OXFORD

Briefings in Bioinformatics, 2025, **26(3)**, bbaf233
https://doi.org/10.1093/bib/bbaf233
Problem Solving Protocol

Influenza virus reassortment patterns exhibit preference and continuity while uncovering cross-species transmission events

Xiao Ding[1,2,‡], Yun Ma[1,2,‡], Shicheng Li[3,‡], Jingze Liu[1,2], Luyao Qin[1,2], Aiping Wu[1,2,*]

[1]State Key Laboratory of Common Mechanism Research for Major Diseases, Suzhou Institute of Systems Medicine, Chinese Academy of Medical Sciences & Peking Union Medical College, 100 Chongwen Road, Industrial Park District, Suzhou 215123, Jiangsu, China
[2]Key Laboratory of Pathogen Infection Prevention and Control (Peking Union Medical College), Ministry of Education, 16 Tianrong Street, Daxing District, Beijing 102629, China
[3]Center for Cancer Diagnosis and Treatment, The Second Affiliated Hospital of Soochow University, 1055 Sanxiang Road, Gusu District, Suzhou 215004, Jiangsu, China
[*]Corresponding author: Aiping Wu, State Key Laboratory of Common Mechanism Research for Major Diseases, Suzhou Institute of Systems Medicine, Chinese Academy of Medical Sciences & Peking Union Medical College, 100 Chongwen Road, Industrial Park District, Suzhou 215123, Jiangsu, China.
E-mail: wap@ism.cams.cn
[‡]Xiao Ding, Yun Ma and Shicheng Li contributed equally to this work.

Abstract

Genomic reassortment is a key driver of influenza virus evolution and a major factor in pandemic emergence, as reassorted strains can exhibit significantly altered antigenicity. However, due to technical and ethical constraints, research on reassortment patterns (RPs) has been limited, impeding effective surveillance and control strategies. To address this gap, we developed FluRPId, a framework for identifying RPs based on the genetic diversity of influenza viruses. FluRPId integrates principles of reassortment diversity maximization, dominance, and epidemiological likelihood to assess the credibility of detected reassortment events. Applying FluRPId, we constructed a comprehensive reassortment landscape of influenza viruses, encompassing widespread reassortment events with high credibility, which also include most previously reported reassortment events. Our analysis revealed that the NS gene frequently reassorts with PA and NA, while reassortment involving HA, NA, and NS occurs more frequently than expected. Furthermore, we identified specific loci combinations that exhibit strong linkage during reassortment, providing insights into segment association preferences. Additionally, extensive reassortment chains were observed across all subtypes, underscoring the continuity of reassortment in influenza virus evolution. Notably, we identified significant cross-species reassortment events and characterized host adaptation changes in cross-species-transmitted viruses. Our study provides the most comprehensive reassortment landscape of influenza viruses to date, uncovering key patterns, preferences, and evolutionary continuity. These findings bridge a critical gap in macro-scale reassortment studies and offer insights for future research and control efforts.

Keywords: influenza virus; reassortment pattern; genomic diversity; reassortment preference; cross-species transmission

Introduction

Influenza viruses, members of the *Orthomyxoviridae* family, possess segmented RNA genomes that enable genetic reassortment when multiple strains coinfect a host cell [1]. This process drives influenza virus evolution and has contributed to pandemics such as the 1957 H2N2, 1968 H3N2, and 2009 H1N1 outbreaks [1–4]. Reassortment also allows viruses to cross species barriers, making its identification crucial for epidemic control [5–7]. Consequently, the accurate identification of influenza virus reassortment patterns (RPs) and preferences is crucial for effective epidemic prevention and control.

Traditionally, reassortment detection relies on phylogenetic methods, which compare segment trees for incongruences but are labor-intensive and computationally demanding [8]. For instance, phylogenetic studies were instrumental in elucidating the reassortment origin of the 2009 H1N1 pandemic virus [4, 9]. Automated tools like FluReF, FluResort, and GiRaF improve efficiency but depend heavily on accurate phylogenies [10–12]. To overcome these limitations, nonphylogenetic methods such as CCV-based and HopPER algorithms broaden applicability but struggle with large genomic datasets, limiting RPs analysis [13, 14].

Recent studies have revealed that influenza virus reassortment is not entirely random but is influenced by host species, viral subtypes, and segment combination biases [15–22]. For instance, reassortment events predominantly occur within wild waterfowl populations, reflecting host-specific dynamics [15]. Furthermore, while seasonal human influenza subtypes H1 and H3 co-circulate, reassortment events between these subtypes are notably rare [18, 19]. Despite these findings, most insights remain observational and rely heavily on computational analyses, providing only a partial understanding of reassortment trends. Moreover, ethical and biosafety concerns associated with viral experiments [23, 24], coupled with the limitations of current computational methods, highlight the need for a systematic and comprehensive analysis of influenza virus RPs.

Received: March 7, 2025. **Revised:** April 16, 2025. **Accepted:** May 1, 2025

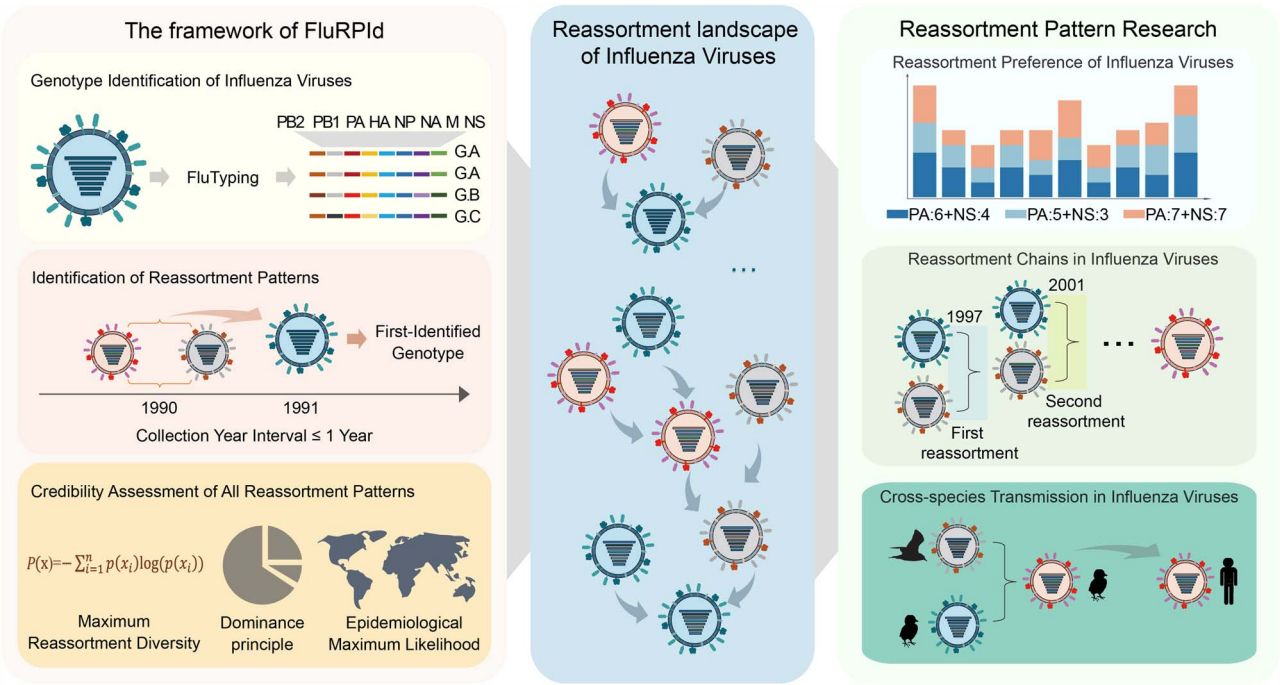

Figure 1. Schematic diagram of this study. The FluRPId framework, built upon the FluTyping tool, comprises three core components: Genotype identification, RP detection, and credibility assessment. Using FluRPId, we constructed a comprehensive reassortment landscape of influenza viruses, providing a systematic approach for advanced studies, including reassortment preference analysis, the investigation of reassortment chains, and the study of cross-species transmission in influenza viruses.

