Journal Pre-proof

Variations outside the conserved motifs of PB1 catalytic active site may affect replication efficiency of the RNP complex of influenza A virus

Kaitlyn Waters, Hamilton J. Wan, Lei Han, Jianli Xue, Matthew Ykema, Yizhi J. Tao, Xiu-Feng Wan

PII: S0042-6822(21)00092-1

DOI: https://doi.org/10.1016/j.virol.2021.04.001

Reference: YVIRO 9472

To appear in: *Virology*

Received Date: 30 January 2021

Revised Date: 31 March 2021

Accepted Date: 5 April 2021

Please cite this article as: Waters, K., Wan, H.J., Han, L., Xue, J., Ykema, M., Tao, Y.J., Wan, X.-F., Variations outside the conserved motifs of PB1 catalytic active site may affect replication efficiency of the RNP complex of influenza A virus, *Virology* (2021), doi: https://doi.org/10.1016/j.virol.2021.04.001.

CRediT author statement

Kaitlyn Waters: Data Curation, Data analyses, Original draft preparation; **Hamilton J. Wan**: Methodology, Software and Data analyses, Original draft preparation; **Lei Han**: Methodology; **Jianli Xue**: Methodology; **Matthew Ykema**: Data analyses; **Yizhi J. Tao**: Reviewing and Editing; **Xiu-Feng Wan**: Conceptualization, Original draft preparation, Reviewing, and Editing

Title: Variations outside the conserved motifs of PB1 catalytic active site may affect replication efficiency of the RNP complex of influenza A virus

Running title: PB1 and RNP replication of IAV

Kaitlyn Waters[a,b,c,d], Hamilton J. Wan[a,b,c], Lei Han[d], Jianli Xue[d], Matthew Ykema[e], Yizhi J. Tao[e], and Xiu-Feng Wan[a,b c,d,f*]

[a]Missouri University Center for Research on Influenza Systems Biology (CRISB), University of Missouri, Columbia, MO, USA; [b]Department of Molecular Microbiology and Immunology, School of Medicine, University of Missouri, Columbia, MO 65212; [c]Bond Life Sciences Center, University of Missouri, Columbia, MO, USA; [d]Department of Basic Sciences, College of Veterinary Medicine, Mississippi State University, Mississippi State, Mississippi 39762, USA; [e]Department of BioSciences, Rice University, Houston, TX 77251; [f]Department of Electrical Engineering & Computer Science, College of Engineering, University of Missouri, Columbia, MO, USA.

Corresponding author: *Address Correspondence to: Dr. Xiu-Feng Wan, Phone: (573):882-8943. E-mail: wanx@missouri.edu.

Word Count for the Abstract: 149
Word Count for the Main Text: 6,332
Number of Tables: 2
Number of Figures: 4
Supplementary Tables: 3
Supplementary Figures: 2

Abstract

PB1 functions as the catalytic subunit of influenza virus RNA polymerase complex and plays an essential role in viral RNA transcription and replication. To determine plasticity in the PB1 enzymatic site and map catalytically important residues, 658 mutants were constructed, each with one to seven mutations in the enzymatic site of PB1. The polymerase activities of these mutants were quantified using a minigenome assay, and polymerase activity-associated residues were identified using sparse learning. Results showed that polymerase activities are affected by the residues not only within the conserved motifs, but also across the inter-motif regions of PB1, and the latter are primarily located at the base of the palm domain, a region that is conserved in avian PB1 but with high sequence diversity in swine PB1. Our results suggest that mutations outside the PB1 conserved motifs may affect RNA replication and could be associated with influenza virus host adaptation.

Key words: polymerase basic 1, polymerase activity, minigenome assay, enzymatic active site, machine learning.

Abbreviations: IAV, influenza A virus; PB1, polymerase basic 1 protein; PB2, polymerase basic 2 protein; PA, polymerase acidic protein; pdm09, A/California/07/2009 (H1N1); mem94, A/Memphis/7/1994 (H3N2); sH4N6, A/swine/Missouri/A01727926/2015 (H4N6); PR8, A/Puerto Rico/8/1934 (H1N1); ep-PCR, error-prone PCR; RdRp, RNA-dependent RNA polymerase; RPER, relative protein expression ratio.

Introduction

Influenza A viruses (IAVs) are segmented, negative-sense, single-stranded RNA viruses belonging to the *Orthomyxoviridae* family that are responsible for infecting humans, avian, and various mammalian species, such as swine, equine, canine, and marine mammals [1-3]. IAVs continue to be a major public health concern because of the virus' ability to cause both seasonal and pandemic outbreaks in the human population. Annually, seasonal influenza results in 290,000 to 650,000 deaths worldwide [4, 5]. Four influenza pandemic strains have been documented, and at least three of those strains contain polymerase proteins of avian origin [6, 7], indicating polymerase proteins play an important role in influenza host adaptation and the emergence of pandemic viruses.

IAVs encode a heterotrimeric RNA-dependent RNA polymerase (RdRp) made up of polymerase basic 2 (PB2), polymerase basic 1 (PB1), and polymerase acidic (PA) proteins. In association with nucleoprotein (NP) and viral RNA, the three polymerase proteins form the ribonucleoprotein (RNP) complex required for the transcription of virally encoded mRNAs as well as the replication of the eight negative-sense, single stranded RNAs known to make up the viral genome [8-11]. Transcription is a primer-dependent process in which 5'- capped oligonucleotides are cleaved from host cell pre-mRNAs as primers for elongation [10, 11]. Replication is a primer-independent process [8] involving two steps in which viral RNAs (vRNAs) first act as a template for the polymerase to generate full-length complementary RNAs (cRNAs), which are then further replicated into nascent vRNAs.

Although each of the three polymerase proteins are required for efficient transcription and replication of the influenza viral genome [12], PB1 acts as the backbone and functions as the catalytic subunit of the polymerase complex. It contains the finger, palm, and thumb subdomains, along with several conserved motifs that are characteristic of all viral RdRps [13, 14]. During catalysis, promoter sequences of vRNAs or cRNAs, in the form of a partially complementary panhandle consisting of both 5' and 3'-terminal sequences, first bind to PB1 [15] as an allosteric regulator [16]. The influenza RdRp acquires primers for viral mRNA transcription via "cap-snatching" [11] in close association with the host RNA polymerase II C-terminal domain [17-19]. During cap-snatching, the endonuclease domain of PA

first cleaves a capped RNA of 10-13 nucleotides long from the host pre-mRNA [20, 21] bound to the cap-binding domain of PB2 [22, 23], and the PB2 cap-binding domain then rotates to position the capped primer into the catalytic active site of PB1 where it undergoes elongation using vRNA as a template [24].

Atomic structures have been reported for several influenza virus polymerase complexes [16, 24-30], providing models of the three polymerase proteins interacting with each other as well as in association with other viral and host factors. The catalytic core of PB1 is made up of highly conserved RdRp motifs known as pre-A/F and A-E [13, 16, 31, 32]. The following residues are located across the PB1 motifs: 229-257 (pre-A/F), 296-314 (motif A), 401-422 (motif B), 436-449 (motif C), 474-486 (motif D), and 487-497 (motif E). Analysis of mutations occurring within those reported motifs of PB1 proteins illustrates an overall reduction of polymerase activity [33]. One exception is the mutation K480R which increases polymerase activity, but does not enhance virulence [33]. However, several other studies have established that amino acid changes in polymerase proteins can not only increase viral polymerase activity, but also contribute to enhanced virulence [34-37]. While investigation into the conserved regions within the catalytic active site of PB1 have been explored, the impact of mutations outside the reported motifs on polymerase activities of influenza viruses remains opaque.

In this study, we mutated residues at the PB1 enzymatic site and attempted to identify the residues affecting polymerase activities to better understand the role of PB1 mutations on viral RNA replication and virus host adaptation.

Results

Human, avian, and swine H1N1 IAVs exhibit sequence diversity at the PB1 enzymatic site. The IAV PB1 enzymatic site is shaped like a right-hand comprising finger, palm, and thumb domains with the catalytic center for template-directed nucleotide addition within the protein formed by highly conserved RdRp motifs known as pre–A/F and A-E (Fig. 1A). To identify potential residues in the PB1 enzymatic site that may be important for host adaptation, we compared PB1 protein sequences of 11,965 H1N1

IAVs from human, swine, or avian origin. Our analyses focused on residues 393 to 490, a region encompassing motifs B, C, D and a portion of motif E within the finger and palm domains (Fig. 1B). Results showed that both avian and human PB1 harbor variations at seven residues: avian at 400, 430, 433, 435, 455, 473, and 486 (motif D) and human at 397, 430, 435, 456, 457, 473, and 486 (motif D). Swine PB1 were the least conserved (Supplementary Fig. 1), and variations were located across eighteen residues, including 393, 397, 398, 400, 429, 430, 433, 435, 451, 454, 455, 456, 457, 464, 469, 473, 480 (motif D), and 486 (motif D). Among the residues, residue 486 in motif D was observed to vary across and within human, avian, and swine PB1. The level of sequence diversity varies significantly depending on both the PB1 position and the host origin of the virus (Table 1).

Taken together, these results suggest that variations are located both inside and outside of the PB1 conserved motifs, and the palm region in swine PB1 is the least conserved.

PB1 mutant library generated by error-prone PCR (ep-PCR) based mutagenesis strategy. To identify polymerase activity-associated residues, we used the PB1 of A/California/07/2009 (H1N1) (abbreviated as pdm09) as the template to construct a mutant library, focusing on residues 393-490, by using an error prone-PCR based mutagenesis method [38]. The pdm09 is a swine-origin IAV, with PB1 from human seasonal subtype H3N2 IAV [39]. In total, 1,385 clones were generated, among which 658 did not have stop codons. Out of the 658 clones without stop codons, 480 are unique mutants with 1 to 7 mutations in the target region, including 200 plasmids with a single mutation, 168 with double, 71 with triple, 25 with quadruple, 13 with quintuple, 1 with six, and 2 with seven mutations (Fig. 2A). Further analysis of the 480 unique mutants indicated that 169 have mutations located only at the conserved motifs (105 with a single motif and 64 across multiple motifs), 139 contain mutations only at inter-motif regions (95 at a single region and 44 across multiple regions), and 172 have mutations across at least one inter-motif and one motif region.

