



Computational predicting the human infectivity of H7N9 influenza viruses isolated from avian hosts

Yeping Sun¹ | Kun Zhang² | Heyuan Qi³ | He Zhang¹ | Shuang Zhang¹ | Yuhai Bi¹  |
Linhuan Wu³ | Lei Sun^{1,4} | Jianxun Qi¹ | Di Liu⁵ | Juncai Ma³ | Po Tien¹ |
Wenjun Liu^{1,4,6,7} | Jing Li^{1,4} 

¹CAS Key Laboratory of Pathogenic Microbiology and Immunology, Institute of Microbiology, Chinese Academy of Sciences, Beijing, China

²Philips Institute for Oral Health Research, School of Dentistry, Virginia Commonwealth University, Richmond, Virginia

³Information Center, Institute of Microbiology, Chinese Academy of Sciences, Beijing, China

⁴Savaid Medical School, University of Chinese Academy of Sciences, Beijing, China

⁵CAS Key Laboratory of Special Pathogens and Biosafety, Wuhan Institute of Virology, Chinese Academy of Sciences, Wuhan, China

⁶Institute of Microbiology, Center for Biosafety Mega-Science, Chinese Academy of Sciences, Beijing, China

⁷State Key Laboratory for Conservation and Utilization of Subtropical Agro-Bioresources & Laboratory of Animal Infectious Diseases, College of Animal Sciences and Veterinary Medicine, Guangxi University, Guangxi, China

Correspondence

Yeping Sun and Jing Li, CAS Key Laboratory of Pathogenic Microbiology and Immunology, Institute of Microbiology, Chinese Academy of Sciences, Beijing, China.

Emails: sunyeping@im.ac.cn (Y.S.); lj418@163.com (J.L.)

Funding information

This work was supported by grants from the National Natural Science Foundation of China (3191101787, 31970153, 31630079), the National Key R&D Program of China (2016YFD0500206), the Strategic Priority Research Program of the Chinese Academy of Sciences (XDB29010000) and the Mega-Project of Guangxi Natural Science Foundation (2015GXNSFEA139002). J.L. is supported by Youth Innovation Promotion Association of CAS (2019091).

Abstract

The genome composition of a given avian influenza virus is the primary determinant of its potential for cross-species transmission from birds to humans. Here, we introduce a viral genome-based computational tool that can be used to evaluate the human infectivity of avian isolates of influenza A H7N9 viruses, which can enable prediction of the potential risk of these isolates infecting humans. This tool, which is based on a novel class weight-biased logistic regression (CWBLR) algorithm, uses the sequences of the eight genome segments of an H7N9 strain as the input and gives the probability of this strain infecting humans (reflecting its human infectivity). We examined the replication efficiency and the pathogenicity of several H7N9 avian isolates that were predicted to have very low or high human infectivity by the CWBLR model in cell culture and in mice, and found that the strains with high predicted human infectivity replicated more efficiently in mammalian cells and were more infective in mice than those that were predicted to have low human infectivity. These results demonstrate that our CWBLR model can serve as a powerful tool for predicting the human infectivity and cross-species transmission risks of H7N9 avian strains.

KEYWORDS

class weight-biased logistic regression, H7N9 influenza virus, human infectivity, prediction, viral genome

This is an open access article under the terms of the Creative Commons Attribution-NonCommercial-NoDerivs License, which permits use and distribution in any medium, provided the original work is properly cited, the use is non-commercial and no modifications or adaptations are made.

© 2020 The Authors. *Transboundary and Emerging Diseases* published by Blackwell Verlag GmbH

1 | INTRODUCTION

Since 2013, H7N9 avian influenza virus has been one of the most serious public health events in China and Southeast Asia (Gao, et al., 2013; Zhang, Luo, & Shen, 2018). During 2013–2019, there were 1,568 confirmed cases in humans, which resulted in 616 deaths (Nations, 2019).

Since its first reemergence in 2013, H7N9 viruses have diversified into multiple HA sequence-based clades and dispersed to varying degrees across China (Petrie & Lauring, 2019; Su, et al., 2017). Substantial genetic diversity in the internal genes among different strains has also been demonstrated (Qi, et al., 2018). This genetic diversity among H7N9 strains raises two key questions: do H7N9 avian isolates with different genome sequences have the same potential to infect humans, and can the infectivity of a specific H7N9 strain be quantitatively predicted?

For an avian influenza virus strain to be able to undergo cross-species transmission from birds to humans, it must first acquire adaptive mutations in its genome that enable it to bind to the human-type influenza virus receptor with high affinity, to escape the host immune system effectively and to replicate efficiently in human cells (Long, Mistry, Haslam, & Barclay, 2019). Many such adaptive mutations have been identified in influenza viruses through gene sequence comparison and phenotype analysis (Schrauwen & Fouchier, 2014; Shi, Wu, Zhang, Qi, & Gao, 2014). Furthermore, Chen et al. discovered 'species-associated' positions in influenza virus genome-encoding proteins using position-specific entropy profiling methods, specific mutations which may enable an avian virus to become a human virus. (Chen, et al., 2006). Together, these findings suggest that the infectivity of a given avian influenza virus strain in humans is determined by its genome and could be predicted using viral genome-based methods.

In this paper, we introduce a novel class weight-biased logistic regression (CWBLR) algorithm that can recognize the human H7N9 strain genome with 100% accuracy and can predict whether an avian H7N9 strain has the potential to infect humans based on its genome.

2 | METHODS

2.1 | Biosafety and ethical approval

All experiments using H7N9 subtype strains were conducted in the BSL-3 level laboratory approved by the Wuhan Institute of Virology, Chinese Academy of Sciences. Animal care and housing were in compliance with ethical guidelines and approved by the Experimental Animal Ethic and Welfare Committee of the Institute of Microbiology, Chinese Academy of Sciences.

2.2 | H7N9 virus strain genome sequence procession

The complete 877 genome segment sequences of H7N9 avian and human isolates were downloaded from the National Institute of

Allergy and Infectious Diseases Influenza Research Database (IRD) (Zhang, et al., 2017). The original FASTA file downloaded from IRD containing all of the gene segments for all of the strains was split into eight FASTA files for each of the gene segments (PB2, PB1, PA, HA, NP, NA, MP and NS). Each segment was then aligned by multiple alignment using fast Fourier transform (MAFFT) (Katoh & Standley, 2013), following which the non-coding regions of each aligned segment were removed and the eight segments of each strain were concatenated to generate the aligned full-genome sequences. Each strain was then assigned a class label of 0 for avian isolates and 1 for human isolates. Finally, the nucleic acid codes 'A', 'T', 'G', 'C' and '-' (where '-' refers to a gap) at each genome position were converted to the integer codes 0, 1, 2, 3 and 4, respectively. The aligned, concatenated and integer-coded nucleic acid genome sequence of each H7N9 strain was used as the input variable x and its class label was used as the output variable Y for *model building*.

2.3 | Comparison of different classification algorithms

Different classification algorithms, including logistic regression (LR), k-nearest neighbour (KNN), random forest (RF), gaussian naive bayes (NB), support vector machine (SVM) and multilayer perceptron (MLP) implemented in the scikit-learn library in Python (Pedregosa, 2011) were applied to the aligned, concatenated, and integer-coded nucleic acid genome sequence data set described above, and the corresponding models were fitted with fivefold cross-validation. The predictive accuracy (A) in each fitting was calculated as follows:

$$\text{Accuracy}(A) = \frac{\text{Number of correct predictions}}{\text{Number of total predictions}} \quad (1)$$

We considered that a predicted probability of >0.5 for a human (or avian) strain being placed in the human (or avian) class was a correct prediction, whereas a predicted probability of ≤ 0.5 was an incorrect prediction.