To address these gaps, we developed FluRPId, a fully automated framework leveraging FluTyping-defined phylogenetic diversity [25]. FluRPId constructs a macro-scale reassortment landscape, validated through independent criteria, revealing reassortment chains and preferences at subtype, gene segment, and phylogenetic clade levels. It retrospectively traced H7N9's cross-species transmission in 2013 and elucidated H5 reassortment's role in host adaptation. These findings provide a robust foundation for influenza research and epidemic preparedness.

Results

Framework and methodological overview

In this study, we developed FluRPId (**R**eassortment **P**attern **Id**entification for In**flu**enza Viruses), an advanced computational framework built upon FluTyping, a genotyping tool for influenza viruses (Fig. 1). FluRPId enhances throughput, efficiency, and accuracy, addressing a critical gap in reassortment research through a data-driven, rapid identification strategy.

Using FluRPId, we constructed a macro-scale landscape of influenza virus reassortment evolution, systematically illustrating how genotypes emerge and evolve through genomic reassortment. This analysis revealed subtype-, gene-, and clade-specific reassortment preferences, as well as reassortment chains linking distinct patterns (Fig. 1). Our findings highlighted significant reassortment preferences and the continuous nature of evolutionary reassortment. Additionally, reassortant genotype distributions across hosts uncovered key cross-species transmission events and adaptive changes following transmission, shedding light on host-specific evolutionary dynamics.

As illustrated in Fig. 1, the FluRPId workflow consists of three main steps. First, influenza virus genotypes were determined using FluTyping [25], incorporating comprehensive epidemiological metadata, including collection year, geographic origin, and host species, forming the foundation of this study. Second, RPs were identified based on two principles: (i) the reassortant genotype must be newly identified, and (ii) the reassortant and parental strains must be collected within 1 year. Third, credibility assessments were performed using three criteria: maximum reassortment diversity, the dominance principle, and epidemiological maximum likelihood. Each pattern was assigned a credibility score, with high credibility defined by various combinations of these criteria. Details are provided in the Methods and Materials section.

Validation and robustness assessment of FluRPId

Credibility score distribution of identified RPs

Using FluRPId, we identified 12 425 RPs and assigned credibility scores (Supplementary Table S1). The distribution of scores, including those from a validation dataset of previously reported events, is shown in Fig. 2a. Most patterns (44.4%) had scores between 0.3 and 0.5, while 27.45% met the high-credibility threshold (≥0.6, see Materials and methods section). Among 62 literature-reported reassortment events, 69.35% were classified as highly credible, with 34% reaching the maximum score of 1. These results indicate that while reassortment is globally distributed and evolutionarily significant, it remains relatively rare in influenza virus evolution.

Impact of biased sampling on the performance of RP identification

Significant sampling bias was observed in subtype, host, location, and year distributions. To evaluate FluRPId's robustness under such bias, we applied two sampling strategies (Methods).

First, we randomly sampled 500–50 000 strains (100 repeats) without considering epidemiological or genomic data. As shown in Supplementary Fig. S1a, credibility score distributions stabilized beyond 5000 strains, with Pearson correlations exceeding 0.762. Second, we sampled within specific categories (e.g. collection year, country, host, subtype), selecting 1 ~ 20 strains

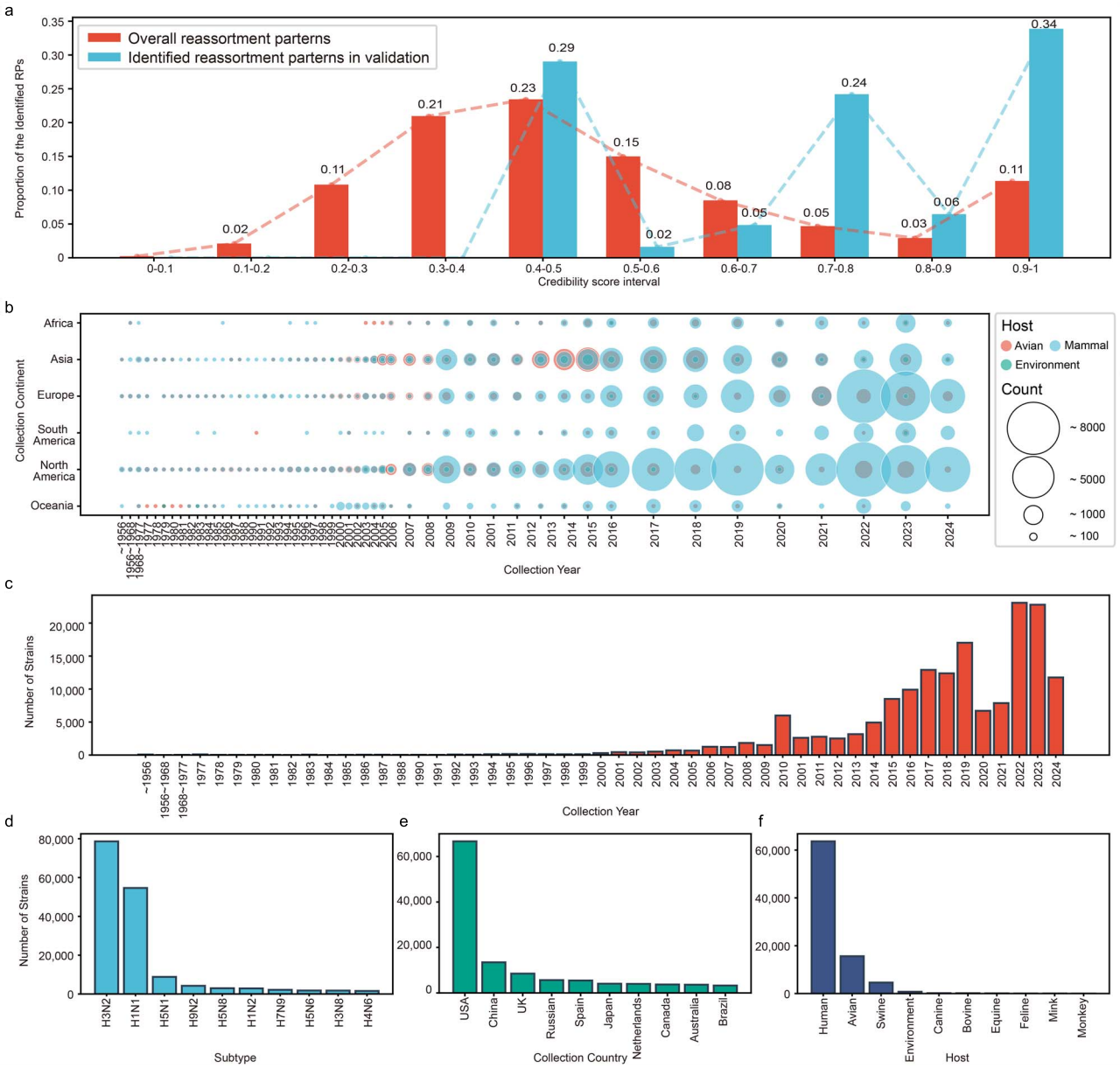

Figure 2. Overview of influenza virus RPs identified by FluRPId and the associated strains. (a) Distribution of credibility scores for all identified RPs (red bars) and the validation reassortment dataset (light blue bars). (b) Spatiotemporal distribution of strains involved in high-credibility RPs. The x-axis represents the collection year, and the y-axis denotes the collection continent. Bubble size corresponds to the number of sampled strains, while color indicates the collection host. (c) Distribution of collection years for the involved strains. (d–f) Distributions of the top 10 strain subtypes, collection countries, and sampling hosts, ranked by the number of samples.

per category (100 repeats). Supplementary Fig. S1b shows strong convergence and high correlation with the full dataset, even at one strain per category (5987 strains, correlation = 0.76). Category-based sampling also outperformed random sampling of similar sizes, with two strains per category (10 051 strains) achieving a higher correlation (0.792) than random sampling (0.762). Overall, sampling bias has minimal impact on FluRPId beyond 5000 strains and is further reduced by incorporating diverse epidemiological and genetic data.

Performance comparisons between FluRPId and other reassortment identification methods

To validate FluRPId, we compared its performance with RDP4 and GiRaf using real reassortment events and two simulated datasets (inter- and intra-subtype reassortment; see Methods).