Polymerase complex activity of wild-type and PB1 mutant library. To determine plasticity in the PB1 enzymatic site, we determined polymerase activities of each PB1 mutant using a minigenome assay, in which cells were transfected with four plasmids expressing PB2, PB1, PA, and NP, each with a human RNA pol I promoter and a human cytomegalovirus (HCMV) pol II promoter [40], a plasmid expressing *Renilla* luciferase with a human RNA pol I promoter, and another plasmid expressing *firefly* luciferase but with a simian virus 40 (SV40) pol II promoter. Through the HCMV pol II promoter, mRNAs for the four viral RNP genes were first synthesized by host polymerase and translated to produce the four RNP proteins, which then functions to package, replicate and transcribe the negative-sense RNAs of the viral PB2, PB1, PA, NP genes and the Renilla luciferase reporter produced from the human RNA pol I promoter [41-43]. The SV40 pol II promoter ensures that the *firefly luciferase* activity is transcribed and translated by host machinery independent of the viral RNP. With the same amount of plasmids and host cells used in the transfection, a higher *Renilla* luciferase activity would indicate a more efficient polymerase activities of the testing RNP complex whereas the *firefly* luciferase activity is expected to be constant. To make the analyses quantifiable, the luciferase activity (L) for a testing RNP complex is calculated by *Renilla/firefly*, and the relative polymerase activity (R) for a mutant by $L_{mutantPB1}/L_{WT\text{-}PB1}$.

Our results show extensive diversity in the replication efficiencies among the 480 PB1 mutants. We categorized these mutants into four groups: high ($R \geq 2.00$; n = 35), moderate ($2.00 > R \geq 0.50$; n = 99), low ($0.50 > R \geq 0.10$; n = 50), or abolished polymerase activity ($R < 0.10$; n = 296) (Fig. 2B). Further analyses showed that the group of the mutants with only mutation(s) at inter-motif regions are more likely to be associated with increasing polymerase activities. For example, 27 out of 35 (77.14%) mutants in the group of high polymerase activities have only mutations at inter-motif regions. However, there is a moderate level of plasticity in harboring mutations within conserved motif regions. For example, among those 35 mutants in the group of high polymerase activities, 8 have mutations in conserved motifs (1 in motif B, 4 in C, and another 3 in D). On the other hand, among the 296 mutants in the group of abolished polymerase, only 50 have only mutations at the inter-motif regions.

To confirm that variations in luciferase activities correlate with the PB1 mutant specific polymerase activities, we examined the NP protein expression using minigenome assays. We expect that an inefficient PB1 would result in an overall inefficient RNP complex, and ultimately in a low level of all RNP proteins, including NP protein. Specifically, four mutants expressing moderate (Y436F, R = 1.449; H456Q, 1.718; S478C, 0.4232; I459T, 0.419) and low and abolished polymerase activities (I475F, 0.0045; M409L, 0.012), were selected for western blot analyses (Fig. 2C). The NP protein expression was first normalized by β-actin, and then we calculated the relative protein expression ratio (RPER) by the ratio of the normalized NP value for a mutant PB1 over that of the WT PB1. Compared with the WT NP (RPER = 1.0), mutants Y436F and H456Q had similar NP protein expression, with a relative protein expression ratio of 1.04 and 0.92, respectively. Mutants S478C (a relative protein expression ratio, 0.48) and I459T (0.29) exhibited significantly decreased NP protein expression. Mutants I475F (0.04) and M409L (0.05) exhibited little to no detectable NP protein expression (Fig. 2D). The negative control group had no detectable protein expression. These results indicate that the polymerase activities of the mutants are associated with efficiency of viral RNA replication and, consequently, the quantities of NP protein in the RNP complex.

Taken together, the 480 PB1 mutants generated by error-prone PCR had a large extent diversity on polymerase activities, and the mutations at both the conserved motifs and the inter-motif regions of the PB1 catalytic active site can increase and decrease polymerase activities. However, the positions at the inter-motif regions had a greater plasticity compared those in the conserved regions.

Multi-task sparse learning identifies residues associated with polymerase activities. To map residues affecting the polymerase activities, we naturally formulated this problem as a machine learning algorithm aiming to select feature residues (input variable) affecting polymerase activities (output variable). Specifically, we defined the luciferase activities as phenotype output variables, and mutations as input variables. Two widely used sparse learning regression methods, Elastic Net [44] and LASSO [45, 46], and three feature scoring functions (binary, 3 groups of amino acids, and Protein-Protein Interactions in

Macromolecular Analysis), were compared (see details in Materials and Methods). The cross-validation analyses suggest the LASSO-PIMA regression model had the best performance (Supplementary Table 1) and, thus, was selected for further analyses as described below.

A total of 20 residues across the PB1 393-490 region were identified to be associated with polymerase activities, including 395, 400 (located in the finger domain), 402 (motif B), 417 (motif B), 419 (motif B), 428, 429, 432, 434, 442 (motif C), 443 (motif C), 446 (motif C), 447 (motif C), 459, 460, 462, 481 (motif D), 482 (motif D), 486 (motif D), and 489 (motif E) (Table 2 and Supplementary Table 2). To better understand how these residues impact viral RNA synthesis, we further modeled a RNA template / a short RNA primer / nucleotide substrate / two divalent metal ions into the influenza A virus polymerase by superimposing the structure of the IAV polymerase with that of a reovirus polymerase elongation complex [47] (Fig. 3).

Among these residues, residues 395 and 400, which are located in or near the finger domain, form the outer rim of the template RNA entry tunnel and, therefore, are likely to participate in RNA template binding (Fig. 3C). S402 is largely buried within the template tunnel and, thus, a mutation at this residue could adversely affect RNA template entry. Residues 428 and 429 are located at the end of the palm domain and distant from the catalytic active center. Our model also shows that S443 forms hydrogen bonds to priming RNA strand (Fig. 3D). Residue 446 is the second aspartic acid of the strictly conserved -SDD- in motif C, and our model shows that it is essential for coordinating one of the divalent metal ions. K481 from motif D is predicted to directly interact with the phosphate moiety on the incoming nucleotide substrate (Fig. 3C), and R486 forms a salt-bridge with residue E677 from the PA subunit. Residue 489 is found at the interface between the palm and thumb domain facing the PA subunit (Fig. 3D). Residues 417, 419, 442, 447, and 482 are at the vicinity of the catalytic active site, but have no direct contact with the RNA/substrate/metal ions (Fig. 3D); residues 432, 434, 459, 460 and 462 are clustered together underneath the four-stranded β-sheet in the palm subdomain (Fig. 3D), and most of these residues are surface-exposed, raising the possibility this region may serve as a binding site for a yet unidentified viral/host factor.

Taken together, these results support our hypothesis that residues outside of those reported motifs may play pivotal roles for polymerase activity, in addition to previous findings that residues within the reported motifs of PB1 are important for replication/transcription of influenza polymerase complexes [16, 33].

Luciferase activity of PB1 mutants in various influenza RNP complex backgrounds. Among those 20 sites we identified (Table 2), residue 400 varies among avian and swine PB1 and residue 486 varies among avian, swine, and human PB1 (Table 1). Therefore, these two residues were selected to evaluate their effects on polymerase activities in two different RNP backgrounds: A/Memphis/7/1994 (H3N2) (abbreviated as mem94), a human seasonal IAV, and A/swine/Missouri/A01727926/2015 (H4N6) (sH4N6), an avian origin IAV.

Results show that, compared to WT mem94 PB1, with three genes (PB2, PB1, and PA) from mem94, pdm09PB1-T400S significantly decreases the polymerase activities whereas K486R increases the polymerase activities instead (Fig. 4B). With all four RNP genes from mem94 (after we introduced T400S and K486R into the PB1 of mem94), both T400S and K486R significantly increases the polymerase activity, compared to WT mem94 PB1 (Fig. 4C). However, neither T400S nor R486K mutations contain significant variations in polymerase activities with sH4N6 RNP background (Fig. 4D and 4E).

Taken together, the effects of those mutations at the PB1 enzymatic site on the polymerase activities are not only position dependent (Table 2), but are also affected by the composition of three other genes (i.e., PB2, PB1, and NP) of the RNP complex.

Discussion

The structure and function of the PB1 protein in the IAV RNP complex has been studied extensively, and it is well known that the catalytic active site of PB1 is made up of highly conserved motifs pre-A/F, A, B, C, D, and E [13, 25]. Studies have demonstrated that mutations occurring in these

reported motifs of PB1 reduce or even abolish the replication efficiency of the RNP complex [16, 33]. However, sequence analyses suggest that variation does occur in nature at residues in the inter-motif regions of PB1 proteins, which could be host-dependent (Table 1). Nevertheless, there is still a lack of systematic studies on how such variation affects the replication efficacy of the RNP complex. In this study, we hypothesize that, in addition to those in the reported motifs, residues outside of the conserved PB1 motifs affect polymerase activities. To test this hypothesis, we constructed a mutant library (n = 658) targeting the PB1 catalytic active site (residues 393 to 490) of pdm09, which includes four reported motifs (B, C, D, and E) and three inter-motif regions with evident sequence variations (Fig. 1). Phenotypic characterization of these PB1 mutants using a minigenome assay demonstrated a large variation in polymerase activities. Results from machine learning showed that, in addition to those at the reported motifs, the mutations at inter-motif regions can enhance polymerase activities, validating our hypothesis. In addition, the diverse phenotypic variations among PB1 mutants in our library and those identified by machine learning suggest that residues within inter-motif regions of PB1 should be explored further for potential roles in viral host adaptation. Furthermore, our results also suggest that there is a moderate level of plasticity in the PB1 enzymatic site, including those residues within the conserved motifs.