2.4 | CWBLR Model building and verification

Binomial LR algorithm (Li, 2012) was used to build a classification model based on the genome sequences of the H7N9 strains. The resulting binomial logistic regression model had the following conditional probability distribution:

$$P(Y=1|x) = \frac{\exp(w \cdot x + b)}{1 + \exp(w \cdot x + b)} \quad (2)$$

$$P(Y=0|x) = \frac{1}{1 + \exp(w \cdot x + b)} \quad (3)$$

where $x \in R^n$ is the input, $Y \in \{0, 1\}$ is the output, $w \in R^n$ is the weight coefficient, and b is the intercept. Let $w = (w^{(1)}, w^{(2)}, \dots, w^{(n)}, b)$ and let $x = (x^{(1)}, x^{(2)}, \dots, x^{(n)}, 1)$, then:

$$P(Y=1|x) = \frac{\exp(w \cdot x)}{1 + \exp(w \cdot x)} \quad (4)$$

$$P(Y=0|x) = \frac{1}{1 + \exp(w \cdot x)}$$

The parameters of the logistic regression model were estimated using the maximum-likelihood method. Let $P(Y=1|x) = \pi(x)$, $P(Y=0|x) = 1 - \pi(x)$, and the likelihood function:

$$\prod_{i=1}^N [\pi(x_i)]^{y_i} [1 - \pi(x_i)]^{1-y_i} \quad (5)$$

Then, the logarithmic likelihood function is.

$$\begin{aligned} \log L(w) &= \sum_i^m [y_i \log \pi(x_i) + (1-y_i) \log(1 - \pi(x_i))] \\ &= \sum_i^m \left[y_i \log \frac{\pi(x_i)}{1 - \pi(x_i)} + (1-y_i) \log(1 - \pi(x_i)) \right] \\ &= \sum_i^m [y_i(w \cdot x_i) - (1-y_i) \log(1 + \exp(w \cdot x_i))] \end{aligned} \quad (6)$$

The partial derivative of the logarithmic likelihood function is.

$$\begin{aligned} \frac{\partial \log L(w)}{\partial w} &= \sum_i^m \left[y_i - \frac{\exp(w \cdot x_i)}{1 + \exp(w \cdot x_i)} \right] \cdot x_i \\ &= \sum_i^m [y_i - \pi(x_i)] \cdot x_i \end{aligned} \quad (7)$$

The iterative equation of gradient ascent is.

$$w = w + \alpha [y - \pi(x)] \cdot x \quad (8)$$

The gradient ascent algorithm finally finds the parameter w which makes the logarithmic likelihood function reach a minimum. The parameter w is a vector containing equal number of elements with the input variables (x). It is also called the coefficient vector. The coefficient value corresponding to each input variable (each genome position in the present study) reflects the importance of this variable to the probability output. We used the logistic regression algorithm implemented in the scikit-learn library in Python (Pedregosa, 2011) to fit the model and calculate the parameter w . Our novel consideration was to give the human class a higher weight so that the predictive accuracy of human virus isolates could reach 100%.

We aim to build a logistic regression model which can 100% accurately predict human isolates. A predictive accuracy of 100% for the human class means that the model can always correctly recognize the genome of a human H7N9 virus isolate and so can also correctly recognize avian H7N9 isolates with a high potential of falling into the

human class. In other words, the ideal model built by our method should work as follows: the sequences of the eight genome segments of an H7N9 strain are input into the model, following which it will calculate the probability that this strain belongs to the human class. If the strain is really a human isolate, the probability will always be >0.5 , whereas if the strain is an avian isolate, the probability will reflect its potential of belonging to the human class, with a probability of >0.5 suggesting that the avian strain has a high risk of infecting humans. Therefore, the probability that is calculated by the model can be considered to reflect the human infectivity of a given H7N9 strain.

We assigned a series of class weights that were >0.5 to the human class and then fitted the logistic regression model using fivefold CV. For each class weight, the entire data set (887 strains) was randomly divided into five subsets of similar size, four of which were used as training data sets for model fitting and one of which was used as a test data set to calculate the accuracy of the fitted model. There were five possible combinations of the training test data sets, so five models could be fitted and five accuracies could be calculated. Therefore, the average of the five accuracies was used as the overall accuracy of the model that corresponded to a particular class weight. The accuracies of the human and avian classes were then calculated for each class weight, and the class weight that gave a human class accuracy of 100% and the largest avian class accuracy was selected. The corresponding model was then used to calculate the human infection probabilities of the H7N9 virus strains that were isolated in 2016–2017. The procedure from genome sequence procession to model selection is illustrated in Figure S1.

2.5 | Cell lines and viruses

The H7N9 strains SD183, YZ30, TJ186, BD1 and TZ45 were generously provided by Professor Xiufan Liu at Yangzhou University, Jiangsu Province, China. Human A/Anhui/1/2013 (H7N9, AH1) virus was initially isolated from a throat-swab specimen of the third case of laboratory-confirmed human A/H7N9 virus, served as positive control. The viruses were inoculated into 10-day-old embryonated chicken eggs for 48 hr at 37°C, following which the allantoic fluid was collected and tested for HA activity using 0.5% chicken red blood cells. The median tissue culture infective dose (TCID₅₀) of each strain was determined using the Reed and Muench method (Cottey, Rowe, & Bender, 2001), based on at least three independent assays. The virus genome sequences were verified using the Sanger method.

The MDCK cell lines were grown in Dulbecco's modified Eagle medium (Gibco) supplemented with 10% foetal bovine serum (Gibco) at 37°C in 5% CO₂.

2.6 | In vitro infections

MDCK cell monolayers were infected with each virus strain at a multiplicity of infection (MOI) of 0.001 in phosphate-buffered saline

(PBS) containing 0.2% bovine serum albumin (BSA) for 1 hr. Unbound viruses were then washed away with PBS containing 0.2% BSA, and serum-free minimum essential medium, and 0.2% BSA was added. The cells were incubated at 37°C under 5% CO₂, and virus titres in the supernatant were periodically monitored using the TCID₅₀.

2.7 | In vivo infections

The TCID₅₀ values of the H7N9 viruses were measured using groups of BALB/c mice (7 weeks old, female). Mice were intranasally (i.n.) inoculated with 50 µL of 100-fold serial dilutions of each indicated influenza virus (SD183, YZ30, TJ186, BD1 and TZ45) in PBS under isoflurane sedation at 10⁴ TCID₅₀/ml. The survival and body weights of the mice were then recorded daily until 14 days post-infection (d.p.i.). Animals that showed signs of severe disease and weight loss that was >25% of their initial body weight were considered moribund and were humanely killed according to animal ethics guidelines.

Mice in each group were euthanized at 3, 5, and 7 d.p.i. and mouse lungs were excised. The lung index was calculated as (wet lung weight/body weight) × 100%, following which lung tissue samples were homogenized in PBS and the viral titres in the supernatants were determined. Tissue samples from the lungs and turbinate bone were homogenized in PBS with antibiotics in a homogenizer and used to determine the viral titres using the plaque assay. In addition, the lungs were fixed in formalin, sectioned at 4 µm and stained with haematoxylin and eosin for inspection by light microscopy.

2.8 | Statistical analysis

Differences in the body weights and virus titres among different treatment groups were analysed by one-way analysis of variance, while the differences between two groups were analysed using

Student's *t* test. A probability value of $p < .05$ was considered statistically significant.

