AMERICAN SOCIETY FOR MICROBIOLOGY | mBio®

RESEARCH ARTICLE

Annotating Protein Functional Residues by Coupling High-Throughput Fitness Profile and Homologous-Structure Analysis

Yushen Du,[a,b] Nicholas C. Wu,[a,c]* Lin Jiang,[d] Tianhao Zhang,[a,c] Danyang Gong,[a] Sara Shu,[a] Ting-Ting Wu,[a] Ren Sun[a,b,c]

Department of Molecular and Medical Pharmacology, University of California Los Angeles, Los Angeles, California, USA[a]; Cancer Institute, Collaborative Innovation Center for Diagnosis and Treatment of Infectious Diseases, ZJU-UCLA Joint Center for Medical Education and Research, Zhejiang University School of Medicine, Zhejiang University, Hangzhou, Zhejiang, China[b]; Molecular Biology Institute, University of California Los Angeles, Los Angeles, California, USA[c]; Department of Neurology, University of California Los Angeles, Los Angeles, California, USA[d]

* Present address: Nicholas C. Wu, Department of Integrative Structural and Computational Biology, The Scripps Research Institute, La Jolla, California, USA.

ABSTRACT **Identification and annotation of functional residues are fundamental questions in protein sequence analysis. Sequence and structure conservation provides valuable information to tackle these questions. It is, however, limited by the incomplete sampling of sequence space in natural evolution. Moreover, proteins often have multiple functions, with overlapping sequences that present challenges to accurate annotation of the exact functions of individual residues by conservation-based methods. Using the influenza A virus PB1 protein as an example, we developed a method to systematically identify and annotate functional residues. We used saturation mutagenesis and high-throughput sequencing to measure the replication capacity of single nucleotide mutations across the entire PB1 protein. After predicting protein stability upon mutations, we identified functional PB1 residues that are essential for viral replication. To further annotate the functional residues important to the canonical or noncanonical functions of viral RNA-dependent RNA polymerase (vRdRp), we performed a homologous-structure analysis with 16 different vRdRp structures. We achieved high sensitivity in annotating the known canonical polymerase functional residues. Moreover, we identified a cluster of noncanonical functional residues located in the loop region of the PB1 β-ribbon. We further demonstrated that these residues were important for PB1 protein nuclear import through the interaction with Ran-binding protein 5. In summary, we developed a systematic and sensitive method to identify and annotate functional residues that are not restrained by sequence conservation. Importantly, this method is generally applicable to other proteins about which homologous-structure information is available.**

IMPORTANCE **To fully comprehend the diverse functions of a protein, it is essential to understand the functionality of individual residues. Current methods are highly dependent on evolutionary sequence conservation, which is usually limited by sampling size. Sequence conservation-based methods are further confounded by structural constraints and multifunctionality of proteins. Here we present a method that can systematically identify and annotate functional residues of a given protein. We used a high-throughput functional profiling platform to identify essential residues. Coupling it with homologous-structure comparison, we were able to annotate multiple functions of proteins. We demonstrated the method with the PB1 protein of influenza A virus and identified novel functional residues in addition to its canonical function as an RNA-dependent RNA polymerase. Not limited to virology, this method is generally applicable to other proteins that can be functionally selected and about which homologous-structure information is available.**

Received 27 September 2016 **Accepted** 7 October 2016 **Published** 1 November 2016

Citation Du Y, Wu NC, Jiang L, Zhang T, Gong D, Shu S, Wu T, Sun R. 2016. Annotating protein functional residues by coupling high-throughput fitness profile and homologous-structure analysis. mBio 7(6):e01801-16. doi:10.1128/mBio.01801-16.

Editor Peter Palese, Icahn School of Medicine at Mount Sinai

Address correspondence to Ren Sun, rsun@mednet.ucla.edu.

Amino acid residues in a protein have two roles: providing a structural framework (structural residues) and mediating interactions with other biomolecules (functional residues). Identification and annotation of functional residues are fundamental goals in protein characterization (1–5). A number of methods have been developed to achieve these goals. Most methods use sequence conservation information, with the assumption that functional residues are often conserved in homologous proteins (6–8). The residues identified are then expected to perform functions similar to those of other homologs. Other methods predict functional residues on the basis of the shapes and properties of three-dimensional protein structures (9–15). Starting from well-known functional domains (ligand binding, catalytic, etc.), these analyses determine similar local structures and key residues that may be related to the function. Conservation-based methods provide valuable information on protein functional residues but are limited by the insufficient sampling of protein sequence space in natural evolution. It is also challenging for conservation-based methods to assess structural and functional constraints and to assign functionality at the single-residue level (Fig. 1) (16). Therefore, a more direct and systematic method needs to be used for the accurate identification and annotation of functional residues.

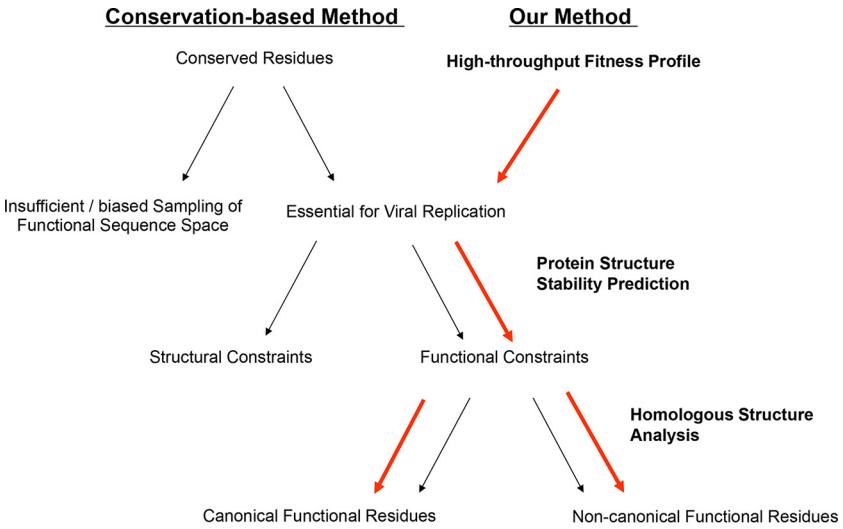

FIG 1 Comparison of the conservation-based method and our method. The conservation-based method is commonly used to identify and annotate functional protein residues, but it has three major limitations. First, it is limited by the insufficient sampling of protein functional space in natural evolution. Second, it is challenging for this method to dissect residues with structural or functional constraints. Lastly, it is limited to distinguishing the diverse functions within the same protein. The method we present here may overcome these limitations and provide a systematic way to annotate functional residues. Using high-throughput fitness profiling, we can identify essential residues for viral replication. Through mutant protein stability prediction, we are able to dissect the structural and functional constraints. Homologous structural analysis is used to further annotate canonical and noncanonical functional residues.

Because of their compact genome, viruses usually encode multifunctional proteins, including viral polymerase proteins. Viral RNA-dependent RNA polymerase (vRdRp) is used by many RNA viruses for transcription and replication. Functions of vRdRp can be grouped into two classes: canonical and noncanonical. The canonical vRdRp functions include template and nucleotide binding, initiation, and elongation (17–19). Among different classes of RNA viruses, these canonical functions and corresponding protein structural features are conserved (17–22). The noncanonical functions of vRdRp, however, are specific to each virus. For example, multimerization of hepatitis C virus (HCV) vRdRp is essential for viral replication. Thus, the interacting residues in HCV vRdRp are noncanonical functional residues specific to HCV (23, 24). Moreover, vRdRp often recruits cellular machinery for replication and plays a role in inhibition of the cellular immune response (25–31). Noncanonical functional residues are usually involved in the performance of those functions and thus are essential for viral replication. Noncanonical functional residues in vRdRp are difficult to determine by commonly used methods and are not as well studied as the key residues for polymerase catalytic functions. However, the noncanonical functional residues are indispensable for thorough protein characterization and may act as drug targets. As a result, it is essential to develop methods that enable the identification of noncanonical functional residues.

We previously developed a method to systematically identify functional residues by coupling experimental fitness measurement with protein stability prediction (16). Here, we extended this method to annotate functional residues in combination with structural comparison of homologous proteins. The method consists of three steps. First, the effect of PB1 mutations on viral replication at single-nucleotide resolution is examined by saturation mutagenesis and high-throughput sequencing. Second, functional PB1 residues that are essential for viral growth but do not affect protein stability are identified by protein stability prediction. Third, homologous structural alignment is used to further annotate specific biological functions (canonical versus noncanonical functions) for each functional residue (Fig. 1). We achieved high sensitivity in identifying and annotating the canonical polymerase functional residues. Moreover, we also identified noncanonical functional residues, which are exemplified by a cluster of residues located in the loop region of the PB1 β-ribbon. These previously uncharacterized residues were shown to be important for PB1 protein nuclear import by interacting with Ran-binding protein 5 (RanBP5) (32).

RESULTS

Fitness profile of influenza A/WSN/33 virus segment 2 at single-nucleotide resolution. High-throughput genetics have been applied to a number of viral, bacterial, and cellular proteins (16, 33–38, 111, 112). Here, point mutations were randomly introduced into segment 2 of influenza A/WSN/33 virus through error-prone PCR. To provide a more accurate quantification of the fitness effect of single mutations, we employed the "small-library" method that we recently developed (16). Nine small libraries were generated to cover all of segment 2 (see Fig. S1A in the supplemental material). Each small library was transfected into 293T cells together with seven plasmids that encoded the other wild-type viral segments (39). Reconstituted mutant virus libraries were used to infect A549 cells at a multiplicity of infection (MOI) of 0.05, and supernatants were collected 24 h postinfection. The input DNA libraries, posttransfection libraries, and postinfection libraries were subjected to Illumina sequencing. To control for technical error and assess library quality, biological duplicates were included in both transfection and subsequent infection steps (Fig. 2A).