Among the residues associated with polymerase activity, nine residues are located at the inter-motif regions and 11 in conserved motifs B, C, and D, and these residues include those reported by others, such as residues 446 [16, 48] and 480 [25] (Table 2). As expected, residues affecting polymerase activities not only depend on position, but also specific mutations at the position. In general, a large change in the biophysical properties of an amino acid can lead to variations in polymerase activity, and many of these changes in the conserved motif can lead to a decrease in polymerase activities [16, 48]. For example, the aspartic acid residing at residue 446 was previously reported to be important for polymerase activity [16, 48] while maintaining protein stability [33]; this particular aspartic acid has been implicated as one of the aspartic acids responsible for coordinating two divalent metal ions required for catalysis [13, 16, 31, 32], as was also supported by our structural modeling results (Fig. 3D). Motif D hosts two

conserved lysine residues at positions 480 and 481 that are important for NTP binding (Fig. 3C) [25], and a change from lysine to arginine at residue 480 has been shown to increase polymerase activity in pandemic H1N1 and H5N1 RNP complexes [33]. Due to the involvement of residues 480 and 481 with NTP binding, it is possible that changes at either residue to a non-basic amino acid could impair binding of incoming NTPs and, thus, subsequent steps of catalysis. Furthermore, these two lysine residues are stabilized by contacts with α–helix 20 of PA (residues 656-663) [25].

In addition to those residues reported, this study mapped a few new polymerase associated residues, including those within the conserved motifs and at the inter-motif regions (Table 2). For example, residues 402, 443, and 486 are located in reported motifs B, C, and D. Structural modeling suggests that S443 forms hydrogen bonds to priming RNA strand (Fig. 3D) and R486 forms a salt-bridge with residue E677 from the PA subunit. The roles of S443 in vRNA synthesis and the potential impacts these residues may have on structure stability should be explored further; variation has been shown to occur at residue 486 among human, swine, and avian hosts (Table 1, Fig. 1C, Supplementary Fig. 1), and this position should be explored further for any potential role it may play in host adaptation. Residues 432, 434, 459, 460, and 462 are located in the inter-motif regions and are found at the base of the palm domain of PB1. Residues 432 and 434 are located in a turn between α-helix 12 of the palm domain and β–sheet 16; whereas, residues 459, 460, and 462 are found in α-helix 14 [25]; of note, most of these residues are located at surface-exposed regions (Fig. 3D), and whether this region interacts with an unknown host factor needs to be further investigated. Structural modeling suggests that residues 395 and 400 at an inter-motif region could affect RNA template binding (Fig. 3C); however, the majority of the residues may not have direct contact with RNA/substrate/metal ions (Fig. 3D) and how these residues influence polymerase activities have remained relatively unknown.

Besides humans, pigs, dogs, horses, and sea mammals (e.g. seals and whales), IAVs have been recovered from variety of bird species, including at least 105 wild bird species of 26 different families [49]. In addition to glycan receptor binding preference, the genetic constellation of RNP complex and

adaptative mutations in RNP genes, including PB1, have been shown to affect the host and tissue tropisms of IAVs [50-53]. Prior studies suggest that an avian RNP complex has defects in its replication efficiency in human cells, and mutations occurring across PB2, PB1, PA, and NP have been demonstrated to facilitate the adaptation of avian IAVs in humans and other mammalian species (reviewed in Mänz et al. [54]). Our results show mutations in PB1 can either increase or decrease polymerase activities. While the majority of residues are conserved among all IAVs, including those from human, swine, and avian IAVs, diverse genetic variation occurs at a small number of PB1 residues, especially those within the inter-motif regions in human and swine IAVs, but not in avian IAVs (Table 1, Supplementary Figure 1). Previous studies suggest that V473L of PB1 contributes to high polymerase activity and affects pdm09 viral replication in mammalian cell lines [55]. In our PB1 mutant library, the diverse polymerase activities also confirm that variations at residues in inter-motif regions that occur in nature such as 400, 433, 435, and 456 (Fig. 1C and Supplementary Fig.1) increase polymerase activities compared to the pdm09 WT PB1 (data not shown). Two residues, 400 and 486, that vary in nature were identified by machine learning to be associated with polymerase activities (Table 2). Residues 400 and 486 are located in an inter-motif region and conserved motif D, respectively. Introduction of T400S and K486R in PB1 proteins enhanced polymerase activities of the human seasonal H3N2 RNP but not of avian origin swine H4N6 RNP (Figure 4), indicating the importance of these two mutations in host adaption from avian to mammals but not *vice versa*. Of note, residue 486 of most human and swine IAVs can harbor both K and R but those of avian IAVs have predominantly K (Fig. 1 and Supplementary Fig. 1). Taken together, our results suggest that residue variations at the PB1 enzymatic site, especially those at the inter-motif regions, can enhance the RNP complex activity in mammals and could be associated with host adaptation when IAVs are transmitted from birds to mammals. Further studies are needed to evaluate how these mutations affect RNP activities in avian systems.

To evaluate how biophysical properties of amino acids alter polymerase activities shown in Table 2, we use machine learning methods and grouped the 20 amino acids into three categories: nonpolar (V, L, I, M, C, F, W, and Y), small nonpolar (G, A, and P), and polar/charged (S, T, N, Q, H, D, E, K, and R).

Results from sparse learning showed that the polymerase activities are amino acid type dependent and that at the same position, some mutations can increase polymerase activities, whereas others decreased or had no effect on polymerase activities (Supplementary Table 3). For example, changes in the pdm09 PB1 from polar/charged to nonpolar amino acids at residue 417 increased polymerase activities, whereas the same changes at residues 402 and 481 decreased polymerase activity. A change from polar/charged amino acid at residue 400 to another polar/charged or nonpolar amino acid at residue 400 can increase activities; whereas, a change from polar/charged to small nonpolar at this residue decreases activities. When switching amino acids from polar/charged to nonpolar at residue 486, polymerase activities are reduced. However, this type of chemical change at residues 429 and 462 result in an increase in activity. Nevertheless, by integrating large-scale mutagenesis, minigenome analyses, and machine learning, this study effectively identified a handful of important residues capable of affecting polymerase activities, while also supporting previous findings [16, 33].

There are several limitations in this study. First, the polymerase activities were determined using a minigenome *in vitro* platform, but not with the entire viral genome, which has four additional genes HA, NA, MP, and NS. Because each of the four plasmids expressing PB2, PB1, PA, and NP gene has a human RNA pol I promoter and a human HCMV pol II promoter [40], we would expect the RNP with a testing PB1 mutant to affect both transcription and replication efficiency, which are ultimately reflected by the luciferase activities we quantified. Nevertheless, we evaluated the protein expression and confirmed that, in the minigenome system we used, the level of NP protein expressions correlated with replication and transcription efficiencies of RNP complexes, which contain heterologous PB1 but identical PB2, PA, and NP (Fig. 2D). To further assess the impacts of testing PB1 mutants within the context of viral genomics, we tested the replication efficiency of PB1-F412L (abolished) and PB1-H456Q (high) along with seven other genes of pdm09, and results showed that PB1-F412L could not be rescued, and that, from growth kinetics analyses, the PB1-H456Q mutant virus replicated more efficiently at each of the three testing temperatures than the WT pdm09 (Supplementary Fig. 2). However, further experiments will be required to evaluate the effects of these mutations on cell, tissue, and host tropisms. Additionally, this study

focuses only on a small region of PB1, and it is known that other residues outside our target region, as well as other genes of the RNP complex, can impact polymerase activities of IAVs. For example, residue 375 has been implicated to be important for host adaptation [56]. In the future, we will extend our study to other regions of PB1 as well as the remaining genes of the RNP complex.

In summary, we report the use of large-scale mutagenesis strategies combined with machine learning to map key residues affecting polymerase activities of influenza viruses. In addition to those reported in the literature, this study identified a set of new residues associated with polymerase activity, and about half of these residues are located at the inter-motif region. The variations at these residues could be associated with virus host adaptation when an IAV switches a host (e.g. from avian to swine).

Materials and Methods

Cells. Madin-Darby Canine Kidney (MDCK) cells and human embryonic kidney (HEK) 293T cells (both from American Type Culture Collection, Manassas, VA) were maintained in Dulbecco's Modified Eagle Medium (DMEM; GIBCO/BRL, Grand Island, NY). The cell medium was supplemented with 10% fetal bovine serum (Atlanta Biologicals, Lawrenceville, GA), and 1% penicillin–streptomycin (100 U/mL and 100 µg/mL, respectively (GIBCO/BRL)). Cells were maintained at 37°C with 5% CO_2.

Viruses. The A/California/07/2009 pandemic H1N1 (abbreviated as pdm09) virus was used as the template for PB1 plasmid generation. Of note, the PB1 of pdm09 has been circulating at the animal-human interface over the past 50 years: this gene was first introduced from avian to human during 1968 H3N2 pandemics, transmitted to swine population in the 1990s, and then back to humans during 2009 H1N1 pandemics [6, 7]. To test the effects of the PB1 mutations on polymerase activities when combined with different RNP background, in addition to the prototype laboratory-adapted strain A/Puerto Rico/8/1934 (H1N1) (abbreviated as PR8), A/Memphis/7/1994(H3N2) (abbreviated as Mem94) and A/swine/Missouri/A01727926/2015 (H4N6) (sH4N6) were used. Mem94 was selected because it has a PB1 gene resembling the pmd09 precursor and also bears other RNP genes of H3N2 seasonal IAVs; and

sH4N6 was a spillover virus from avian and selected to represent an emerging virus at the animal-human interface, posing a potential risk in generating novel reassortants. All viruses were propagated in MDCK cells (the American Type Culture Collection, Virginia, USA) at 37°C with 5% CO_2 in Opti-MEM I Reduced Serum Medium (Thermo Fisher Scientific, Asheville, NC, USA) supplemented with 1 µg/mL of TPCK-trypsin (Gibco, New York, USA) before being used in the molecular cloning.

Molecular cloning. Construction of full length, wild-type protein expression plasmids encoding virus polymerase (PB2, PB1, and PA) and NP proteins was performed as previously described [40]. Briefly, viral RNA was isolated using GeneJET Viral RNA Purification Kit according to the manufacturer's instruction (Thermo Fisher Scientific, Pittsburgh, PA). The reverse transcription was performed using SuperScript III Reverse Transcriptase and a pair of influenza A virus-specific primer Uni12, and gene specific primers for PB2, PB1, PA, or NP was used to amplify the full-length gene fragments using the following PCR protocol: one cycle at 98° C for 30 sec, one cycle at 98°C for 10 sec, followed by 35 cycles at 53°C for 30 sec, 72°C for 2 min, and 72°C for 10 min. PCR products were then purified through gel electrophoresis and extracted using the GeneJET Gel Extraction Kit (Thermo Fisher Scientific, Pittsburgh, PA). The clones were sequenced by Eurofins Genomics (Huntsville, AL) to confirm that no unexpected mutations were introduced.