3 | RESULTS

3.1 | Comparison of different classification algorithms for predicting the hosts of H7N9 strains

At the first stage of this study, we tried to build a classifier which can predict the host of a particular H7N9 strains. The eight cDNA fragments (with non-coding region removed) of each human or avian H7N9 strain were aligned, concatenated, converted to integer codes and labelled, as described in details in the Methods section. Different classification algorithms, including LR, KNN, RF, NB, SVM and MLP were applied to the data set, and the prediction accuracies were used as the evaluation standard. The mean ± standard deviation for the accuracies of the LR, KNN, RF, NB, SVM and MLP are 0.723 ± 0.259, 0.684 ± 0.205, 0.657 ± 0.222, 0.630 ± 0.230, 0.568 ± 0.324 and 0.537 ± 0.329, respectively (Figure 1). These results suggest that LR has the highest performance and stability in differentiating avian and human strains among the tested algorithms. However, the accuracies of these primary algorithms, even the highest one of the LR algorithm, are not satisfying.

3.2 | Building and selecting the optimal CWBLR model which predicts human infectivities of avian strains

To identify H7N9 avian isolates that have the potential to infect humans, we used the novel CWBLR (described in details in the Methods section) to fit a LR model to a data set consisting of 887 H7N9 virus strains that were isolated before 2016. We found that when the weight of the human class was set at 0.83 (and that of the avian class was set at 0.17), both the human and avian classes had the accuracies

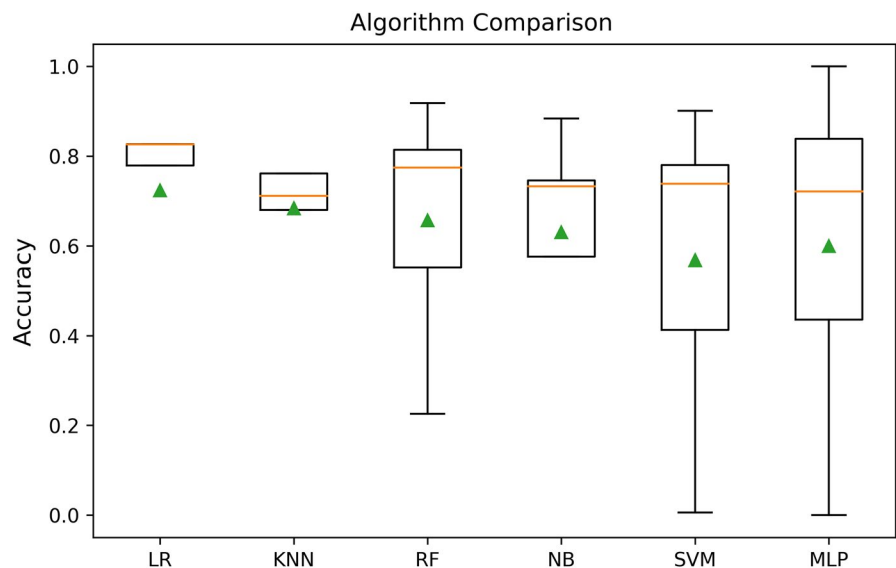


FIGURE 1 Comparison of the accuracies of different classification algorithms (LR, KNN, RF, NB, SVM and MLP) in predicting the hosts of H7N9 strains. The boxplot of accuracies of fivefold cross-validation are shown. The means of the accuracies of the five fitting for each algorithm are shown as green triangles

of 100% (Figure 1). Therefore, the model that was fitted with this class weight was selected as the final CWBLR model and was used to calculate the probability that a given H7N9 strain belonged to the human class (i.e. its human infectivity). An online version of the computational tool is available at <http://124.16.144.116:8099/h7n9/Index/home.do>.

As shown in Figure 2a, the accuracy of this model (with the human class weight of 0.83) for the 2016–2017 human strains was calculated as 100%, while the accuracy for the 2016–2017 avian class was 50%. These results suggest that the final CWBLR model can recognize human strains with 100% accuracy. The 2016–2017 avian strains that had an incorrect prediction using the final CWBLR model were those that had high calculated probabilities (> 0.5) of belonging to the human class, indicating a risk of cross-species transmission and infection of human. In addition, the receiver operating characteristic curve (ROC) plotting shows that the mean ROC has an area under curve (AUC) of 0.99 (Figure 2b), which further confirms the validation of the model.

3.3 | Identification of the critical genome positions for human/avian class determination

Once the model had been fitted, we were able to obtain the coefficients (the parameter w in Equations 2–8 in the Methods section) for every genome position, the absolute values of which reflect their attribution or importance for the human/avian classification. By plotting the coefficients against the amino acid position corresponding to each genome position (Figure 3), we found that residues 627 and 701 had the highest absolute coefficient values and so are the most important positions for human/avian class determination.

Sequence analysis showed that other genome positions may also be relevant to host tropism, despite their low absolute coefficient values. By comparing the sequences of the 48 H7N9 strains with the highest probabilities of belonging to the human class (all

of which are human strains) and the 48 H7N9 strains with the lowest probabilities of belonging to the human class (all of which are avian strains), we found that the residues at position 191 in the PB2 gene were all Glu (E) for the strains with the highest probabilities and Lys (K) for most strains with the lowest probabilities; the residues at position 394 in the PA gene were all Asp (D) for the strains with the highest probabilities and Gln (N) for most strains with the lowest probabilities; the residues at position 256 in the HA protein (position 226 of H3 numbering) were either Leu (L) or Gln (Q) for strains with the highest probabilities and were all L for most strains with the lowest probabilities; and the residues at position 27 in the NS1 gene were either K or L for strains with the highest probabilities and were Met (M) or L for strains with the lowest probabilities. Among these, position 256 in the HA protein (226 of H3 numbering) has been shown to be critical for avian/human receptor specificity, with 226Q preferentially binding to the avian receptor (α -2,3-linked sialic acid) and 226L preferentially binding to the human receptor (α -2,6-linked sialic acid) (Shi, et al., 2013, 2014). Our findings suggest that K191E in PB2, N291D in PA, and K27M in NS1 may be human adaptive mutations.

3.4 | Human infectivity of H7N9 strains according to the CWBLR model

We used the CWBLR model with a human class weight of 0.83 to calculate the human infectivity (i.e. the probability of belonging to the human class) of all of the H7N9 strains that were used to build the model (i.e. the 887 strains isolated before 2016) and 48 strains that were isolated in 2016–2017. We found that the calculated probabilities of belonging to the human class were >0.5 (i.e. 100% human class accuracy) for all of the human strains that were isolated before 2016 and <0.5 (i.e. nearly 100% avian class accuracy) for nearly all of the avian strains that were isolated