The distribution of the number of mutations in the input DNA library was examined. Thirty to 35% of the input DNA library plasmids contained the desired single nucleotide mutations (see Fig. S1B in the supplemental material). We achieved at least 20,000× sequencing coverage for each nucleotide position (see

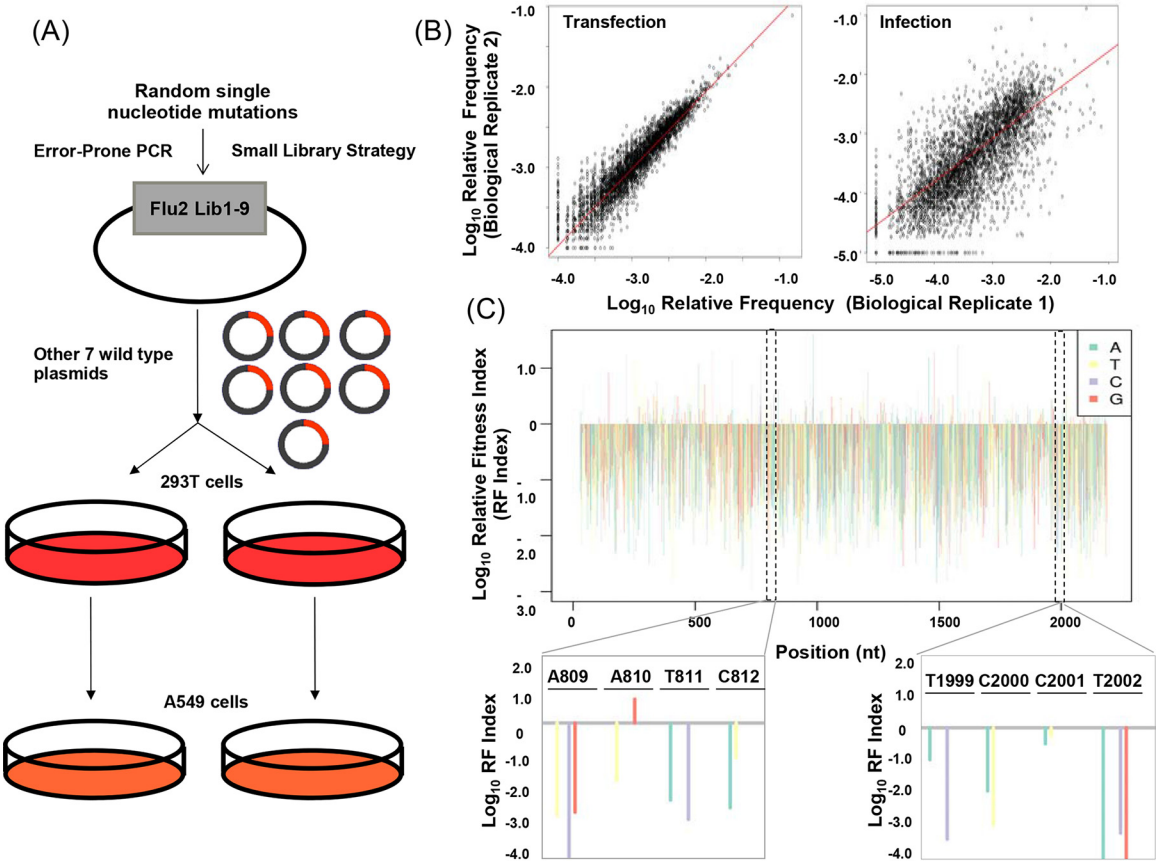

FIG 2 Fitness profile of influenza A virus segment 2 at single-nucleotide resolution. (A) Schematic representation of the experimental flow of high-throughput fitness profiling. Random single nucleotide mutations were introduced into influenza A/WSN/33 virus segment 2. Mutant viral libraries were generated by cotransfecting a mutant DNA library with seven plasmids encoding the other wild-type viral fragments. Viral libraries were then passaged in A549 cells. High-throughput sequencing of the plasmid mutant libraries and posttransfection and postinfection viral libraries was performed. (B) Correlation of the relative frequency of each single-nucleotide mutation between biological duplicates. (C) RF scores of individual mutations of influenza A/WSN/33 virus segment 2 in $\log_{10}$. Two representative regions are zoomed in on to show the single nucleotide change.

Fig. S1C). The library covered 94.9% of the nucleotides in segment 2 and included 98.2% of the single nucleotide mutations of observed positions (see Fig. S2A and C). To further improve the accuracy of fitness quantification, we focused on the mutations that make up >0.1% of the plasmid mutant library. After this quality control, we were still able to observe 94.2% of the nucleotide positions with 63.9% of the single nucleotide mutations. More than 82% of the nucleotide positions were covered with two or three nucleotide mutations (see Fig. S2B and D in the supplemental material). To assess the quality and reproducibility of our mutant library, we compared the relative frequencies of single mutations between biological replicates. We obtained a strong Spearman correlation coefficient of 0.93 for two independent transfections and 0.75 for infections (Fig. 2B). A relative fitness (RF) index was calculated for individual mutations as the ratio of relative frequency in the infection library to that in the input DNA library. The profiling data of all of segment 2 are shown in Fig. 2C, where most of the mutations had a fitness cost ($\log_{10}$ RF index of <0).

Systematic identification of deleterious mutations of the PB1 protein. Segment 2 of influenza A virus encoded three proteins: PB1, PB1-F2, and N40. N40 was a truncated form of the PB1 protein that lacked the first 39 amino acids. PB1-F2 is not essential for viral replication *in vitro*, as completely abolishing PB1-F2 expression had no effect on viral growth (40, 41) (see Fig. S3 in the supplemental material). So we focused on the PB1 protein for downstream analysis. The RF indexes of silent mutations were considered an internal quality control since most, if not all, of them were expected to have a growth capacity comparable to that of the wild type. In the fitness profile of the PB1 protein, the RF indexes of silent mutations followed a normal distribution with a mean of 0.9 and were significantly higher than those of nonsense mutations (two-tailed *t* test, $P = 4.6E\text{-}21$) (see Fig. S4 in the supplemental material). This result confirms the presence of fitness selection and validates the data quality.

To systematically identify deleterious mutations, we chose a stringent RF index cutoff of ≤0.1. A total of 2.4 percentage points of silent mutations fell below the cutoff, which represented type I error. A total of 43.1 percentage points of missense mutations that satisfied this cutoff were identified as deleterious mutations (Fig. 3A). We randomly selected 14 deleterious mutations and reconstructed them individually. Rescue experiments were performed, and the resultant viral titers were quantified by 50% tissue culture infective dose ($TCID_{50}$) assay. Thirteen of 14 mutant viruses had at least a 10-fold drop in the viral titer compared to that of the wild type. The other mutant also showed a more-than-6-

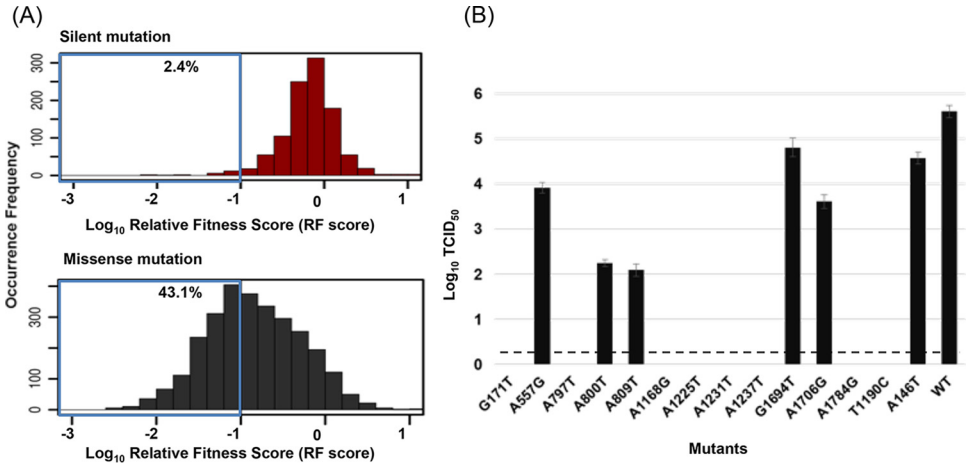

FIG 3 Systematic identification of deleterious mutations of the PB1 protein. (A) Histogram illustrations of the RF distribution (RF index in $\log_{10}$) of silent and missense mutations. Mutations with an RF index of ≤0.1 were identified as deleterious mutations. The percentages of silent and missense mutations that fall below this cutoff are boxed in blue. (B) Fourteen deleterious mutations were selected and reconstructed in the viral genome. The $TCID_{50}$s of selected single nucleotide mutations are shown. The dashed line represents the detection limit of the $TCID_{50}$ assay. Data are presented as mean values ± standard deviations of biological duplicates. WT, wild type.

fold titer decrease (Fig. 3B). These results validated the approach we used to systematically quantify the RF and identify deleterious mutations of the PB1 protein.

Identifying functional residues by dissecting structural and functional constraints. A mutation might be deleterious because of structural or functional constraints (16, 42). We have recently demonstrated that coupling high-throughput genetics with mutant stability predictions can identify residues that are dominated by functional constraints (16). Briefly, deleterious mutations that do not destabilize the protein are identified as functional residues. Here, we modeled protein stability by using two computational tools: I-Mutant and Rosetta ddg monomer (see Data Set S1 in the supplemental material).

I-Mutant is a supporter vector machine-based software used to predict the effect of single-site mutations on protein stability (ΔΔG) (43–45). On the basis of the predicted ΔΔG, mutations can be classified as destabilizing (ΔΔG, ≤ -0.5), neutral ($-0.5 <$ ΔΔG < 0.5), or stabilizing (ΔΔG, ≥ 0.5). We applied I-Mutant predictions for all missense mutations in PB1 with the structure resolved from the bat influenza A virus polymerase complex (Protein Data Bank [PDB] code 4WSB) (46, 47). Of the mutations for which structure information is available, 64.5% were shown to be destabilizing, 33.5% were neutral, and 2% were stabilizing (see Fig. S5A in the supplemental material). As expected, destabilizing mutations had a significantly small solvent-accessible surface area (SASA) (48–50) (see Fig. S5B). To further reduce the rate of false-negative functional residue identification, we performed protein stability prediction with Rosetta for all deleterious mutations (16, 42, 44). Unlike the machine learning algorithm used by I-mutant, Rosetta generated structural models for single amino acid mutations based on a preoptimized wild-type structure. With a high-resolution protocol, 50 models of wild-type and mutant protein structures were generated and the three lowest ΔΔG values were averaged on the basis of optimized rotamers. The absolute correlation coefficient of the predictions that resulted from these two methods was 0.3 (see Fig. S5C). Aiming at getting a conserved classification of functional residues, we classified a residue as functional if it had one or more missense mutations satisfying both the deleterious RF index cutoff and nondestabilizing criteria of ΔΔG predictions from either software. We identified 297 residues as functional.

To examine the sensitivity of our method of identifying functional residues in PB1, we performed a thorough literature search, compiled 31 residues that were reported to be functional in PB1 (32, 51–54), and compared the performance of our method with that of four other methods: FireStar, Frpreq, Consurf, and Concavity (6, 10, 55–58) (Table 1). Our method was able to identify 21 of the 31 residues and thus had a sensitivity of ~68%. FireStar failed to identify any of them. Frprep, Concavity, and Consurf identified 4 (Frprep score, ≥8), 7 (Concavity score, >0.1), and 17 (Consurf score, 9) residues, respectively. Notably, our method was the only one that identified functional residues related to noncanonical polymerase functions (four of the eight residues) that were not conserved in sequence or structure. Overall, these results validated our method of combining high-throughput genetics with mutant stability prediction to identify functional residues in PB1 in a sensitive and unbiased manner (16, 42, 44).