Error-prone PCR (ep-PCR) based mutagenesis method to generate PB1 mutant library. The GeneMorph II Random Mutagenesis Kit (Agilent Technologies, Santa Clara, CA) was used to introduce random mutations at the location aa 393-491 using primers (mutation F: 5'-TAAGGCCTCTTCTAATAGATGG-3'; mutation R: 5'-GTGAATTCAAATGTCCCTGTC-3'). The PCR product was used as primer to perform site directed mutagenesis. The epPCR amplification mixture consisted of: 41.5 µl of distilled water, 5 µl of 10x Mutazyme II buffer, 1.25 µl of 10mM deoxyribonucleotide triphosphates, 1 µl of 10 µM primer F, 1 µl of 10 µM primer R, 1 µl of Mutazyme II DNA polymerase (Agilent Technologies, Santa Clara, CA), and 1 µl of polymerase basic 1 plasmid of

CA/07/09 (20 pg/ µl). The following epPCR parameters were used: one cycle at 95°C for 5 min, followed by 30 cycles at 94°C for 1 min, 50°C for 1 min, and 72°C for 2 min, and finally one cycle at 72°C for 10 min. QuickChange II Site-Directed Mutagenesis Kit (Stratagene, La Jolla, CA) was used to perform site directed mutagenesis per the manufacturer's instructions. The PCR product (2µL) was digested with *Dpn*I at 37°C for 1 hour and transformed into XL-1blue competent cells (Agilent Technologies, Santa Clara, CA). The transformed cells were directly inoculated on Luria Bertani (LB) agar with 50 µg/ml Ampicillin. Single colonies were grown, and plasmid DNA were extracted using the GeneJET Plasmid Miniprep Kit (Thermo Fisher Scientific).

Validation of mutations through sequencing. Plasmid DNAs were sent to Eurofins Genomics for sequencing. Samples were prepared per manufacturer's instructions. DNASTAR LaserGene® and BioEdit Sequence Alignment Editor were used to analyze sequence data for mutations.

Luciferase activity to quantify ribonucleoprotein complex expression of PB1 mutants *in vitro*. A total of 4×10^4 293T cells were transfected with polymerase protein expression pHW2000 plasmids PB2, PA, NP, and mutant or wild-type pdm09 PB1plasmids (40 ng) in Corning® 96 well plates. The pHW2000 plasmid was kindly provided by Dr. Richard Webby from St. Jude Children's Research Hospital, and expressed a human RNA polymerase I promoter and a pol II promoter of the human cytomegalovirus [40]. 40 ng of plasmid phPOLI-RLUC expressing *Renilla* luciferase controlled by the human RNA polymerase I promoter and 4 ng of pGL4.13 [*luc2*/SV40] expressing *firefly* luciferase (Promega, Madison, WI) were also co-transfected in 293T cells. After transfection, cells were incubated at 37°C for 48 hours. Luciferase activities were measured in cell lysates after an incubation period using the Dual-Luciferase Reporter Assay System (Promega, Madison, WI) per the manufacturer's instructions. PB1 mutant and wild-type pdm09 PB1 activities were measured in triplicate. Luciferase activity levels were expressed as normalized values to the wild-type pdm09 PB1 values.

Site-directed mutagenesis. The QuickChange Lightning Site-Directed Mutagenesis Kit (Agilent Technologies, Santa Clara, CA) was used to introduce mutations at residues of PB1 proteins from A/Memphis/7/1994 (H3N2) and A/swine/Missouri/A01727926 (H4N6) IAVs. In total, 2 mutations were introduced into each of the PB1 proteins. We used forward primer 5'-TCCAGGGCTCAATGATGCGGAGCCATCTATTAGAAGAGG-3' and reverse primer 5'-CCTCTTCTAATAGATGGCTCCGCATCATTGAGCCCTGGA-3' to generate mutation T400S into the A/Memphis/7/1994 PB1. Forward primer 5'-TGAATTCAAATGTCCCTGTCCTATTTATATAGGACTTCTTTTTGCTCATGTTGATTCC-3' and reverse primer 5'-GGAATCAACATGAGCAAAAAGAAGTCCTATATAAATAGGACAGGGACATTTGAATTCA-3' were used to generate the K486R mutation into the A/Memphis/7/1994 PB1. The forward primer 5'-GGGCTTAATGAGGCGGAACCATCTATTAGAAGAGGCCTTAT-3' and reverse primer 5'-ATAAGGCCTCTTCTAATAGATGGTTCCGCCTCATTAAGCCC-3' were used to generate the T400S mutation into the A/swine/Missouri/A01727926 (H4N6) PB1. Finally, the forward primer 5'-AATTCAAATGTTCCTGTCTTGTTTATGTAGGACTTCTTCTTGCTCATGTTGA-3' and reverse primer 5'-TCAACATGAGCAAGAAGAAGTCCTACATAAACAAGACAGGAACATTTGAATT-3' were used to introduce the R486K mutation into the A/swine/Missouri/A01727926 (H4N6) PB1.

Western blot analysis. 293T cells in 6-well culture plates were transfected with polymerase protein expression plasmids PB2, PA, NP, and wild-type or PB1 mutants. Cell supernatants were treated with RIPA buffer and then prepared for SDS-PAGE analysis. NP expression in the cell lysates was detected by primary antibodies specific for NP (diluted 1:500, BEI Resources, USA). The secondary antibody anti-mouse IgG HRP (diluted 1:5,000, available in our group) was used. β-actin (diluted 1:5000, Sigma-Aldrich, St. Louis, MO) was used as an internal control. HRP presence was detected using the Pierce DAB Substrate Kit (Thermo Fisher Scientific, USA) following the manufacturer's protocol.

Multi-task sparse learning. Sparse learning is a machine learning technique used frequently to extract a small number of significant features from a set of experimental data. The advantage of sparse learning over other conventional machine learning approaches is that its efficiency and generalizability generates accurate models using a small number of non-zero elements. Sparse learning also takes advantage of the sparsity of predominant features in influenza proteins. This is important, because high dimensional features can be redundant and noisy, resulting in poor performance generalization [57]. Previously, we have successfully applied sparse learning in determining antigenicity-associated features for influenza viruses [46, 58-60]. In this study, sparse learning was used to determine the most important residues affecting polymerase activity in the target region of the pdm09 PB1 sequence. Our sparse learning approach addresses the redundancy and noise levels present in replication efficiency data. Thus, we expect the sparse learning method to increase performance in feature selection and facilitate data interpretation.

Specifically, two sparse learning regression methods were used: the L1-norm regularized method (LASSO) and the L1- and L2-norm regularized method (Elastic Net). Briefly, the LASSO regression seeks to minimize the following: $\|y-(x\cdot w+b)\|_2^2+\lambda_1\|w\|_1$ while the Elastic Net regression seeks to minimize: $\|y-(x\cdot w+b)\|_2^2+\lambda_1\|w\|_1+\lambda_2\|w\|_2^2$, where y is the matrix of actual response values, w is the matrix of weights, x is the matrix of explanatory values, b is the bias term, λ_1 and λ_2 are constraint parameters, and $\|\cdot\|_1$ and $\|\cdot\|_2$ are the L1- and L2-norms.

To determine distances among the residues, three approaches were used: 1) the binary method, 2) the Protein-Protein Interactions in Macromolecular Analysis (PIMA) method, and 3) the 3 groups of amino acids method. In the binary method, amino acid sequence *i* was encoded into a vector of binary values, x_i, by comparison to a wild-type sequence. Element *j* (representing the *j*th residue) of x_i was encoded to 0 if the residue in position *j* matched the wild-type residue in position *j* and 1 otherwise. In the PIMA method, amino acid sequences were encoded into vectors x_i by comparison to a wild-type sequence

as previously demonstrated [59]. Element *j* (representing the *j*th residue) was encoded based on the type of residue mutation at the *j*th residue. Mutations between different pairs of residues were given a weight between 0 and 5, inclusive. In the 3 groups method, amino acid sequences were encoded into vectors x_i by comparison to a wild-type sequence. Each amino acid residue was assigned to one of 3 groups: nonpolar, small nonpolar, and polar/charged. Each position was given several states to indicate mutations between amino acid groups from the wild-type to the sequence in question, of which exactly one, the state indicating a mutation from amino acid group X to amino acid group Y, is set equal to 1 (if no mutation occurred, then the state indicating a mutation from amino acid group X to itself is set equal to 0). All other states for the position are set to 0 since those mutations did not occur. Since there are 3 groups, there are 9 different possible combinations (with a mutation from X to Y being counted as a distinct mutation from Y to X), meaning a total length of 900 for each vector x_i.

For each of the above three methods, vectors x_i were thus constructed for all amino acid sequences for a total of 658 plasmid vectors (the wild-type sequence was also compared to itself). These vectors were all put into a matrix of explanatory values, *x*, with 658 rows and 100 columns (or 900 columns for the 3 groups method). Similarly, the vector *y* of response values was constructed by comparison to the wild-type response values. Element *j* was computed by dividing the response of viral strain *j* by the response of the wild-type strain (the wild-type response was also compared to itself). Vector *y*, therefore, had 658 rows and 1 column.

Applying LASSO or Elastic Net regression provided the weight vector *w* with dimension 1 x 100 (or 900 for the 3 groups method) and the bias vector *b* with dimension 658 x 1 (with all elements being the same). For the binary and PIMA methods, element *i* of *w* represented the weight for position *i* of the amino acid sequence. Therefore, a greater magnitude of weight indicated that position *i* was more significant than a position with a lower magnitude of weight. For the 3 groups method, each position *i* was given 9 weights representing each of the 9 possible mutations states. For instance, a higher magnitude of weight at position N with a mutation from X to Y indicated that that specific mutation at that position

may be significant. Similarly, lower magnitude of weight of a mutation at a specific position would indicate that that mutation at that specific position is not significant. Therefore, this method allowed for the identification of not only the important positions, but also the significant mutations at those positions.

Sequence, data analyses, structural modeling and visualization. A total of 11,965 PB1 protein sequences from avian, swine, and human H1N1 IAVs were downloaded from Influenza Research Database (https://www.fludb.org). After removing the duplicate protein sequences, the sequences were aligned using MUSCLE [61]. The visualization of the three-dimensional PB1 structure (PDB ID 4WSB) and the overall polymerase structure (PDB ID 6QNW) was performed using PyMol (https://pymol.org). Template/primer/substrate/metal ions are modeled into the structure by superimposing it with the reovirus polymerase structure (PDB ID 1N38) using Dali [62].