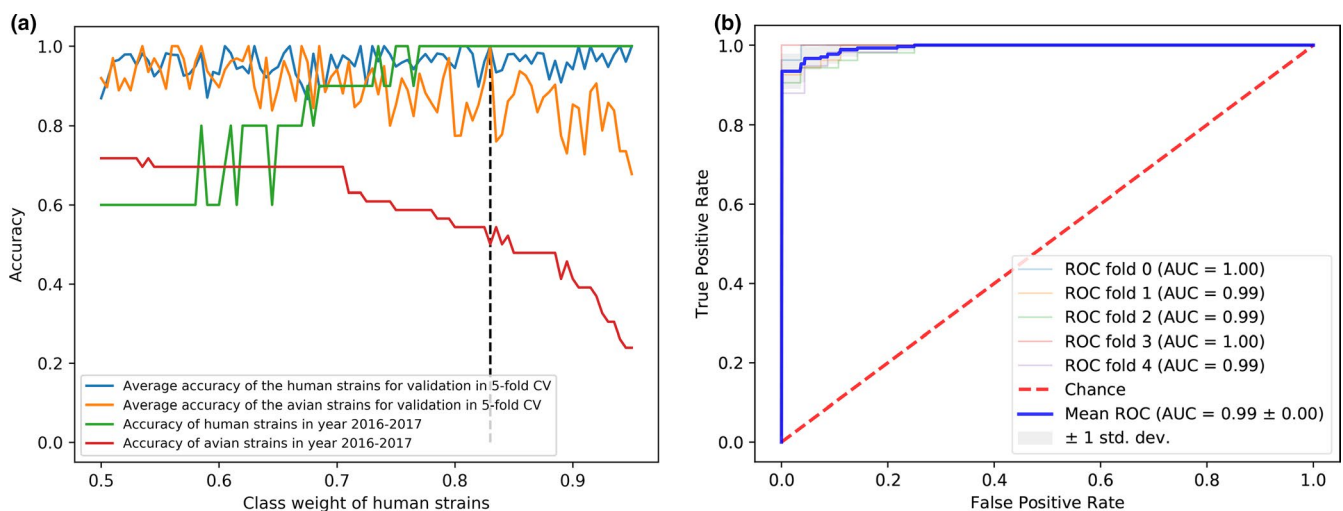


FIGURE 2 Accuracy of the logistic regression models with different human class weights for the avian and human strains. Each model was built using genome sequences of avian H7N9 influenza viruses that were isolated before 2016 using a different human strain class weight. The accuracy of each model for avian and human strains was then calculated using fivefold cross-validation (CV). In addition, the accuracy of each of these models for avian and human strains isolated during 2016–2017 was calculated

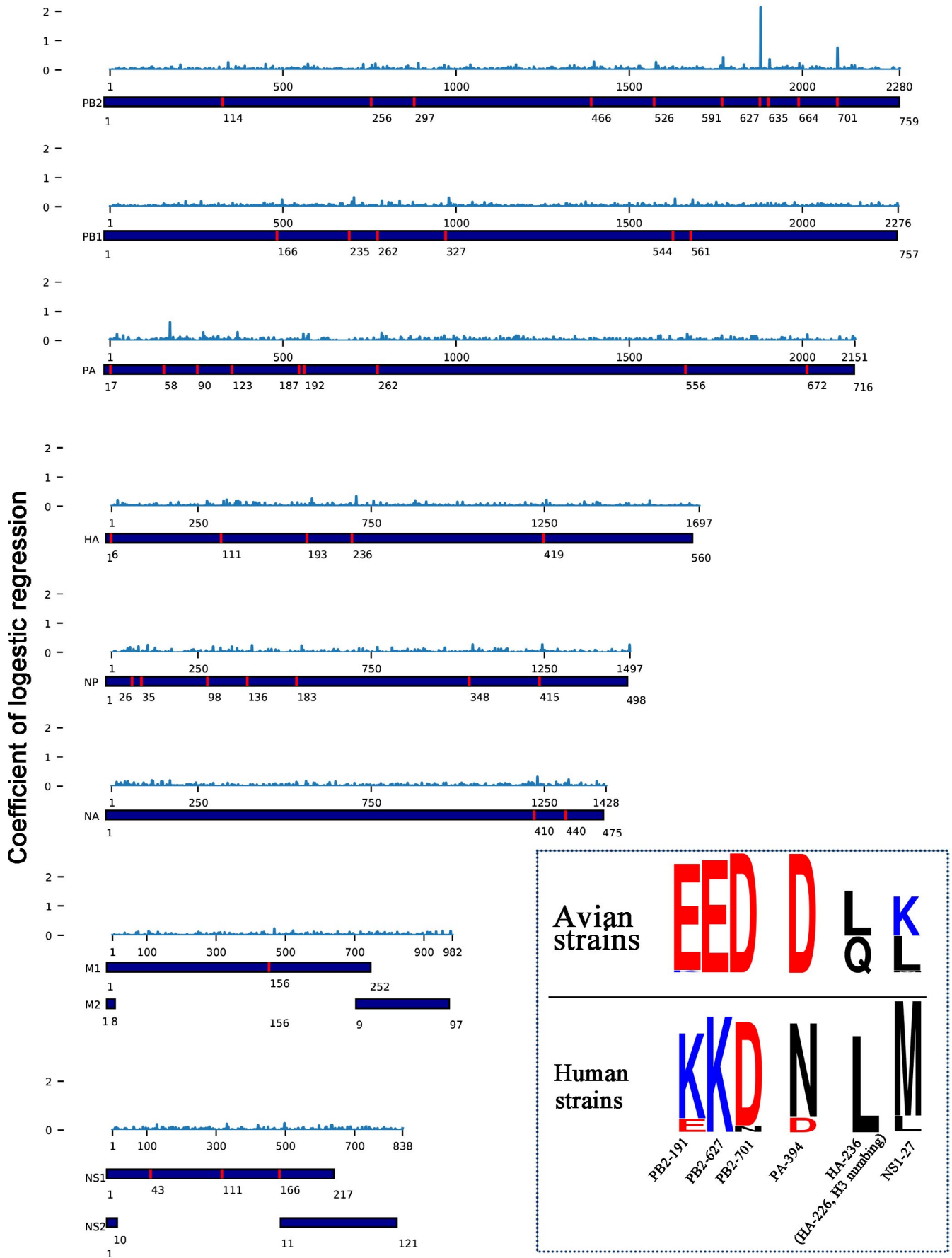


FIGURE 3 Absolute coefficient for each genome nucleotide and the corresponding amino acid position. Amino acid positions with absolute coefficients of >0.2 are indicated by a red perpendicular line

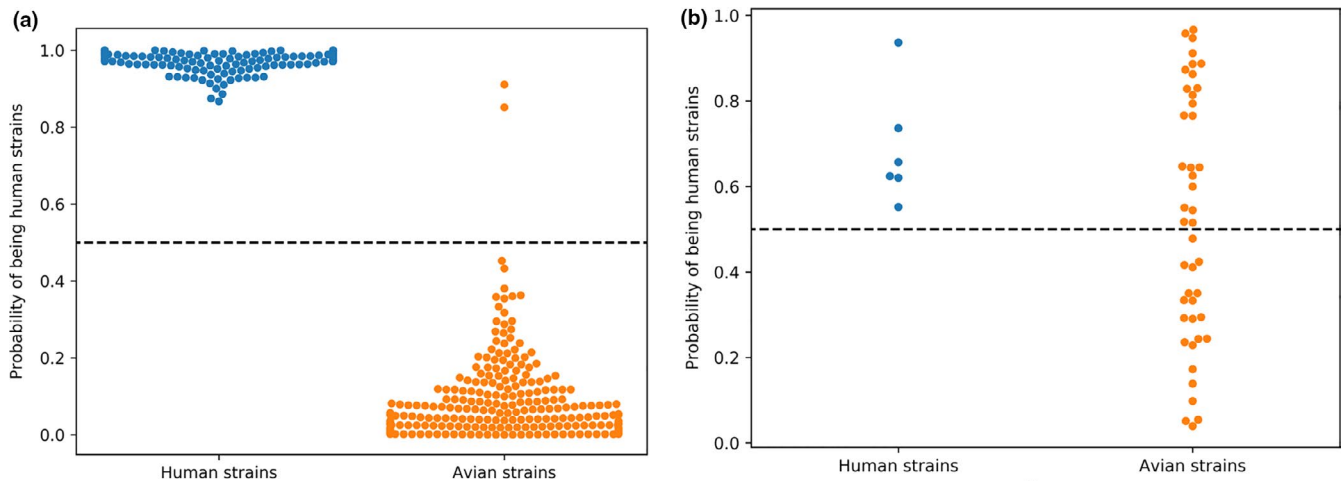


FIGURE 4 Estimated probability of each H7N9 strain belonging to the human class using the final class weight-biased logistic regression (CWBLR) model. (a) Strains isolated before 2016, which were used to build the model. (b) Strains isolated in 2016–2017. Each point represents a different strain

before 2016 (Figure 4a). Similarly, the calculated probabilities of belonging to the human class were also >0.5 (i.e. 100% human class accuracy) for all of the human strains that were isolated in 2016–2017 (Figure 4b and Table 1). However, the probabilities of belonging to the human class were evenly distributed between 0 and 1 for the avian strains that were isolated in 2016–2017. Therefore, we hypothesize that those avian strains that had predicted human infectivity (i.e. probabilities of belonging to the human class) of $> .5$ have genome characteristics that are similar to human strains and may have the potential to infect humans, with a higher probability indicating a higher risk of human infection.