Annotating functional residues by homologous structural alignment. The vRdRp family has a conserved "right-handed" structure. It consists of three major conserved domains (finger, palm, and thumb) and six motifs (pre-A/F and A to E) (20). Since canonical vRdRp functional residues of the PB1 protein are expected to be structurally conserved, they aligned well with other protein structures from the vRdRp family. Therefore, homologous structural alignment might enable us to further annotate PB1 residues by distinguishing canonical and noncanonical vRdRp functional residues. The recent improvement of algorithms provides opportunities for more accurate structure comparison. Here we used TM-align and 3DCOMB for pairwise and multiple structure alignments (MSAs) (59–61). Both softwares use TM-score to quantify protein structural similarity, which is robust to local structural variation and is protein length independent (59, 60). Moreover, 3DCOMB takes into account both local and global

TABLE 1 Comparison of methods of identification of known functional PB1 residues

Mutation	Functional annotation	Our method	FireStar	Frpred	Consurf	ConCavity
L8	Interact with PA	0	0	1	3	0
F9	Interact with PA	0	0	1	3	1.40E-6
L10	Interact with PA	0	0	1	6	0
K11	Interact with PA	1	0	1	5	0
M179	Polymerase activity	0	0	2	4	4.40E-8
K188	Nuclear localization	1	0	2	6	0
R189	Nuclear localization	1	0	1	3	0
R208	Nuclear localization	1	0	1	1	0
K209	Nuclear localization	0	0	2	3	0
K229	Polymerase activity	1	0	7	9	0.288
R233	Polymerase activity	0	0	7	9	0.044
K235	Polymerase activity	1	0	7	9	0.682
R238	Polymerase activity	1	0	7	9	0.201
R239	Polymerase activity	0	0	7	9	0.187
K278	Polymerase activity	1	0	6	9	0.022
K279	Polymerase activity	1	0	6	9	1.08E-5
N306	Polymerase activity	1	0	6	8	0.437
K308	Polymerase activity	1	0	6	9	0.027
M409	Polymerase activity	1	0	9	9	0.829
Q442	Polymerase activity	1	0	4	9	0.653
S444	Polymerase activity	1	0	7	9	0.009
D445	Polymerase activity	1	0	6	9	0.001
D446	Polymerase activity	1	0	8	9	5.25E-6
N476	Polymerase activity	1	0	7	9	0.008
S478	Polymerase activity	0	0	7	9	0.011
K481	Polymerase activity	1	0	8	9	0
Y483	Polymerase activity	1	0	4	8	0
E491	Polymerase activity	1	0	8	9	0.028
F492	Polymerase activity	1	0	6	8	0.001
F496	Polymerase activity	0	0	5	8	0.001

features, which is suitable for alignment of distantly related protein structures (61).

Twenty representative vRdRp structures were selected from positive single-stranded RNA (ssRNA) viruses, negative ssRNA viruses, and double-stranded RNA (dsRNA) virus families on the basis of previously stated criteria (20). Briefly, representative structures were selected from each of the Baltimore classes that encoded vRdRp, including positive ssRNA viruses (*Caliciviridae*, *Flaviviridae*, *Picornaviridae*, *Cystoviridae*), dsRNA viruses (*Birnaviridae*, *Cystoviridae*, and *Reoviridae*), and negative ssRNA viruses (*Bunyaviridae*) (62–81). Structures with no mutations and with a bound substrate were preferred. PDB files with the highest resolution were picked for each protein (see Table S1 in the supplemental material).

To ensure sufficient structural similarity, a pairwise structural comparison was performed with the selected protein and PB1 by using TM-align. The structures with TM scores of >0.5 were kept for multiple structural alignment, which generally indicated similar protein folding (43). Figure S6A in the supplemental material provides an example superimposition of the PB1 protein with HCV NS5B (PDB code 2XI3) with decent alignment in major protein domains (67). A total of 16 proteins were included for MSA with PB1 by using 3DCOMB (see Table S1 in the supplemental material).

The root mean square deviation (RMSD), the measurement of the average distance between the atoms and superimposed proteins, was reported by 3DCOMB for each residue as the representative of structure conservation. As the reported aligned residues had RMSD scores ceiled at 9, we assigned the residues that did not align among structures with an RMSD value of 10 (Fig. 4A). Low RMSDs meant that the residues were conserved in the vRdRp family and thus more likely to have canonical vRdRp functions. As expected, the structurally conserved residues were less tolerant of mutations. The average RF index of structurally conserved residues was significantly lower than that of nonconserved residues (two-tailed *t* test, $P = 0.0006$, Fig. 4B). The RMSDs of all of the identified functional residues of the PB1 protein were plotted. A smooth curve of RMSDs was fitted by local polynomial (loess) regression. We could clearly identify the six conserved domains (pre-A/F and A to E) of vRdRp as valleys on the smooth curve (Fig. 4C). These results demonstrated the feasibility of using homologous structural alignment to identify canonical vRdRp residues.

Identification of noncanonical functional residues, ones involved in nuclear import of the PB1 protein. Forty-three percent of the functional residues identified could not be aligned with other protein structures from the vRdRp family. Although this could be due to poor alignment quality, it is also possible that these residues have noncanonical functions that are essential for viral growth. Interestingly, 62% of these residues belong to the protein interface between PB1 and PB2 or PA, as identified by the change in SASA upon complex formation by using Sppider (residues with at least a 4% decrease in SASA and >5 Å^2 of exposed surface area upon complex formation) (82) (see Fig. S6B in the supplemental material). These interface residues also accounted for some of the peaks (residue 50 to 80, residues 350 to 400, and residues at the C terminus of PB1) in the smooth RMSD curve of functional residues in Fig. 3C.

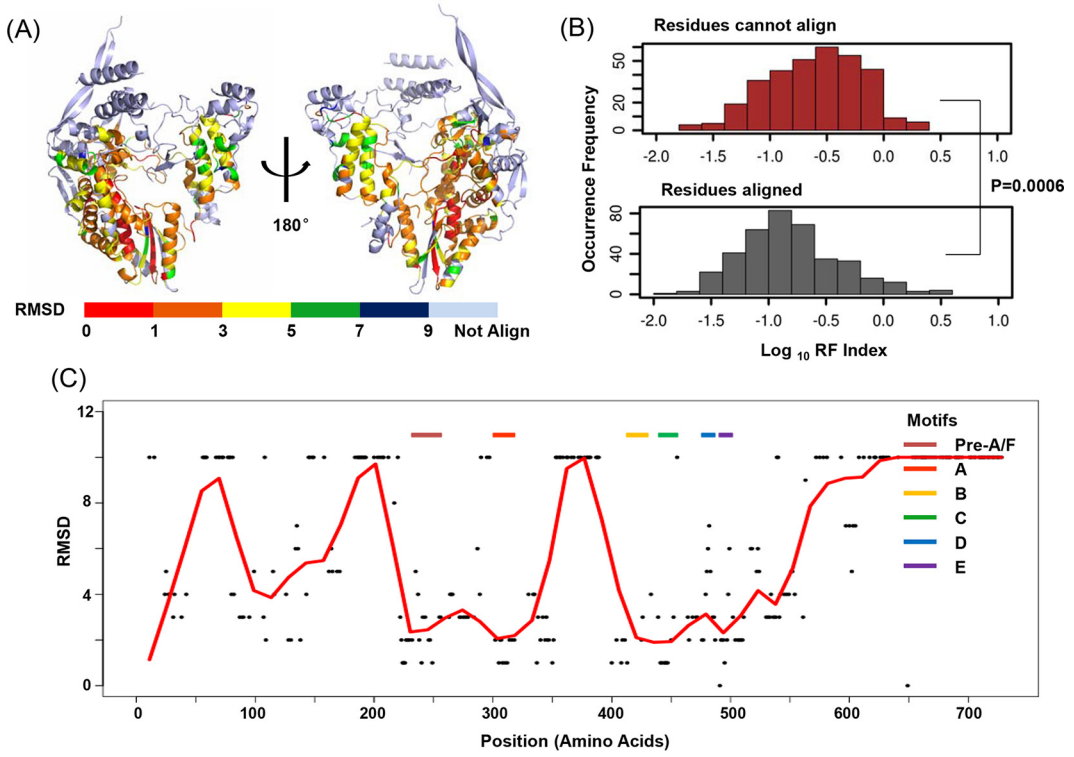

FIG 4 Annotation of PB1 functional residues with homologous structural alignment. (A) An MSA was performed with PB1 and 16 other homologous structures in the vRdRp family. The PB1 structure is rainbow colored according to the RMSD of each residue. (B) Histograms of the RF indexes are shown for residues that cannot be aligned (red) and residues that can be aligned with other structures in the vRdRp family. The RF indexes of residues that cannot be aligned were significantly higher (two-tailed *t* test, $P = 0.0006$). (C) RMSDs of functional residues. A smooth curve was fitted by loess regression. Conserved vRdRp domains (pre-A/F and A to E) are labeled and shown as valleys on the smooth RMSD curve.

We then performed a detailed analysis of the noncanonical functional residues that were not located in the heterotrimer-forming interface. When mapped onto the protein structure, some of them (residues 180 to 220) formed a noticeable cluster (Fig. 5A and B). This clustered region is unique to the PB1 protein, which consists of a long twisted β-ribbon connected by a non-structured loop (47). It protrudes from the polymerase complex structure and is fully solvent exposed. Two nuclear localization signals (NLSs) were reported in the β-ribbon region (amino acids 187 to 190 and 207 to 210) to mediate PB1 nuclear import through interaction with RanBP5 (32, 83). Nonetheless, the function of this loop region is not completely clear. It is suspected to interact with the viral genome in the resolved influenza B and C virus structures (46, 47, 84), and K198 of influenza A virus was suggested to be related to host adaptation (85). As the density of the loop region (residues 195 to 198) is missing from the influenza A virus polymerase crystal structure, we used kinematic loop modeling in the Rosetta software to computationally reconstruct the loop region (86). From the above-described analysis, D193 in the loop region was identified as a noncanonical functional residue. Interestingly, it was the only negatively charged residue located within a highly positively charged environment. It was 100% conserved among all of the human influenza A virus PB1 sequences from the Influenza Research Database under purifying selection (ratio of nonsynonymous to synonymous evolutionary changes [*dN*/*dS* ratio] of 0.015) (87–89, 113) (see Table S2 in the supplemental material). Two positively charged residues (K197, K198) located on the side opposite D193 in the loop region were also highly conserved in human influenza A viruses (>99%) and possibly interact with D193. Although they were not classified as essential residues according to our high-throughput fitness profile, their mutations in charges (K197E, K198E) resulted in a >6-fold drop in the RF index. To examine if the loop region has possible noncanonical functions, we introduced single substitutions (D193G, K197E, K198E) and double substitutions (D193G-K197E, G193G-K198E, K197E-K198E) into the PB1 protein. We also constructed mutant versions with substitutions in the NLS region (K188A-R189A, R208A-K209A) and mutant versions that decreased the polymerase activity (W55R, H184R, H47L, Q268L) as controls. Of note, all of the controls were identified as deleterious in our high-throughput fitness profile. Viral production of all of the mutant versions was measured by $TCID_{50}$ assay with viral rescue experiments. D193G, D193G-K197E, G193G-K198E, K197E-K198E, and the reported substitutions in the NLS region (K188A-R189A) had severe impacts on viral production, with no detectable viral titer posttransfection (Fig. 5C) (32). Consistently, these mutations also resulted in a significantly lower viral growth rate in A549 cells (Fig. 5D). To examine the vRdRp function of these mutations, we used a minigenome replicon assay by cotransfecting a virus-inducible luciferase reporter and polymerase segments (PB2, PB1, PA, NP) in 293T cells. The reported mutant NLS (K188A-R189A), which was highly deleterious for viral replication, still had ~50% polymerase activity in the minigenome replicon assay. Similarly, D193G and all of the double substitutions