Acknowledgments

This project was supported by grants R01AI116744 and R21AI135820 from the National Institutes of Health to XFW, C-1565 from the Welch Foundation to YJT. KW was supported partially by the United States Department of Agriculture Animal Plant Health Inspection Service's National Bio- and Agro-defense Facility Scientist Training Program. We would like to thank Dr. Zhu Guo from the Centers of Disease Control and Prevention for his assistance in developing and optimizing the minigenome assay and Cynthia Tang for her assistance in preparing this manuscript.

References

[1] Webster RG, Bean WJ, Gorman OT, Chambers TM, Kawaoka Y. Evolution and ecology of influenza A viruses. Microbiol Rev. 1992;56:152-79.

[2] Lipatov AS, Govorkova EA, Webby RJ, Ozaki H, Peiris M, Guan Y, et al. Influenza: emergence and control. J Virol. 2004;78:8951-9.

[3] Crawford PC, Dubovi EJ, Castleman WL, Stephenson I, Gibbs EP, Chen L, et al. Transmission of equine influenza virus to dogs. Science. 2005;310:482-5.

[4] Iuliano AD, Roguski KM, Chang HH, Muscatello DJ, Palekar R, Tempia S, et al. Estimates of global seasonal influenza-associated respiratory mortality: a modelling study. Lancet. 2018;391:1285-300.

[5] Paget J, Spreeuwenberg P, Charu V, Taylor RJ, Iuliano AD, Bresee J, et al. Global mortality associated with seasonal influenza epidemics: New burden estimates and predictors from the GLaMOR Project. J Glob Health. 2019;9:020421.

[6] Smith GJ, Vijaykrishna D, Bahl J, Lycett SJ, Worobey M, Pybus OG, et al. Origins and evolutionary genomics of the 2009 swine-origin H1N1 influenza A epidemic. Nature. 2009;459:1122-5.

[7] Kawaoka Y, Krauss S, Webster RG. Avian-to-human transmission of the PB1 gene of influenza A viruses in the 1957 and 1968 pandemics. J Virol. 1989;63:4603-8.

[8] Shapiro GI, Krug RM. Influenza virus RNA replication in vitro: synthesis of viral template RNAs and virion RNAs in the absence of an added primer. J Virol. 1988;62:2285-90.

[9] Portela A, Digard P. The influenza virus nucleoprotein: a multifunctional RNA-binding protein pivotal to virus replication. J Gen Virol. 2002;83:723-34.

[10] Bouloy M, Plotch SJ, Krug RM. Globin mRNAs are primers for the transcription of influenza viral RNA in vitro. Proc Natl Acad Sci U S A. 1978;75:4886-90.

[11] Plotch SJ, Bouloy M, Ulmanen I, Krug RM. A unique cap(m7GpppXm)-dependent influenza virion endonuclease cleaves capped RNAs to generate the primers that initiate viral RNA transcription. Cell. 1981;23:847-58.

[12] Huang TS, Palese P, Krystal M. Determination of influenza virus proteins required for genome replication. J Virol. 1990;64:5669-73.

[13] Poch O, Sauvaget I, Delarue M, Tordo N. Identification of four conserved motifs among the RNA-dependent polymerase encoding elements. EMBO J. 1989;8:3867-74.

[14] Ulmanen I, Broni BA, Krug RM. Role of two of the influenza virus core P proteins in recognizing cap 1 structures (m7GpppNm) on RNAs and in initiating viral RNA transcription. Proc Natl Acad Sci U S A. 1981;78:7355-9.

[15] Klumpp K, Ruigrok RW, Baudin F. Roles of the influenza virus polymerase and nucleoprotein in forming a functional RNP structure. EMBO J. 1997;16:1248-57.

[16] Biswas SK, Nayak DP. Mutational analysis of the conserved motifs of influenza A virus polymerase basic protein 1. J Virol. 1994;68:1819-26.

[17] Engelhardt OG, Smith M, Fodor E. Association of the influenza A virus RNA-dependent RNA polymerase with cellular RNA polymerase II. J Virol. 2005;79:5812-8.

[18] Martinez-Alonso M, Hengrung N, Fodor E. RNA-Free and Ribonucleoprotein-Associated Influenza Virus Polymerases Directly Bind the Serine-5-Phosphorylated Carboxyl-Terminal Domain of Host RNA Polymerase II. J Virol. 2016;90:6014-21.

[19] Lukarska M, Fournier G, Pflug A, Resa-Infante P, Reich S, Naffakh N, et al. Structural basis of an essential interaction between influenza polymerase and Pol II CTD. Nature. 2017;541:117-21.

[20] Dias A, Bouvier D, Crepin T, McCarthy AA, Hart DJ, Baudin F, et al. The cap-snatching endonuclease of influenza virus polymerase resides in the PA subunit. Nature. 2009;458:914-8.

[21] Yuan P, Bartlam M, Lou Z, Chen S, Zhou J, He X, et al. Crystal structure of an avian influenza polymerase PA(N) reveals an endonuclease active site. Nature. 2009;458:909-13.

[22] Blaas D, Patzelt E, Kuechler E. Identification of the cap binding protein of influenza virus. Nucleic Acids Res. 1982;10:4803-12.

[23] Guilligay D, Tarendeau F, Resa-Infante P, Coloma R, Crepin T, Sehr P, et al. The structural basis for cap binding by influenza virus polymerase subunit PB2. Nat Struct Mol Biol. 2008;15:500-6.

[24] Reich S, Guilligay D, Pflug A, Malet H, Berger I, Crepin T, et al. Structural insight into cap-snatching and RNA synthesis by influenza polymerase. Nature. 2014;516:361-6.

[25] Pflug A, Guilligay D, Reich S, Cusack S. Structure of influenza A polymerase bound to the viral RNA promoter. Nature. 2014;516:355-60.

[26] Hengrung N, El Omari K, Serna Martin I, Vreede FT, Cusack S, Rambo RP, et al. Crystal structure of the RNA-dependent RNA polymerase from influenza C virus. Nature. 2015;527:114-7.

[27] Chang S, Sun D, Liang H, Wang J, Li J, Guo L, et al. Cryo-EM structure of influenza virus RNA polymerase complex at 4.3 A resolution. Mol Cell. 2015;57:925-35.

[28] Thierry E, Guilligay D, Kosinski J, Bock T, Gaudon S, Round A, et al. Influenza Polymerase Can Adopt an Alternative Configuration Involving a Radical Repacking of PB2 Domains. Mol Cell. 2016;61:125-37.

[29] Pflug A, Lukarska M, Resa-Infante P, Reich S, Cusack S. Structural insights into RNA synthesis by the influenza virus transcription-replication machine. Virus Res. 2017;234:103-17.

[30] Fan H, Walker AP, Carrique L, Keown JR, Serna Martin I, Karia D, et al. Structures of influenza A virus RNA polymerase offer insight into viral genome replication. Nature. 2019;573:287-90.

[31] Muller R, Poch O, Delarue M, Bishop DH, Bouloy M. Rift Valley fever virus L segment: correction of the sequence and possible functional role of newly identified regions conserved in RNA-dependent polymerases. J Gen Virol. 1994;75 (Pt 6):1345-52.

[32] Bruenn JA. A structural and primary sequence comparison of the viral RNA-dependent RNA polymerases. Nucleic Acids Res. 2003;31:1821-9.

[33] Chu C, Fan S, Li C, Macken C, Kim JH, Hatta M, et al. Functional analysis of conserved motifs in influenza virus PB1 protein. PLoS One. 2012;7:e36113.

[34] Rolling T, Koerner I, Zimmermann P, Holz K, Haller O, Staeheli P, et al. Adaptive mutations resulting in enhanced polymerase activity contribute to high virulence of influenza A virus in mice. J Virol. 2009;83:6673-80.

[35] Suzuki Y, Uchida Y, Tanikawa T, Maeda N, Takemae N, Saito T. Amino Acid Substitutions in PB1 of Avian Influenza Viruses Influence Pathogenicity and Transmissibility in Chickens. J Virol. 2014;88:11130-9.

[36] Hulse-Post DJ, Franks J, Boyd K, Salomon R, Hoffmann E, Yen HL, et al. Molecular changes in the polymerase genes (PA and PB1) associated with high pathogenicity of H5N1 influenza virus in mallard ducks. J Virol. 2007;81:8515-24.

[37] Zhu W, Zhu Y, Qin K, Yu Z, Gao R, Yu H, et al. Mutations in polymerase genes enhanced the virulence of 2009 pandemic H1N1 influenza virus in mice. PLoS One. 2012;7:e33383.

[38] Ye J, Wen F, Xu Y, Zhao N, Long L, Sun H, et al. Error-prone pcr-based mutagenesis strategy for rapidly generating high-yield influenza vaccine candidates. Virology. 2015;482:234-43.

[39] Garten RJ, Davis CT, Russell CA, Shu B, Lindstrom S, Balish A, et al. Antigenic and genetic characteristics of swine-origin 2009 A(H1N1) influenza viruses circulating in humans. Science. 2009;325:197-201.

[40] Hoffmann E, Neumann G, Kawaoka Y, Hobom G, Webster RG. A DNA transfection system for generation of influenza A virus from eight plasmids. Proc Natl Acad Sci U S A. 2000;97:6108-13.

[41] Biswas SK, Boutz PL, Nayak DP. Influenza virus nucleoprotein interacts with influenza virus polymerase proteins. J Virol. 1998;72:5493-501.

[42] Newcomb LL, Kuo RL, Ye Q, Jiang Y, Tao YJ, Krug RM. Interaction of the influenza a virus nucleocapsid protein with the viral RNA polymerase potentiates unprimed viral RNA replication. J Virol. 2009;83:29-36.

[43] Portela A, Digard P. The influenza virus nucleoprotein: a multifunctional RNA-binding protein pivotal to virus replication. J Gen Virol. 2002;83:723-34.

[44] Zou H, Hastie T. Regularization and variable selection via the elastic net. Journal of the Royal Statistical Society (Series B). 2005;67:301-20.