3.5 | Testing the predictions in mammalian cells (in vitro model)

To test the validity of our CWBLR model, we selected five H7N9 avian strains that were predicted to have the lowest and highest human infectivity (i.e. probabilities of belonging to the human class) and compared their replication capacities in cells of avian or mammalian origin. The selected strains with low predicted human infectivity included A/chicken/Jiangsu/YZ30/2017 (YZ30) and A/chicken/Shandong/SD183/2016 (SD183), and the selected strains with high predicted human infectivity included A/chicken/Jiangsu/TZ45/2017 (TZ45), A/chicken/Jiangsu/TJ186/2017 (TJ186) and A/chicken/Jiangsu/0116/2017 (BD1). We also included the human strain A/Anhui/01/2013 (AH1), which had a predicted human infectivity of 0.999, as a control (Figure 5a). The replication efficiencies of these strains were compared in three cell lines: Madin–Darby canine kidney (MDCK; mammalian origin), A549 (human origin), and Douglas Foster-1 (DF1; chicken origin). Interestingly, in the human A549 cells, AH1, BD1, TJ186 and TZ45 exhibited similar replication kinetics, while the viral titres of SD183 and YZ30 were significantly lower than those of the other viruses at 24–48 hr post-infection (Figure 5b). Similarly, in the mammalian MDCK cells, the AH1 and TJ86 viruses had the highest replication efficiencies, while SD183

and YZ30 exhibited the lowest replication efficiencies, particularly at 48–60 hr post-infection (Figure 5c). By contrast, in the avian DF-1 cells, SD183 and YZ30 replicated faster and reached significantly higher levels than the other viruses. These findings suggest that the BD1, TJ186 and TZ45 viruses have higher replication efficiencies in mammalian cells and have the potential to infect humans (Figure 5d), which is consistent with our predictive results.

3.6 | Testing the predictions in mice (in vivo model)

To further verify that H7N9 avian isolates with higher predicted probabilities of belonging to the human class have higher infectivity in humans, we tested the infectivity of the YZ30, SD183, TZ45, TJ186, BD1 and AH1 strains of H7N9 in mice. The average body weights of all six groups of mice that were infected with the six virus strains decreased from day 0 to day 7 (Figure 6a). However, the average body weights of the groups that were infected with the virus strains with low predicted human infectivity (i.e. YZ30 and SD183) gradually recovered after day 8, with no mice dying in these groups, whereas all mice in the groups that were infected with the virus strains with high predicted human infectivity (i.e. TZ45, TJ186 and BD1) and the human strain AH1, died on days 7–9 (Figure 6a,b). In addition, the virus titres in the lungs were significantly higher in mice that had been infected with TZ45, TJ186 and BD1 than in those that had been infected with YZ30 and SD183 on days 3 and 5 (Figure 6c), suggesting that the virus strains with high predicted human infectivity replicated more effectively than those with low predicted human infectivity in mouse lungs. The lung index was also significantly higher in mice that had been infected with TZ45, TJ186 and BD1 than in those that had been infected with YZ30 and SD183 on days 3, 5 and 7 (Figure 6d), suggesting that the virus strains with high predicted human infectivity cause more serious inflammation in mouse lungs.

Histopathological examinations revealed that the lung tissues of mice infected with SD183 and YD30 had a multifocal mild or

TABLE 1 The 2016–2017 avian H7N9 strains' probability of being human strains

No.	Strain name	Probability of being human strain
1	A/chicken/Guangdong/SD1433/2016	0.024834663
2	A/chicken/Shandong/SD183/2016	0.029829032
3	A/chicken/Jiangsu/YZ30/2017	0.036512287
4	A/chicken/Zhejiang/S1074/2016	0.0391336
5	A/chicken/Liaoning/LN1/2016	0.042916638
6	A/chicken/Jiangsu/S1045/2016	0.08222006
7	A/duck/Jiangsu/S1220/2016	0.110399144
8	A/chicken/Guangdong/GD20/2017	0.142770386
9	A/chicken/Jiangsu/S1441/2016	0.193405124
10	A/chicken/Jiangsu/JS11/2016	0.206321077
11	A/chicken/Jilin/SD009/2016	0.236471889
12	A/duck/Zhejiang/S1375/2016	0.241795902
13	A/chicken/Zhejiang/JH16/2017	0.246470525
14	A/chicken/Guangdong/Q1/2016	0.257789959
15	A/chicken/Heilongjiang/BQC01/2017	0.259065607
16	A/chicken/Guangdong/Q26/2017	0.268599063
17	A/chicken/Hunan/S12753/2016	0.293752638
18	A/chicken/Guangdong/30/2017	0.303606044
19	A/chicken/Heinan/ZZ01/2017	0.31582883
20	A/chicken/Jiangsu/S1460/2016	0.371993123
21	A/chicken/Guangdong/J1/2017	0.378423664
22	A/chicken/Guangdong/SD010/2017	0.382724485
23	A/chicken/Zhejiang/SD001/2016	0.406821438
24	A/chicken/Guangdong/J2/2017	0.464087989
25	A/chicken/Guangdong/GD15/2016	0.481478747
26	A/chicken/Longquan/LQ78/2016	0.494869684
27	A/chicken/Ganzhou/GZ79/2016	0.508499539
28	A/chicken/Shandong/SD216/2016	0.535164193
29	A/duck/Jiangsu/S1700/2016	0.587207668
30	A/chicken/Guangdong/SD031/2017	0.608101557
31	A/chicken/Guangdong/SD008/2017	0.608244892
32	A/chicken/Guangdong/SD032/2017	0.608834928
33	A/chicken/Zhejiang/ZJ19/2017	0.720285739

(Continues)

TABLE 1 (Continued)

No.	Strain name	Probability of being human strain
34	A/chicken/Guangdong/Q39/2017	0.732032151
35	A/chicken/Guangdong/SD028/2017	0.738615034
36	A/chicken/Jiangsu/LY246/2017	0.764368019
37	A/chicken/Zhejiang/ZJ14/2017	0.765102218
38	A/chicken/Jiangsu/JT156/2016	0.793610445
39	A/chicken/Guangdong/SD027/2017	0.801158389
40	A/chicken/Hebei/HB13/2016	0.833050936
41	A/chicken/Jiangsu/JT164/2017	0.840450603
42	A/chicken/Hebei/S1257/2016	0.845365849
44	A/chicken/Zhejiang/1.10 HZ142/2017	0.874358
43	A/chicken/Guangdong/SD034/2017	0.902831284
45	A/chicken/Liaoning/05.12 SY059/2017	0.905891
46	A/chicken/Jiangsu/0116/2017	0.930523603
47	A/chicken/Jiangsu/JT186/2017	0.934247297
48	A/chicken/Jiangsu/TZ45/2017	0.947953526

moderate inflammation and consolidation. By contrast, the lung tissues of mice infected with BD1 and YZ40 showed extensive consolidation and caseous necrosis, and the lung tissues of mice infected with TJ186 also showed serious consolidation, as shown in Figure 7. Therefore, the test results in mice are highly consistent with the predicted results from our CWBLR model.