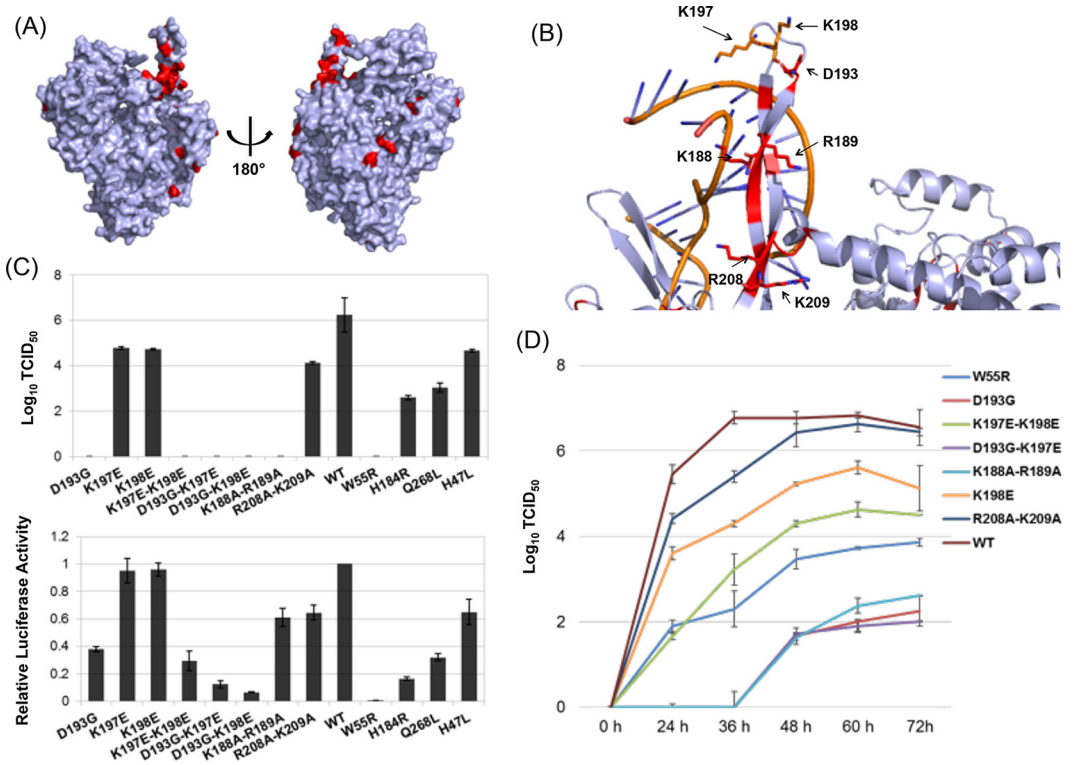

FIG 5 Identification of noncanonical functional residues of the PB1 protein. (A and B) Noncanonical noninterface functional residues of the PB1 protein are red. A cluster of residues is located in the long twisted β-ribbon region. The nonstructured loop region (amino acids 195 to 198) was reconstructed with Rosetta. (C) $TCID_{50}$s (top) and relative polymerase activities (bottom) of the mutations indicated. The data are presented as mean values ± standard deviations of four independent biological replicates. (D) Growth curves of the mutations indicated. A549 cells were infected with the mutant viruses indicated at an MOI of 0.1. Viruses were collected at the time points indicated, and $TCID_{50}$s were measured. WT, wild type.

(D193G-K197E, D193G-K198E, K197E-K198E) showed discordance between vRdRp function and viral growth capacity. Compared with W55R, H184R, H47L, and Q268L, which remained at ~0.1 to 65% polymerase activity, the fitness drop caused by these newly identified loop mutations was much more severe, indicating that they might have a noncanonical polymerase function of PB1 (Fig. 5C).

Unlike other RNA viruses, the genome replication and transcription of influenza virus are performed inside the nucleus. Nuclear localization function is thus specific to influenza virus and belongs to noncanonical functions of the PB1 protein. We tested if the mutations identified in the loop region (D193G, D193G-K197E, K197E-K198E) had effects on protein nuclear import. A549 cells were infected with wild-type and mutant viruses at an MOI of 0.1. Cells were fixed and subjected to immunofluorescence analysis (IFA) at 18 h postinfection. As expected, the PB1 proteins of the wild-type virus were localized mostly in the nucleus. However, the PB1 proteins of mutant viruses were significantly enriched in the cytoplasm, suggesting that these mutations were defective in PB1 protein nuclear import (Fig. 6A and B). More severe defects were observed for double mutations (D193G-K197E, K197E-K198E). Similar results were observed at earlier time points (8 h postinfection) at an MOI of 0.5 (see Fig. S7 in the supplemental material). Interestingly, for those PB1 mutant versions, the nuclear import of PA protein was also delayed, which is consistent with the notion that PA and PB1 are imported into the nucleus as a complex (32, 83, 90, 91) (see Fig. S7).

RanBP5 belongs to the importin-β family, which has a nonclassical nuclear import function (92, 93). RanBP5 has been shown to be important for influenza A virus PB1 nuclear import. The NLS mutations affected protein nuclear import by decreasing binding to RanBP5 (32, 83, 92). Thus, we further tested if mutations in the loop region (D193G, D193G-K197E, and K197E-K198E) would also affect the interaction between PB1 and RanBP5. Immunoprecipitation (IP) was performed by cotransfecting the FLAG-tagged PB1 protein and the hemagglutinin (HA)-tagged RanBP5 protein into 293T cells. Two days later, the total cell lysate was collected and subjected to IP with anti-HA antibody-conjugated beads or IgG-conjugated beads. As shown in Fig. 6B, all three mutant proteins showed decreased binding with RanBP5. Consistent with our IFA results, double mutations (D193G-K197E, K197E-K198E) produced a greater reduction in protein binding. The above-described results indicate that the residues in the loop region are important for nuclear import of the influenza A virus PB1 protein through interaction with RanBP5, which is a noncanonical function in the vRdRp family.

DISCUSSION

For a comprehensive characterization of protein function, identification and annotation of functional residues are the fundamental tasks. Here we present a systematic approach to these tasks by using influenza A virus PB1 as the target protein. Our approach combines high-throughput fitness profiling with mutant stability prediction and homologous structural alignment to identify and

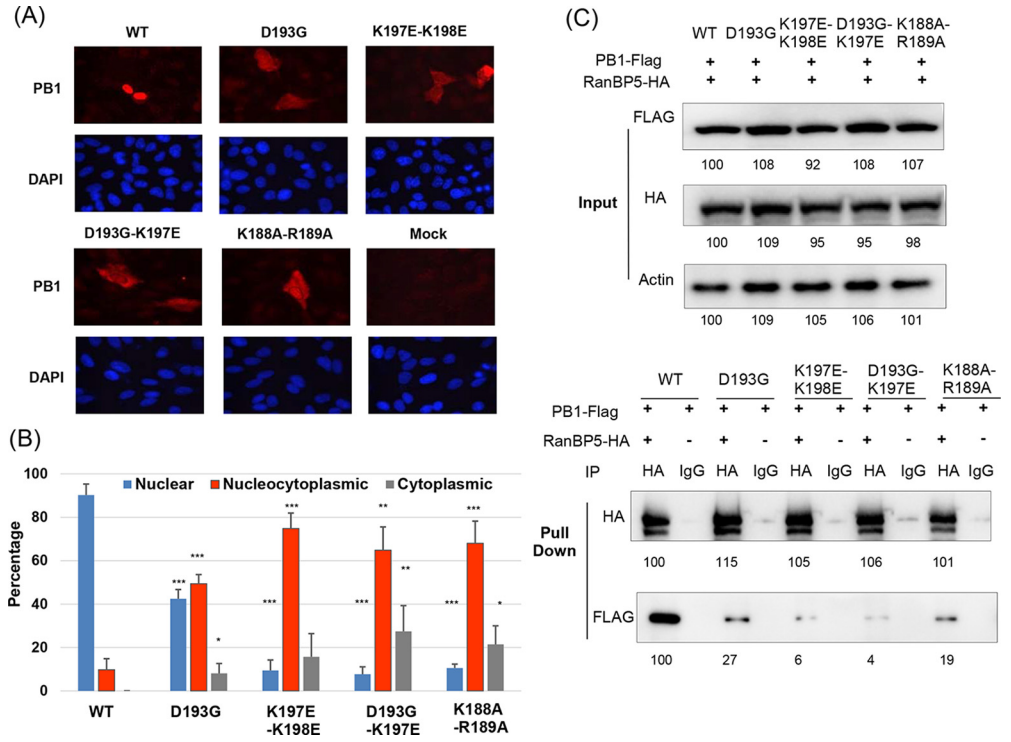

FIG 6 The noncanonical functional residues identified may be involved in nuclear import of the PB1 protein by interaction with RanBP5. (A) Cellular localizations of wild-type (WT) and mutant PB1 proteins determined by IFA. (B) Percentages of cells with different PB1 localizations. Data are presented as mean values ± standard deviations of three independent biological replicates. At least 50 cells of each replicate were analyzed with ImageJ. *, $P < 0.05$; **, $P < 0.01$; ***, $P < 0.001$ (two-tailed t test). (C) Interactions between PB1 proteins and RanBP5 were examined by IP. The value below each band is the intensity quantification measured by Image Lab.

annotate canonical and noncanonical vRdRp functional residues (Fig. 1). Interestingly, we identified a cluster of mutations that were highly deleterious for viral replication but resulted in relatively intact vRdRp function. These mutations were located in the loop region of the PB1 β-ribbon and were shown to be important for PB1 nuclear import. The combination of high-throughput fitness profiling and structural analysis provided a general approach to the identification and annotation of functional residues that can be applied to a wide range of proteins about which homologous structural information is available.

In the context of evolutionary biology, proteins from the same homolog family have an ancestor in common and possess significant sequence and structural similarities (94–97). Structural similarities are postulated to be maintained by functional constraints (98, 99). vRdRps probably evolved from a common ancestor (100). Although their sequence identity is ~20%, they have adopted similar structural domains and use similar catalytic mechanisms (20). Throughout evolution, different proteins also evolved diverse functions to satisfy the needs of specific organisms. Thus, the specific structural motifs that differentiate one protein from homologous proteins may have organism-specific functions. Here we used homologous protein structure information to further annotate the diverse protein functions. Therefore, a multifunctional protein might harbor both canonical (evolutionarily conserved) and noncanonical (organism-specific) functions. The combination of high-throughput genetic screening with homologous-structure analysis enabled us to systematically understand functional residues and important single nucleotide polymorphisms.

Here we show that the residues in the loop region of the PB1 β-ribbon are important for PB1 nuclear import. Unlike other RNA viruses, influenza A virus performs its genome replication inside the nucleus. Thus, the polymerase complex needs to be translocated into the nucleus to perform its function. It is known that PB1 and PA are translocated together as a complex, while PB2 can be translocated by itself (101). RanBP5 is important for the nuclear import of PB1 and PA through direct interaction with PB1. Besides the two reported NLSs, we show that the mutations in the loop region also impact the interaction between PB1 and RanBP5, thus causing the defect in PB1 nuclear import. We do not have direct evidence that the loop region works as a direct NLS or by affecting the nearby NLS regions, but on the basis of the sequence of the loop region, it did not fall into any of the six classes of NLSs (32, 102). Thus, we suspected that this region affected PB1 nuclear import by affecting the nearby NLS regions. In agreement with previous observations, there seems to be no clear consensus sequence that is responsible or important for RanBP5 binding (32, 103). The detailed mechanism needs to be further defined.

Genetic studies are greatly facilitated by the improvement of sequencing capacity and the growing number of protein structures being resolved. Large amounts of information generated with current technologies demand more effective approaches to determine structure-function relationships. Coupling mutagenesis with high-throughput sequencing, high-throughput fitness profiling provides a sensitive and unbiased way to identify the essential residues of targeted proteins (16, 33–37, 104–107). The same principle applies to other proteins/organisms, as long as the proper functional measurement can be made (37). For example,

we can study the proteins related to cell proliferation by using the cell growth rate as a readout. By using saturated mutagenesis, we can learn which mutation is related to an abnormal cell growth rate and can further use flow cytometry to differentiate cells in different phases. We can also investigate the roles of mutant proteins in cancer metastasis through transwell migration assays *in vitro* or by using mouse xenograft models *in vivo*. The structures of target or homologous proteins can be linked to a genetic profile and further facilitate the understanding of biomolecular functions related to each functional residue. We foresee that this approach will become more powerful as more protein structures are determined at an accelerated rate by crystallography and cryoelectron microscopy and the escalating sequencing technology.

In summary, we have developed a systematic and sensitive method to identify and annotate functional residues. More importantly, the method presented here is generally applicable to other proteins with structural information of homologous proteins.

MATERIALS AND METHODS

Construction of influenza A virus segment 2 mutant libraries. Influenza A/WSN/33 virus segment 2 mutant libraries were generated with the eight-plasmid transfection system (39). In brief, the entire influenza virus gene was separated into nine small 240-bp segments. Random mutagenesis was performed with error-prone polymerase Mutazyme II (Stratagene). For each small library, mutagenesis was performed separately and the amplified segment was gel purified, BsaI digested, ligated to the vector, and transformed with MegaX DH10B T1R cells (Life Technologies). As each small library was expected to have ~1,000 single mutations, ~50,000 bacterial colonies were collected to cover the entirety. Plasmids from collected bacteria were midiprepped as the input DNA library.

Transfection, infection, and viral titer. To generate the mutant viral library, ~30 million 293T cells were transfected with 32 μg of DNA. Transfections were performed with Lipofectamine 2000 (Life Technologies). Virus was collected at 72 h posttransfection. $TCID_{50}$s were measured with A549 cells. To passage viral libraries, ~10 million A549 cells were infected at an MOI of 0.05. Cells were washed with phosphate-buffered saline (PBS) three times at 2 h postinfection. Virus was collected 24 h postinfection from supernatant.

Individual mutant viral plasmids were generated with a quick-change system. To generate mutant virus, ~2 million 293T cells were transfected with 10 μg of DNA. To measure the growth curve, ~1 million A549 cells were infected at an MOI of 0.1 and supernatants were collected at the times indicated.

Sequencing library construction and data analysis. Viral RNA was extracted with the QIAamp Viral RNA Minikit (Qiagen Sciences). DNase I (Life Technologies) treatment was performed, followed by reverse transcription with the SuperScript III system (Life Technologies). At least 10^6 viral copies were used to amplify the mutated segment. The amplified segment was then digested with BpuEI and ligated with the sequencing adaptor, which had three nucleotides multiplexing ID to distinguish between different samples.

Deep sequencing was performed with Illumina sequencing MiSeq PE250. Raw sequencing reads were demultiplexed by using the three-nucleotide ID. Sequencing error was corrected by filtering unmatched forward and reverse reads. Mutations were called by comparing sequencing reads with the wild-type sequence. Clones containing two or more mutations were discarded. The RF index was calculated for individual point mutations, and only mutations that had a frequency of >0.1% in the DNA library were reported. The formula used was $RF\ index_{mutant\ i} = Relative\ Frequency\ of\ Mutant\ i_{infection}/Relative\ Frequency\ of\ Mutant\ i_{plasmid}$, where $Relative\ Frequency\ of\ Mutant\ i = Reads\ of\ Mutant\ i/Reads\ of\ wild\ type$.

All data processing and analysis was performed with customized python scripts, which are available upon request.

Protein structural analysis. Chain B (PB1 protein) of PDB code 4WSB was used for protein ΔΔG prediction with single amino acid mutations (46, 47). ΔΔG predictions were performed with both the I-Mutant 2.0 package and ddg_monomer in the Rosetta software (43, 108). Default parameters (temperature of 25°C, pH 7.0) were used in the I-Mutant package. The parameters used for Rosetta were the same as those previous described (16, 109). A ΔΔG of <0 in I-Mutant and a ΔΔG of >0 in Rosetta mean destabilization.

The DSSP tool was used to calculated SASA, which was then normalized to the empirical scale as previously described (48–50). Sppider was used to identify the protein-protein interface. Residues with at least a 4% reduction and a >5-Å^2 reduction in SASA upon complex formation were identified as protein-protein interface residues (82).

TM-align and 3DCOMB were used for pairwise structural alignment and multiple structural alignment (59, 61). TM-score normalized to the PB1 protein was used.

Protein loop modeling. In the loop region of the PB1 β-ribbon, electron density for residues 195 to 198 is missing from the X-ray crystal structure (PDB code 4WSB). Rosetta software was used to computationally reconstruct the loop region, which was based on Monte Carlo sampling with exact kinematic loop closure (86). After energy optimization, each model was ranked by Rosetta full atom energy function (80). The lowest-energy model with a hairpin-like loop was selected.

Polymerase activity assay. One hundred nanograms each of PB2, PB1 (wild type and indicated mutations), PA, and NP; 50 ng of a virus-inducible luciferase reporter; and 5 ng of PGK-*Renilla* luciferase were transfected into 293T cells in 24-well plates (110). Cells were lysed at 24 h posttransfection, and luciferase assay was measured with the Dual-Luciferase Assay kit (Promega).

IFA. The localizations of wild-type PB1 and mutant PB1 proteins were determined by Immunofluorescence analysis (IFA). Infected A549 cells were fixed in 2% paraformaldehyde, permeabilized with 0.1% Triton X-100, and then blocked with 3% bovine serum albumin and 10% fetal bovine serum. Viral PB1 protein was detected with anti-PB1 antibody (GeneTex GTX125923). Hoechst 33342 dye was used for nucleic acid staining.

IP. Immunoprecipitation (IP) experiments were performed with HA- and FLAG-tagged proteins expressed in 293T cells. Briefly, cells were transfected with corresponding expression plasmids with Lipofectamine 2000 reagents (Invitrogen) and lysed at 2 days posttransfection with radioimmunoprecipitation assay (RIPA) buffer (50 mM Tris-HCl [pH 7.4], 0.5% NP-40, 150 mM KCl, 1 mM EDTA, protease inhibitor). Cell lysates were incubated with 1 μg of anti-HA antibody for 4 h at 4°C with constant agitation, washed with RIPA buffer five times, and eluted with 60 μl of SDS-PAGE sample buffer. All samples were subjected to SDS-PAGE and Western blotting.

Western blotting. Proteins in SDS-PAGE sample buffer were heated at 95°C, resolved by SDS-PAGE, and then transferred onto polyvinylidene difluoride membrane. Proteins were detected with antibodies against FLAG-epitope, HA-epitope, or actin.

Phylogenetic analysis. PB1 coding sequences were downloaded from the Influenza Research Database (87). Multiple sequence alignment was performed with MUSCLE (88). We randomly sampled 3,000 sequences for *dN*/*dS* calculation by Fubar with HyPhy (89).

Accession number(s). Raw sequencing data have been submitted to the NIH Short Read Archive under accession number PRJNA318707.

SUPPLEMENTAL MATERIAL

Supplemental material for this article may be found at http://mbio.asm.org/lookup/suppl/doi:10.1128/mBio.01801-16/-/DCSupplemental.

Figure S1, TIF file, 22.9 MB.
Figure S2, TIF file, 22.9 MB.
Figure S3, TIF file, 22.9 MB.
Figure S4, TIF file, 22.9 MB.

Figure S5, TIF file, 23.7 MB.
Figure S6, TIF file, 23.7 MB.
Figure S7, TIF file, 23.6 MB.
Table S1, DOCX file, 0.03 MB.
Table S2, DOCX file, 0.01 MB.
Data Set S1, XLS file, 0.8 MB.

ACKNOWLEDGMENTS

Y.D. was supported by a Philip Whitcome Predoctoral Fellowship and a UCLA Dissertation Year Fellowship. N.C.W. was supported by a Philip Whitcome Predoctoral Fellowship and an Audree Fowler Fellowship in Protein Science.

FUNDING INFORMATION

This work, including the efforts of Yushen Du, was funded by Philip Whitcome Pre-Doctoral Fellowship. This work, including the efforts of Yushen Du, was funded by UCLA Dissertation Year Fellowship. This work, including the efforts of Nicholas C. Wu, was funded by Philip Whitcome Pre-Doctoral Fellowship. This work, including the efforts of Nicholas C. Wu, was funded by Audree Fowler Fellowship in Protein. This work, including the efforts of T Wu and R Sun, was funded by NIH (CA177322).

The funders had no role in study design, data collection and interpretation, or the decision to submit the work for publication.

REFERENCES

1. **Mills CL, Beuning PJ, Ondrechen MJ**. 2015. Biochemical functional predictions for protein structures of unknown or uncertain function. Comput Struct Biotechnol J **13:**182–191. http://dx.doi.org/10.1016/j.csbj.2015.02.003.
2. **Gonzalez-Perez A, Mustonen V, Reva B, Ritchie GR, Creixell P, Karchin R, Vazquez M, Fink JL, Kassahn KS, Pearson JV, Bader GD, Boutros PC, Muthuswamy L, Ouellette BF, Reimand J, Linding R, Shibata T, Valencia A, Butler A, Dronov S, Flicek P, Shannon NB, Carter H, Ding L, Sander C, Stuart JM, Stein LD, Lopez-Bigas N**. 2013. Computational approaches to identify functional genetic variants in cancer genomes. Nat Methods **10:**723–729. http://dx.doi.org/10.1038/nmeth.2562.
3. **Aloy P, Querol E, Aviles FX, Sternberg MJ**. 2001. Automated structure-based prediction of functional sites in proteins: applications to assessing the validity of inheriting protein function from homology in genome annotation and to protein docking. J Mol Biol **311:**395–408. http://dx.doi.org/10.1006/jmbi.2001.4870.
4. **Betancourt AJ, Bollback JP**. 2006. Fitness effects of beneficial mutations: the mutational landscape model in experimental evolution. Curr Opin Genet Dev **16:**618–623. http://dx.doi.org/10.1016/j.gde.2006.10.006.
5. **Calabrese R, Capriotti E, Fariselli P, Martelli PL, Casadio R**. 2009. Functional annotations improve the predictive score of human disease-related mutations in proteins. Hum Mutat **30:**1237–1244. http://dx.doi.org/10.1002/humu.21047.
6. **Glaser F, Pupko T, Paz I, Bell RE, Bechor-Shental D, Martz E, Ben-Tal N**. 2003. ConSurf: identification of functional regions in proteins by surface-mapping of phylogenetic information. Bioinformatics **19:**163–164. http://dx.doi.org/10.1093/bioinformatics/19.1.163.
7. **Sankararaman S, Kolaczkowski B, Sjölander K**. 2009. INTREPID: a web server for prediction of functionally important residues by evolutionary analysis. Nucleic Acids Res **37:**W390–W395. http://dx.doi.org/10.1093/nar/gkp339.
8. **Wilkins AD, Bachman BJ, Erdin S, Lichtarge O**. 2012. The use of evolutionary patterns in protein annotation. Curr Opin Struct Biol **22:**316–325. http://dx.doi.org/10.1016/j.sbi.2012.05.001.
9. **Panchenko AR, Kondrashov F, Bryant S**. 2004. Prediction of functional sites by analysis of sequence and structure conservation. Protein Sci **13:**884–892. http://dx.doi.org/10.1110/ps.03465504.
10. **Capra JA, Laskowski RA, Thornton JM, Singh M, Funkhouser TA**. 2009. Predicting protein ligand binding sites by combining evolutionary sequence conservation and 3D structure. PLoS Comput Biol **5:**e1000585. http://dx.doi.org/10.1371/journal.pcbi.1000585.
11. **Tong W, Williams RJ, Wei Y, Murga LF, Ko J, Ondrechen MJ**. 2008. Enhanced performance in prediction of protein active sites with THEMATICS and support vector machines. Protein Sci **17:**333–341. http://dx.doi.org/10.1110/ps.073213608.
12. **Xie L, Bourne PE**. 2007. A robust and efficient algorithm for the shape description of protein structures and its application in predicting ligand binding sites. BMC Bioinformatics **8**(Suppl 4):S9. http://dx.doi.org/10.1186/1471-2105-8-S4-S9.
13. **Skolnick J, Brylinski M**. 2009. FINDSITE: a combined evolution/structure-based approach to protein function prediction. Brief Bioinform **10:**378–391. http://dx.doi.org/10.1093/bib/bbp017.
14. **Pazos F, Sternberg MJ**. 2004. Automated prediction of protein function and detection of functional sites from structure. Proc Natl Acad Sci U S A **101:**14754–14759. http://dx.doi.org/10.1073/pnas.0404569101.
15. **Petrova NV, Wu CH**. 2006. Prediction of catalytic residues using Support Vector Machine with selected protein sequence and structural properties. BMC Bioinformatics **7:**312. http://dx.doi.org/10.1186/1471-2105-7-312.
16. **Wu NC, Olson CA, Du Y, Le S, Tran K, Remenyi R, Gong D, Al-Mawsawi LQ, Qi H, Wu T-T, Sun R**. 2015. Functional constraint profiling of a viral protein reveals discordance of evolutionary conservation and functionality. PLoS Genet **11:**e1005310. http://dx.doi.org/10.1371/journal.pgen.1005310.
17. **te Velthuis AJ**. 2014. Common and unique features of viral RNA-dependent polymerases. Cell Mol Life Sci **71:**4403–4420. http://dx.doi.org/10.1007/s00018-014-1695-z.
18. **Shatskaya GS, Dmitrieva TM**. 2013. Structural organization of viral RNA-dependent RNA polymerases. Biochemistry (Mosc) **78:**231–235. http://dx.doi.org/10.1134/S0006297913030036.
19. **Ortín J, Parra F**. 2006. Structure and function of RNA replication. Annu Rev Microbiol **60:**305–326. http://dx.doi.org/10.1146/annurev.micro.60.080805.142248.
20. **Černý J, Černá Bolfíková B, Valdés JJ, Grubhoffer L, Růžek D**. 2014. Evolution of tertiary structure of viral RNA dependent polymerases. PLoS One **9:**e96070. http://dx.doi.org/10.1371/journal.pone.0096070.
21. **Bruenn JA**. 2003. A structural and primary sequence comparison of the viral RNA-dependent RNA polymerases. Nucleic Acids Res **31:**1821–1829. http://dx.doi.org/10.1093/nar/gkg277.
22. **Campagnola G, McDonald S, Beaucourt S, Vignuzzi M, Peersen OB**. 2015. Structure-function relationships underlying the replication fidelity of viral RNA-dependent RNA polymerases. J Virol **89:**275–286. http://dx.doi.org/10.1128/JVI.01574-14.
23. **Wang QM, Hockman MA, Staschke K, Johnson RB, Case KA, Lu J, Parsons S, Zhang F, Rathnachalam R, Kirkegaard K, Colacino JM, Al WET, Irol JV**. 2002. Oligomerization and cooperative RNA synthesis activity of hepatitis C virus RNA-dependent RNA polymerase. J Virol **76:**3865–3872. http://dx.doi.org/10.1128/JVI.76.8.3865-3872.2002.
24. **Gao L, Aizaki H, He J-W, Lai MM**. 2004. Interactions between viral nonstructural proteins and host protein hVAP-33 mediate the formation of hepatitis C virus RNA replication complex on lipid raft. J Virol **78:**3480–3488. http://dx.doi.org/10.1128/JVI.78.7.3480-3488.2004.
25. **König R, Stertz S, Zhou Y, Inoue A, Hoffmann H-H, Bhattacharyya S, Alamares JG, Tscherne DM, Ortigoza MB, Liang Y, Gao Q, Andrews SE, Bandyopadhyay S, De Jesus P, Tu BP, Pache L, Shih C, Orth A, Bonamy G, Miraglia L, Ideker T, García-Sastre A, Young JAT, Palese P, Shaw ML, Chanda SK**. 2010. Human host factors required for influenza virus replication. Nature **463:**813–817. http://dx.doi.org/10.1038/nature08699.
26. **Brass AL, Dykxhoorn DM, Benita Y, Yan N, Engelman A, Xavier RJ, Lieberman J, Elledge SJ**. 2008. Identification of host proteins required for HIV infection through a functional genomic screen. Science **319:**921–926. http://dx.doi.org/10.1126/science.1152725.
27. **Karlas A, Machuy N, Shin Y, Pleissner K-P, Artarini A, Heuer D, Becker D, Khalil H, Ogilvie LA, Hess S, Mäurer AP, Müller E, Wolff T, Rudel T, Meyer TF**. 2010. Genome-wide RNAi screen identifies human host factors crucial for influenza virus replication. Nature **463:**818–822. http://dx.doi.org/10.1038/nature08760.
28. **Varga ZT, Grant A, Manicassamy B, Palese P**. 2012. Influenza virus protein PB1-F2 inhibits the induction of type I interferon by binding to MAVS and decreasing mitochondrial membrane potential. J Virol **86:**8359–8366. http://dx.doi.org/10.1128/JVI.01122-12.
29. **Varga ZT, Ramos I, Hai R, Schmolke M, García-Sastre A, Fernandez-Sesma A, Palese P**. 2011. The influenza virus protein PB1-F2 inhibits the induction of type I interferon at the level of the

MAVS adaptor protein. PLoS Pathog **7:**e1002067. http://dx.doi.org/10.1371/journal.ppat.1002067.

30. **Menachery VD, Eisfeld AJ, Schäfer A, Josset L, Sims AC, Proll S, Fan S, Li C, Neumann G, Tilton SC, Chang J, Gralinski LE, Long C, Green R, Williams CM, Weiss J, Matzke MM, Webb-Robertson BJ, Schepmoes AA, Shukla AK, Metz TO, Smith RD, Waters KM, Katze MG, Kawaoka Y, Baric RS.** 2014. Pathogenic influenza viruses and coronaviruses utilize similar and contrasting approaches to control interferon-stimulated gene responses. mBio **5:**e01174–e01114. http://dx.doi.org/10.1128/mBio.01174-14.
31. **Aevermann BD, Pickett BE, Kumar S, Klem EB, Agnihothram S, Askovich PS, Bankhead A, Bolles M, Carter V, Chang J, Clauss TRW, Dash P, Diercks AH, Eisfeld AJ, Ellis A, Fan S, Ferris MT, Gralinski LE, Green RR, Gritsenko Ma, Hatta M, Heegel Ra, Jacobs JM, Jeng S, Josset L, Kaiser SM, Kelly S, Law GL, Li C, Li J, Long C, Luna ML, Matzke M, McDermott J, Menachery V, Metz TO, Mitchell H, Monroe ME, Navarro G, Neumann G, Podyminogin RL, Purvine SO, Rosenberger CM, Sanders CJ, Schepmoes AA, Shukla AK, Sims A, Sova P, Tam VC, Tchitchek N, et al.** 2014. A comprehensive collection of systems biology data characterizing the host response to viral infection. Sci Data **1:**140033. http://dx.doi.org/10.1038/sdata.2014.33.
32. **Hutchinson EC, Orr OE, Man Liu S, Engelhardt OG, Fodor E.** 2011. Characterization of the interaction between the influenza A virus polymerase subunit PB1 and the host nuclear import factor Ran-binding protein 5. J Gen Virol **92:**1859–1869. http://dx.doi.org/10.1099/vir.0.032813-0.
33. **Wu NC, Young AP, Al-Mawsawi LQ, Olson CA, Feng J, Qi H, Luan HH, Li X, Wu T-T, Sun R.** 2014. High-throughput identification of loss-of-function mutations for anti-interferon activity in the influenza A virus NS segment. J Virol **88:**10157–10164. http://dx.doi.org/10.1128/JVI.01494-14.
34. **Qi H, Olson CA, Wu NC, Ke R, Loverdo C, Chu V, Truong S, Remenyi R, Chen Z, Du Y, Su S-Y, Al-Mawsawi LQ, Wu T-T, Chen S-H, Lin C-Y, Zhong W, Lloyd-Smith JO, Sun R.** 2014. A quantitative high-resolution genetic profile rapidly identifies sequence determinants of hepatitis C viral fitness and drug sensitivity. PLoS Pathog **10:**e1004064. http://dx.doi.org/10.1371/journal.ppat.1004064.
35. **Wu NC, Young AP, Al-Mawsawi LQ, Olson CA, Feng J, Qi H, Chen S-H, Lu I-H, Lin C-Y, Chin RG, Luan HH, Nguyen N, Nelson SF, Li X, Wu T-T, Sun R.** 2014. High-throughput profiling of influenza A virus hemagglutinin gene at single-nucleotide resolution. Sci Rep **4:**4942. http://dx.doi.org/10.1038/srep04942.
36. **Stiffler MA, Hekstra DR, Ranganathan R.** 2015. Evolvability as a function of purifying selection in article evolvability as a function of purifying selection in TEM-1 β-lactamase. Cell **160:**882–892. http://dx.doi.org/10.1016/j.cell.2015.01.035.
37. **Fowler DM, Fields S.** 2014. Deep mutational scanning: a new style of protein science. Nat Methods **11:**801–807. http://dx.doi.org/10.1038/nmeth.3027.
38. **Heaton NS, Sachs D, Chen C-J, Hai R, Palese P.** 2013. Genome-wide mutagenesis of influenza virus reveals unique plasticity of the hemagglutinin and NS1 proteins. Proc Natl Acad Sci U S A **110:**20248–20253. http://dx.doi.org/10.1073/pnas.1320524110.
39. **Hoffmann E, Neumann G, Kawaoka Y, Hobom G, Webster RG.** 2000. A DNA transfection system for generation of influenza A virus from eight plasmids. Proc Natl Acad Sci U S A **97:**6108–6113. http://dx.doi.org/10.1073/pnas.100133697.
40. **Chen W, Calvo PA, Malide D, Gibbs J, Schubert U, Bacik I, Basta S, O'Neill R, Schickli J, Palese P, Henklein P, Bennink JR, Yewdell JW.** 2001. A novel influenza A virus mitochondrial protein that induces cell death. Nat Med **7:**1306–1312. http://dx.doi.org/10.1038/nm1201-1306.
41. **Zamarin D, Ortigoza MB, Palese P.** 2006. Influenza A virus PB1-F2 protein contributes to viral pathogenesis in mice. J Virol **80:**7976–7983. http://dx.doi.org/10.1128/JVI.00415-06.
42. **Cheng G, Qian B, Samudrala R, Baker D.** 2005. Improvement in protein functional site prediction by distinguishing structural and functional constraints on protein family evolution using computational design. Nucleic Acids Res **33:**5861–5867. http://dx.doi.org/10.1093/nar/gki894.
43. **Capriotti E, Fariselli P, Casadio R.** 2005. I-Mutant2.0: predicting stability changes upon mutation from the protein sequence or structure. Nucleic Acids Res **33:**W306–W310. http://dx.doi.org/10.1093/nar/gki375.
44. **Potapov V, Cohen M, Schreiber G.** 2009. Assessing computational methods for predicting protein stability upon mutation: good on average but not in the details. Protein Eng Des Sel **22:**553–560. http://dx.doi.org/10.1093/protein/gzp030.
45. **Thiltgen G, Goldstein RA.** 2012. Assessing predictors of changes in protein stability upon mutation using self-consistency. PLoS One **7:**e46084. http://dx.doi.org/10.1371/journal.pone.0046084.
46. **Reich S, Guilligay D, Pflug A, Malet H, Berger I, Crépin T, Hart D, Lunardi T, Nanao M, Ruigrok RW, Cusack S.** 2014. Structural insight into cap-snatching and RNA synthesis by influenza polymerase. Nature **516:**361–366. http://dx.doi.org/10.1038/nature14009.
47. **Pflug A, Guilligay D, Reich S, Cusack S.** 2014. Structure of influenza A polymerase bound to the viral RNA promoter. Nature **516:**355–360. http://dx.doi.org/10.1038/nature14008.
48. **Joosten RP, Te Beek TA, Krieger E, Hekkelman ML, Hooft RW, Schneider R, Sander C, Vriend G.** 2011. A series of PDB related databases for everyday needs. Nucleic Acids Res **39:**D411–D419. http://dx.doi.org/10.1093/nar/gkq1105.
49. **Kabsch W, Sander C.** 1983. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. Biopolymers **22:**2577–2637. http://dx.doi.org/10.1002/bip.360221211.
50. **Tien MZ, Meyer AG, Sydykova DK, Spielman SJ, Wilke CO.** 2013. Maximum allowed solvent accessibilities of residues in proteins. PLoS One **8:**e80635. http://dx.doi.org/10.1371/journal.pone.0080635.
51. **Chu C, Fan S, Li C, Macken C, Kim JH, Hatta M, Neumann G, Kawaoka Y.** 2012. Functional analysis of conserved motifs in influenza virus PB1 protein. PLoS One **7:**e36113. http://dx.doi.org/10.1371/journal.pone.0036113.
52. **Li C, Wu A, Peng Y, Wang J, Guo Y, Chen Z, Zhang H, Wang Y, Dong J, Wang L, Qin FX, Cheng G, Deng T, Jiang T.** 2014. Integrating computational modeling and functional assays to decipher the structure-function relationship of influenza virus PB1 protein. Sci Rep **4:**7192. http://dx.doi.org/10.1038/srep07192.
53. **Perez DR, Donis RO.** 2001. Functional analysis of PA binding by influenza A virus PB1: effects on polymerase activity and viral infectivity. J Virol **75:**8127–8136. http://dx.doi.org/10.1128/JVI.75.17.8127-8136.2001.
54. **Jung TE, Brownlee GG.** 2006. A new promoter-binding site in the PB1 subunit of the influenza A virus polymerase. J Gen Virol **87:**679–688. http://dx.doi.org/10.1099/vir.0.81453-0.
55. **López G, Valencia A, Tress ML.** 2007. Firestar—prediction of functionally important residues using structural templates and alignment reliability. Nucleic Acids Res **35:**W573–W577. http://dx.doi.org/10.1093/nar/gkm297.
56. **Fischer JD, Mayer CE, Söding J.** 2008. Prediction of protein functional residues from sequence by probability density estimation. Bioinformatics **24:**613–620. http://dx.doi.org/10.1093/bioinformatics/btm626.
57. **Ashkenazy H, Erez E, Martz E, Pupko T, Ben-Tal N.** 2010. ConSurf 2010: calculating evolutionary conservation in sequence and structure of proteins and nucleic acids. Nucleic Acids Res **38:**W529–W533. http://dx.doi.org/10.1093/nar/gkq399.
58. **Lopez G, Maietta P, Rodriguez JM, Valencia A, Tress ML.** 2011. Firestar—advances in the prediction of functionally important residues. Nucleic Acids Res **39:**W235–W241. http://dx.doi.org/10.1093/nar/gkr437.
59. **Zhang Y, Skolnick J.** 2005. TM-align: a protein structure alignment algorithm based on the TM-score. Nucleic Acids Res **33:**2302–2309. http://dx.doi.org/10.1093/nar/gki524.
60. **Zhang Y, Skolnick J.** 2004. Scoring function for automated assessment of protein structure template quality. Proteins **57:**702–710. http://dx.doi.org/10.1002/prot.20264.
61. **Wang S, Peng J, Xu J.** 2011. Alignment of distantly related protein structures: algorithm, bound and implications to homology modeling. Bioinformatics **27:**2537–2545. http://dx.doi.org/10.1093/bioinformatics/btr432.
62. **Collins PJ, Haire LF, Lin YP, Liu J, Russell RJ, Walker PA, Skehel JJ, Martin SR, Hay AJ, Gamblin SJ.** 2008. Crystal structures of oseltamivir-resistant influenza virus neuraminidase mutants. Nature **453:**1258–1261. http://dx.doi.org/10.1038/nature06956.
63. **Mastrangelo E, Pezzullo M, Tarantino D, Petazzi R, Germani F, Kramer D, Robel I, Rohayem J, Bolognesi M, Milani M.** 2012. Structure-based inhibition of norovirus RNA-dependent RNA poly-

merases. J Mol Biol **419:**198–210. http://dx.doi.org/10.1016/j.jmb.2012.03.008.
64. **Zamyatkin DF, Parra F, Alonso JM, Harki DA, Peterson BR, Grochulski P, Ng KK.** 2008. Structural insights into mechanisms of catalysis and inhibition in Norwalk virus polymerase. J Biol Chem **283:**7705–7712. http://dx.doi.org/10.1074/jbc.M709563200.
65. **Fullerton SW, Blaschke M, Coutard B, Gebhardt J, Gorbalenya A, Canard B, Tucker PA, Rohayem J.** 2007. Structural and functional characterization of sapovirus RNA-dependent RNA polymerase. J Virol **81:**1858–1871. http://dx.doi.org/10.1128/JVI.01462-06.
66. **Noble CG, Lim SP, Chen Y-L, Liew CW, Yap L, Lescar J, Shi P-Y.** 2013. Conformational flexibility of the dengue virus RNA-dependent RNA polymerase revealed by a complex with an inhibitor. J Virol **87:**5291–5295. http://dx.doi.org/10.1128/JVI.00045-13.
67. **Harrus D, Ahmed-El-Sayed N, Simister PC, Miller S, Triconnet M, Hagedorn CH, Mahias K, Rey FA, Astier-Gin T, Bressanelli S.** 2010. Further insights into the roles of GTP and the C terminus of the hepatitis C virus polymerase in the initiation of RNA synthesis. J Biol Chem **285:**32906–32918. http://dx.doi.org/10.1074/jbc.M110.151316.
68. **Choi KH, Groarke JM, Young DC, Kuhn RJ, Smith JL, Pevear DC, Rossmann MG.** 2004. The structure of the RNA-dependent RNA polymerase from bovine viral diarrhea virus establishes the role of GTP in de novo initiation. Proc Natl Acad Sci U S A **101:**4425–4430. http://dx.doi.org/10.1073/pnas.0400660101.
69. **Ferrer-Orta C, Arias A, Pérez-Luque R, Escarmís C, Domingo E, Verdaguer N.** 2007. Sequential structures provide insights into the fidelity of RNA replication. Proc Natl Acad Sci U S A **104:**9463–9468. http://dx.doi.org/10.1073/pnas.0700518104.
70. **Love RA, Maegley KA, Yu X, Ferre RA, Lingardo LK, Diehl W, Parge HE, Dragovich PS, Fuhrman SA.** 2004. The crystal structure of the RNA-dependent RNA polymerase from human rhinovirus: a dual function target for common cold antiviral therapy. Structure **12:**1533–1544. http://dx.doi.org/10.1016/j.str.2004.05.024.
71. **Gruez A, Selisko B, Roberts M, Bricogne G, Bussetta C, Jabafi I, Coutard B, De Palma AM, Neyts J, Canard B.** 2008. The crystal structure of coxsackievirus B3 RNA-dependent RNA polymerase in complex with its protein primer VPg confirms the existence of a second VPg binding site on *Picornaviridae* polymerases. J Virol **82:**9577–9590. http://dx.doi.org/10.1128/JVI.00631-08.
72. **Gong P, Peersen OB.** 2010. Structural basis for active site closure by the poliovirus RNA-dependent RNA polymerase. Proc Natl Acad Sci U S A **107:**22505–22510. http://dx.doi.org/10.1073/pnas.1007626107.
73. **Wright S, Poranen MM, Bamford DH, Stuart DI, Grimes JM.** 2012. Noncatalytic ions direct the RNA-dependent RNA polymerase of bacterial double-stranded RNA virus 6 from de novo initiation to elongation. J Virol **86:**2837–2849. http://dx.doi.org/10.1128/JVI.05168-11.
74. **Tao Y, Farsetta DL, Nibert ML, Harrison SC.** 2002. RNA synthesis in a cage—structural studies of reovirus polymerase lambda3. Cell **111:**733–745. http://dx.doi.org/10.1016/S0092-8674(02)01110-8.
75. **Lu X, McDonald SM, Tortorici MA, Tao YJ, Vasquez-Del Carpio R, Nibert ML, Patton JT, Harrison SC.** 2008. Mechanism for coordinated RNA packaging and genome replication by rotavirus polymerase VP1. Structure **16:**1678–1688. http://dx.doi.org/10.1016/j.str.2008.09.006.
76. **Gerlach P, Malet H, Cusack S, Reguera J.** 2015. Structural insights into bunyavirus replication and its regulation by the vRNA promoter. Cell **161:**1267–1279. http://dx.doi.org/10.1016/j.cell.2015.05.006.
77. **Lu G, Gong P.** 2013. Crystal structure of the full-length Japanese encephalitis virus NS5 reveals a conserved methyltransferase-polymerase interface. PLoS Pathog **9:**e1003549. http://dx.doi.org/10.1371/journal.ppat.1003549.
78. **Takeshita D, Tomita K.** 2012. Molecular basis for RNA polymerization by Qβ replicase. Nat Struct Mol Biol **19:**229–237. http://dx.doi.org/10.1038/nsmb.2204.
79. **Graham SC, Sarin LP, Bahar MW, Myers RA, Stuart DI, Bamford DH, Grimes JM.** 2011. The N-terminus of the RNA polymerase from infectious pancreatic necrosis virus is the determinant of genome attachment. PLoS Pathog **7:**e1002085. http://dx.doi.org/10.1371/journal.ppat.1002085.
80. **Garriga D, Navarro A, Querol-Audí J, Abaitua F, Rodríguez JF, Verdaguer N.** 2007. Activation mechanism of a noncanonical RNA-dependent RNA polymerase. Proc Natl Acad Sci U S A **104:**20540–20545. http://dx.doi.org/10.1073/pnas.0704447104.
81. **Ferrer-Orta C, Arias A, Perez-Luque R, Escarmís C, Domingo E, Verdaguer N.** 2004. Structure of foot-and-mouth disease virus RNA-dependent RNA polymerase and its complex with a template-primer RNA. J Biol Chem **279:**47212–47221. http://dx.doi.org/10.1074/jbc.M405465200.
82. **Porollo A, Meller J.** 2007. Prediction-based fingerprints of protein-protein interactions. Proteins **66:**630–645. http://dx.doi.org/10.1002/prot.21248.
83. **Deng T, Engelhardt OG, Thomas B, Akoulitchev AV, Brownlee GG, Fodor E.** 2006. Role of Ran binding protein 5 in nuclear import and assembly of the influenza virus RNA polymerase complex. J Virol **80:**11911–11919. http://dx.doi.org/10.1128/JVI.01565-06.
84. **Hengrung N, El Omari K, Martin Is VFT, Cusack S, Rambo RP, Vonrhein C, Bricogne G, Stuart DI, Grimes JM, Fodor E.** 2015. Crystal structure of the RNA-dependent RNA polymerase from influenza C virus. Nature **527:**114–117. http://dx.doi.org/10.1038/nature15525.
85. **Arai Y, Kawashita N, Daidoji T, Ibrahim MS, El-Gendy EM, Takagi T, Takahashi K, Suzuki Y, Ikuta K, Nakaya T, Shioda T, Watanabe Y.** 2016. Novel polymerase gene mutations for human adaptation in clinical isolates of avian H5N1 influenza viruses. PLoS Pathog **12:**e1005583. http://dx.doi.org/10.1371/journal.ppat.1005583.
86. **Mandell DJ, Coutsias EA, Kortemme T.** 2009. Sub-angstrom accuracy in protein loop reconstruction by robotics-inspired conformational sampling. Nat Methods **6:**551–552. http://dx.doi.org/10.1038/nmeth0809-551.
87. **Squires RB, Noronha J, Hunt V, García-Sastre A, Macken C, Baumgarth N, Suarez D, Pickett BE, Zhang Y, Larsen CN, Ramsey A, Zhou L, Zaremba S, Kumar S, Deitrich J, Klem E, Scheuermann RH.** 2012. Influenza research database: an integrated bioinformatics resource for influenza research and surveillance Influenza Other Respir Viruses **6:**404–416. http://dx.doi.org/10.1111/j.1750-2659.2011.00331.x
88. **Edgar RC.** 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res **32:**1792–1797. http://dx.doi.org/10.1093/nar/gkh340.
89. **Pond SL, Frost SD, Muse SV.** 2005. HyPhy: hypothesis testing using phylogenies. Bioinformatics **21:**676–679. http://dx.doi.org/10.1093/bioinformatics/bti079.
90. **Broadbent AJ, Santos CP, Godbout RA, Subbarao K.** 2014. The temperature-sensitive and attenuation phenotypes conferred by mutations in the influenza virus PB2, PB1, and NP genes are influenced by the species of origin of the PB2 gene in reassortant viruses derived from influenza A/California/07/2009 and A/WSN/33 viruses. J Virol **88:**12339–12347. http://dx.doi.org/10.1128/JVI.02142-14.
91. **Da Costa B, Sausset A, Munier S, Ghounaris A, Naffakh N, Le Goffic R, Delmas B.** 2015. Temperature-sensitive mutants in the influenza A virus RNA polymerase: alterations in the PA linker reduce nuclear targeting of the PB1-PA dimer and result in viral attenuation. J Virol **89:**6376–6390. http://dx.doi.org/10.1128/JVI.00589-15.
92. **Deane R, Schäfer W, Zimmermann HP, Mueller L, Görlich D, Prehn S, Ponstingl H, Bischoff FR.** 1997. Ran-binding protein 5 (RanBP5) is related to the nuclear transport factor importin-beta but interacts differently with RanBP1. Mol Cell Biol **17:**5087–5096. http://dx.doi.org/10.1128/MCB.17.9.5087.
93. **Yaseen NR, Blobel G.** 1997. Cloning and characterization of human karyopherin beta3. Proc Natl Acad Sci U S A **94:**4451–4456. http://dx.doi.org/10.1073/pnas.94.9.4451.
94. **Betts MJ, Guigó R, Agarwal P, Russell RB.** 2001. Exon structure conservation despite low sequence similarity: a relic of dramatic events in evolution? EMBO J **20:**5354–5360. http://dx.doi.org/10.1093/emboj/20.19.5354.
95. **Naim HY, Niermann T, Kleinhans U, Hollenberg CP, Strasser AW.** 1991. Striking structural and functional similarities suggest that intestinal sucrase-isomaltase, human lysosomal alpha-glucosidase and Schwanniomyces occidentalis glucoamylase are derived from a common ancestral gene. FEBS Lett **294:**109–112. http://dx.doi.org/10.1016/0014-5793(91)81353-A.
96. **Lee D, Redfern O, Orengo C.** 2007. Predicting protein function from sequence and structure. Nat Rev Mol Cell Biol **8:**995–1005. http://dx.doi.org/10.1038/nrm2281.
97. **Dalal A, Atri A.** 2014. An introduction to sequence and series. Int J Res **1:**1286–1292.
98. **Russell RB, Sasieni PD, Sternberg MJ.** 1998. Supersites within superfolds. Binding site similarity in the absence of homology. J Mol Biol **282:**903–918. http://dx.doi.org/10.1006/jmbi.1998.2043.

99. **Russell RB**. 1998. Detection of protein three-dimensional side-chain patterns: new examples of convergent evolution. J Mol Biol **279:** 1211–1227. http://dx.doi.org/10.1006/jmbi.1998.1844.
100. **Hansen JL, Long AM, Schultz SC**. 1997. Structure of the RNA-dependent RNA polymerase of poliovirus. Structure **5:**1109–1122.
101. **Hutchinson EC, Fodor E**. 2012. Nuclear import of the influenza A virus transcriptional machinery. Vaccine **30:**7353–7358. http://dx.doi.org/10.1016/j.vaccine.2012.04.085.
102. **Kosugi S, Hasebe M, Matsumura N, Takashima H, Miyamoto-Sato E, Tomita M, Yanagawa H**. 2009. Six classes of nuclear localization signals specific to different binding grooves of importin alpha. J Biol Chem **284:**478–485. http://dx.doi.org/10.1074/jbc.M807017200.
103. **Chook YM, Süel KE**. 2011. Nuclear import by karyopherin-βs: recognition and inhibition. Biochim Biophys Acta **1813:**1593–1606. http://dx.doi.org/10.1016/j.bbamcr.2010.10.014.
104. **Guo HH, Choe J, Loeb LA**. 2004. Protein tolerance to random amino acid change. Proc Natl Acad Sci U S A **101:**9205–9210. http://dx.doi.org/10.1073/pnas.0403255101.
105. **Jacquier H, Birgy A, Le Nagard H, Mechulam Y, Schmitt E, Glodt J, Bercot B, Petit E, Poulain J, Barnaud G, Gros PA, Tenaillon O**. 2013. Capturing the mutational landscape of the beta-lactamase TEM-1. Proc Natl Acad Sci U S A **110:**13067–13072. http://dx.doi.org/10.1073/pnas.1215206110.
106. **McLaughlin RN, Poelwijk FJ, Raman A, Gosal WS, Ranganathan R**. 2012. The spatial architecture of protein function and adaptation. Nature **491:**138–142. http://dx.doi.org/10.1038/nature11500.
107. **Melnikov A, Rogov P, Wang L, Gnirke A, Mikkelsen TS**. 2014. Comprehensive mutational scanning of a kinase in vivo reveals substrate-dependent fitness landscapes. Nucleic Acids Res **42:**e112. http://dx.doi.org/10.1093/nar/gku511.
108. **Das R, Baker D**. 2008. Macromolecular modeling with Rosetta. Annu Rev Biochem **77:**363–382. http://dx.doi.org/10.1146/annurev.biochem.77.062906.171838.
109. **Kellogg EH, Leaver-Fay A, Baker D**. 2011. Role of conformational sampling in computing mutation-induced changes in protein structure and stability. Proteins **79:**830–838. http://dx.doi.org/10.1002/prot.22921.
110. **Lutz A, Dyall J, Olivo PD, Pekosz A**. 2005. Virus-inducible reporter genes as a tool for detecting and quantifying influenza A virus replication. J Virol Methods **126:**13–20. http://dx.doi.org/10.1016/j.jviromet.2005.01.016.
111. **Bloom JD**. 2014. An experimentally determined evolutionary model dramatically improves phylogenetic fit. Mol Biol Evol **31:**1956–1978.
112. **Doud MB, Bloom JD**. 2016. Accurate measurement of the effects of all amino-acid mutations on influenza hemagglutinin. Viruses **8:**155.
113. **Murrell B, Moola S, Mabona A, Weighill T, Sheward D, Pond SLK, Scheffler K**. 2013. FUBAR : A Fast, Unconstrained Bayesian AppRoximation for inferring selection. Mol Biol Evol **30:**1196–1205.