[45] Tibshirani R. Regression shrinkage and selection via the LASSO. Journal of the Royal Statistical Society (Series B). 1996;58:267-88.

[46] Cai Z, Ducatez MF, Yang J, Zhang T, Long LP, Boon AC, et al. Identifying Antigenicity-Associated Sites in Highly Pathogenic H5N1 Influenza Virus Hemagglutinin by Using Sparse Learning. J Mol Biol. 2012;422:145-55.

[47] Tao Y, Farsetta DL, Nibert ML, Harrison SC. RNA synthesis in a cage--structural studies of reovirus polymerase lambda3. Cell. 2002;111:733-45.

[48] Zamyatkin DF, Parra F, Alonso JM, Harki DA, Peterson BR, Grochulski P, et al. Structural insights into mechanisms of catalysis and inhibition in Norwalk virus polymerase. J Biol Chem. 2008;283:7705-12.

[49] Olsen B, Munster VJ, Wallensten A, Waldenstrom J, Osterhaus AD, Fouchier RA. Global patterns of influenza a virus in wild birds. Science. 2006;312:384-8.

[50] Snyder MH, Betts RF, DeBorde D, Tierney EL, Clements ML, Herrington D, et al. Four viral genes independently contribute to attenuation of live influenza A/Ann Arbor/6/60 (H2N2) cold-adapted reassortant virus vaccines. J Virol. 1988;62:488-95.

[51] Jin H, Zhou H, Lu B, Kemble G. Imparting temperature sensitivity and attenuation in ferrets to A/Puerto Rico/8/34 influenza virus by transferring the genetic signature for temperature sensitivity from cold-adapted A/Ann Arbor/6/60. J Virol. 2004;78:995-8.

[52] Jin H, Lu B, Zhou H, Ma C, Zhao J, Yang CF, et al. Multiple amino acid residues confer temperature sensitivity to human influenza virus vaccine strains (FluMist) derived from cold-adapted A/Ann Arbor/6/60. Virology. 2003;306:18-24.

[53] Herlocher ML, Clavo AC, Maassab HF. Sequence comparisons of A/AA/6/60 influenza viruses: mutations which may contribute to attenuation. Virus Res. 1996;42:11-25.

[54] Manz B, Schwemmle M, Brunotte L. Adaptation of avian influenza A virus polymerase in mammals to overcome the host species barrier. J Virol. 2013;87:7200-9.

[55] Xu C, Hu WB, Xu K, He YX, Wang TY, Chen Z, et al. Amino acids 473V and 598P of PB1 from an avian-origin influenza A virus contribute to polymerase activity, especially in mammalian cells. J Gen Virol. 2012;93:531-40.

[56] Taubenberger JK, Reid AH, Lourens RM, Wang R, Jin G, Fanning TG. Characterization of the 1918 influenza virus polymerase genes. Nature. 2005;437:889-93.

[57] Lai H, Pan Y, Liu C, Lin L, Wu J. Sparse Learning-to-Rank via an Efficient Primal-Dual Algorithm. IEEE TRANSACTIONS ON COMPUTERS. 2011.

[58] Han L, Zhang Y, Wan X-F, Zhang T. Generalized Hierarchical Sparse Model for Arbitrary-Order Interactive Antigenic Sites Identification in Flu Virus Data. Proceedings of the 22nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD). San Francisco2016. p. 865-74.

[59] Sun H, Yang J, Zhang T, Long LP, Jia K, Yang G, et al. Inferring influenza virus antigenicity using sequence data. mBio. 2013;4:4.

[60] Yang J, Zhang T, Wan XF. Sequence-based antigenic change prediction by a sparse learning method incorporating co-evolutionary information. PLoS One. 2014;9:e106660.

[61] Edgar RC. MUSCLE: a multiple sequence alignment method with reduced time and space complexity. BMC Bioinformatics. 2004;5:113.

[62] Holm L. Benchmarking fold detection by DaliLite v.5. Bioinformatics. 2019;35:5326-7.

Tables

Table 1. Variations between the PB1 enzymatic site (393-490) of H1N1 IAV and that of pdm09.

Host[a]	No. of Sequences	Location of varied residues[b]							Total mutations
		393-400 (Inter-motif)	401-422 (Motif B)	423-435 (Inter-motif)	436-449 (Motif C)	450-473 (Inter-motif)	474-486 (Motif D)	487-490 (Motif E)	
Human	9,443	3,441	15	5,208	11	1,684	1,502	2	11,863 (76.7%)[d]
Avian	485	19	3	74	2	28	11	0	137 (0.9%)
Swine	2,037	553	30	1,384	15	675	808	1	3,466 (22.4%)
	Total mutations	4,013 (25.95%)[c]	48 (0.31%)	6,666 (43.10%)	28 (0.18%)	2,387 (15.43%)	2,321 (15.01%)	3 (0.02%)	15,466

Note: [a]the host origin of H1N1 IAV; [b] the region covers three inter-motif region and four motifs (B, C, D, and E); [c]the total number of mutations, and the number in the parenthesis denotes the percentile of those mutations in the specific region; [d]the total number of mutations, and the number in the parathesis denotes the percentile of those mutations found in human, avian, or swine IAVs.

Table 2. Residues in the PB1 enzymatic site identified by machine learning to be associated with polymerase activities.

Location	Residue	Weight[a]	Mutant (R)[b]
Inter-motif	395	0.01018743	**L395I** (6.555), **L395P** (0.026)
	400	0.0124385	**T400S** (0.779), **T400A**-K430T-Y431F-Q460R (0.517), **T400I** (0.291), **T400S**-M408L-N425K-W437R (0.023), **T400S**-R468W (0.002)
Motif B	402	0.0081042	**S402T**-I450L (1.134), **S402P** (0.027), **S402P**-Y467F-K471M-M477V (0.01), **S402T**-D446V (0.006)
	417	0.009831	**T417S** (1.125), A401T-**T417A**-D439N (0.049), **T417S**-G420R-L424Q (0.03), **T417S**-G440W-M477V (0.025), **T417S**-A453T (0.022), L403P-**T417A**-E457V (0.008), **T417M**-E457D-I459L (0.007), **T417S**-I423K (0.006), **T417M** (0.005), S416T-**T417S**-M477L (0.005), P405H-**T417R**-Q428H (0.004)
	419	0.0093204	**L419S**-I423V (2.186), **L419M** (1.708), M414V-**L419F**-V463M (0.873), L419S-**L449I** (0.155), M414I-**L419S**-I435L-A453V-V473E-G474R-N476T (0.016), A401T-N413D-**L419S**-S443L-Q460R (0.014), G410C-**L419M** (0.012), **L419S**-Q428R-K433M-A448T (0.007), **L419W**-K429R-R468W-N485I (0.006)
Inter-motif	428	0.01272691	**Q428H** (1.337), **Q428R**-K429N (0.876), **Q428E**-R468M (0.087), M414K-**Q428H**-Y431F-A453S (0.037), **Q428L**-Q442L-D445V (0.026), **Q428H**-Y467N (0.024), **Q428L**-Y436H-Y467H (0.01), N425D-**Q428H**-A461S-N476S-K479E (0.01), M409I-**Q428R**-Y436N (0.008), L419S-**Q428R**-K433M-A448T (0.007), L396P-S416N-**Q428L**-W438L (0.006), M409L-Q428H (0.005), P405H-T417R-**Q428H** (0.004), **Q428R**-P454S (0.003)
	429	0.01861988	**K429Q** (1.669), **K429N** (1.224), **K429R**-H456Q (1.2), **K429R**-T432S (1.161), Q428R-**K429N** (0.876), **K429M** (0.866), **K429N**-N455I (0.22), **K429T**-K471E (0.152), M408T-M414I-**K429E**-D439V-S478N (0.059), G410D-**K429N** (0.016), I423V-**K429M**-D446E-I450V (0.009), L419W-**K429R**-R468W-N485I (0.006)
	432	0.0100022	**T432A** (1.358), K429R-**T432S** (1.161), **T432S** (0.973), **T432I** (0.794), **T432A**-E457V (0.56), T432I-D439G (0.47), I423L-L424M-**T432I** (0.194), **T432I**-A448D (0.017), M409L-L415I-**T432A**-V451L-Y467F (0.015), N413K-L426R-**T432I** (0.01), G427E-**T432I**-V451L (0.008), **T432A**-S482F (0.004), F412Y-**T432S**-D445N-V463A-Y467H (0.002)
	434	0.0274777	**T434R** (2.316), **T434S** (1.577), **T434P**-F447I (0.129), **T434I** (0.044), **T434I**-G440V-V451M (0.017), M407I-**T434S**-D439E (0.016)
Motif C	442	0.00908953	**Q442H** (2.218), **Q442L**-V451E-A461T (0.725), **Q442L** (0.525), **Q442H**-T469I (0.476), **Q442L**-T469I-L472I (0.094), **Q442K**-I450L-Q460R (0.028), Q428L-**Q442L**-D445V (0.026), M409N-S416I-**Q442R** (0.023), N413S-**Q442H**-N476K-K479N-K486R (0.019), L415Q-**Q442E**-N476H (0.008), **Q442K** (0.007), S416G-V421D-L441I-**Q442H**-I450V-N452I-Y467N (0.005), **Q442P**-I475V (0.003)
	443	0.0083673	Y436F-**S443Q**-N485D (0.042), **S443T** (0.022), A401T-N413D-L419S-**S443L**-Q460R (0.014), **S443P** (0.012), K433T-**S443L** (0.009), S416C-**K433N**-S443L (0.007), Y431N-**S443P**-I450M-N452Y-A461S (0.006)
	446	0.0091995	**D446E** (0.028), G420E-**D446E**-G474V (0.023), G440E-**D446E** (0.016), L396I-**D446V** (0.015), **D446V** (0.013), G420R-**D446E**-F466L (0.012), **D446Y** (0.011), I423V-K429M-**D446E**-I450V (0.009), G440R-**D446Y**-P454L (0.008), K430E-**D446N**-L449P-N485Y (0.008), **D446E**-N485H (0.006), S402T-**D446V** (0.006), **D446N**-N476D (0.005), F412L-**D446S** (0.003), S404T-P405A-W437R-**D446E** (0.002)
	447	0.01115407	**F447I** (0.924), **F447Y** (0.888), **F447L** (0.515), W437C-**F447L**-C470S (0.358), T434P-**F447I** (0.129), F412L-**F447Y** (0.109), Y436H-**F447Y** (0.046), **F447C** (0.045), F412I-**F447S**-N455S (0.029), **F447Y**-N452K-Y483N (0.016), **F447Y**-D464G (0.013), G406A-**F447L** (0.012), **F447Y**-A448D-L472I-I475V (0.011), M408L-**F447C** (0.008), **F447L**-I450T (0.004), **F447Y**-N476I (0.003)
Inter-motif	459	0.0098647	H456P-**I459L** (2.006), **I459L** (1.689), N452K-**I459L** (1.295), **I459T** (0.419), P454R-E457V-**I459T** (0.208), L426H-**I459M** (0.027), S444T-**I459L** (0.025), G406E-S422P-S444P-**I459L**-K479I (0.01), A401S-**I459L**-K471M-K480M (0.008), T417M-E457D-**I459L** (0.007)
	460	0.008106	K433N-**Q460H** (5.241), **Q460H** (0.767), L441H-**Q460K** (0.626), T400A-K430T-Y431F-**Q460R** (0.517), E457D-**Q460L**-V473M (0.339), Q442K-I450L-**Q460R** (0.028), **Q460L**-Y467C (0.015), A401T-N413D-L419S-S443L-**Q460R** (0.014), M409V-**Q460P**-S478N (0.01)
	462	0.0108798	**G462V** (0.022), **G462R**-V473E (0.016), **G462E**-M477I (0.013), **G462R** (0.006), F412L-**G462R** (0.004)
Motif D	481	0.0096242	N413I-**K481M** (0.066), N413D-D464E-**K481E** (0.053), M409R-**K481M** (0.025), W438R-**K481R** (0.017), **K481R** (0.007), D464G-**K481E** (0.003)
	482	0.0092459	V418D-**S482F** (0.022), **S482Y** (0.012), T432A-**S482F** (0.004), K433M-N452S-**S482F** (0.002)
	486	0.0193654	**K486R** (2.122), H456Y-**K486M** (2.085), **K486M** (1.744), **K486E** (0.28), A448D-K471M-**K486M** (0.04), N413S-Q442H-N476K-K479N-**K486R** (0.019), N413D-R465G-T469A-**K486M** (0.011), L426P-L472F-**K486R** (0.006), M411R-D464V-**K486M** (0.006), G427S-D439H-N452I-M477I-**K486R** (0.001)

