics of the isolate and helps in understanding the evolutionary patterns of influenza A viruses.

The clustering and genotyping steps are implemented using self-written Perl scripts, and the details of the methodology have been published in https://github.com/dingxiao8715/FluTyping. These scripts are designed to handle large-scale genomic data efficiently and accurately identify the genotypes of IAVs based on the combined information from the clustering and phylogenetic calibration steps.

2.3. Clustering pipeline in FluTyping

The clustering step in FluTyping involves three procedures: the epidemiological combination, the MCU-based combination, and the distance-based clustering.

In the epidemiological combination, the genomic sequences are grouped into multiple clusters based on their collection years and countries. To reduce bias and computational complexity, representative sequences are obtained for each cluster using CD-HIT (Li and Godzik, 2006) with specific sequence similarity cutoffs for surface genes (0.5%) and internal genes (0.1%). Phylogenetic trees are then constructed for each of the eight genes using FastTree (Price et al., 2009) with the representative sequences.

In the MCU-based combination step, a bottom-up strategy is employed, defining inner nodes in the phylogenetic trees as “units” if they only contain leaf nodes. A node is considered a minimum clade unit (hereinafter referred to as MCU) if it consists of two units or one unit and an outlier belonging to that unit. The MCUs are identified in the phylogenetic tree, and an MCU can be combined with another MCU if specific criteria are met (Supplementary Fig. S1), such as high average sequence similarity (>0.99), minimal change in entropy (<0.01), and no shared specific genomic loci between different units. The parameters and their cutoff values for these criteria are determined through “Quantifying Genetic Heterogeneity Among MCUs” and “Optimizing Cutoff Value Selection” sections. The MCU-based combination continues until no further combination is possible. The converged MCUs are then clustered

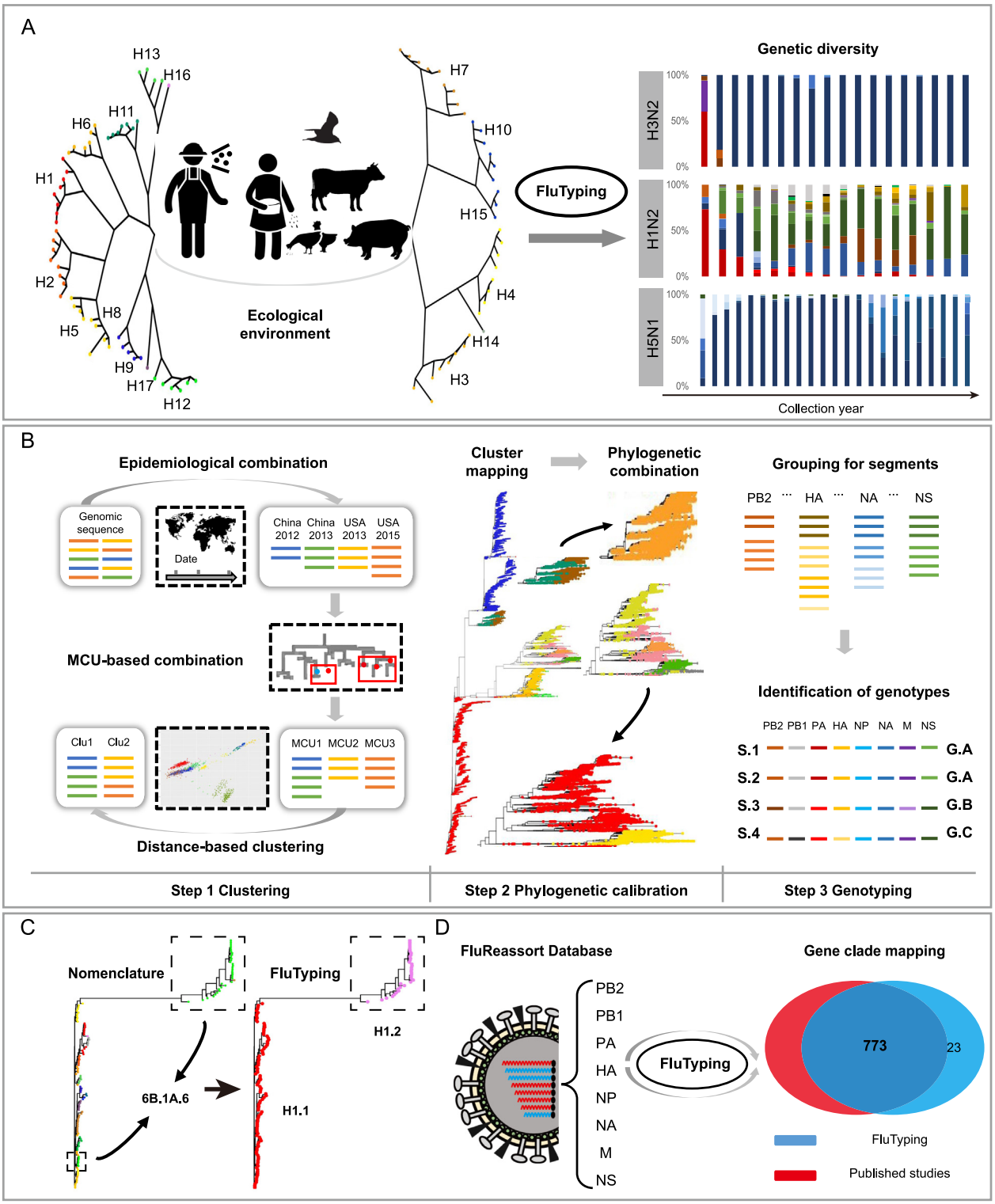

Fig. 1. Overall methodological framework and performance evaluation of FluTyping. **A** The research motivation of this study. We developed FluTyping and established a comprehensive genotype targeting all subtypes of influenza A virus. Using this genotype, we explored the genetic patterns of influenza A virus at a macroscopic level, such as the temporal distribution of different genotypes for specific subtype shown in **A**. **B** Framework for FluTyping. Input consists of sequences from all genomic segments of influenza A viruses, which undergo clustering, phylogenetic calibration, and genotyping steps. Finally, a specific genotype will be assigned to each isolate. **C** Comparison of classification between of nomenclature and in FluTyping of the H1 subtype in phylogenetic trees. **D** Validating FluTyping by identifying reassortment events using the FluReassort database.

based on their genetic distance, and the optimal number of clusters is estimated using the Bayesian Information Criterion (BIC) via the R package mclust (Scrucca et al., 2016). Finally, hierarchical clustering is applied to obtain the distance-based clusters with the estimated number.

After the hierarchical clustering, mixed clusters (containing strains from different clusters) and outliers (clades with only one strain) may occur. To optimize the clusters based on phylogenetic relationships, manual combination of these abnormal types of clusters is performed using information from the phylogenetic tree. The counts of remaining categories for various genomic segments after each process in FluTyping were presented in the supplementary material (Supplementary Table S2).

2.4. Quantifying genetic heterogeneity among MCUs

In the MCU-based combination process of FluTyping, each MCU represents a phylogenetic clade containing a set of genetic sequences. The genetic heterogeneity within each MCU is evaluated using three parameters: the average intra-MCU sequence similarity, the entropy change after combining MCUs, and the overlap of unit-specific genomic loci.

The average intra-MCU sequence similarity is calculated as the mean genetic distance between each pair of sequences within the MCU. This measure gives an indication of how closely related the sequences are to each other within the clade.

For the entropy change, the entropy of each unit within the MCU is calculated using equation 1:

$$E_{unit} = \frac{1}{l}\sum_{i=1}^{l} E_{site} = -\frac{1}{l}\sum_{i=1}^{l}\sum_{b\in\{a,t,c,g\}} p_{i_s}\log_2 p_{i_s}$$

the length of the alignment, denoted as "l", represents the genomic sites utilized for calculating entropy. The proportions of specific nucleotides (a, t, c, and g) at site "i" are represented by p_{i_s}, where "s" corresponds to each nucleotide type.

The change in entropy, denoted as ΔE, is then computed as the sum of the differences between the entropy of the combined MCU and the entropy of each unit within the MCU, according to equation 2. This value represents how much the genetic diversity changes when the MCUs are combined.

$$\Delta E = \sum_{i=1}^{n} E_{MCU} - E_{Unit(i)}$$

Finally, the overlap of unit-specific genomic loci is assessed by determining the identical genomic loci shared among different units within the MCUs. This parameter helps assess whether the MCUs contain sequences that are distinct from each other or if there is significant overlap in their genetic content.

By evaluating these three parameters, FluTyping determines whether to combine MCUs to optimize the clustering process and improve the accuracy of genotype assignment for each strain based on its genetic characteristics.

2.5. Optimizing cutoff value selection

To determine the cutoff values for the three parameters (average intra-MCU sequence similarity, entropy change, and overlap of unit-specific genomic loci), a comprehensive assessment was conducted using the nomenclature of the H1, H3, and H5 subtypes of influenza viruses.

Representative strains of the H1 and H3 subtypes were obtained from the nextstrain webserver until November 1, 2022 (https://nextstrain.org/) (Hadfield et al., 2018), and their genomic sequences were downloaded from the GISAID database (Elbe and Buckland-Merrett, 2017). Similarly, the H5 genomic sequences were obtained from the GenBank database based on previous studies (Benson et al., 2013; Smith and Donis, 2015; WHO/OIE/FAO H5N1 Evolution Working Group, 2012, 2009; World Health Organization/World Organisation for Animal Health/Food and Agriculture Organization (WHO/OIE/FAO) H5N1 Evolution Working Group, 2014).

To ensure the universal applicability of cutoff values and to evaluate the computational complexity of our assessments, we selected the clade with the highest number of sequences at each level of the nomenclature hierarchy (Supplementary Table S3). We conducted two hundred calculations for each parameter, utilizing randomly selected sets of sequences with an increasing sequence number ranging from 5 to 50. Both intra-clade and inter-clade parameters were calculated two hundred times. Our objective was to identify an optimal threshold for different parameters that could differentiate between calculated parameters in datasets of intra-clade and inter-clades with varying sequence numbers.

As an illustration, consider the 6B clade of the H1 subtype. Intra-clade parameter assessments were performed using two randomly selected nonredundant sequence sets within the 6B clade. Conversely, inter-clade assessments involved two randomly selected sequence sets: one from the 6B clade and the other from different selected clades, each with varying sequence amounts. This comprehensive approach enabled us to thoroughly evaluate performance across diverse scenarios and clades.

Based on the assessments, it was observed that two sets of sequences from the same clade had no specific genomic loci, while all clades had specific loci, primarily with different sequence amounts (Supplementary Figs. S2–S4). As a result, the cutoff value for the overlap of unit-specific genomic loci was set to 0, indicating that for two sets of sequences to be considered as part of the same clade if they share more than one specific genomic locus.

Furthermore, to determine the optimal cutoff combination for the sequence similarity and the entropy change, the accuracy of all inter-clades and intra-clade assessments was calculated for various cutoff values. The sequence similarity was varied from 0.96 to 0.99 with an interval of 0.01, while the change of entropy was varied from 0.05 to 0 with the same interval. Among the combinations tested, the cutoff combination of a sequence similarity of 0.99 and an entropy change of 0.01 yielded the highest accuracy of 99.68% (Supplementary Table S4).

Additionally, the distribution of assessment results for different clades with varying numbers of sequences was also shown in Supplementary Figs. S2–S10. Based on the analysis, the final cutoff values for the average intra-MCU sequence similarity and the entropy change after MCU combining were determined to be 0.99 and 0.01, respectively.

Through these comprehensive assessments, the cutoff values for the three parameters were determined to be as appropriate as possible, ensuring the rationality and universality of the MCU combination process. These cutoff values play a crucial role in ensuring the accuracy and reliability of the genotype assignments made by FluTyping for each strain based on its genetic characteristics.

2.6. Genotype geographic distribution analysis

For the genetic diversity pattern shown multi-genotypes co-circulation, we have selected two representative subtypes, H1N2 and H9N2. The former primarily infects swine, and swine farming practices exhibit significant variations across different continents. The latter has the widest range of avian hosts. Therefore, these two subtypes are suitable for studying the correlation between regional transmission constraints and the formation of genetic diversity pattern. For these two subtypes, we initially calculated the proportion of each genotype present in different collection countries. Subsequently, based on this distribution, we utilize chi-square tests to analyze the differential enrichment distribution of various genotypes in the ccollection regions. This statistical analysis was performed using the R language.

2.7. Genomic evolution analysis

For various representative subtypes of influenza viruses, we initially calculated the dynamic changes in the genomic mutation proportion of eight gene segments over the years of collection. The objective was to observe differences in mutation rates among different gene segments across various influenza virus types. Additionally, we analyzed variations in selection pressures on nine proteins among different influenza virus types. This analysis utilized KaKs Calculator 3.0 with the ML evolutionary mode (Zhang, 2022). Visualization of the analysis results was achieved using the ggplot 2 v3.4.2 package in the R language v4.3.1 (Wickham, 2009).

2.8. Phylogenetic tree construction and visualization

In this study, the construction of all phylogenetic trees was performed using FastTree v2.1.10 (Price et al., 2009). Prior to tree construction, the alignment of the analyzed genomic sequences was carried out using mafft v7.505 (Katoh et al., 2002), ensuring that the sequences were properly aligned for accurate tree inference. The resulting phylogenetic trees were visualized using the R package GGtree v3.8.2 (Yu et al., 2017).

3. Results

3.1. Overview of FluTyping

FluTyping defines a comprehensive set of genotypes for all IAVs based on both genetic distance and phylogenetic relationships between isolates. The process involves three main steps: clustering, phylogenetic calibration, and genotyping. In the clustering step, distinct phylogenetic classes of each genomic segment are semiautomatically clustered using epidemiological information and evolutionary measures between isolates. The phylogenetic calibration step manually combines mixed clusters and outliers based on the phylogenetic topological structure. Finally, each isolate is assigned a specific genotype by combining optimized clusters of eight genes in the genotyping step. The algorithm implementation and parameter selection of FluTyping are detailed in the “Methods and Materials” section.

To validate the genotypes for studying the genetic diversity of IAV, the performance of the classifications of each gene in FluTyping was evaluated. On the one hand, the classifications of viral HA genomic sequences of seasonal influenza A viruses (H1N1 and H3N2 subtypes) and avian influenza virus (H5N1 subtype) were compared to the corresponding nomenclatures. As shown in Fig. 1C and Supplementary Fig. S11, the sequences of H1, H3 and H5 subtypes were classified into only two, three and one categories in FluTyping. Although FluTyping provided rougher classifications compared to the specific nomenclatures, it corrected unreasonable phylogenetic classifications in the nomenclatures. For instance, the 6B.1A.6 clade of H1 subtype was in the bottom of the corresponding phylogenetic tree constructing with all H1 nomenclative clades. Nonetheless, a cluster of the H1N1 isolates was also identified as the 6B.1A.6 clade distributing in the top of the phylogenetic tree, which likely due to the genomic sequences of these isolates had most mutations of the specific loci of 6B.1A.6 clade. Whereas, FluTyping accurately identified abnormal clades as individual clades, unlike the nomenclature-based classifications.

Additionally, the determination of homology between genomic sequences was employed to assess the validity of the genomic classifications in FluTyping. About 796 pairs of homologous genomic sequences from the FluReassort database, representing comprehensive reassortment events of IAV, were used for this evaluation (Ding et al., 2020). FluTyping classified 773 pairs of sequences into the same class, only 23 pairs of sequences showed different phylogenetic relationships compared to previous studies, confirming its ability to characterize distinct phylogenetic relationships of IAV genomic sequences effectively (Fig. 1D). These two assessments demonstrated the reliability of FluTyping in assigning genotypes to IAVs.

Overall, FluTyping proves to be a valuable tool for studying the genetic diversity of IAVs, allowing for a more comprehensive understanding of their evolutionary patterns, potential sources, and early detection of emerging viruses.

3.2. Landscape of genetic diversity of influenza A virus

With the distinct classes of eight genomic segments, FluTyping identified a total of 1698 genotypes for all IAV subtypes. However, 1061 of these genotypes had no more than three isolates, likely due to the emerging genotypes lacking a competitive advantage in the competition, leading to their elimination and failure to become prevalent. Apart from the potential low-adaptation genotypes, the majority genotypes (208/637) contained between 10 and 50 isolates, as shown in the Supplementary Tables S5–S6. The genotypes belonging to different subtypes were clustered in distinct clades in the phylogenetic tree constructed using the whole genome of all IAVs, confirming the rationality of genetic differences between genotypes identified by FluTyping (Fig. 2A).

For clarity, only the genotypes containing more than ten isolates were utilized to describe the landscape of genotypes, which involved 64 influenza subtypes. The isolates in these genotypes were collected from 1905 to 2022 and originated from 175 countries across six continents. As depicted in Fig. 2B, Supplementary Fig. S12, and Supplementary Table S7, the H3N2 subtype had the highest number of isolates (63,915) but only 20 identified genotypes, ranking fourth among all subtypes. In contrast, avian influenza H9N2 and H3N8 had both recognized 22 genotypes, with only a few thousand isolates each. The years 2015 and 2016 exhibited the highest genotype diversity, with 165 genotypes collected during this period. However, from 2016 to 2022, the diversity of genotypes decreased steadily. In 2022, a total of 23,118 isolates were collected, but only 41 genotypes were identified. In terms of collection regions, the United States and China took the top two positions, exhibiting the highest diversity in genetic types and the largest pool of isolates. Notably, the greater diversity of avian hosts led to a higher count of identified genotypes from avian sources (274) compared to other hosts. This underscores the significant role of avians in influencing the overall genomic diversity of IAVs. Detailed distribution statistics for all genotypes can be found in Supplementary Table S7.

The genetic diversity of all IAVs was quantitatively assessed using various measures, including the proportions of genotypes, the overall count of identified genotypes, along with the count of newly emerging genotypes, and the entropy of all genotypes over time. As shown in Fig. 2C, the genotypes exhibited increasing diversity since the year 2000. The 2009H1N1 pandemic led to the dominance of a novel genotype, causing a transient decline in genomic diversity. However, after 2009, the genotypes continued to become more diverse until around 2015. Fig. 2D depicts the trend of involving genotypes, showing an overall increase followed by a continuous decrease. Multiple inconspicuous peaks in genotype diversity also occurred in certain years, such as 1986, 2002, and 2007. The entropy, which is a measure of diversity, exhibited a similar trend to the number of genotypes over time. Notably, the entropy of overall genotypes reached its peak in 2008 and sharply declined following the 2009H1N1 pandemic. Another peak in entropy occurred in 2020, whereas the number of all genotypes had been continuously decreasing since its highest peak in 2015.

To determine whether the tendency of genetic diversity change was influenced by the high proportion of seasonal influenza H1N1 and H3N2 subtypes (which accounted for approximately 70% of all isolates), the entropy of genotypes without these seasonal influenza human-infection isolates was also calculated. As shown in Fig. 2D, the difference between the entropy calculated with (red line) and without (orange line) the seasonal influenza isolates increased with the proportion of seasonal influenza isolates. However, the change trend of the two sets of entropy values remained similar over time, indicating that there was a specific

change pattern in the genomic diversity of influenza viruses that was not significantly influenced by sampling bias.

3.3. Three distinct genetic diversity patterns of influenza A virus

To validate the determination of genetic diversity patterns, we focused on subtypes with more than 500 isolates collected across all years. This analysis led to the discovery of three distinguishable patterns, which are presented in Fig. 3: (1) One genotype domination pattern, (2) Multi-genotypes co-circulation pattern, and (3) Hybrid-circulation pattern. In general, representative subtypes formed distinct clades in the phylogenetic tree constructed using the whole genome (Fig. 3D).

The first pattern, characterized by one genotype domination, included only the seasonal influenza viruses of H1N1 and H3N2 subtypes. This pattern showed that a single genotype consistently dominated over a period of years, as depicted in Fig. 3A. The second pattern, represented by multiple genotypes co-circulation, included the swine H1N2 subtype and several representative subtypes of avian influenza, such as

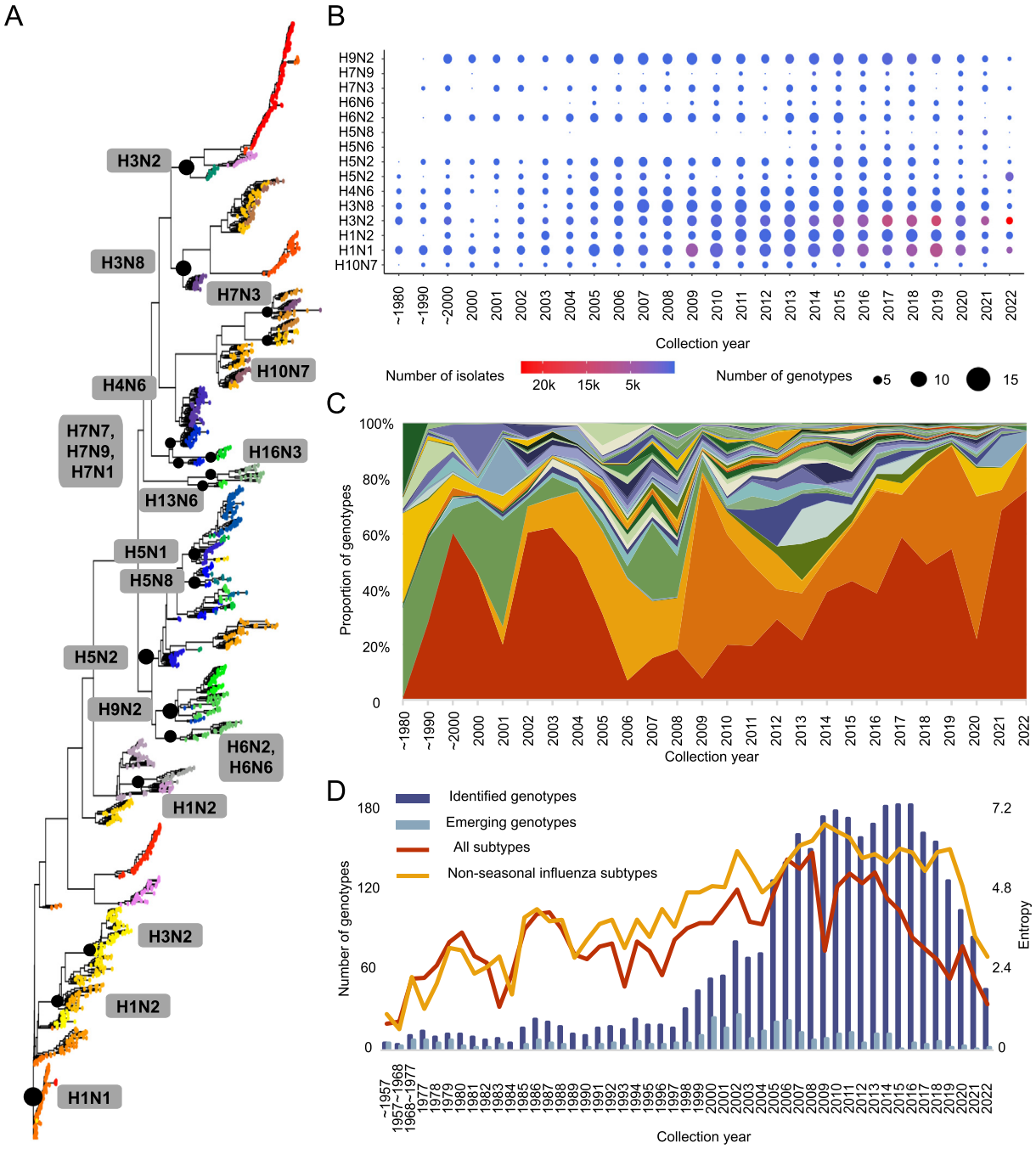

Fig. 2. Overview of genotypes identified for IAVs by FluTyping. **A** Phylogenetic relationships and subtype distributions based on the whole genome of all influenza viruses. **B** Temporal distribution of genotypes containing more than ten isolates for representative subtype. Circle size represents the number count of identified genotypes in different years, and circle color reflects the variety of sampled strains count. **C** Temporal analysis of genetic diversity in influenza viruses based on the proportion of genotypes over time. Different genotypes are represented by different colors. **D** Quantitative assessment of genetic diversity dynamics in all influenza A viruses. The assessments were conducted using the number of identified genotypes (dark blue), the number of emerging genotypes (light blue), and the entropy of all influenza viruses with and without seasonal influenza based on the proportion of genotypes (red and orange lines, respectively).

H9N2, H3N8, H4N6, etc. In this pattern, multiple genotypes were prevalent simultaneously, and there was no individual genotype dominating the subtypes, as shown in Fig. 3B. The third pattern was a mixture pattern of prevalence and included subtypes such as H7N9, H5N1, H5N6, and H5N8. For instance, Fig. 3C illustrates the case of the H5N1 subtype, where a dominant genotype circulated until 2013, followed by the emergence of different genotypes that co-circulated up to 2022. The temporal distribution of genotypes for various influenza subtypes within

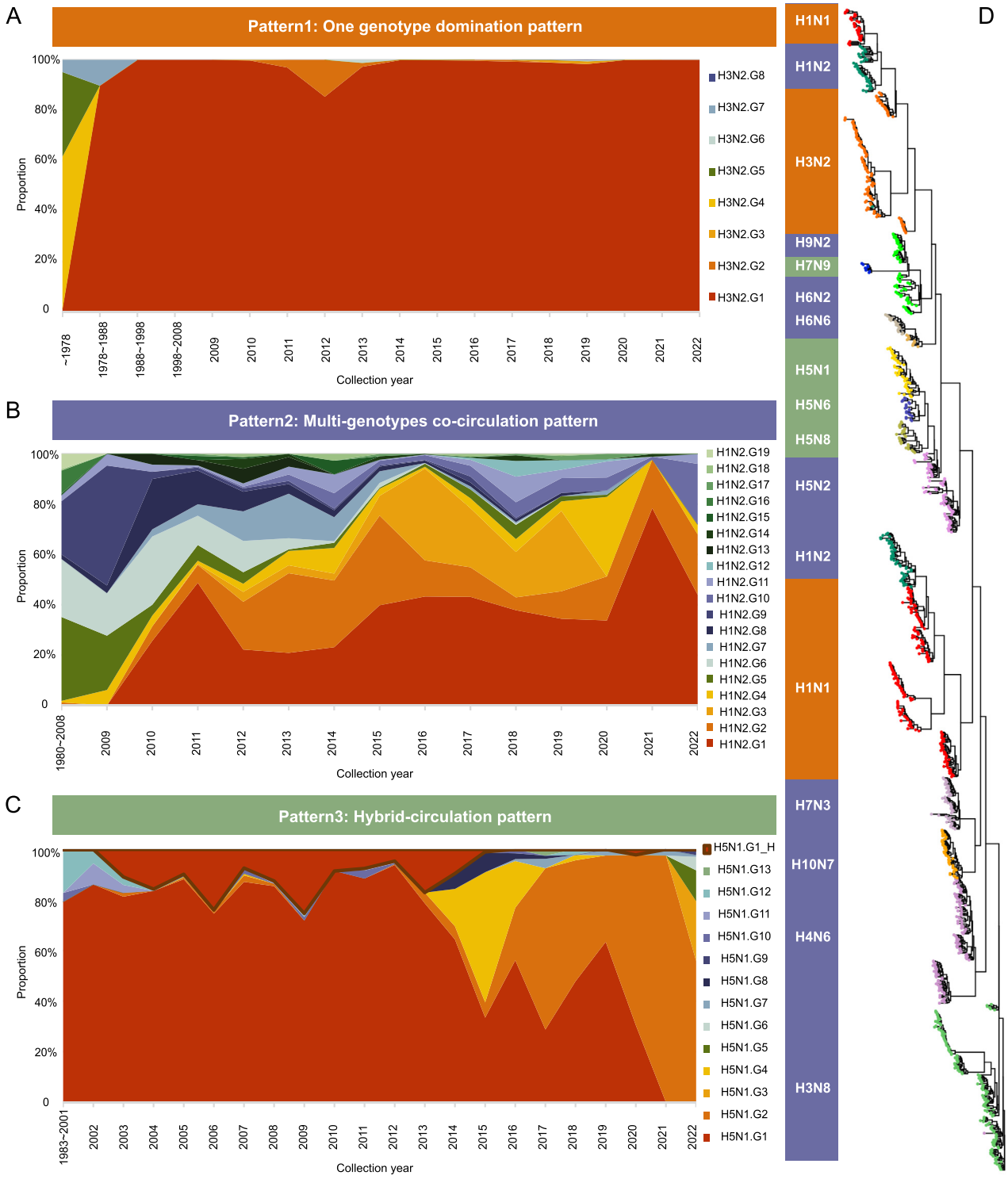

Fig. 3. Three distinct genetic patterns of influenza A virus. **A** Genetic diversity of H3N2 subtype, exhibiting the first pattern of one genotype domination, includes only seasonal influenza viruses of H1N1 and H3N2 subtypes. **B** Genetic diversity of swine H1N2 subtype, exhibiting the second pattern of multiple genotypes co-circulation. This pattern includes the swine H1N2 subtype and several representative subtypes of avian influenza, such as H9N2, H3N8, H4N6, etc. **C** Genetic diversity of H5N1 subtype, exhibiting the third pattern of multi-patterns mixture. This pattern includes subtypes such as H7N9, H5N1, H5N6, and H5N8. **D** Distribution of virus subtypes with different genetic diversity patterns on the phylogenetic tree constructed from the whole genome.

each genetic diversity pattern is illustrated in Supplementary Figs. S13 and S14.

3.4. Transmission restriction on genetic diversity

Certainly, identifying the factors that contribute to the distinct patterns of genomic diversity in IAVs is a significant scientific challenge. One key observation is that seasonal influenza viruses are epidemic globally, while swine and avian influenza viruses tend to be endemic. Furthermore, these two classes of influenza viruses exhibit different genetic diversity patterns. Based on this, we hypothesized that the genetic diversity of specific influenza viruses is influenced by their mode of transmission. To validate this hypothesis, we conducted statistical analyses on the distributions of genotypes in representative swine and avian influenza viruses according to their collection regions.

As shown in Fig. 4A and C, for some genotypes, there was a significant prevalence in a single country, with the proportion of the genotype being more than 75%, and the chi-square test resulting in a *P*-value less than 0.01. For example, the G4 genotype of the H1N2 subtype and the G1 genotype of the H9N2 subtype showed strong prevalence in specific countries. However, there were also genotypes, like the G1 of the H1N2 viruses, that were epidemic in multiple countries. To investigate whether different communities of viruses circulating in specific countries were present within these genotypes, we examined the distribution of isolates collected from different countries in the phylogenetic tree based on the whole genome. As seen in Fig. 4B, the isolates belonging to the G1 genotype of the H1N2 subtype primarily circulated in four countries (France, Germany, Spain, and the United Kingdom), and they formed distinct phylogenetic clades in the tree. Similarly, the isolates of the G10 genotype of H9N2 virus (Fig. 4D) were clustered in different phylogenetic clades based on the countries from which they were collected. As is well known, both pigs and poultry are part of confined farming methods. Additionally, wild birds only inhabit specific areas, with migratory processes existing solely in certain routes and times, limited to migratory birds. Therefore, the living areas of most non-human influenza host species are restricted. In contrast to human communal living and convenient transportation, the spread of influenza viruses is greatly constrained. As a result, some viruses with genotypes infecting pigs and poultry may only prevail in specific regions, unable to disseminate to other areas, thereby forming a model of genetic diversity with multiple genotypes co-circulation. These findings indicate that viral transmission in swine and avians limits the spread of specific genotypes, likely leading to the genetic diversity pattern of multiple genotypes co-circulating simultaneously.

In summary, our results support the hypothesis that the mode of transmission influences the genetic diversity patterns observed in IAVs. The restriction of viral transmission within specific regions contributes to the prevalence of certain genotypes in particular countries, which in turn shapes the genetic diversity pattern of multiple genotypes co-circulating within these virus subtypes.

3.5. Genomic evolution underlying genetic diversity

The study delved into the genomic evolutionary driving forces behind different genetic diversity patterns by analyzing the mutation proportions of different genes and the selection pressure on different proteins over time. Fig. 5A illustrates that the dominant genotype of the H3N2 subtype exhibited a significant linear increase in mutation proportion across all eight genomic segments over the years. Notably, the viral surface genes, namely the *HA* and *NA* genes, displayed more mutations compared to the internal genes (*PB2*, *PB1*, *PA*, *NP*, *M*, and *NS* genes). Similar trends were observed for the major genotypes of the H1N1 subtype, both before and after the 2009 epidemic (Supplementary Fig. S15).

To investigate whether these mutation rules were specific to the genetic diversity pattern of one genotype domination, the major genotypes of the H1N2, H7N9, H5N1, and H9N2 subtypes were also studied. Fig. 5B and Supplementary Fig. S15 demonstrate that three genotypes of the H1N2 subtype showed a similar change trend in mutation proportion compared to seasonal influenza viruses, but the mutations in the eight genes of H1N2 viruses were milder over time when compared to H1N1 and H3N2 subtypes. Additionally, Fig. 5C and D reveal that the *NP* and *PB1* internal genes of the H5N1 subtype had more mutations than the surface genes, which significantly differed from the H1N2 swine influenza virus and the seasonal influenza viruses. However, the mutation proportions of all eight genes of the H5N1 subtype increased steadily over time, similar to the H1N2 subtype. Similar mutation change patterns were also observed in major genotypes of the H7N9 and H9N2 subtypes (Supplementary Fig. S15). Furthermore, when comparing the same genotype from human and avian hosts, the mutation rates displayed a similar change trend over time for different genes, as seen in H5N1 and H7N9 subtypes (Fig. 5C and D, and Supplementary Fig. S15).

The study also measured the selection pressures on viral proteins using the dN/dS ratio calculated by the software of KaKs Calculator 3.0 (Zhang, 2022). Fig. 5E shows that, overall, the surface proteins (HA and NA) experienced greater selection pressures than the polymerase proteins (PB1 and PB2) and the nucleocapsid protein (NP) across all influenza subtypes. Additionally, the M2 protein in avian influenza viruses had a higher dN/dS ratio than in H1N2 swine influenza viruses and seasonal influenza viruses, while the opposite result was observed for the NS1 protein. Thus, based on genomic evolution analysis, no specific rules were identified for distinct genetic diversity patterns. However, significant differences were observed between influenza viruses with human/swine hosts and avian hosts.

4. Discussion

Our research differs from previous studies that focused on specific influenza viruses. Instead, we aimed to uncover common evolutionary patterns of all IAVs. To address this, we developed FluTyping, a robust method for identifying each influenza isolate's genotype within the entire virus population. The genotypes in FluTyping were assigned based on the classifications of eight genomic segments of overall IAVs, considering both genomic distance and phylogenetic relationships between isolates.

In general, the broad evolution analysis may be significantly affected by sampling bias. To alleviate the problem, an epidemiological combination of all influenza viruses was primarily employed. In this process, the genomic sequences of all isolates were clustered based on their collection years and countries. To further reduce sampling bias, representative sequences in each cluster were obtained based on the sequence similarity.

The classification of genomic sequences is influenced by a variety of complex factors, and there is currently no universally recognized standard for assessing its validity. However, in the case of influenza viruses, a widely accepted nomenclature system in the field involves the classification of subtypes H1, H3, and H5. These classification nomenclatures are primarily based on the distinct phylogenetic partitions, coupled with sequence similarities, to evaluate differences. Consequently, we have compared and assessed our genotype classification against these three established nomenclatures. Comparisons with the nomenclatures of the *HA* gene of H1N1, H3N2, and H5N1 subtypes showed FluTyping's classifications corrected unreasonable clades in phylogenetic trees, likely due to overly intricate partitions of specific subtypes. Additionally, the identification of reassortment events in influenza viruses inherently includes the recognition of homologous gene fragments. Therefore, we have also conducted a comparative analysis of our classification with previously acknowledged reassortment events in previous studies to evaluate the effectiveness of our classification. Results indicated the classifications of different genomic segments effectively identified most homologous genomic sequences (773/796) in influenza reassortment

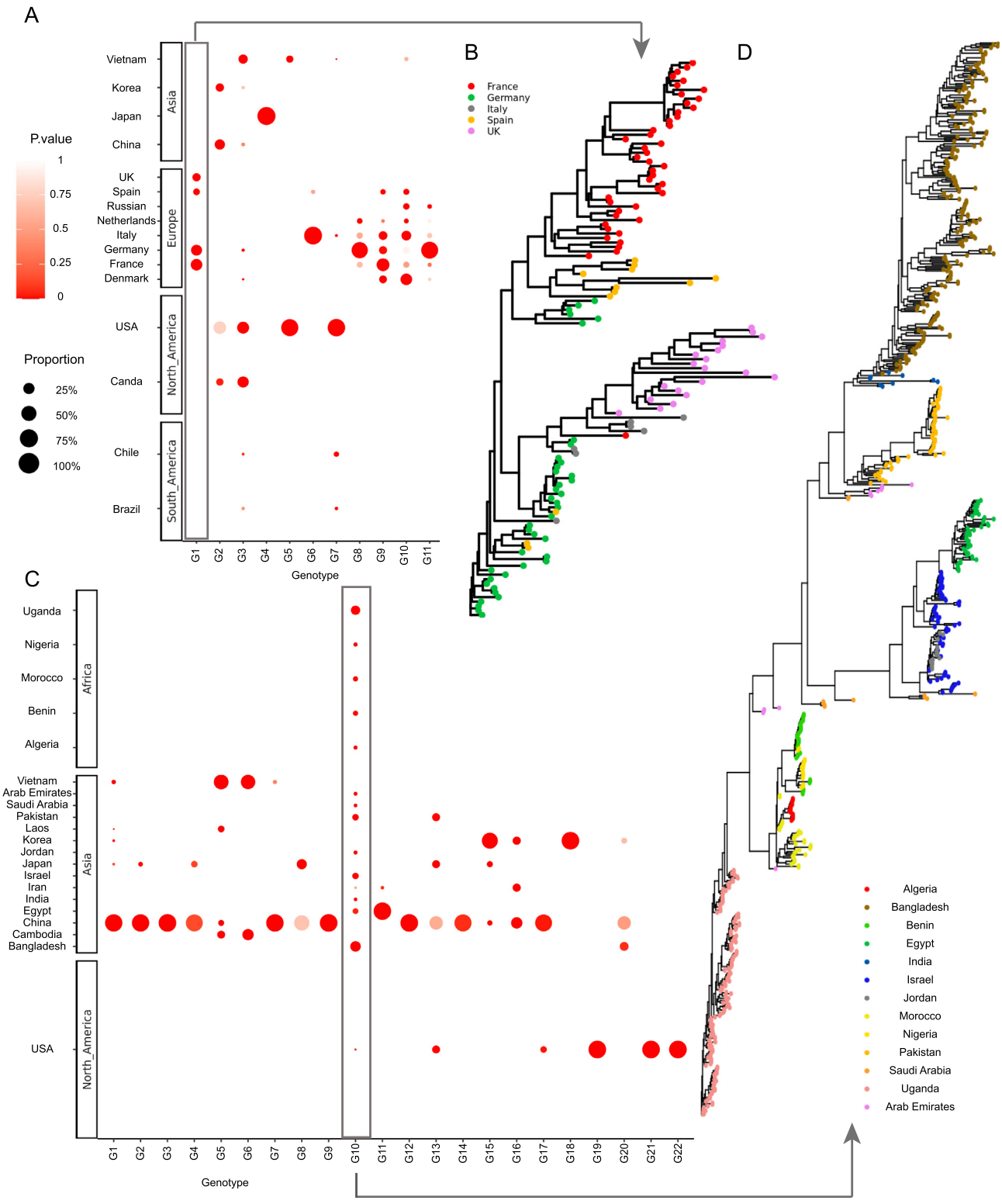

Fig. 4. Geographical circulation in genotypes of influenza A viruses from non-human hosts. **A** Geographical distribution enrichment analysis of H1N2 swine influenza viruses from different countries. The circle size represents the proportion of sampled strains for each genotype in the respective collection country. The color indicates the *P*-value from the chi-square test, where a deeper red color corresponds to a smaller *P*-value, indicating a higher enrichment of the genotype in that collection country. **B** Distribution of the G1 genotype of the H1N2 swine subtype from different countries on the phylogenetic tree constructed from the whole genome. **C** Geographical distribution enrichment analysis of the H9N2 avian influenza viruses from different countries, with the same quantitative indicators as in **A**. **D** Distribution of the G10 genotype of the H9N2 avian subtype from different countries on the phylogenetic tree constructed from the whole genome.

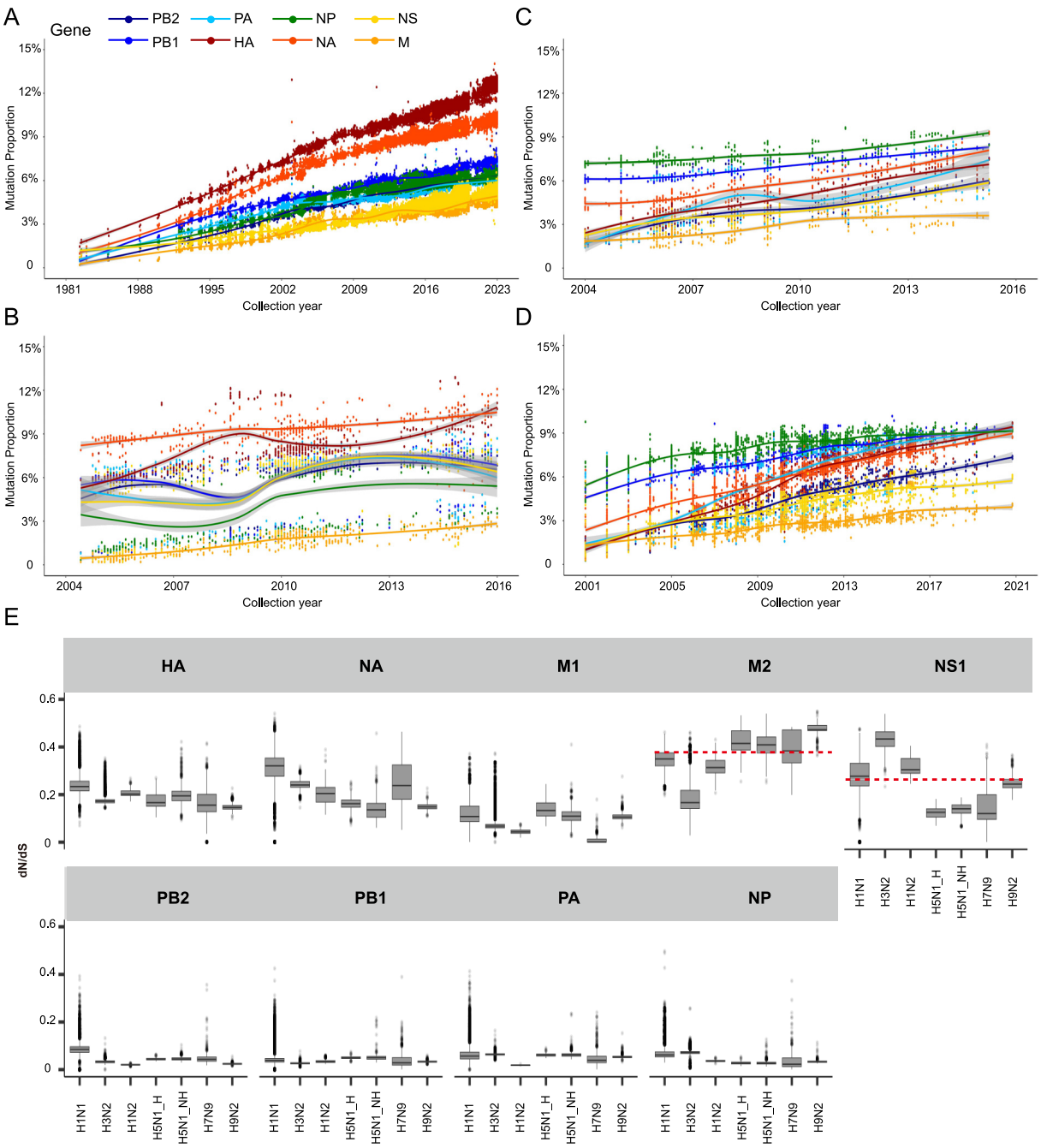

Fig. 5. Molecular evolution characteristics of different influenza A viruses. **A-D**, Temporal changes in mutation proportions across eight genomic segments of the major genotype in H3N2 (**A**), H1N2 (**B**), and the two major genotypes of the H5N1 subtype (**C** and **D**), respectively. **E**, Selection pressure on nine proteins of representative influenza subtypes. The selection pressures of viral proteins for different influenza virus subtypes were assessed by the dN/dS ratio. H5N1_H and H5N1_NH represent the H5N1 strains sampled from human and non-human hosts, respectively.

events from credible studies. These results demonstrate FluTyping's ability to accurately characterize distinct phylogenetic relationships among influenza isolates.

For isolates containing the complete genome, i.e., the eight genomic segments, a specific genotype was assigned based on the gene classifications in FluTyping. However, a larger number of genotypes (1061/1698) were related to no more than three isolates, resulting from that these genotypes may lack a competitive advantage in the competition, leading to their elimination and failure to become prevalent. Therefore, to ensure the validity of the obtained results, subsequent studies of the evolutionary pattern of all influenza viruses would employ genotypes containing enough isolates.

In summary, genotypes were assembled from various subtypes in the whole-genome phylogenetic tree. Different genotypes of the same subtype were clustered in distinct clades (Fig. 2A), demonstrating that these genotypes effectively characterize isolate phylogenetic relationships by influenza subtype. The H3N2 subtype had the highest isolates (63,915) with 20 genotypes, while the H9N2 subtype had 22 genotypes with 3808 isolates (Supplementary Table S7). This suggests diverse genetic patterns for different IAV subtypes.

The proportions of different genotypes for subtypes with more than 500 isolates were analyzed over time, revealing three distinct genetic diversity patterns: (1) One genotype domination pattern, (2) Multi-genotypes co-circulation pattern, and (3) Hybrid-circulation pattern (Fig. 3). The genetic

diversity pattern dominated by one genotype exclusively comprises seasonal influenza H1N1 and H3N2, while the other two patterns involve H1N2 swine flu and avian influenza viruses. Notably, swine and avian influenza viruses exhibit restricted transmission compared to seasonal influenza viruses. Consequently, we posit that viral transmission plays a crucial role in shaping distinct genetic diversity patterns across all IAVs, prompting us to conduct a restriction analysis of viral transmission.

To support this hypothesis, representative avian influenza subtypes and the H1N2 swine IAV were analyzed to study genotype distribution preferences across regions. Fig. 4 revealed not all genotypes were country-specific epidemics for the H1N2 and H9N2 subtypes. For the genotypes circulating in multiple countries, phylogenetic relationships between isolates from different countries were explored, showing genotypes forming distinct sub-groups circulating in specific countries. In summary, non-seasonal influenza genotypes exhibited high adaptability to regions, facilitating regional epidemics. In contrast, seasonal flu viruses were prevalent globally. Thus, varying viral transmission restrictions among different IAVs led to diverse genetic diversity patterns.

Moreover, we also explored key factors in IAV genetic diversity, focusing on molecular evolution. Fig. 5A–D showed higher mutation rates in *HA* and *NA* genes of seasonal IAVs and H1N2 swine IAV compared to avian influenza viruses. Fig. 5E revealed greater selection pressure on the NS1 protein in H1N2 and seasonal IAVs, opposite for the M2 protein. These results suggest the genomic evolution characteristics of all IAVs are more influenced by host-specific factors than by differences between genetic diversity patterns. Additionally, Fig. 5C and D showed similar molecular evolution patterns in avian viruses infecting humans and avians, indicating the human-infecting avian viruses shared genotypes with dominant circulating subtypes. This result establishes a theoretical foundation for understanding and raising awareness about human infections caused by avian influenza viruses.

Accessing comprehensive genotypes through FluTyping enables systematic origin tracing and evolutionary analysis of novel and circulating IAVs. For example, based on the identified genotypes, Supplementary Fig. S16 analyzed the temporal distribution of different genotypes of seasonal influenza H1N1 and H3N2 viruses, elucidating their genetic diversity. In this case, the identification of reassortment events of IAVs is a crucial scientific problem. In addition, there are many computational methods developed for identifying the reassortment events of IAVs (Ding et al., 2021). Based on the genotypes identified by FluTyping, rational reassortment criteria are formulated, incorporating constraints from epidemiology, Occam's razor principle, and other factors. This allows us to infer potential reassortment events for each genotype. Not only can we retrospectively identify past reassortment events through analytical analysis, but we can also match genotypes for newly emerging viruses, thereby speculating on the occurrence of reassortment events. Furthermore, we can identify the molecular characteristics of genes associated with these dominant genotypes, including specific residues in loci (Supplementary Fig. S17). Additionally, these findings support developing models to understand avian influenza evolution in non-human hosts worldwide, incorporating epidemiological and genomic data. Enhanced prevention and control measures, especially for avian virus spillover to humans, can result. Understanding factors driving avian influenza evolution and transmission dynamics will improve outbreak prevention and mitigation on human populations.

5. Conclusions

In conclusion, this study investigated the genetic diversity pattern of all IAVs using our self-developed genotype identification method called FluTyping. Unlike previous studies that focused on specific individual subtypes or groups of IAVs, we comprehensively analyzed the assigned genotypes of each isolate from FluTyping to recognize distinct genetic diversity patterns across all IAVs. Our in-depth analysis revealed that viral transmission emerged as the most crucial factor driving the variations in genetic diversity patterns among overall IAVs. By shedding light on the common evolutionary patterns of influenza A viruses, this research contributes to a deeper understanding of their genetic dynamics and transmission dynamics, which can be invaluable for devising effective strategies to combat future influenza outbreaks.

Data availability

Code for the analyses conducted in the study is available at https://github.com/dingxiao8715/FluTyping.

Ethics statement

No human or animal subjects were involved in this study.

Author contributions

Xiao Ding: conceptualization, investigation, data curation, methodology, formal analysis, visualization, writing original draft, writing-review, editing and funding acquisition. Jingze Liu: writing original draft and visualization. Taijiao Jiang: supervision and funding acquisition. Aiping Wu: conceptualization, supervision and funding acquisition.

Conflict of interest

The authors of this study declared that they have no conflict of interest.

Acknowledgements

We would like to express our gratitude to all those who have contributed sequences to the GISAID database (https://www.gisaid.org/). A comprehensive list of acknowledgments for the GISAID genome sequences utilized in our research can be found in the supplementary material (Supplementary Table S1).

This work was supported by the National Key Plan for Scientific Research and Development of China (2021YFC2301305 and 2021YFC2302001), the National Natural Science Foundation of China (32370703, 92169106, 9216910042 and 32070678), the CAMS Innovation Fund for Medical Science (2022-I2M-1-021, 2021-I2M-1-051), the Non-profit Central Research Institute Fund of Chinese Academy of Medical Sciences (2021-PT180-001) and the Emergency Key Program of Guangzhou Laboratory (grant EKPG21-12).

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.virs.2024.02.005.

References

Ali, S.T., Cowling, B.J., 2021. Influenza virus: tracking, predicting, and forecasting. Annu. Rev. Publ. Health 42, 43–57.

Barrat-Charlaix, P., Huddleston, J., Bedford, T., Neher, R.A., 2021. Limited predictability of amino acid substitutions in seasonal influenza viruses. Mol. Biol. Evol. 38, 2767–2777.

Benson, D.A., Cavanaugh, M., Clark, K., Karsch-Mizrachi, I., Lipman, D.J., Ostell, J., Sayers, E.W., 2013. GenBank. Nucleic Acids Res. 41, D36–D42.

Borkenhagen, L.K., Allen, M.W., Runstadler, J.A., 2021. Influenza virus genotype to phenotype predictions through machine learning: a systematic review. Emerg. Microb. Infect. 10, 1896–1907.

Carter, D.M., Darby, C.A., Lefoley, B.C., Crevar, C.J., Alefantis, T., Oomen, R., Anderson, S.F., Strugnell, T., Cortés-Garcia, G., Vogel, T.U., Parrington, M., Kleanthous, H., Ross, T.M., 2016. Design and characterization of a computationally optimized broadly reactive hemagglutinin vaccine for H1N1 influenza viruses. J. Virol. 90, 4720–4734.

Ciminski, K., Chase, G.P., Beer, M., Schwemmle, M., 2021. Influenza A viruses: understanding human host determinants. Trends Mol. Med. 27, 104–112.

Ding, X., Yuan, X., Mao, L., Wu, A., Jiang, T., 2020. FluReassort: a database for the study of genomic reassortments among influenza viruses. Briefings Bioinf. 21, 2126–2132.

Ding, X., Qin, L., Meng, J., Peng, Y., Wu, A., Jiang, T., 2021. Progress and challenge in computational identification of influenza virus reassortment. Virol. Sin. 36, 1273–1283.
Dong, Z.-L., Gao, G.F., Lyu, F., 2020. Advances in research of HIV transmission networks. Chinese Med J 133, 2850–2858.
Du Toit, A., 2023. Avian influenza takes flight in humans by evading restriction. Nat. Rev. Microbiol. 21, 551.
Elbe, S., Buckland-Merrett, G., 2017. Data, disease and diplomacy: GISAID's innovative contribution to global health. Global Challenges 1, 33–46.
Ganti, K., Bagga, A., Carnaccini, S., Ferreri, L.M., Geiger, G., Joaquin Caceres, C., Seibert, B., Li, Yonghai, Wang, L., Kwon, T., Li, Yuhao, Morozov, I., Ma, W., Richt, J.A., Perez, D.R., Koelle, K., Lowen, A.C., 2022. Influenza A virus reassortment in mammals gives rise to genetically distinct within-host subpopulations. Nat. Commun. 13, 6846.
Gao, R., Cao, B., Hu, Y., Feng, Z., Wang, D., Hu, W., Chen, J., Jie, Z., Qiu, H., Xu, K., Xu, X., Lu, H., Zhu, W., Gao, Z., Xiang, N., Shen, Y., He, Z., Gu, Y., Zhang, Z., Yang, Y., Zhao, X., Zhou, L., Li, Xiaodan, Zou, S., Zhang, Ye, Li, Xiyan, Yang, L., Guo, J., Dong, J., Li, Q., Dong, L., Zhu, Y., Bai, T., Wang, S., Hao, P., Yang, W., Zhang, Yanping, Han, J., Yu, H., Li, D., Gao, G.F., Wu, G., Wang, Y., Yuan, Z., Shu, Y., 2013. Human infection with a novel avian-origin influenza A (H7N9) virus. N. Engl. J. Med. 368, 1888–1897.
Hadfield, J., Megill, C., Bell, S.M., Huddleston, J., Potter, B., Callender, C., Sagulenko, P., Bedford, T., Neher, R.A., 2018. Nextstrain: real-time tracking of pathogen evolution. Bioinformatics 34, 4121–4123.
Han, Alvin X., Parker, E., Maurer-Stroh, S., Russell, C.A., 2019a. Inferring putative transmission clusters with Phydelity. Virus Evol 5, vez039.
Han, Alvin X., Parker, E., Scholer, F., Maurer-Stroh, S., Russell, C.A., 2019b. Phylogenetic clustering by linear integer programming (PhyCLIP). Mol. Biol. Evol. 36, 1580–1595.
Ji, C., Han, N., Cheng, Y., Shang, J., Weng, S., Yang, R., Zhou, H.-Y., Wu, A., 2022. sitePath: a visual tool to identify polymorphism clades and help find fixed and parallel mutations. BMC Bioinf. 23, 504.
Katoh, K., Misawa, K., Kuma, K., Miyata, T., 2002. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. Nucleic Acids Res. 30, 3059–3066.
Krammer, F., Smith, G.J.D., Fouchier, R.A.M., Peiris, M., Kedzierska, K., Doherty, P.C., Palese, P., Shaw, M.L., Treanor, J., Webster, R.G., García-Sastre, A., 2018. Influenza. Nat. Rev. Dis. Prim. 4, 3.
Li, W., Godzik, A., 2006. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics 22, 1658–1659.
Long, J.S., Mistry, B., Haslam, S.M., Barclay, W.S., 2019. Host and viral determinants of influenza A virus species specificity. Nat. Rev. Microbiol. 17, 67–81.
Medina, R.A., García-Sastre, A., 2011. Influenza A viruses: new research developments. Nat. Rev. Microbiol. 9, 590–603.
Müller, N.F., Stolz, U., Dudas, G., Stadler, T., Vaughan, T.G., 2020. Bayesian inference of reassortment networks reveals fitness benefits of reassortment in human influenza viruses. Proc. Natl. Acad. Sci. USA 117, 17104–17111.
Naguib, M.M., Verhagen, J.H., Mostafa, A., Wille, M., Li, R., Graaf, A., Järhult, J.D., Ellström, P., Zohari, S., Lundkvist, Å., Olsen, B., 2019. Global patterns of avian influenza A (H7): virus evolution and zoonotic threats. FEMS Microbiol. Rev. 43, 608–621.
Patrono, L.V., Vrancken, B., Budt, M., Düx, A., Lequime, S., Boral, S., Gilbert, M.T.P., Gogarten, J.F., Hoffmann, L., Horst, D., Merkel, K., Morens, D., Prepoint, B., Schlotterbeck, J., Schuenemann, V.J., Suchard, M.A., Taubenberger, J.K., Tenkhoff, L., Urban, C., Widulin, N., Winter, E., Worobey, M., Schnalke, T., Wolff, T., Lemey, P., Calvignac-Spencer, S., 2022. Archival influenza virus genomes from Europe reveal genomic variability during the 1918 pandemic. Nat. Commun. 13, 2314.
Pineo, R., 2021. Four flu pandemics: lessons that need to Be learned. J. Develop. Soc. 37, 398–448.
Ping, X., Hu, W., Xiong, R., Zhang, X., Teng, Z., Ding, M., Li, L., Chang, C., Xu, K., 2018. Generation of a broadly reactive influenza H1 antigen using a consensus HA sequence. Vaccine 36, 4837–4845.
Poon, A.F.Y., 2016. Impacts and shortcomings of genetic clustering methods for infectious disease outbreaks. Virus Evolution 2, vew031.
Price, M.N., Dehal, P.S., Arkin, A.P., 2009. FastTree: computing large minimum evolution trees with profiles instead of a distance matrix. Mol. Biol. Evol. 26, 1641–1650.
Prosperi, M.C.F., Ciccozzi, M., Fanti, I., Saladini, F., Pecorari, M., Borghi, V., Di Giambenedetto, S., Bruzzone, B., Capetti, A., Vivarelli, A., Rusconi, S., Re, M.C., Gismondo, M.R., Sighinolfi, L., Gray, R.R., Salemi, M., Zazzi, M., De Luca, A., on behalf of the ARCA collaborative group, 2011. A novel methodology for large-scale phylogeny partition. Nat. Commun. 2, 321.
Ragonnet-Cronin, M., Hodcroft, E., Hué, S., Fearnhill, E., Delpech, V., Brown, A.J.L., Lycett, S., 2013. Automated analysis of phylogenetic clusters. BMC Bioinf. 14, 317.
Scrucca, L., Fop, M., Murphy, T.B., Raftery, A.E., 2016. mclust 5: clustering, classification and density estimation using Gaussian finite mixture models. R J 8, 289–317.
Shi, J., Deng, G., Ma, S., Zeng, X., Yin, X., Li, M., Zhang, B., Cui, P., Chen, Y., Yang, H., Wan, X., Liu, L., Chen, P., Jiang, Y., Guan, Y., Liu, J., Gu, W., Han, S., Song, Y., Liang, L., Qu, Z., Hou, Y., Wang, X., Bao, H., Tian, G., Li, Y., Jiang, L., Li, C., Chen, H., 2018. Rapid evolution of H7N9 highly pathogenic viruses that emerged in China in 2017. Cell Host Microbe 24, 558–568.e7.
Smith, G.J.D., Donis, R.O., World Health Organization/World Organisation for Animal Health/Food and Agriculture Organization (WHO/OIE/FAO) H5 Evolution Working Group, 2015. Nomenclature updates resulting from the evolution of avian influenza A(H5) virus clades 2.1.3.2a, 2.2.1, and 2.3.4 during 2013-2014. Influenza Other Respir Viruses 9, 271–276.
Tan, M., Long, H., Liao, B., Cao, Z., Yuan, D., Tian, G., Zhuang, J., Yang, J., 2019. QS-Net: reconstructing phylogenetic networks based on quartet and sextet. Front. Genet. 10, 607.
Waters, K., Gao, C., Ykema, M., Han, L., Voth, L., Tao, Y.J., Wan, X.-F., 2021. Triple reassortment increases compatibility among viral ribonucleoprotein genes of contemporary avian and human influenza A viruses. PLoS Pathog. 17, e1009962.
WHO Working Group, 2023. Influenza (Seasonal). https://www.who.int/en/news-room/fact-sheets/detail/influenza-(seasonal)/. (Accessed 15 July 2023).
WHO/OIE/FAO H5N1 Evolution Working Group, 2009. Continuing progress towards a unified nomenclature for the highly pathogenic H5N1 avian influenza viruses: divergence of clade 2.2 viruses. Influenza Other Respir Viruses 3, 59–62.
WHO/OIE/FAO H5N1 Evolution Working Group, 2012. Continued evolution of highly pathogenic avian influenza A (H5N1): updated nomenclature. Influenza Other Respir Viruses 6, 1–5.
Wickham, H., 2009. ggplot2: Elegant Graphics for Data Analysis. Springer, New York, NY. https://doi.org/10.1007/978-0-387-98141-3.
World Health Organization/World Organisation for Animal Health/Food and Agriculture Organization (WHO/OIE/FAO) H5N1 Evolution Working Group, 2014. Revised and updated nomenclature for highly pathogenic avian influenza A (H5N1) viruses. Influenza Other Respir Viruses 8, 384–388.
Wu, A., Su, C., Wang, D., Peng, Y., Liu, M., Hua, S., Li, T., Gao, G.F., Tang, H., Chen, J., Liu, X., Shu, Y., Peng, D., Jiang, T., 2013. Sequential reassortments underlie diverse influenza H7N9 genotypes in China. Cell Host Microbe 14, 446–452.
Yamayoshi, S., Kawaoka, Y., 2019. Current and future influenza vaccines. Nat. Med. 25, 212–220.
Yang, H., Dong, Y., Bian, Y., Xu, N., Wu, Y., Yang, F., Du, Y., Qin, T., Chen, S., Peng, D., Liu, X., 2022. The influenza virus PB2 protein evades antiviral innate immunity by inhibiting JAK1/STAT signalling. Nat. Commun. 13, 6288.
Yu, G., Smith, D.K., Zhu, H., Guan, Y., Lam, T.T.-Y., 2017. ggtree: an r package for visualization and annotation of phylogenetic trees with their covariates and other associated data. Methods Ecol. Evol. 8, 28–36.
Zaraket, H., Baranovich, T., Kaplan, B.S., Carter, R., Song, M.-S., Paulson, J.C., Rehg, J.E., Bahl, J., Crumpton, J.C., Seiler, J., Edmonson, M., Wu, G., Karlsson, E., Fabrizio, T., Zhu, H., Guan, Y., Husain, M., Schultz-Cherry, S., Krauss, S., McBride, R., Webster, R.G., Govorkova, E.A., Zhang, J., Russell, C.J., Webby, R.J., 2015. Mammalian adaptation of influenza A(H7N9) virus is limited by a narrow genetic bottleneck. Nat. Commun. 6, 6553.
Zhang, Z., 2022. KaKs_Calculator 3.0: calculating selective pressure on coding and non-coding sequences. Dev. Reprod. Biol. 20, 536–540.