In the real dataset, FluRPId, RDP4, and GiRaf detected 52, 143, and 73 events, respectively, all identifying three known events with high sensitivity. Among FluRPId's 49 unique events, only six had credibility scores above 0.6, ensuring minimal false positives-134 fewer than RDP4 and 64 fewer than GiRaf. As shown in Supplementary Fig. S2, FluRPId achieved superior precision and F1-score, surpassing RDP4 by 0.312 and 0.459 and GiRaf by 0.292 and 0.421, while all three methods maintained a recall of 1. Accuracy was not used as a primary metric due to dataset imbalance.

For the inter-subtype simulation, FluRPId achieved perfect precision, recall, and F1-score, correctly identifying all 4410 reassortment events. In contrast, RDP4 and GiRaf detected only 11 and 5 events, with recalls of 0.0025 and 0.0011, and F1-scores of 0.005 and 0.0023. Both maintained perfect precision. In the intra-subtype simulation, FluRPId again achieved perfect scores, while

RDP4 detected five events (recall: 0.0011, F1-score: 0.0023) and GiRaf failed to detect any events.

These results highlight FluRPId's robustness, particularly in complex cases with mutated parental strains, and demonstrate its superior performance across both real and simulated datasets, establishing it as a reliable reassortment analysis tool.

Comprehensive reassortment-driven evolution of influenza viruses identified by FluRPId

Using high-credibility RPs identified by FluRPId, we constructed the most comprehensive evolutionary landscape of influenza virus reassortment to date. Spanning 1932–2024, the dataset covers 177 countries, 16 host species, and 66 viral subtypes (Supplementary Table S2). Reassortment events surged post-2009 due to intensified surveillance following the H1N1 pandemic but declined after 2019, likely due to the impact of COVID-19 on influenza monitoring. Despite global distribution, distinct regional clusters persist, reflecting disparities in surveillance and sequencing capacity. Overall, statistical analyses confirm that RPs identified by FluRPId are globally widespread, persist over an extended evolutionary timescale, and involve diverse hosts and subtypes (Fig. 2b–f).

To visualize reassortment dynamics, we built a high-credibility reassortment network (credibility ≥0.6) under strict epidemiological constraints (Fig. 3a), deriving from the top 20 most frequent RP connections. Nodes represent reassortant genotypes, while directed edges indicate evolutionary links, with edge colors denoting reassortant subtype origins. The network is dominated by five subtypes-H5N1, H3N2, H3N8, H6N1, and H6N2-forming an interconnected reassortment system, alongside smaller intra-subtype clusters like H3N2 and H5N6.

A key structural feature of this network is the presence of "hub nodes," indicated by small squares in Fig. 3a. These hubs serve as central points in reassortment evolution: Source hub nodes (red squares) represent parent genotypes that have contributed to multiple reassortment events, acting as progenitors in the evolutionary process. Target hub nodes (blue squares) denote reassortant genotypes derived from a diverse array of parental strains, exemplified by an H1N8 genotype in the blue square of Fig. 3a.

Overall, FluRPId reconstructs a detailed reassortment landscape, offering new insights into reassortment dynamics, cross-species transmission, and adaptive evolution, providing a robust foundation for future influenza research.

Analysis of influenza virus reassortment preferences

The reassortment preferences for high credibility patterns were analyzed from three perspectives: the subtype-specific level, genomic segment level, and phylogenetic clade level.

First, as shown in Supplementary Fig. S3a, inter-subtype RPs (3317 in total) significantly outnumber intra-subtype RPs (94 in total) across all influenza virus subtypes. Moreover, reassortment events that did not lead to the emergence of new subtypes were more common than those that did. A more detailed statistical analysis of intra-subtype reassortment revealed that H3N8, H9N2, and H1N1 were the three most frequently involved subtypes (Supplementary Fig. S3b). Notably, the number of H3N8 strains amounted to 1571, ranking only ninth among all subtype strains. This suggests that H3N8 viruses exhibit high genetic diversity and may serve as a crucial gene pool for influenza virus reassortment, much like H9N2 viruses. This observation is further corroborated by Supplementary Fig. S3c, which presents the distribution of the top 10 RP combinations among different subtypes. Strikingly, all these combinations involve H3N8, either as the reassortant strain or as a parental strain.

Further analysis of these 10 combinations showed minimal variation in the number of RPs they involved. The most frequently observed RPs were between H6N8 and H3N8, leading to the generation of H3N8 reassortant strains occurring 32 times. Even the least frequent combination was observed 20 times. Two particular combinations warrant special attention: H6N2 reassorting with H3N8 and H3N2, resulting in the emergence of H3N2 and H3N8 subtypes, respectively. Since these events generate new subtype viruses, they carry the potential risk of triggering pandemics.

Next, an analysis of individual genomic segments was conducted to identify genes that are more prone to reassortment. As illustrated in Fig. 3b, the neuraminidase (NA) gene was the most frequently reassorted, appearing in 658 RPs—at least 1.27 times more than any other gene segment. Additionally, the hemagglutinin (HA), polymerase acidic (PA), and nonstructural (NS) genes also exhibited high reassortment frequencies, significantly surpassing other segments. In contrast, the nucleoprotein (NP) gene had the lowest reassortment frequency, with only 29 occurrences, indicating its conserved nature during genomic evolution.

Regarding dual-gene combinations, PA-NS was the most frequent, occurring 198 times-1.22 times more than the second most common combination, HA-NS. Given the high reassortment frequencies of NA, HA, PA, and NS, combinations such as NA-NS, PA-NA, and HA-NA each exhibited reassortment frequencies exceeding 100, significantly higher than other combinations. In contrast, combinations involving fewer than 10 reassortment events were predominantly associated with genes such as M and NP. Notably, despite their individual segments being involved in frequent reassortments, certain combinations, such as PB1-NS (9 occurrences) and PB1-NA (1 occurrence), displayed low reassortment frequencies.

For three-gene combinations, HA-NA-NS was the most frequent, occurring 2.84 times more often than the second most common combination, PA-HA-NA. Apart from PA-HA-NS, which appeared 12 times, all other three-gene combinations were observed in fewer than 10 reassortment events. This suggests that high reassortment frequencies of individual segments contribute significantly to the prevalence of three-segment combinations.

Finally, reassortment tendencies at the phylogenetic clade level were examined (Fig. 3c–e). Clade 6 of PA and clades 4 and 3 of NS exhibited the highest reassortment occurrences, while the N2.2 clade of NA ranked fourth, with 193 reassortment events. Other NA clades, including N8.2, N1.5, N3.1, N6.1, and N9, each had fewer than 100 reassortments. These findings, in conjunction with segment-level analysis, highlight the NA gene's dominant role in reassortment, significantly contributing to influenza virus genetic diversity. For dual-gene reassortments at the phylogenetic clade level, the most frequent combinations involved PA and NS genes: PA|3-NS|3, PA|6-NS|4, and PB2|4-NS|4. These findings align with the high reassortment frequencies of individual clades such as PA|6, NS|4, and NS|3. A similar trend was observed for three-gene combinations, where three out of the top five involved the HA-NA-NS combination, reinforcing its dominant role in reassortment.

To further investigate the molecular mechanisms underlying reassortment tendencies at the phylogenetic clade level, we identified specific sites within different gene clades where reassortment occurred. A site was defined as clade-specific if an amino acid at that position appeared in over 90% of sequences within a given clade but in less than 20% of sequences across other clades.

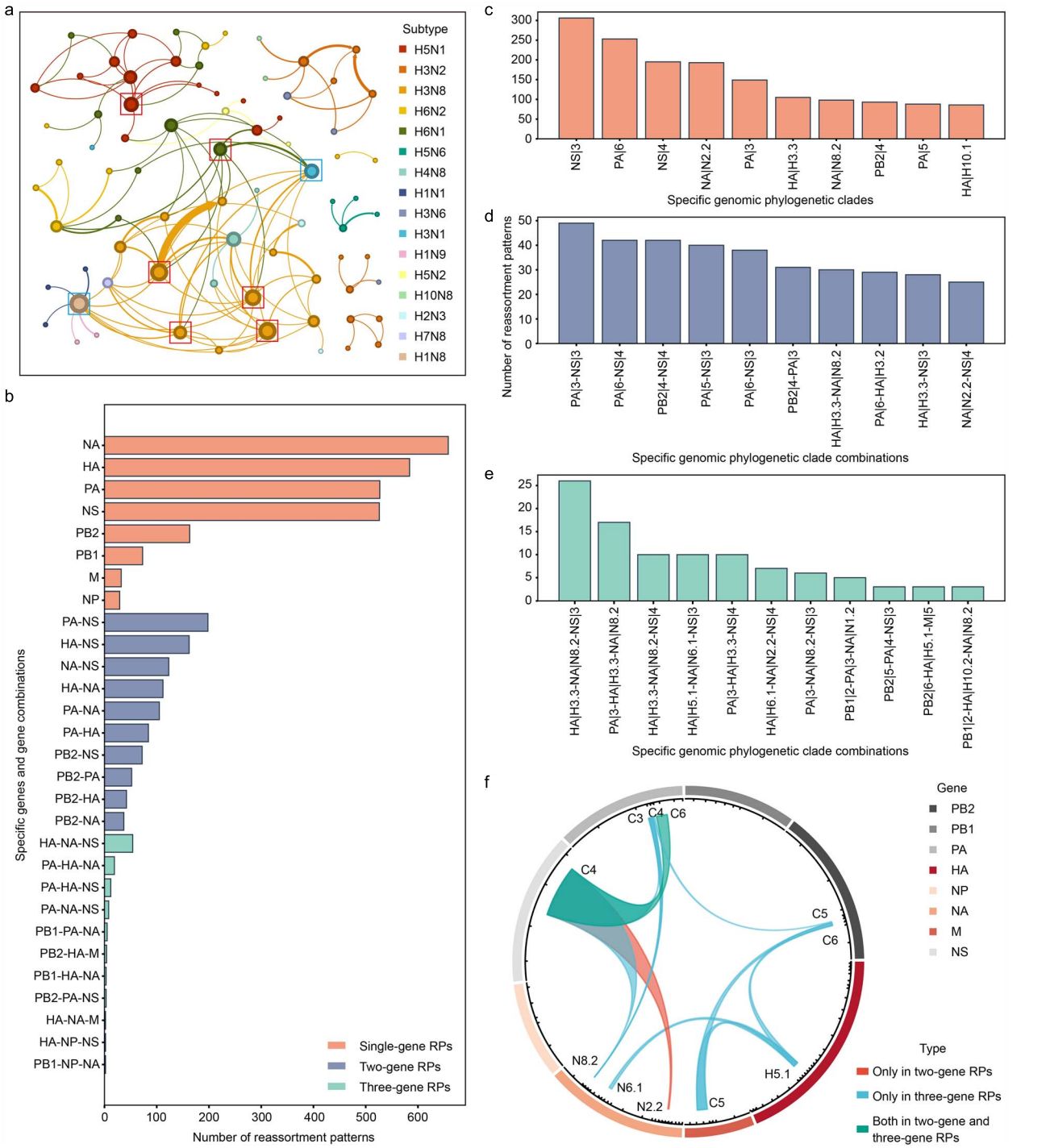

Figure 3. Construction of the reassortment network and analysis of reassortment preferences based on high-credibility RPs identified by FluRPId. (a) Reassortment network constructed using the top 20 most frequently occurring, highly credibility RPs in influenza viruses. This network visually represents the predominant RPs and their connections among these patterns. (b) Gene-level reassortment preferences analyzed by quantifying the frequency of reassortment events. The analysis includes the number of genes involved in reassortment and the reassortment tendencies of different gene types, providing insights into gene-specific reassortment dynamics. (c–e) Reassortment preferences examined from a phylogenetic clade perspective. (c) Reassortment tendencies of individual genes. (d) Pairwise reassortment preferences between gene segments. (e) RPs involving three-gene combinations. (f) Distribution of specific gene loci combinations that preferentially co-segregate during reassortment.

As illustrated in Fig. 3f, a total of 11 pairs of co-occurring clade-specific site combinations were identified, spanning six gene segments, excluding PB1 (Polymerase Basic Protein 1) and NP (Nucleoprotein). Most of these site combinations were found in three-gene reassortment events, with only one appearing in a two-gene reassortment event. Notably, the specific site combination PA|6 and NS|4 appeared in both two-gene and three-gene RPs. Detailed clade-specific site information is provided in Supplementary Table S3.

In summary, influenza virus reassortment exhibits distinct preferences at the subtype, genomic segment and phylogenetic clade levels. However, reassortment tendencies of specific gene

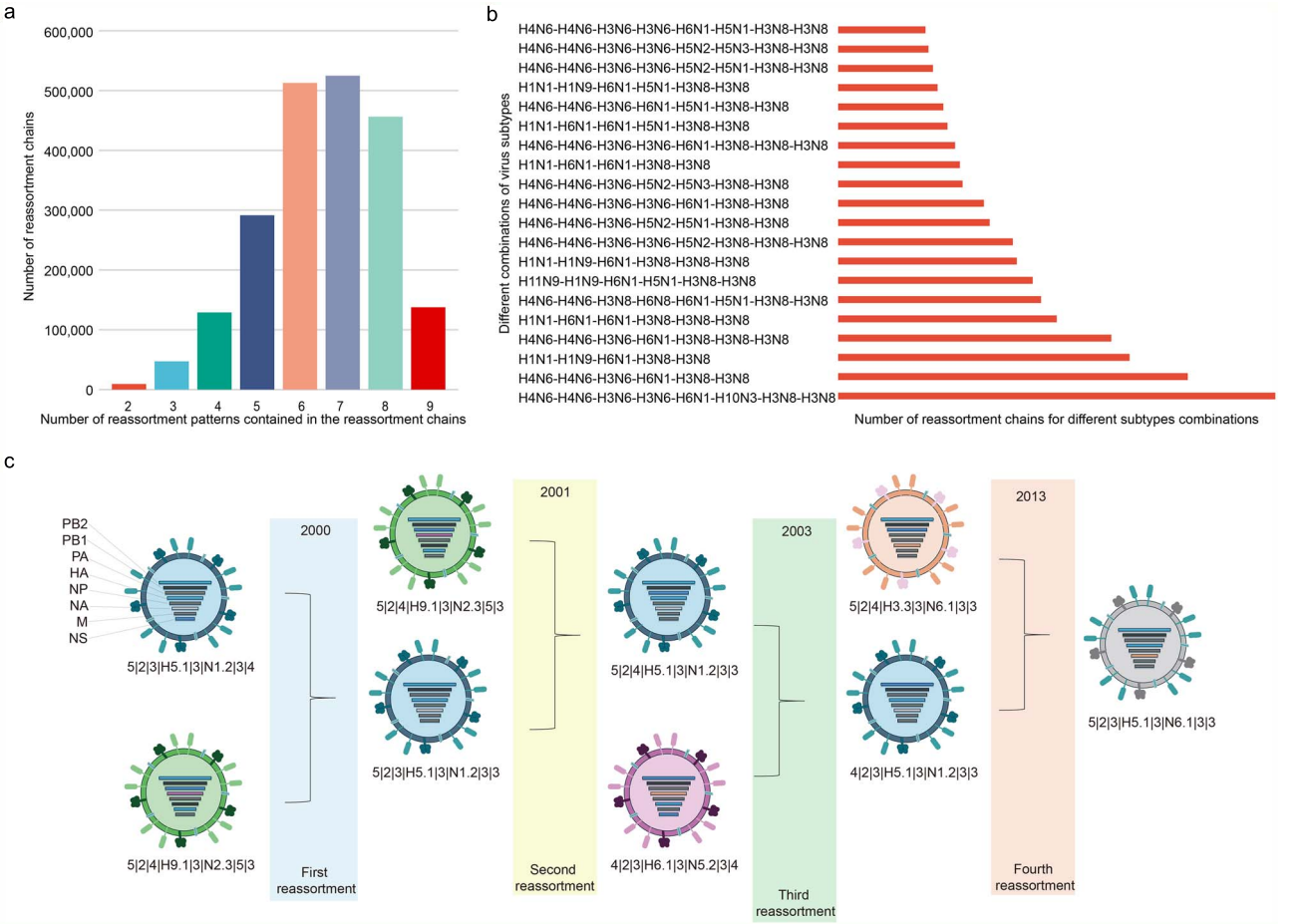

Figure 4. Analysis of influenza virus reassortment chains based on high-credibility RPs. (a) Statistical analysis of the distribution of reassortment chains based on the number of RPs they contain. (b) Statistical analysis of the distribution of connections between reassortant strain subtypes within reassortment chains for different high-credibility RPs (showing only the top 20 most frequently occurring combinations). (c) A representative reassortment chain in which each reassortment event has been previously documented in the scientific literature. This case serves as a validation, demonstrating the alignment of identified RPs with existing research findings.

segment combinations do not always align with patterns observed at the clade level, underscoring the role of reassortment in expanding the genetic diversity of influenza viruses.

Reassortment chains in influenza viruses

Reassortment chains based on credible RPs were identified across all influenza virus subtypes. We identified 2 106 889 nonredundant reassortment chains across all influenza virus subtypes, with the majority consisting of six or seven reassortment events (49.2%). Chains with fewer events (two or three) were less frequent (2.66%), as summarized in Fig. 4a. The longest reassortment chain observed in 2024 involved nine events, but chains with nine events made up only 6.52%. This suggests that reassortment-driven evolution in influenza viruses favors the inheritance of advantageous genes. A statistical analysis of the top 20 frequent chains revealed that the three most frequent chains each contained eight events, with six events sharing identical genotypes, indicating possible evolutionary regularities.

We analyzed a reassortment chain involving H5N1, which included four consecutive events with credibility scores of 1, 0.63, 1, and 0.8 (Fig. 4c). The chain began with an H5N1 strain (genotype 5|2|3|*H5.1*|3|*N1.2*|3|4) reassorting with an H9N2 strain, producing a new H5N1 strain (genotype 5|2|3|*H5.1*|3|*N1.2*|3|3), exemplified by the strain A/ck/hebei/718/01 [26]. This strain then underwent additional reassortment within the same year with H9N2 viruses, producing an H5N1 strain of genotype 5|2|4|*H5.1*|3|*N1.2*|3|3, such as A/ck/shangtou/5738/01 [26]. Two years later, in 2003, this reassortant genotype interacted with an H6N5 virus (genotype 4|2|3|*H6.1*|3|*N5.2*|3|4), yielding an H5N1 virus of genotype 4|2|3|*H5.1*|3|*N1.2*|3|3, represented by A/duck/hunan/689/2006 [27]. A decade later, in 2013, this H5N1 strain reassorted with an H3N6 virus (genotype 5|2|4|*H3.3*|3|*N6.1*|3|3), forming an H5N6 strain of genotype 5|2|3|*H5.1*|3|*N6.1*|3|3, exemplified by A/duck/guangzhou/018/2014 [28]. All reassortment events in this chain involved inter-subtype reassortment.

This analysis highlights the continuity of reassortment in influenza viruses, with significant variation in the frequency and complexity of events between human and avian viruses. It underscores the ability of influenza viruses to retain beneficial genetic material, particularly through inter-subtype reassortments, which are more common in evolutionary processes.

Analysis of cross-species transmission of influenza viruses

Genetic reassortment is a key factor in the cross-species transmission of influenza viruses [29–32]. To understand its role in host adaptation, we analyzed the annual distribution of reassortant strains from credible RPs, focusing on cross-species transmission from avian to human hosts. We identified six RPs linked to such events, with three of these patterns showing a clear correlation with cross-species transmission, as shown in Fig. 5.

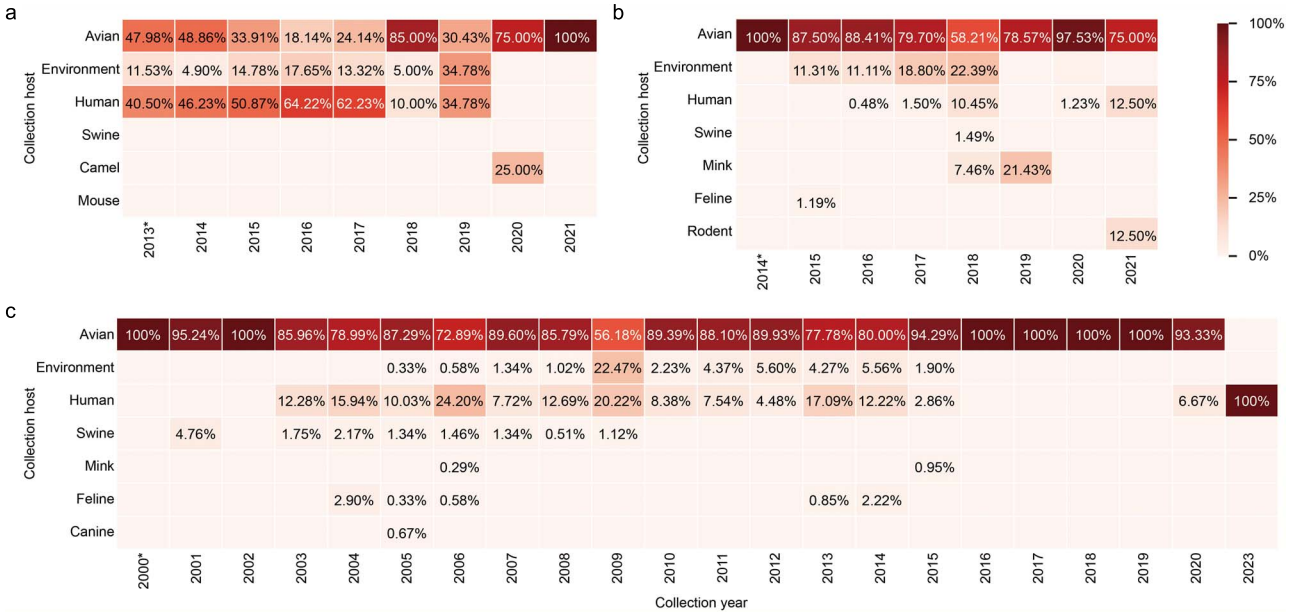

Figure 5. Analysis of cross-species transmission events influenza viruses based on high-credibility RPs. (a)–(c) present distinct reassortant genotypes that are strongly implicated in cross-species transmission for one H7N9, H5N6, and H5N1 subtype, respectively, which leveraged the alterations in the collection hosts of reassortant genotypes across different collection years.

Figure 5a shows the annual distribution of sampling hosts for the genotype of the reassortant strain that led to the H7N9 virus, generated by the reassortment of H7N3 and H9N9 from 2013 to 2021. This strain, responsible for the 2013 H7N9 outbreak in China, was capable of cross-species transmission. In 2013, 40.5% of the virus strains infected humans, and this proportion rose steadily over time, while avian strains declined. After the fifth outbreak in 2017, the proportion of human strains decreased, and the H7N9 genotype disappeared from the human population by 2019. These findings suggest the reassortment occurred before 2013, and the H7N9 virus adapted to humans before its eventual decline due to evolving immunity.

Figure 5b and c shows RPs for H5N6 and H5N1, respectively, linked to cross-species transmission events. Unlike H7N9, human infections for these strains emerged one to two years later, with cases reported in most subsequent years but without a clear upward trend. These reassortments likely enabled human infectivity, but the viruses remained primarily avian, limiting large-scale human infections. However, their gradual adaptation to humans underscores the need for continued vigilance, as these genotypes may pose a future pandemic risk similar to H7N9.

Overall, the highly credible RPs identified using FluRPId provide valuable insights into significant cross-species transmission events and the dynamic changes in host adaptation throughout the evolutionary history of influenza viruses. This information offers strong support for the prevention and control of potential pandemics caused by the cross-species transmission of influenza viruses in the future.

Conclusion and discussion

We developed FluRPId, a RP identification framework leveraging FluTyping's genetic classification. Using high-credibility RPs, we reconstructed the most comprehensive reassortment landscape to date, revealing reassortment chains, subtype-specific preferences, and cross-species transmission events, providing key insights into influenza virus evolution.

Despite these advancements, FluRPId has several limitations and areas for future development. Its accuracy depends on FluTyping's genotyping, making it a coarse-grained approach that identifies RPs rather than individual events. It also requires related large datasets (≥5000 strains) for reliable analysis. Notably, as shown in our analysis (Supplementary Fig. S1), when the number of sampled strains is insufficient, there is a marked increase in low-credibility RPs, underscoring the method's sensitivity to data volume. This highlights the importance of comprehensive and representative sampling in ensuring robust reassortment detection and reliable evolutionary insights. Additionally, the field lacks a standardized reassortment dataset, currently relying on FluReassort. Future efforts aim to establish a comprehensive benchmark for reassortment evaluation.

Our findings sometimes diverge from experimental studies. For instance, while Schulman et al. suggested the NS gene rarely reassorts, our analysis indicates frequent NS-PA reassortment [21]. This discrepancy may arise from strain-limited experimental settings versus our large-scale dataset. Additionally, reassortment events observed in controlled experimental settings, whether in vitro or in vivo, may not necessarily reflect natural environments. For example, the mutations detected in SARS-CoV-2 viral quasispecies were not consistently observed in consensus genomes at the population level [33, 34].

In addition, our results indicate that NP gene are the least prone to reassortment, which is likely attributable to their high level of conservation. The core functions of NP proteins—such as vRNA binding, viral genome transport, transcription, and replication—are essential for the survival of the virus. Consequently, the critical regions of NP genes are highly constrained against significant mutations, making them less likely to be exchanged during reassortment events.

Past studies focused on individual reassortment events, but successive reassortments can drive pandemics [35]. Many reassortment chains we identified include high-credibility events documented in literature, underscoring their role in influenza evolution and pandemic prediction.

The recent cross-species transmission events of H5N1 avian influenza in the United States—particularly the reported case of human infection likely linked to dairy cows—also drew our attention. Our constructed reassortment landscape revealed that the virus involved in this case belongs to genotype

3|3|3|H5.1|4|N1.2|3|3, which was first identified in 2022 and associated with 19 high-credibility RPs. Host distribution trends for this genotype showed a clear shift: while all detected viruses in 2022 were isolated from birds, by 2024, 73.7% of the strains were from dairy cows and only 22.25% from avian hosts, with a small proportion (0.58%) isolated from humans. Although direct evidence of cow-to-human transmission remains inconclusive, this shift strongly suggests an ongoing adaptation of the virus toward mammalian hosts. These findings highlight the importance of continuous monitoring of reassortant genotypes that may pose increasing risks at the animal–human interface.

Future improvements include integrating coinfection dynamics, ecological factors, and host-specific interactions to refine reassortment predictions. AI-driven multimodal analysis could further enhance risk assessment, enabling real-time surveillance and informed public health responses, advancing influenza research and pandemic preparedness.

Materials and methods

The framework of FluTyping

In the FluTyping framework, a total of 170 119 whole genomes of influenza A viruses were preprocessed and genotyped using data collected prior to 1 June 2024, from the Global Initiative on Sharing All Influenza Data (GISAID) database [36]. It organizes genomic sequences into phylogenetic classes per segment, refining them for accuracy. Each isolate's genotype consolidates segment clusters (PB2, PB1, PA, HA, NP, NA, M, NS) into a standardized format (e.g. 5|2|6|*H7.1*|3|*N7.2*|3|1). FluTyping effectively captures genetic characteristics, facilitating the identification of evolutionary patterns.

Identification of the RPs

Reassortment events were identified by analyzing genotypes produced by FluTyping in conjunction with epidemiological metadata. A reassortant strain was defined as one possessing a novel genotype—a unique combination of gene segments not previously observed. For each such genotype, plausible parental strains were sought whose gene segments could collectively reconstruct the reassortant genotype.

To ensure biological feasibility, two constraints were applied: (i) the reassortant genotype must be novel in the dataset, and (ii) both the reassortant and each of its inferred parental strains must have been collected within a one-year time window.

We exhaustively scanned all compatible genotype combinations across strains, leading to the identification of 472 377 reassortment events. To reduce redundancy and identify representative RPs, we grouped events by year, country, host species, and subtype. Only one event per group was retained, yielding a final set of 12 425 unique RPs.

This two-step approach—detection of novel genotype combinations and subsequent deduplication—provides a robust framework for studying the reassortment preferences and continuity of influenza viruses across different temporal and spatial contexts.

Credibility assessment of the RPs

To rigorously evaluate the reliability of the identified RPs, we established three rational assessment criteria: the Maximum Reassortment Diversity Principle, the Dominance Principle, and the Epidemiological Maximum Likelihood Principle.

Maximum reassortment diversity principle

Previous research suggests that a single reassortment event in influenza viruses can give rise to multiple reassortant strains [30, 37]. To quantify the diversity of new reassortant strains produced by a reassortment event, we employed entropy as a metric. Higher entropy values indicate greater diversity, thereby suggesting a higher likelihood of the reassortment event occurring. For each RP, we defined a reassortant strain system comprising all genotypes generated through reassortment from the parental genotypes within the past two years (i.e. when the sampling year interval between the reassortant strain and its parental strains is less than or equal to one year). The diversity of the RP was then assessed by calculating the entropy of this system using the following formula:

$$E_{pattern} = -\sum_{i=1}^{n} p\left(g_i\right) \log p\left(g_i\right) \tag{1}$$

Here, g_i represents the genotype generated by the parental genotypes within the RP, $p\left(g_i\right)$ denotes the probability of observing genotype g_i in the reassortant strain community, and n is the total number of distinct reassortant genotype categories produced.

Dominance principle

In general, the more prevalent the circulation of two influenza virus genotypes, the higher the likelihood of coinfection and, consequently, the greater probability of reassortment. To evaluate the reliability of a RP, we defined the *Dominance Score (DS)* as the product of the proportions of its two parental genotypes during the sampling year of the reassortant strain and the preceding year. This score serves as an indicator of the RP's reliability. The calculation formula is as follows:

$$DS_{pattern} = P\left(pg_1\right) \cdot P\left(pg_2\right) \tag{2}$$

Here, pg_1 and pg_2 represent the two parental genotypes involved in the RP and $P\left(pg_1\right)$ and $P\left(pg_2\right)$ denote their respective proportions in the specified time frame.

Epidemiological maximum likelihood principle

For a reassortment event involving two parental influenza viruses sampled in the same year, country, and host, the likelihood of coinfection increases, thereby elevating the probability of reassortment. An RP may involve multiple combinations of different epidemiological information for the parental strains. An RP can encompass various combinations of epidemiological attributes of the parental strains, which include collection year, country, and host species. To evaluate the probability of a given RP, we computed the proportion of parental strain combinations sharing identical epidemiological characteristics among all possible parental strain combinations. This proportion is defined as the *Epidemiological Score (ES)*. The calculation is performed using the following formula:

$$ES_{pattern} = \frac{N_s}{\sum_{i=1}^{n} N_i} \tag{3}$$

where N_s denotes the total number of parental strain combinations with identical epidemiological information and $\sum_{i=1}^{n} N_i$ represents the total number of unique epidemiological parental strain combinations within the RP.

Overall credibility score based on the three principles

We assessed the performance of each individual credibility parameter and their various combinations by analyzing the proportion of high-credibility RPs (defined as those with a credibility score ≥ 0.6) within the reassortment validation dataset.

As shown in Supplementary Fig. S4, among all evaluated strategies, the top three in terms of identifying high-credibility RPs were: (i) using the maximum value across all three evaluation principles (87.5%), (ii) the maximum of the Dominance Principle and the EML Principle (81.2%), and (iii) the maximum of the MRD Principle and the Dominance Principle (78.1%). Notably, selecting the maximum of all three principles yielded a significantly higher proportion of high-credibility patterns compared to other combinations. Based on this comparative analysis, we ultimately adopted the maximum value of the three principles as the final credibility score in this study.

Threshold selection of high-credibility RPs

To identify high-credibility RPs, we set the credibility score threshold at 0.6. This selection was guided by the distribution of credibility scores shown in Fig. 2a. Specifically, although the majority of RPs are distributed within the lower score intervals (e.g. 0.3–0.5), the proportion of patterns successfully validated is significantly higher in the higher score intervals. For example, only 2% of validated RPs fall into the 0.5–0.6 interval, but this percentage increases dramatically to 24% in the 0.7–0.8 interval and peaks at 34% in the 0.9–1.0 interval. This clear enrichment of validated RPs at higher score intervals indicates that credibility scores are strongly correlated with biological plausibility.

Choosing a threshold of 0.6 allows us to retain patterns with moderate to high credibility, excluding the majority of low-score patterns that are less likely to be reliable, while still preserving sufficient quantity for downstream analysis. Notably, patterns with scores above 0.6 account for only a small fraction (27%) of all RPs, highlighting the relative rarity of reassortment compared to other mechanisms like mutation. At the same time, these high-score patterns represent the majority of successfully validated events, ensuring the specificity of our selection. This threshold thus achieves a balanced trade-off between sensitivity and specificity, maximizing the biological relevance of our findings.

Methods for sampling datasets

We evaluated the impact of sampling bias on RP identification using two sampling methods by comparing the credibility score distributions.

Random Sampling: Strains were randomly selected from the full dataset at various sizes (500, 1000, 5000, 10 000, and 50 000), repeated 100 times. This unbiased method aimed to assess the impact of random selection on reassortment reliability.

Epidemiological and Genotype-Based Sampling: Strains were grouped by combinations of collection year, geography, host, subtype, and genotype. We then randomly selected strains from each group, with varying numbers chosen, repeated 100 times, to evaluate the impact of specific biases.

The selection of reassortment identification methods for performance comparisons

We first selected RDP4 (Recombination Detection Program), a widely used recombination detection tool [38]. RDP4 identifies evolutionary events at the species level based on sequence similarity, including genomic reassortment. It integrates multiple recombination detection methods, such as RDP, BootScan, GENECONV, and MaxChi. Additionally, the software provides tools for phylogenetic tree construction to support in-depth analysis of RPs in viral genomes. In addition, several influenza-specific reassortment algorithms exist, but few provide accessible code [39]. FluRPId, a coarse framework for detecting RPs among different influenza genotypes, is most comparable to FluGenome, though this database is no longer available [40]. Among accessible algorithms with verified functionality, we chose GiRaf [11], which detects influenza virus reassortment by comparing phylogenetic trees of different genome segments. It identifies tree topology inconsistencies, which suggest reassortment events. The method uses a probabilistic framework to assess the confidence of these events.

Construction of evaluation datasets and comparison protocol

Validation dataset with reported reassortment events

Using the FluReassort database [41], we mapped documented influenza reassortment events with FluRPId. After excluding incomplete events or those lacking sequence data, 130 events remained. Of these, 92 matched patterns identified by FluRPId, covering key subtypes like H5N1, H1N2, and H7N9. These 92 events formed the validation dataset for assessing FluRPId's reliability.

Dataset for reassortment identification performance comparison

We created three datasets to compare FluRPId with existing methods: one real dataset of influenza reassortment events and two simulated datasets for inter-subtype and intra-subtype reassortments. The real dataset included 96 strains, combining three high-credibility reassortment events from 2001, 2002, and 2013 with 87 randomly selected strains from the same years. For the simulated datasets, we randomly selected 10 high-credibility reassortment events involving either two H5N1 parental strains or H5N1 and H1N1 parental strains. While the reassortant strains in these events varied, the two parental strains in each event were identical. To introduce genetic variability, we generated 20 mutated genomes for each parental strain using a mutation rate of 0.001. These mutated strains were combined to create two simulated datasets, each comprising 52 strains (12 original strains and 40 mutated strains). The inclusion of mutated strains allowed us to assess the robustness of different reassortment identification methods under varying genetic conditions.

Comparison across reassortment identification methods

To evaluate and compare reassortment identification performance, we applied our proposed framework, FluRPId, alongside two widely used tools—GiRaF and RDP4—on all constructed evaluation datasets, including both real and simulated sets. Each method was used to identify potential reassortment events within the datasets, and the results were assessed based on their concordance with known reassortant genotype combinations.

A prediction was classified as a true positive (TP) if the identified genotype combination exactly matched a documented or predefined reassortment event. Predictions that did not correspond to any known event were considered false positives (FP). Known reassortment events that were not detected by the method were counted as false negatives (FN). Strains correctly identified as nonreassortant were considered true negatives (TN). For simulated datasets involving low-mutation strains, all mutated variants were assumed to belong to the same reassortment event as their original parental strains, allowing us to evaluate each method's robustness to minor genetic divergence.

To quantitatively assess method performance, we calculated the following four evaluation metrics based on TP, FP, TN, and FN:

$$Precision = \frac{TP}{(TP + FP)} \tag{4}$$

$$Recall = \frac{TP}{(TP + FN)} \quad (5)$$

$$Accuracy = \frac{(TP + TN)}{(TP + FP + TN + FN)} \quad (6)$$

$$F1 - measure = \frac{2 \times (\text{Precision} \times \text{Recall})}{(\text{Precision} + \text{Recall})} \quad (7)$$

These metrics were used to generate comparative performance plots (Supplementary Fig. S2), enabling a comprehensive evaluation of each method's accuracy and generalization ability under both real and simulated conditions.

Identification of reassortment chains

We identified reassortment chains for all influenza subtypes based on high-credibility patterns, applying strict epidemiological criteria. Reassortant strains from earlier events served as parental strains in subsequent events, with both parental genotypes originating from the same country and host species within the same year. By mapping these linkages, we reconstructed sequential reassortment pathways and annotated each node with its year of occurrence, revealing the temporal dynamics and evolutionary continuity of reassortment.

Construction of reassortment network

We constructed a reassortment network by integrating epidemiologically constrained chains, focusing on the top 20 most frequent RP connections. This network captures key reassortment continuity in influenza evolution. Using Gephi software, we visualized the network with nodes representing genotypes and edges indicating evolutionary relationships [42]. Directed edges reflect reassortment pathways, with colors representing the subtype of the reassortant strain. For example, if RP A produces an H5N1 reassortant strain, which subsequently undergoes reassortment with an N6-containing virus in RP B to generate an H5N6 strain, the directed edge would point from A to B. The edge color would correspond to the H5N1 subtype, reflecting the origin of the reassortant strain. This network offers a comprehensive view of reassortment dynamics, helping to analyze the connectivity and evolutionary implications of influenza virus reassortment.

Key Points

- We developed FluRPId, a fully automated framework that leverages FluTyping-defined phylogenetic diversity.
- FluRPId constructed a comprehensive reassortment landscape, which was validated through independent criteria.
- The reassortment landscape revealed reassortment chains and preferences of influenza viruses at multiple levels.
- The landscape successfully traced the cross-species transmission of H7N9 in 2013 retrospectively and clarified the role of H5 reassortment in host adaptation.

Acknowledgements

We gratefully acknowledge the authors from the originating laboratories and the submitting laboratories where genetic sequence data were generated and shared via GISAID, enabling this research.

Supplementary data

Supplementary data is available at *Briefings in Bioinformatics* online.

Conflict of interest: None declared.

Funding

This work is also supported by the National Natural Science Foundation of China (32370703, 92169106), the Non-profit Central Research Institute Fund of Chinese Academy of Medical Sciences (2021-PT180-001), the CAMS Innovation Fund for Medical Sciences (CIFMS) (2022-I2M-1-021, 2021-I2M-1-061, 2023-PT330-01, 2023-I2M-2-005, 2022-I2M-2-004), the National Key Plan for Scientific Research and Development of China (2021YFC2301305), the Major Project of Guangzhou National Laboratory (GZNL2024A01015) and the Suzhou Applied Basic Research Program (General Program in Medical and Health Sciences) (SYW2024065).

References

1. Webster RG, Bean WJ, Gorman OT. *et al.* Evolution and ecology of influenza A viruses. *Microbiol Rev* 1992;**56**:152–79. https://doi.org/10.1128/mr.56.1.152-179.1992
2. Cox NJ, Subbarao K. Global epidemiology of influenza: past and present. *Annu Rev Med* 2000;**51**:407–21. https://doi.org/10.1146/annurev.med.51.1.407
3. Kilbourne ED. Influenza pandemics of the 20th century. *Emerg Infect Dis* 2006;**12**:9–14. https://doi.org/10.3201/eid1201.051254
4. Smith GJ, Vijaykrishna D, Bahl J. *et al.* Origins and evolutionary genomics of the 2009 swine-origin H1N1 influenza A epidemic. *Nature* 2009;**459**:1122–5. https://doi.org/10.1038/nature08182
5. Chen Y, Liang W, Yang S. *et al.* Human infections with the emerging avian influenza A H7N9 virus from wet market poultry: clinical analysis and characterisation of viral genome. *Lancet* 2013;**381**:1916–25. https://doi.org/10.1016/S0140-6736(13)60903-4
6. Peacock TP, Moncla L, Dudas G. *et al.* The global H5N1 influenza panzootic in mammals. *Nature* 2024;**637**:304–13. https://doi.org/10.1038/s41586-024-08054-z
7. Sevilla N, Lizarraga W, Jimenez-Vasquez V. *et al.* Highly pathogenic avian influenza A (H5N1) virus outbreak in Peru in 2022-2023. *Infect Med (Beijing)* 2024;**3**:100108. https://doi.org/10.1016/j.imj.2024.100108
8. Rambaut A, Pybus OG, Nelson MI. *et al.* The genomic and epidemiological dynamics of human influenza A virus. *Nature* 2008;**453**:615–9. https://doi.org/10.1038/nature06945
9. Garten RJ, Davis CT, Russell CA. *et al.* Antigenic and genetic characteristics of swine-origin 2009 A(H1N1) influenza viruses circulating in humans. *Science* 2009;**325**:197–201. https://doi.org/10.1126/science.1176225
10. Lun AT, Wong JW, Downard KM. FluShuffle and FluResort: new algorithms to identify reassorted strains of the influenza virus by mass spectrometry. *BMC Bioinformatics* 2012;**13**:208.
11. Nagarajan N, Kingsford C. GiRaF: robust, computational identification of influenza reassortments via graph mining. *Nucleic Acids Res* 2011;**39**:e34. https://doi.org/10.1093/nar/gkq1232
12. Yurovsky A, Moret BM. FluReF, an automated flu virus reassortment finder based on phylogenetic trees, *BMC Genomics* 2011;**12** Suppl 2:S3. https://doi.org/10.1186/1471-2164-12-S2-S3.
13. Wan XF, Chen G, Luo F. *et al.* A quantitative genotype algorithm reflecting H5N1 Avian influenza niches. *Bioinformatics* 2007;**23**:2368–75. https://doi.org/10.1093/bioinformatics/btm354

14. Yin R, Zhou X, Rashid S. *et al.* HopPER: an adaptive model for probability estimation of influenza reassortment through host prediction. *BMC Med Genomics* 2020;**13**:9.
15. Dugan VG, Chen R, Spiro DJ. *et al.* The evolutionary genetics and emergence of avian influenza viruses in wild birds. *PLoS Pathog* 2008;**4**:e1000076. https://doi.org/10.1371/journal.ppat.1000076
16. Fulvini AA, Ramanunninair M, Le J. *et al.* Gene constellation of influenza A virus reassortants with high growth phenotype prepared as seed candidates for vaccine production. *PloS One* 2011;**6**:e20823. https://doi.org/10.1371/journal.pone.0020823
17. Marsh GA, Rabadan R, Levine AJ. *et al.* Highly conserved regions of influenza a virus polymerase gene segments are critical for efficient viral RNA packaging. *J Virol* 2008;**82**:2295–304. https://doi.org/10.1128/JVI.02267-07
18. Xu X, Smith CB, Mungall BA. *et al.* Intercontinental circulation of human influenza A(H1N2) reassortant viruses during the 2001-2002 influenza season. *J Infect Dis* 2002;**186**:1490–3. https://doi.org/10.1086/344738
19. Ghedin E, Fitch A, Boyne A. *et al.* Mixed infection and the genesis of influenza virus diversity. *J Virol* 2009;**83**:8832–41. https://doi.org/10.1128/JVI.00773-09
20. Rabadan R, Levine AJ, Krasnitz M. Non-random reassortment in human influenza A viruses. *Influenza Other Respi Viruses* 2008;**2**:9–22. https://doi.org/10.1111/j.1750-2659.2007.00030.x
21. Lubeck MD, Palese P, Schulman JL. Nonrandom association of parental genes in influenza A virus recombinants. *Virology* 1979;**95**:269–74. https://doi.org/10.1016/0042-6822(79)90430-6
22. Octaviani CP, Goto H, Kawaoka Y. Reassortment between seasonal H1N1 and pandemic (H1N1) 2009 influenza viruses is restricted by limited compatibility among polymerase subunits. *J Virol* 2011;**85**:8449–52. https://doi.org/10.1128/JVI.05054-11
23. Butler D. Fears grow over lab-bred flu. *Nature* 2011;**480**:421–2. https://doi.org/10.1038/480421a
24. Fouchier RA. Studies on influenza virus transmission between ferrets: the public health risks revisited. *MBio* 2015;**6**. https://doi.org/10.1128/mbio.02560-14
25. Ding X, Liu J, Jiang T. *et al.* Transmission restriction and genomic evolution co-shape the genetic diversity patterns of influenza A virus. *Virol Sin* 2024;**39**:525–36. https://doi.org/10.1016/j.virs.2024.02.005
26. Neumann G, Green MA, Macken CA. Evolution of highly pathogenic avian H5N1 influenza viruses and the emergence of dominant variants. *J Gen Virol* 2010;**91**:1984–95. https://doi.org/10.1099/vir.0.020750-0
27. Chen J, Fang F, Yang Z. *et al.* Characterization of highly pathogenic H5N1 avian influenza viruses isolated from poultry markets in Central China. *Virus Res* 2009;**146**:19–28. https://doi.org/10.1016/j.virusres.2009.08.010
28. Jiao P, Cui J, Song Y. *et al.* New reassortant H5N6 highly pathogenic avian influenza viruses in southern China, 2014. *Front Microbiol* 2016;**7**:754. https://doi.org/10.3389/fmicb.2016.00754
29. Guo X, Zhou Y, Yan H. *et al.* Molecular markers and mechanisms of influenza A virus cross-species transmission and new host adaptation. *Viruses* 2024;**16**:883. https://doi.org/10.3390/v16060883
30. Ganti K, Bagga A, Carnaccini S. *et al.* Influenza A virus reassortment in mammals gives rise to genetically distinct within-host subpopulations. *Nat Commun* 2022;**13**:6846. https://doi.org/10.1038/s41467-022-34611-z
31. Lee CY. Exploring potential intermediates in the cross-species transmission of influenza A virus to humans. *Viruses* 2024;**16**:1129. https://doi.org/10.3390/v16071129
32. Yan S, Wu G. Possibility of cross-species/subtype reassortments in influenza A viruses: an analysis of nonstructural protein variations. *Virulence* 2013;**4**:716–25.
33. Lau BT, Pavlichin D, Hooker AC. *et al.* Profiling SARS-CoV-2 mutation fingerprints that range from the viral pangenome to individual infection quasispecies. *Genome Med* 2021;**13**:62. https://doi.org/10.1186/s13073-021-00882-2
34. Bader W, Delerce J, Aherfi S. *et al.* Quasispecies analysis of SARS-CoV-2 of 15 different lineages during the first year of the pandemic prompts scratching under the surface of consensus genome sequences. *Int J Mol Sci* 2022;**23**:15658. https://doi.org/10.3390/ijms232415658
35. Wu A, Su C, Wang D. *et al.* Sequential reassortments underlie diverse influenza H7N9 genotypes in China. *Cell Host Microbe* 2013;**14**:446–52. https://doi.org/10.1016/j.chom.2013.09.001
36. Shu Y, McCauley J. GISAID: global initiative on sharing all influenza data – from vision to reality. *Euro Surveill* 2017;**22**:30494. https://doi.org/10.2807/1560-7917.ES.2017.22.13.30494
37. Taylor KY, Agu I, Jose I. *et al.* Influenza A virus reassortment is strain dependent. *PLoS Pathog* 2023;**19**:e1011155. https://doi.org/10.1371/journal.ppat.1011155
38. Martin DP, Murrell B, Golden M. *et al.* RDP4: detection and analysis of recombination patterns in virus genomes. *Virus Evol* 2015;**1**:vev003. https://doi.org/10.1093/ve/vev003
39. Ding X, Qin L, Meng J. *et al.* Progress and challenge in computational identification of influenza virus reassortment. *Virol Sin* 2021;**36**:1273–83. https://doi.org/10.1007/s12250-021-00392-w
40. Dong C, Ying L, Yuan D. Detecting transmission and reassortment events for influenza A viruses with genotype profile method. *Virol J* 2011;**8**:395. https://doi.org/10.1186/1743-422X-8-395
41. Ding X, Yuan X, Mao L. *et al.* FluReassort: a database for the study of genomic reassortments among influenza viruses. *Brief Bioinform* 2020;**21**:2126–32. https://doi.org/10.1093/bib/bbz128
42. Jacomy M, Venturini T, Heymann S. *et al.* ForceAtlas2, a continuous graph layout algorithm for handy network visualization designed for the Gephi software. *PloS One* 2014;**9**:e98679. https://doi.org/10.1371/journal.pone.0098679

Briefings in Bioinformatics, 2025, **26(3)**, bbaf233
https://doi.org/10.1093/bib/bbaf233
Problem Solving Protocol