[a]The absolute value of the weight derived from machine learning; [b] Based on the related polymerase activity (R), we categorized these mutants into four groups: high ($R \geq 2.00$), moderate ($0.50 \leq R < 2.00$), low ($0.10 \leq R < 0.50$), or abolished polymerase activity ($R < 0.10$). The mutant with high polymerase activities are underlined, and those with abolished were labeled in red.

Figure captions

Figure 1. The PB1 enzymatic site to be mapped in this study. (A) Ribbon diagram of PB1 protein color-coded according by those reported domain structures, and the target mutation region is denoted by line format instead of ribbon. (B) Ribbon diagram of the target mutation region with color scheme denoting location of conserved motifs. (C) the target mutation region in color scheme denoting location of conserved motifs and variations observed in H1N1 IAVs recovered from human, swine, and avian. The visualization of three dimensional structure of PB1 (PDB ID 4WSB) was performed using PyMol (https://pymol.org).

Figure 2. Profile of the PB1 mutants generated using ep-PCR. (A) Frequency of the mutations of the 480 unique mutants at each residue within the target mutation region. (B) Distribution of polymerase activities for the 480 unique PB1 mutants, which were classified based on the relative polymerase activities (R) as high ($R \geq 2.00$; n = 35), moderate ($0.50 \leq R < 2.00$; n = 99), low ($0.10 \leq R < 0.50$; n = 50), or abolished polymerase activity ($R < 0.10$; n = 296). To make the analyses to be quantifiable, the luciferase activity (L) for a testing RNP complex is calculated by *Renilla/firefly*, and R for a testing mutant by $L_{mutantPB1}/L_{WT\text{-}PB1}$. (C) Luciferase activity of six PB1 mutants at 37°C selected for Western blot analysis. Data are expressed as the mean ± standard deviation of three replicates normalized to wild-type. (D) Western blot analysis showing NP expression level for RNP complex containing wild-type pdm09 PB1 and mutants. The numbers below the Western blot were the relative protein expression ratio (RPER); the NP protein expression was first normalized by β-actin, and RPER for a testing mutant is calculated by the ratio of the normalized NP value for a mutant PB1 over that of the WT PB1.

Figure 3. Structural modeling of the mapped residues associated with polymerase activities (Table 2). (A) The overall structure of the polymerase (PDB ID 6QNW). Template/primer/substrate/metal ions are modeled into the structure by superimposing it with the reovirus polymerase structure (PDB ID 1N38) using Dali [62]. The residues associated with increasing polymerase activities are shown in green and

those associated with decreasing polymerase activities in red. The PA is colored in pink, PB1 in light blue, PB2 in light green, RNA/NTP/Mn in magenta. (B) the PB1 mutation region (aa393-490). (C) and (D) Magnified views of the boxed regions in (A). Template and priming RNA are labeled. RNA and nucleotide are shown in sticks while the two metal ions are shown by spheres. The color scheme in (B-D) is the same as in (A).

Figure 4. The impacts of T400S and K486R on polymerase activity of human seasonal H3N2 or avian origin swine H4N6 RNP. Polymerase activity of 2009 pandemic H1N1 RNP complex with (A) pdm09-PB1 harboring T400S and K486R; polymerase activity of human H3N2 RNP complex with (B) pdm09-PB1 harboring T400S and K486R and (C) mem94-PB1 harboring T400S and K486R; polymerase activity of avian origin swine H4N6 RNP complex with (D) pdm09-PB1 harboring T400S and K486R and (E) sH4N6-PB1 harboring T400S and R486K. A/California/07/2009 (H1N1) is abbreviated as pdm09, A/Memphis/7/1994 (H3N2) as mem94, and A/swine/Missouri/A01727926 (H4N6) as sH4N6. The data were normalized to the RNP complex without mutations. Data were expressed as the mean ± standard deviation of three replicates. *P* values were calculated by using Student's t-test, and comparing wild-type PB1 with the respective PB1 mutant. ***, $P < 0.001$; **, $P < 0.05$.