4 | DISCUSSION

In this paper, we have introduced a novel computational method that can be used to predict the human infectivity of avian H7N9 isolates. Our CWBLR model can use information from the viral genome to recognize H7N9 human isolates with 100% accuracy, allowing it to distinguish H7N9 avian isolates with high potential for infecting humans from those with low potential.

We use viral nucleotide sequences rather than protein sequences to build the model for the following reasons. First of all, nucleotide sequences contain more potential information than protein sequences. The genome of influenza viruses encodes at least 12 proteins (PB2, PB1, PB1-F2, PA, PA-X, HA, NP, NA, M1, M2, NS1, NS2). It is quite possible that there are unknown proteins encoded in the influenza virus genome. That means if we use protein sequences rather than nucleotide sequence as the input for building the model, the information in the nucleotide sequences encoding the unknown proteins would be missed. Secondly, some proteins such as PB1-F2 are not encoded by all influenza virus strains, so using protein sequence to build the model will

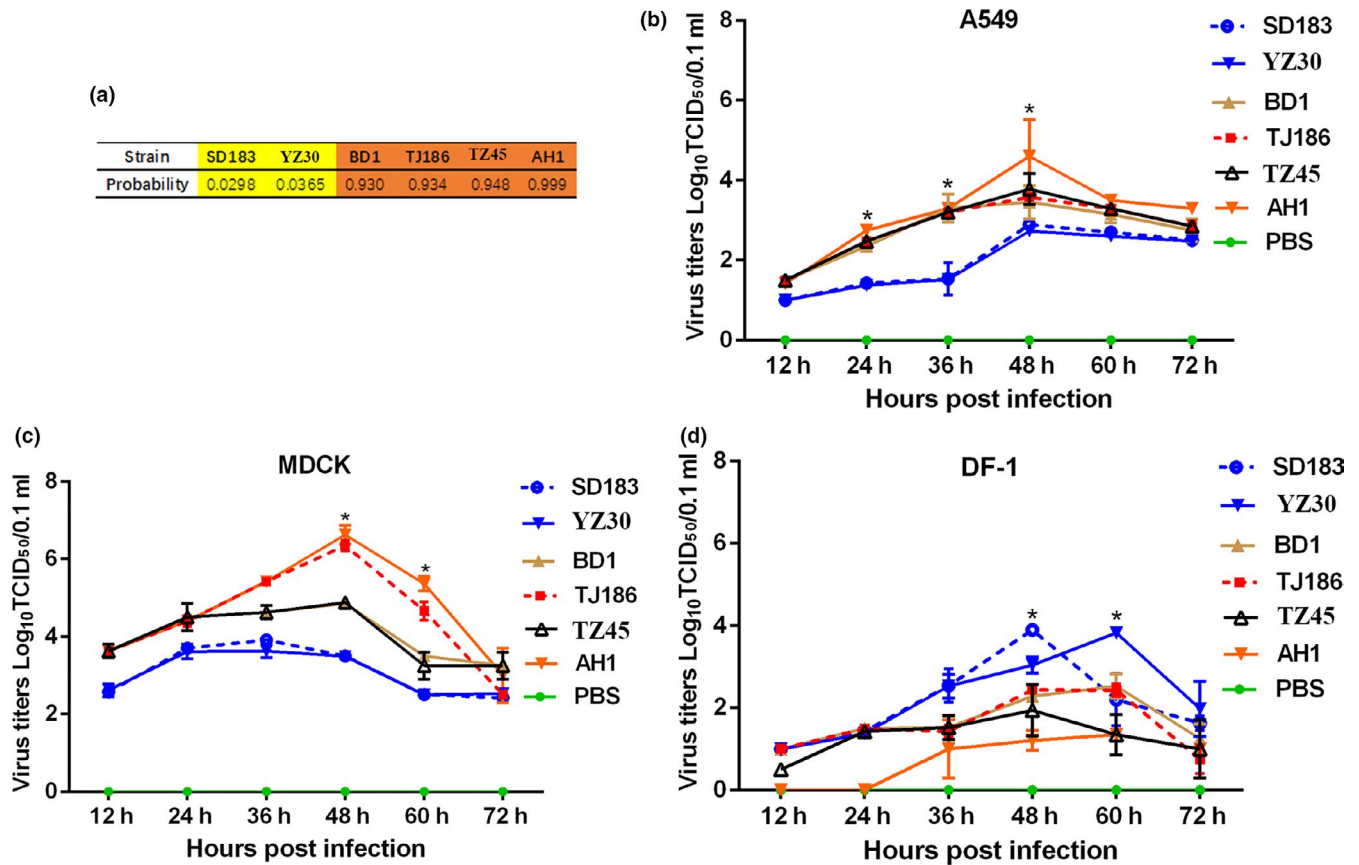


FIGURE 5 Growth curves for the six H7N9 virus strains in various cell lines. (a) Probability that each of the six strains belongs to the human class according to the final class weight-biased logistic regression (CWBLR) model. (b) Infection of the human lung cell line A415 by the six strains. (c) Infection of the Madin–Darby canine kidney (MDCK) by the six strains. (d) Infection of the Douglas Foster-1 (DF-1) cell lines by the six strains. The cells were infected with PBS or the H7N9 strains TZ45, TJ186, BD1, YZ30, SD183 or AH1 at a multiplicity of infection (MOI) of 0.001. The virus titres were monitored at the indicated time points. Statistically significant differences are indicated, * $p < .05$ (Student's *t* test)

increase complexity of the operation. Thirdly, there are only four types of nucleotides (A, T, C and G) while 20 amino acids (A, R, N, D, C, Q, E, G, H, I, L, K, M, S, P, F, T, W, Y and V), so using amino acid coding would increase the computational costs. Actually, the nucleotide positions important to cross-species transmission identified in the model can be conveniently translated into amino acid positions. Using our model, we calculated the human infectivity of a number of H7N9 avian isolates and tested their predicted human infectivity through both in vitro and in vivo experiments. The experimental results were very consistent with our predictive model, demonstrating its reliability. Therefore, we believe that use of the same model building procedure would allow prediction of the human infectivity of other avian influenza viruses, such as H5N1.

In addition to predicting the human infectivity of H7N9 strains based on their genome, our model can quantitatively identify the genome positions that are important to human/avian class determination or host tropism based on the coefficient corresponding to each genome position in the model. Our model explicitly revealed that the genome positions corresponding to PB2 protein residues 627 and 701 are the most important to host tropism, which is consistent with the findings of previous experimental studies that E627K and D701N in PB2 are critical mutations associated with the mammalian

adaptation of avian influenza viruses (Russell & Webster, 2005; Steel, Lowen, Mubareka, & Palese, 2009; Weber, et al., 2015).

Interestingly, all of the H7N9 strains we experimentally tested have the same avian-like residues at positions 627 and 701 in PB2 (627E and 701D), regardless of whether they had high predicted human infectivity (BD1, TJ186 and TZ45) or low predicted human infectivity (SD183 and YZ30). However, these strains do have many differences in their encoding proteins (Table S1), the collective effects of which likely produce the genetic basis for the differences in their infectivity. Therefore, our results suggest that avian influenza viruses may have high infectivity in mammals even if they do not possess the widely accepted mammal-adaptive mutations.

Virologists have long sought the ability to predict the emergence of high risk of influenza viruses that pose a threat to both farm animals and humans. Long-term prospective surveillance has revealed the coincidence of activity of some influenza virus subtypes (such as H7) in wild aquatic birds, poultry and humans and can therefore provide useful, predictive, early-warning information (Krauss, et al., 2007; Krauss & Webster, 2012). Some influenza virus risk assessment frameworks have been proposed to risk of avian influenza virus cross-species transmission and causing a pandemic, and the genetic composition

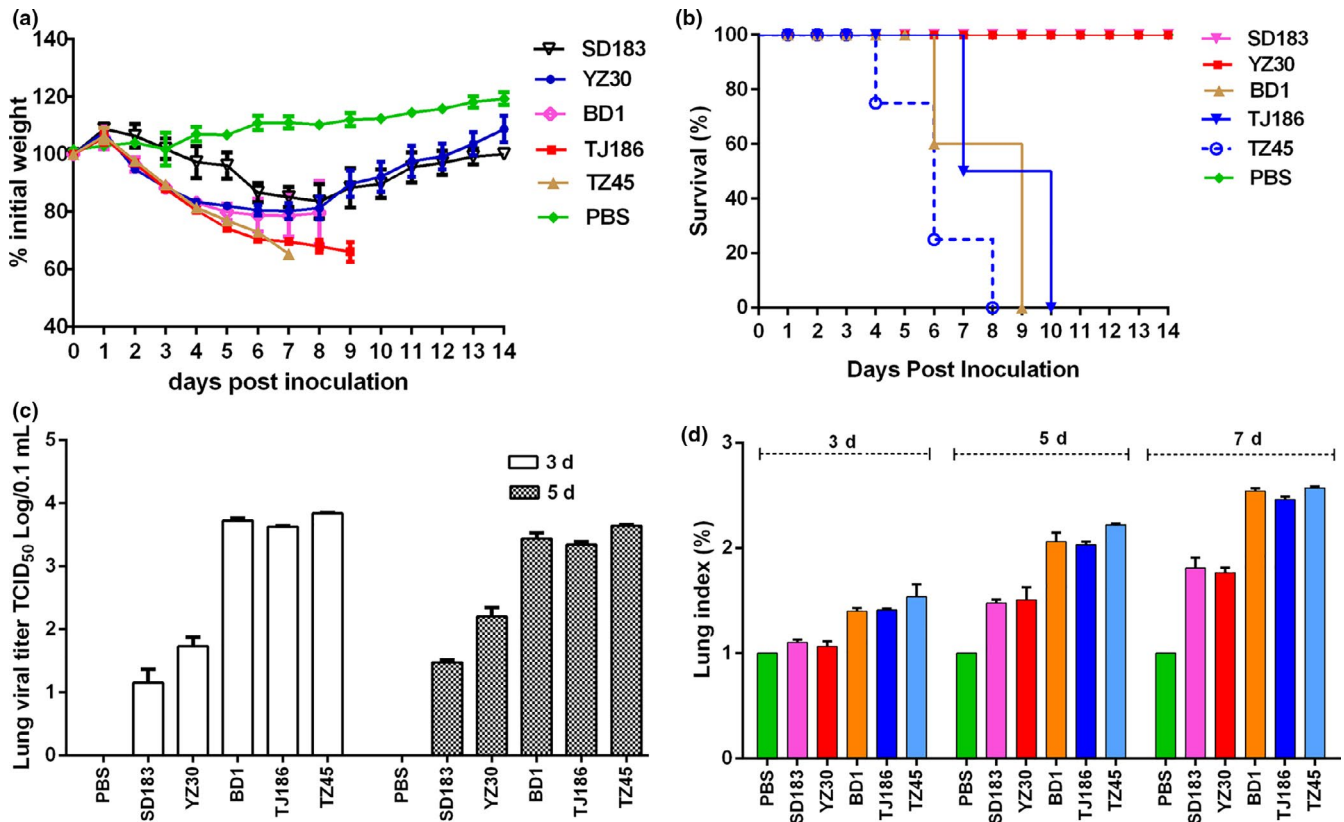


FIGURE 6 Infectivity of the different H7N9 virus strains in mice. Mice were infected with PBS or $50 \mu\text{l } 10^4 \times$ the median tissue culture infective dose (TCID_{50}) per ml of TZ45, TJ186, BD1, YZ30, SD183 or AH1 on day 0, and their survival and body weights were recorded daily until 14 days post-infection. (a) Changes in the body weights of the mice as a percentage of their initial body weights on day 0. (b) Percentage survival each day after infection. Data are shown as mean \pm SD ($n = 8$ mice in each group). (c) Virus titres in the lungs of mice on days 3 and 5 after infection. (d) Lung index of the mice on days 3, 5 and 7. Data are shown as mean \pm SD ($n = 4$ mice in each group)

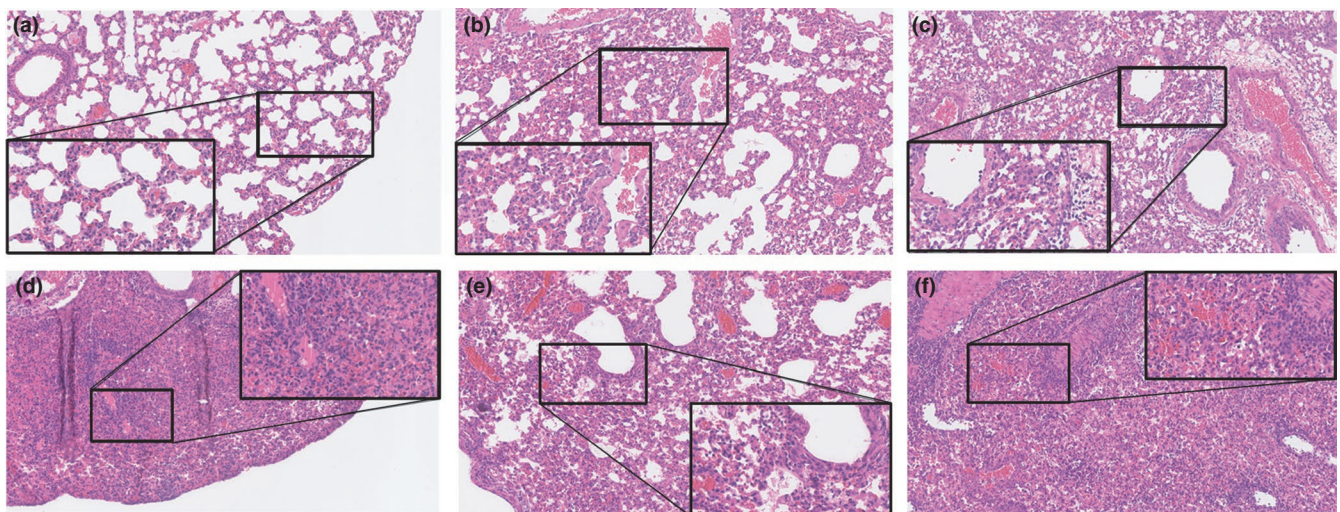


FIGURE 7 Histopathology of the lungs of mice infected with different H7N9 viruses. Mice were infected with PBS or one of the five H7N9 virus strains, and sections of the lungs were subsequently stained with haematoxylin and eosin (H-E). (a) PBS, (b) SD183, (c) YZ30, (d) BD1, (e) TJ186 and (f) TZ45. The main images are $\times 10$ magnification and the insets are $\times 40$ magnification

of influenza viruses is invariably considered to be the top risk factor included in these frameworks (Neumann & Kawaoka, 2019; Trock, Burke, & Cox, 2015). Our viral genome-based computational tool can

directly and conveniently predict the risk of H7N9 strains infecting humans and will therefore be useful for strengthening the influenza virus risk assessment frameworks that are currently used.

CONFLICT OF INTEREST

The authors declare no competing financial interests.

AUTHOR CONTRIBUTIONS

Y.S., J.L. and W.L. conceived and designed the research. Y.S. designed and performed the computational scheme. J.L., H.Z. and S.Z. performed the viral replication ability tests, histopathology and immunology analyses. J.L., Y.S. and K.Z. performed other experimental data analysis and wrote the manuscript. J.L. and Y.S. prepared the manuscript and completed its revision. Y.B., Y.S., J.L., J.Q., D.L. and L.S. suggested many of the experiments in this study. H.Q., L.W. and J.M. prepared the online version of the computational tool. All authors read and approved the final manuscript.

ETHICAL APPROVAL

All experiments using H7N9 subtype strains were conducted in the BSL-3 level laboratory approved by the Wuhan Institute of Virology, Chinese Academy of Sciences. Animal care and housing were in compliance with ethical guidelines and approved by the Experimental Animal Ethic and Welfare Committee of the Institute of Microbiology, Chinese Academy of Sciences.

DATA AVAILABILITY STATEMENT

All data generated or analysed during this study are included in this published article.

ORCID

Yuhai Bi  <https://orcid.org/0000-0002-5595-363X>

Jing Li  <https://orcid.org/0000-0001-9588-5291>

REFERENCES

- Chen, G.-W., Chang, S.-C., Mok, C.-K., Lo, Y.-L., Kung, Y.-N., Huang, J.-H., ... Shih, S.-R. (2006). Genomic signatures of human versus avian influenza A viruses. *Emerging Infectious Diseases*, 12(9), 1353–1360. <https://doi.org/10.3201/eid1209.060276>
- Cotter, R., Rowe, C. A., & Bender, B. S. (2001). *Influenza virus*. *Curr Protoc Immunol* 2001;Chapter 19: Unit 19 11.
- Gao, H.-N., Lu, H.-Z., Cao, B., Du, B., Shang, H., Gan, J.-H., ... Li, L.-J. (2013). Clinical findings in 111 cases of influenza A (H7N9) virus infection. *New England Journal of Medicine*, 368(24), 2277–2285. <https://doi.org/10.1056/NEJMoa1305584>
- Katoh, K., & Standley, D. M. (2013). MAFFT multiple sequence alignment software version 7: Improvements in performance and usability. *Molecular Biology and Evolution*, 30(4), 772–780. <https://doi.org/10.1093/molbev/mst010>
- Krauss, S., Obert, C. A., Franks, J., Walker, D., Jones, K., Seiler, P., ... Webster, R. G. (2007). Influenza in migratory birds and evidence of limited intercontinental virus exchange. *PLoS Path*, 3(11), e167. <https://doi.org/10.1371/journal.ppat.0030167>
- Krauss, S., & Webster, R. G. (2012). Predicting the next influenza virus. *Science*, 337(6095), 644. <https://doi.org/10.1126/science.337.6095.644-a>
- Li, H. (2012). *Statistical learning methods*. Beijing, China: Tsinghua University Press.
- Long, J. S., Mistry, B., Haslam, S. M., & Barclay, W. S. (2019). Host and viral determinants of influenza A virus species specificity. *Nature Reviews Microbiology*, 17(2), 67–81. <https://doi.org/10.1038/s41579-018-0115-z>

- Nations, F.a.A.O.o.t.U (2019). *H7N9 situation update*. In.
- Neumann, G., & Kawaoka, Y. (2019). Predicting the next influenza pandemic. *Journal of Infectious Diseases*, 219(Supplement_1), S14–S20. <https://doi.org/10.1093/infdis/jiz040>
- Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... Vanderplas, J. (2011). Scikit-learn: Machine learning in python. *The Journal of Machine Learning Research*, 12, 2825–2830.
- Petrie, J. G., & Luring, A. S. (2019). Influenza A(H7N9) virus evolution: Which genetic mutations are antigenically important? *Journal of Infectious Diseases*, 219(1), 3–5.
- Qi, X., An, X., Jiao, Y., Yu, H., Xu, K. E., Cui, L., ... Bao, C. (2018). Co-circulation of multiple genotypes of influenza A (H7N9) viruses in eastern China, 2016–2017. *Archives of Virology*, 163(7), 1779–1793. <https://doi.org/10.1007/s00705-018-3800-3>
- Russell, C. J., & Webster, R. G. (2005). The genesis of a pandemic influenza virus. *Cell*, 123(3), 368–371. <https://doi.org/10.1016/j.cell.2005.10.019>
- Schrauwen, E. J., & Fouchier, R. A. (2014). Host adaptation and transmission of influenza A viruses in mammals. *Emerging Microbes & Infections*, 3(2), e9. <https://doi.org/10.1038/emi.2014.9>
- Shi, Y. I., Wu, Y., Zhang, W., Qi, J., & Gao, G. F. (2014). Enabling the 'host jump': Structural determinants of receptor-binding specificity in influenza A viruses. *Nature Reviews Microbiology*, 12(12), 822–831. <https://doi.org/10.1038/nrmicro3362>
- Shi, Y., Zhang, W., Wang, F., Qi, J., Wu, Y., Song, H., ... Gao, G. F. (2013). Structures and receptor binding of hemagglutinins from human-infecting H7N9 influenza viruses. *Science*, 342(6155), 243–247. <https://doi.org/10.1126/science.1242917>
- Steel, J., Lowen, A. C., Mubareka, S., & Palese, P. (2009). Transmission of influenza virus in a mammalian host is increased by PB2 amino acids 627K or 627E/701N. *PLoS Path*, 5(1), e1000252. <https://doi.org/10.1371/journal.ppat.1000252>
- Su, S., Gu, M., Liu, D. I., Cui, J., Gao, G. F., Zhou, J., & Liu, X. (2017). Epidemiology, evolution, and pathogenesis of H7N9 influenza viruses in five epidemic waves since 2013 in China. *Trends in Microbiology*, 25(9), 713–728. <https://doi.org/10.1016/j.tim.2017.06.008>
- Trock, S. C., Burke, S. A., & Cox, N. J. (2015). Development of framework for assessing influenza virus pandemic risk. *Emerging Infectious Diseases*, 21(8), 1372–1378. <https://doi.org/10.3201/eid2108.141086>
- Weber, M., Sediri, H., Felgenhauer, U., Binzen, I., Bänfer, S., Jacob, R., ... Weber, F. (2015). Influenza virus adaptation PB2-627K modulates nucleocapsid inhibition by the pathogen sensor RIG-I. *Cell Host & Microbe*, 17(3), 309–319. <https://doi.org/10.1016/j.chom.2015.01.005>
- Zhang, X., Luo, T., & Shen, Y. (2018). Deciphering the sharp decrease in H7N9 human infections. *Trends in Microbiology*, 26(12), 971–973. <https://doi.org/10.1016/j.tim.2018.10.002>
- Zhang, Y., Aevermann, B. D., Anderson, T. K., Burke, D. F., Dauphin, G., Gu, Z., ... Scheuermann, R. H. (2017). Influenza Research Database: An integrated bioinformatics resource for influenza virus research. *Nucleic Acids Research*, 45(D1), D466–D474. <https://doi.org/10.1093/nar/gkw857>

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section.

How to cite this article: Sun Y, Zhang K, Qi H, et al.

Computational predicting of the human infectivity of H7N9 influenza viruses isolated from avian hosts. *Transbound Emerg Dis*. 2021;68:846–856. <https://doi.org/10.1111/tbed.13750>