Fig. 1

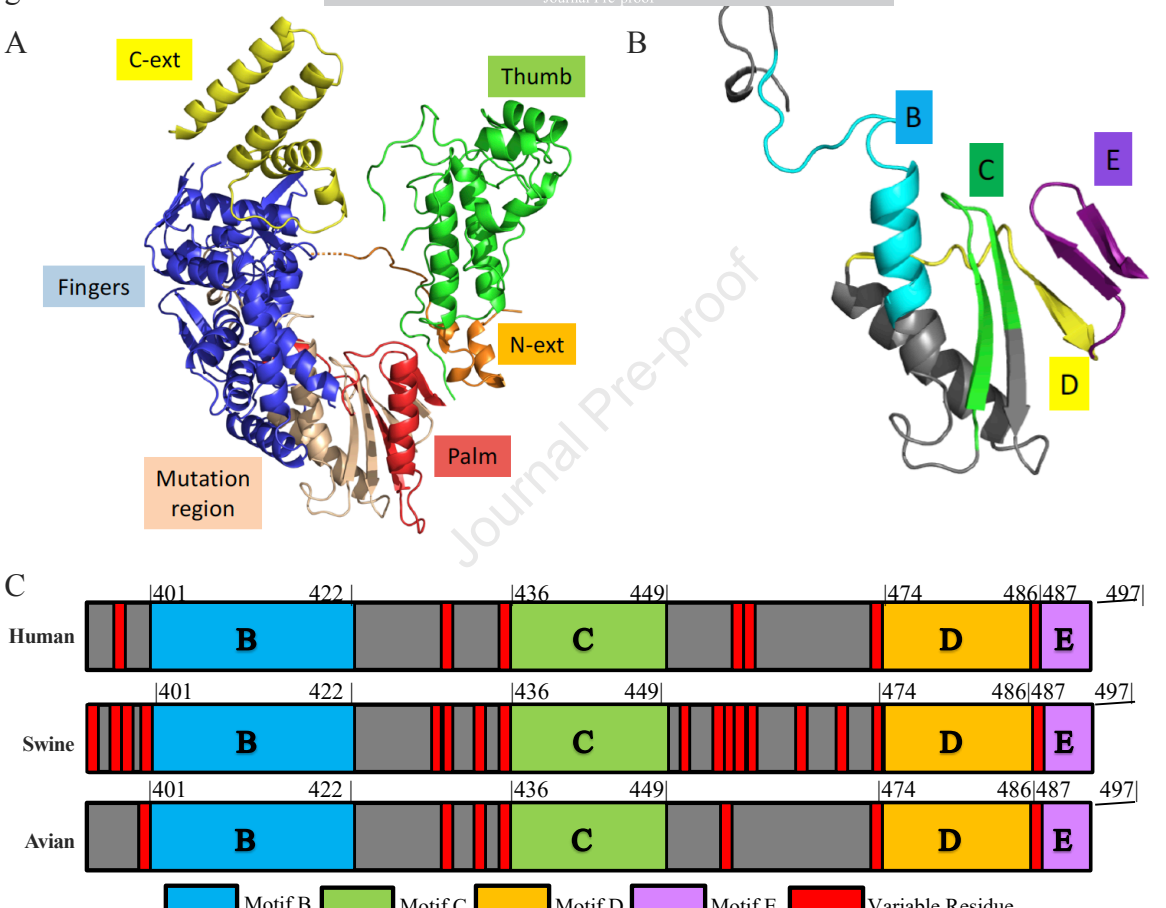


Figure 2

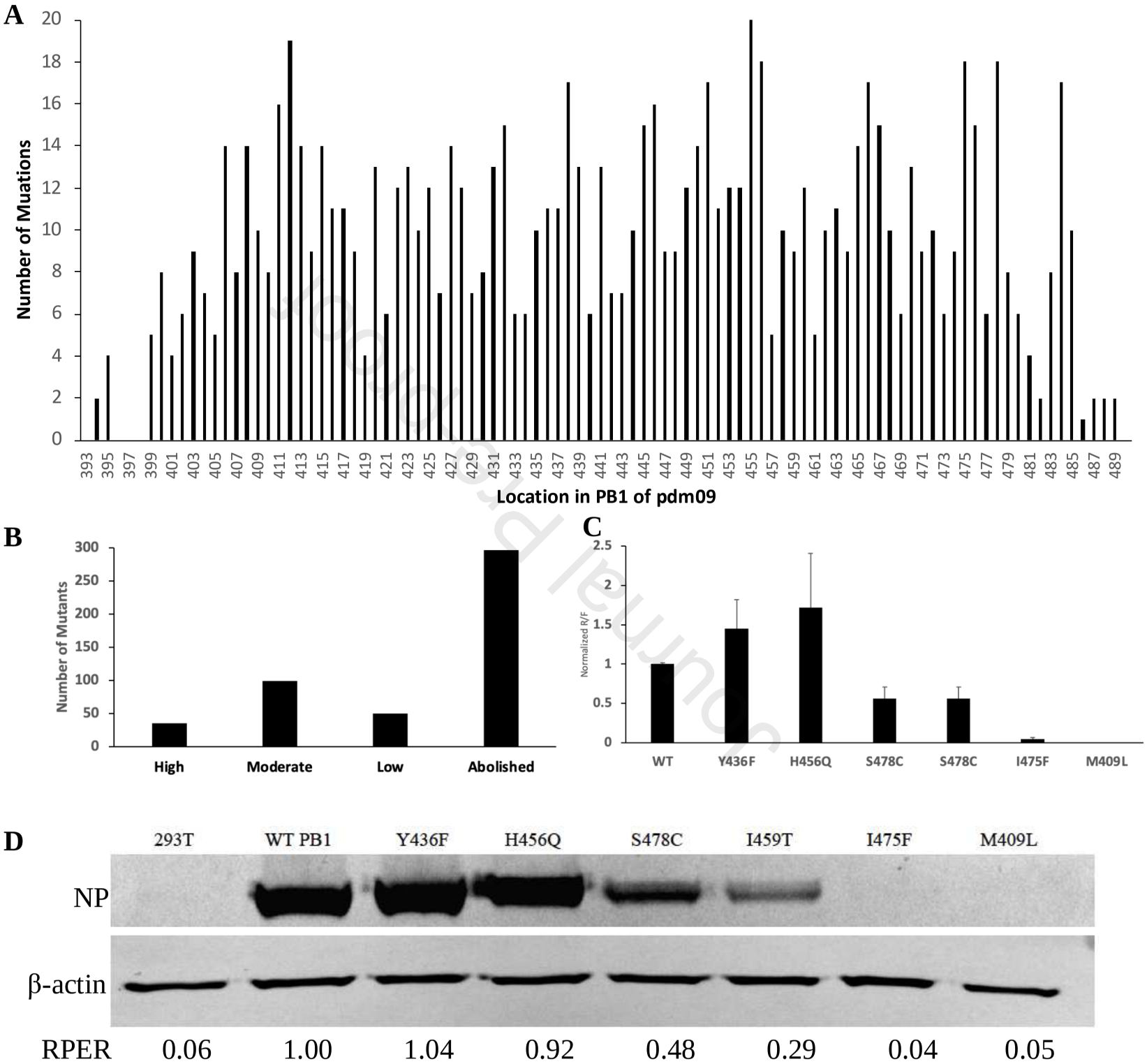


Figure 3

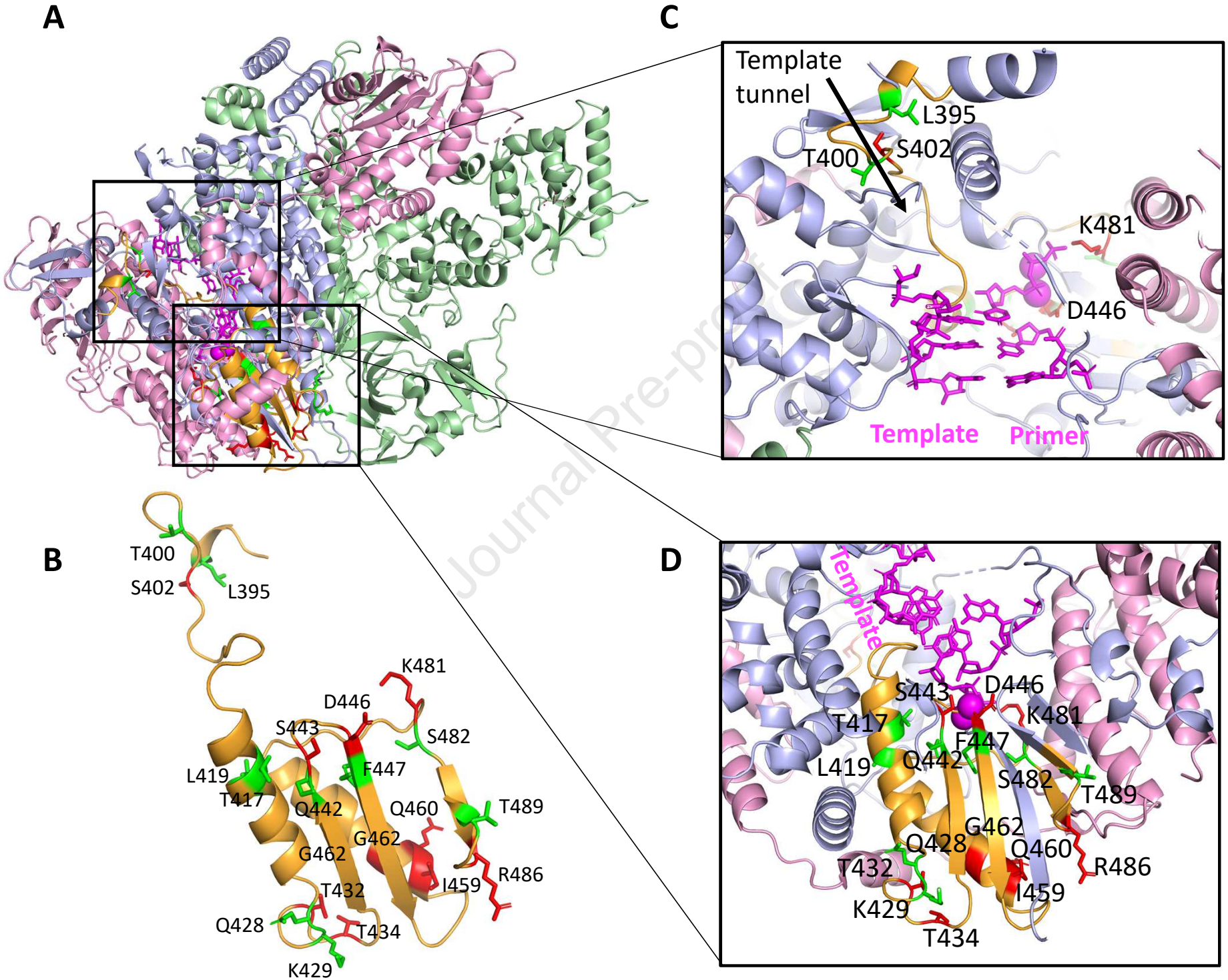
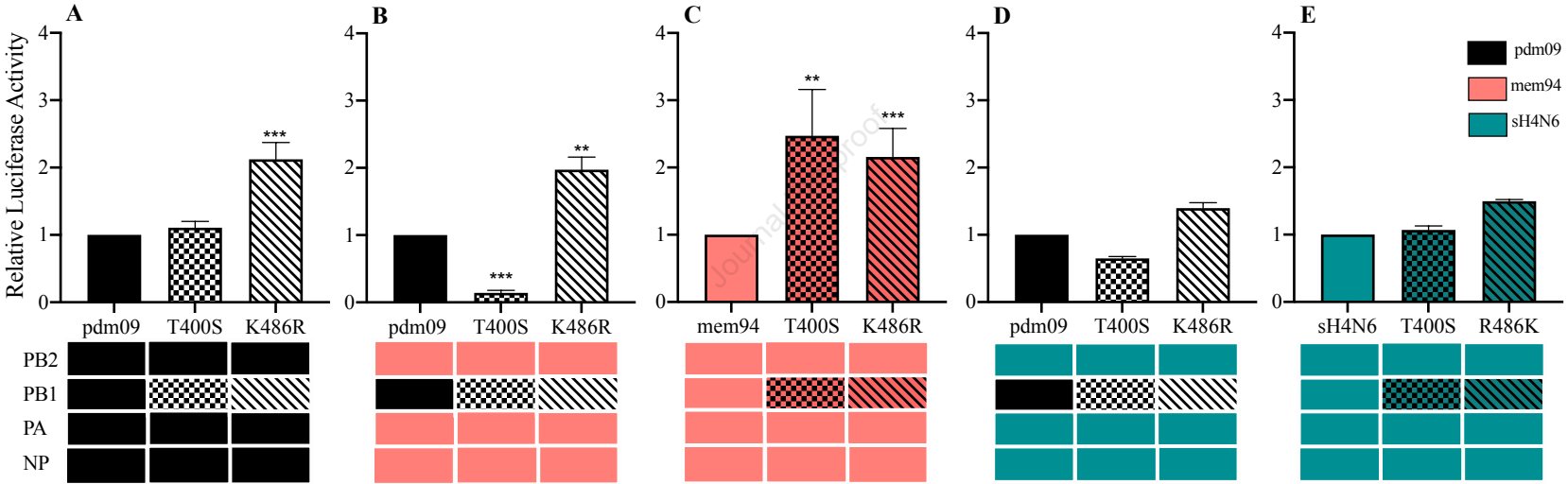


Figure 4



Highlights

- We systematically mapped polymerase associated residues in the PB1 enzymatic active site by integrating a large-scale mutagenesis strategy and computational analyses.
- Polymerase activities are affected by residues widely distributed in the inter-motif and conserved motif regions of the PB1 active site.
- The inter-motif region associated with polymerase activities are at the base of the palm domain with high sequence diversity among those of swine and human origin viruses but not those of avian origin viruses.
- Our results suggest that mutations at those residues outside the conserved motifs of PB1 could be associated with virus-host adaptation when an IAV switches hosts from avian to mammals.

Declaration of interests

☒ The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

☐The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: