UNIVERSITY OF CALIFORNIA

Los Angeles

Quantitative high-throughput genomics in RNA viruses

A dissertation submitted in partial satisfaction of the

requirements for the degree Doctor of Philosophy

in Molecular and Medical Pharmacology

by

Yuchen Du

2017

ABSTRACT OF THE DISSERTATION

Quantitative high-throughput genomics in RNA viruses

by

Yuchen Du

Doctor of Philosophy in Molecular and Medical Pharmacology

University of California, Los Angeles, 2017

Professor Ren Sun, Chair

The high mutation rate and rapid genome replication of RNA viruses drive their adaptation to diverse selection pressures. The emergence of drug resistant or immune escape viral strains is always a major concern to public health. A comprehensive understanding of the mutation tolerability of viral genome is thus crucial to understand the evolution potential of viruses and guild the accurate risk assessments.

Traditional genetics has proven to be a powerful tool for virology studies. Including forward genetics – determine the genetic basis responsible for a phenotype, and reverse genetics – determine the phenotype of a genetic change, it reveals the functional role of many important mutations. However, traditional genetics is usually restricted by limited and biased sampling, and is time and money consuming. To overcome these limitations, we have developed a qantatative high-throughput genomic system that enables us to quantify the phenotype of thousands to millions of mutations as a massive parallel process. Using random mutagenesis or satuated mutagenesis, we can generate a diverse pool of viral library containing desired mutations. The library can be used to assess the function of every amino acid/nucleotide in a variety of protein functional assays as well as viral growth assay, with the frequency of each mutant changed according to their competitive strength. We were able to quantify the relative frequency change of each variant pre and post selection by high-throughput sequencing, which represented their "relative fitness" under the particular selection condition.

Since the first inception of the system, we have optimized and successfully applied it *to* human immunodeficiency virus *(*HIV*),* Hepatitis C Virus (HCV) *an*d influenza A virus. We also explored the applications of the system to a variety of biological questions, with a specicial focus in the following 4 areas:

Firstly, a direct application of the system is to better understand the distribution of fitness effect (DFE), which is fundamental to a variety of evolution theories. We systematically quantified the DFE of single amino acid substitutions (86 amino acids total) in the drug-targeted region of NS5A protein of Hepatitis C Virus (HCV). We found that the majority of non-synonymous substitutions incur large fitness costs, suggesting that NS5A protein is highly optimized in natural conditions. Furthermore, we characterized the evolutionary potential of HCV by subjecting the mutant viruses to varying concentrations of an NS5A inhibitor Daclatasvir. As the selection pressure increases, the DFE of beneficial mutations shifts from an exponential distribution to a heavy-tailed distribution with a disproportionate number of exceptionally fit mutants. The number of available beneficial mutations and the selection coefficient both increase at higher levels of antiviral drug concentration, as predicted by a pharmacodynamics model describing viral fitness as a function of drug concentration. Our large-scale fitness data of mutant viruses also provide insights into the biophysical basis of evolutionary constraints and the role of the genetic code in protein evolution.

Secondly, we explored the usage of fitness profiling to identify and annotate protein functional residues. Using influenza A virus PB1 protein as an example, we developed an approach to achieve this task: Firstly, the effect of PB1 point mutations on viral replication was examined by saturation mutagenesis and high-throughput sequencing. Secondly, functional PB1 residues that are essential for viral growth but do not affect protein stability were identified by

protein stability prediction. Lastly, homologous structural alignment was utilized to further annotate specific biological functions (canonical versus non-canonical functions) for each functional residue. We achieved high sensitivity in identifying and annotating the canonical polymerase functional residues. Moreover, we identified non-canonical functional residues, which are exemplified by a cluster of residues located in the loop region of 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).

Thirdly, the system was shown to be valuable for the identification of drug resistant mutations and the design of personalized therapy. Using influenza NA protein as an example, we characterized the fitness effects of single nucleotide mutations of neuraminidase (NA) and systematically identified resistant mutations for three neuraminidase inhibitors (NAIs): zanamivir, oseltamivir and AV5080. We observed that both the numbers and the effects of resistant mutations of AV5080 are smaller than those of zanamivir and oseltamivir, but so are their fitness costs. We used population genetic models to estimate the rate of increase in fitness under drug selection as a function of drug dosage. AV5080 showed a higher rate of increase in fitness at low drug concentrations due to the low fitness cost of resistant mutations, but also exhibited a steep drop with high drug concentrations because of lower strength of resistance. Our approach also enabled the systematic analyses of cross-resistance against different drugs, which showed to be uncommon between AV5080 and zanamivir.

Lastly and importantly, the system can be utilized to explore new functions of viral proteins. To this end, we systematically identified type I interferon sensitive mutations across the entire influenza A viral genome. We have identified novel IFN-sensitive mutations on PB2, PA, PB1 and M1, in addition to NS1, which provides a foundation to determine multiple anti-IFN mechanisms encoded in different viral segments. Moreover, this quantitative functional information of every amino acid in the genome enabled us to rationally design vaccine to increase the safety and immunogenicity. By selecting and combining 8 mutations into one viral genome, we successfully generated a *d*eficient in *a*nti-*i*nterferon (*DAI)* influenza strain as a live attenuated vaccine candidate. *DAI* is replication-competent in IFN-deficient host, but able to induce transient IFN response and highly attenuated in IFN competent host. Impressively, DAI is capable of inducing a robust humoral response and a strong T cell response, which collectively leads to broad protection. The superior property of *DAI* strain demonstrated the capacity of our approach to construct a safe, effective and broadly protecting live attenuated influenza vaccine. Thus we proposed a novel and generally applicable approach for vaccine design: systematically identifying and eliminating immune evasion functions on the virus genome, while maintaining the replication fitness *in vitro* for vaccine production.

In summary, we have developed the quantitative high-throughput genomic system, and applied it to a variety of biological questions. It is proven to be a powerful system to investigate fundamental evolution problems, identify functional residues and new functions of target proteins, and facilitate drug development. With the maturation of DNA systhesis technology and ever increasing sequencing power, we foresee the further improvement and more broad applications of this system to address foundamental mechanistic questions and practical applications.

The dissertation of Yuchen Du is approved.

Jamie Lloyd-Smith

Lin Jiang

Steven Bensinger

Samson Chow

Ren Sun, Committee Chair

University of California, Los Angeles

2017

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

Four years earlier, when Dr. Ren Sun and I initiated this 4+4 “MD-PhD” program between UCLA and Zhejiang University, not many people believed that I could possibly finish the program on time, including myself. But as I move to this chapter of dissertation, it might indicate that I finished the PhD part of the long training, within 4 years. Did a good job or not, I am not quite sure, but I know these 4 years opened a new “research world” for me that I become very fancy about. Going through all the doubts, troubles and self-denials, I am certain that I will continue to pursue science for the rest of my life.

So many people helped me throughout this journey, without whom I cannot make it today. First of all, I want to express my greatest gratitude to my PhD adviser, Dr. Ren Sun. He is a very sincere person, maintaining all the passion of a naïve soal as a child. That is why he is brave enough to open up many new educational programs, without thinking too much about the down sides or what he can get from the tedious work. The summer training program (CSST) has become one of the most popular programs in China. Each year, over 90 junior undergrads get the chance to come to UCLA and experience serious scientific training. The program became the turning point of a lot of students, who then decided to apply for graduate schools and pursue their careers as researchers. He introduced the MD carriculum system from UCLA to ZJU and actively participates in the current medical reform of China. As one of the students in the third session of that program, I am lucky enough to experience the UCLA curriculum, and constantly think about what is a better way for medical training. It also led me to the joint MD/PhD program. When I first came to UCLA as a PhD student, I have zero experience in research. Thank him a lot for the free-spirited research environment he offeres and importantly, the patience he has with me when I screw up many experiments. He is not a mentor who gives you step-to-step guidance and is very off-handed, but he can always give valuable suggestions at critical points. Thanks to this, I can make most of the decisions myself with proper guidance, which make me grow up much faster.

I also want to express my special thanks to Dr. Ting-ting Wu. She is an amazing mentor, a great scientist, a trustable friend and a very considerate person. Her personality is so different and so complementary to Ren's, which make them a good team together. Although she is not my direct PI, she is always willing to help with any of my questions. I can always learn and get inspiration from discussing with her.

I want to thank Dr. Sam Chow, who is always there to support me. His smile makes me calm and gives me confidence. He is also a very good thinker with lots of excellent suggestions. I am always wondering why Dr. Steven Bensigner can know so many things. After any talk, he can ask questions to the point and seems to have answers for all my concerns. I wish I can be so knowledgeable sometime soon. Dr. James Lloyd-Smith is an expert on evolutionary biology. Although I do not have any background in this area, I am always interested in those topics. He gives me a lot of guidance on how to pursue related ideas and how to think/analyze data in a proper way. Thanks to all his patience for my naïve questions. I also want to thank Dr. Lin Jiang, a genius in structural biology. He taught me so much about structure biology and provided great help on structural analysis. He is at the early stages of setting up his own lab and I am sure he is a rising star in his field.

There are so many other people helped me in these precious 4 years. I want to thank Dr. Nicholas Wu in particular, who guided and taught me pretty much everything I know in the lab. Dr. Lei Dai, my coach in evolution biology. I learned so much from him, about being a sincere scientist and extreme cautious about data analyses. Dr. Yuan Shi and Tianhao Zhang are always good companions, without whom my grad school life will be so short of laughs. Dr. Danyang Gong, our king of molecular biology, who always has protocols, media and can magically make experiments work so beautifully. Dr. Anders Olson, I regret that I have not discussed with him that much until he was planning to leave. I cannot possibly be as good as him as a quantitative biochemist, but he sets up a standard for me to try to rearch. And so many other people in the lab, Dr. Hangfei Qi, Dr. Jun Feng, Dr. Xinghong Dai, Dr. Travis Chapa, Nizar Fahat, Preet Bara, Sara Shu and Dr. Shin Hsin Chu, I am truly grateful for all of your help during the past 4 years.

I also want to thank our collaborator, Dr. Sumit Chada, Dr. Navan Krogen, Dr. Jieru Wang and Dr. William Reiley, et al, it is really a pleasure working with them.

Lastly, I want to thank my husband who always been there throughout my life. He commutes 3 hours each day, going back and forth from UCLA and Irvine so that we can live together. Thank all of my dear friends for the support and caring, and my parents for the unconditional love. Thank our upcoming baby, who stays happy and healthy so that I can focus on my work. Without them, I would not have been who I am today.

Notes about the chaptors in the thesis:

Chapter 2 is a version of the following manuscript: Dai L*, Du Y*, Qi H., Wu NC, Lloyd-Smith JO, Sun R. (2016). Quantifying the evolutionary potential and constraints of a drug-targeted viral protein. bioRxiv, 078428. *Equal contributions.

Chapter 3 is a version of the following manuscript: Du Y, Wu NC, Zhang TH, Gong DY, Shu S, Wu T-T, Sun R. (2016). Annotating Protein Functional Residues by Coupling High-Throughput Fitness Profile and Homologous-Structure Analysis. mBio, 7(6), e01801-16.

Chapter 4 is a version of the following manuscript: Du Y*, Feng J*, Dai L, Wu NC, Wu TT, Sun R. Systematic fitness profiling revealed distinct resistant barriers for multiple neuraminidase inhibitors.

Chapter 5 is a version of the following manuscript: Du Y*, Zheng X*, Zhang TH*,Dai L*, Gorin A, Detels R, Oishi J, Wu TT, Sun R. CTL escape of HIV-1 Gag epitopes is determined by mutational effects on replicative fitness and MHC-I binding

Chapter 6 is a version of the following manuscript: Du Y, Shi Y, Zhang TH, Wu NC, Dai L, Gong D, Brar G, Shu S, Luo J, Reiley W, T-T Wu, Wang J, Sun R, Genome wide identification of anti-interferon function of influenza A virus enabling rational vaccine design

VITA

2011	B.S. Department of psychology Zhejiang University, China
2011-2013	Department of Medical School Zhejiang Univeristy, China
2015	Advancement to Candidacy for Ph. D. Molecular and Medical Pharmacology University of California, Los Angeles

Selected Publications:

Du Y, Wu NC, Zhang TH, Gong DY, Shu S, Wu T-T, Sun R. (2016). Annotating Protein Functional Residues by Coupling High-Throughput Fitness Profile and Homologous-Structure Analysis. mBio, 7(6), e01801-16.

Wu NC*, **Du Y***, Le S, Young AP, Zhang T-H, Wang Y, Zhou J, Yoshizawa JM, Dong L, Li X, others. Coupling high-throughput genetics with phylogenetic information reveals an epistatic interaction on the influenza A virus M segment. BMC Genomics. BioMed Central; 2016; 17(1):1. *Equal contributions.

Dai L*, **Du Y***, Qi H., Wu NC, Lloyd-Smith JO, Sun R. (2016). Quantifying the evolutionary potential and constraints of a drug-targeted viral protein. bioRxiv, 078428. *Equal contributions

Gong D, Kim YH, Xiao Y, **Du Y**, Xie Y, Lee K.K, Feng J, Farhat N, Zhao D, Shu S, Dai X, Sun R, Wu TT, 2016. A Herpesvirus Protein Selectively Inhibits Cellular mRNA Nuclear Export. Cell Host & Microbe, 20(5), pp.642-653.

Dai X, Li ZH, Lai M, Shu S, **Du Y,** Zhou H, Sun R, In situ structures of the genome and genome-delivery apparatus in an ssRNA virus. Nature 541.7635 (2017): 112-116.

Thai M, Thaker S , Feng J, **Du Y**, Hu H, Wu TT, Graeber GT, Braas D, Christofk RH, MYC-induced reprogramming of glutamine catabolism supports optimal virus replication. Nature communications. 2015; 6(12): 8873.

Wu NC, Olson CA, **Du Y**, Le S, Tran K, Remenyi R, Gong D, Al-Mawsawi LQ, Qi H, Wu T-T, Sun R. Functional Constraint Profiling of a Viral Protein Reveals Discordance of Evolutionary Conservation and Functionality. PLoS Genet. 2015; 11(7):e1005310.

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. A quantitative high-resolution genetic profile rapidly identifies sequence determinants of hepatitis C viral fitness and drug sensitivity. PLoS Pathog. 2014; 10(4):e1004064.

Selected Presentations:

Annotating Protein Functional Residues by Coupling High-Throughput Fitness Profile and Homologous-Structure Analysis. American association of virology meeting, 2016

Quantifying the evolutionary potential and constraints of a drug-targeted viral protein. American association of virology meeting, 2016

Quantifying perinatal transmission of Hepatitis B viral quasispecies by tag linkage deep sequencing, American association of virology meeting, 2016

High-throughput functional annotation of influenza A virus genome at single-nucleotide resolution, ISIRV, 2016

CHAPTER 1

INTRODUCTION

1.1 Quantitative high-throughput genetic platform

Viruses, especially RNA viruses, are characterized by highly variable genomes and high mutation rates (1–4). This property enables rapid antigenic drift and genetic shift, and drives emergence of drug resistant and immune escape strains with pandemic risks. Thus, it is essential to understand the property of possible viral mutations, or to understand the evolution potential of viruses for accurate risk assessment (5, 6).

Traditional genetics, including forward genetics – determine the genetic basis responsible for a phenotype, and reverse genetics – determine the phenotype of a genetic change, is a powerful tool for virology study. However, focusing on the relationship between one genotype and one phenotype, traditional genetics is time and money consuming. Combining the concept of both forward and reverse genetics, we developed and optimized a quantitative high throughput system that can quantify the phenotype of mutations at single nucleotide resolution (7–15). Using random mutagenesis or saturated mutagenesis, we can generate up to millions of mutations in the target protein or genome. This pool of diverse mutants can be used to assess the functions of each amino acid of a protein or viral genome in a variety of functional assays as well as viral growth assays. Through these functional assays, the diverse pool of mutants can go through a defined selection condition with the frequency of each mutant changes according to their competitive strength. We are able to quantify the relative frequency change of each variant pre and post selection by high throughput sequencing, which represents their "relative fitness" under the particular selection condition.

The quantitative high throughput system provide a systematic way to evaluate all possible mutations in the target protein or viral genome, thus solving the problem of our limited ability to choose and analyze mutations. Currently, the most common way to analyze mutation properties are through natural occurring sequences. Conserved residues are considered to carry critical function, thus were kept unchanged throughput evolution (16–18). Although this method is powerful and provides critical insight into protein function, it suffers from limited and biased sampling of functional protein space. The high throughput system can provide a complete sequence – function map, which enables us to comprehensively examine the residues with known function, understand the mechanism of sequence conservation, and reveal unknown functions.

The key element in this system is the selection step. Depending on the aspect of protein property to be assessed, different functional assays can be designed and applied (9, 19–24). Here we listed three examples to demonstrate the breadth of possible design.

Firstly, we can examine the protein binding ability to its target. A delicate system has been developed in our lab by incorporating mRNA display and high-throughput screening. The target protein can be conjugated onto beads and interact with the protein library of a large number of variants linked with corresponding mRNA through puromycin. By reading the enrichment, we can quantify the relative binding efficiency for all the varients in the library (10). Another system that facilitates binding affinity measurement is named as “Tite-Seq”. It is capable of quantifying binding titration curves and thus calculating affinity (kd) value for mutants. Stabilizing or destabilizing mutations can also be identified through the ability to rescue or sensitize other mutations (25).

Secondly, the catalytic functions of protein mutations can be directly assayed. A florescence reporter can be linked to the catalytic function, thus the mutant with stronger or weaker function can be sorted out and analyzed (23, 25, 26).

Thirdly and importantly, the function of a protein can be investigated at the organismal growth level. With the establishment of viral reverse genetic system, we can incorporate these mutations into viral genome. The viral population can be selected in biological relevant cell lines or many different desired conditions (7–15), revealing the overall mutational effects on viral growth.

Furthermore, by combining different screening methods, we could establish the linkage between sequence genotype to protein functional phenotype, and then to the organismal fitness.

The establishment of such linkage represents a central step to a variety of evolution theories and a critical focus of systems biology.

In the following chapters, I will describe some of the applications of this quantitative high throughput genetic platform that I have explored during my graduate research. Three major model systems were utilized: influenza A virus, Hepatitis C virus and Human immunodeficiency virus. I will introduce each of them briefly.

1.1.1 INFLUENZA A VIRUS

Influenza A virus belongs to the family of orthomyxoviridae. Among the four classes (influenza A, B, C and D), influenza A virus is the most common human pathogen and can causes mild to severe respiratory diseases. Common symptoms of influenza A infection include high fever, cough, headache, muscle and joint pain, et al (27). For high risk populations who are younger than age 2 years, older than age 65 years, pregnant or under specific medical conditions, influenza infection can cause hospitalization or death. ~36,000 deaths per year in the U.S. and 300,000-500,000 globally are attributed to influenza viral infection (27, 28). Seasonal influenza causes epidemic in the winter season almost every year. Moreover, the threat of future pandemics is a major concern. Four large scale pandemics occur in the past 100 years, including 1918 “Spanish Flu”, 1957 and 1968 pandemic and 2009 “Swine Flu”, each caused 10,000 to 50 million death. Currently, two classes of anti-viral drug are available: M2 inhibitors and neuraminidase (NA) inhibitors. However, due to widespread drug resistance, M2 inhibitors are no longer used clinically (29). Resistant barrier for NA inhibitors is also low, which permits the rapid development of drug resistance (30–33). Vaccination is another widely used strategy to prevent influenza infection and decrease death rate post infection. The seasonal vaccine is updated every year and globally used, but the safety and efficacy still needs to be improved (34–38). Thus there is an urgent need for the development of new drug and new vaccine strategy against influenza infection.

1.1.2 HEPATITIS C VIRUS

Hepatitis C virus (HCV) is a positive single strand RNA virus that belongs to the family of flavirvirus. An estimate of 130-170 million people are infected with HCV virus worldwide, which is one of the leading causes for liver diseases, including fibrosis, cirrhosis and hepatocellular carcinoma.

Recent developed direct acting antivirals (DAAs) were proven to be potent HCV inhibitors that cure chronic HCV infection with drug combinations(39).

1.1.3 HUMAN IMMUNODEFICIENCY VIRUS

Human immunodeficiency virus (HIV) is a member of the genus Lentivirus, which can infect human immune cells and integrate into human genome. Two types of HIV have been characterized, HIV-1 and HIV2. HIV-1 is the major cause of global HIV infection, while HIV-2 is mostly confined to West Africa.

Once an individual is infected with HIV there is currently no way to completely eradicate the virus. Infected individuals can successfully maintain undetectable viral loads by routinely taking combination ART. However, viral replication can quickly resurge upon the stop of ART. The virus emerges from long-lived latent CD4+ T cell populations that harbor stably integrated copies of the HIV provirus. This reservoir of cells with latently integrated provirus is one of the major barriers to complete eradication of the virus, and hence a complete cure of HIV (40, 41).

1.2 Applications of quantitative high-throughput genetic platform

1.2.1 UNDERSTANDING THE FUNDAMENTAL PROPERTIES AND EVOLUTION POTENTIALS OF VIRAL GENOMES

Even since Wright brought up the concept of “fitness landscape”, it has been widely accepted by different fields. Natural evolution can then be viewed as climbing uphill to find a fitter point in the landscape. However, we still know little about its properties in real biological systems.

Previous empirical studies of fitness landscapes have very limited sampling of sequence space. Mutants are generated by site-directed mutagenesis and assayed for growth rate individually.

As the high-throughput system provides a comprehensive genotype-fitness relationship, thus it is useful in elucidating the shape and property of “fitness landscape”. Depending on the size of protein and the strategy of introducing mutations, we can look at the landscape in a “entire”, a “splotlight”, or a “scattered” view (42). Each “view” provide a different aspect and can collectively deepen our understanding of the landscape and possible evolutionary trajectory in different biological systems.

Here I listed three applications of defining the fitness landscape:

Firstly, it can be utilized to better understand the distribution of fitness effect (DFE). DFE of mutations is a fundamental entity in genetics and reveals the local structure of a fitness landscape (2, 43–49). The deleterious mutations are usually abundant and impose severe constraints on the accessibility of fitness landscapes. In contrast, beneficial mutations are rare and provide the raw materials of adaptation. Quantifying the DFE of RNA viruses is crucial for understanding how the pathogens evolve to acquire drug resistance, evade the host immune system, cross species barrier, etc. A complete genotype-fitness map can provide a unique opportunity to investigate the distribution of fitness effect with huge datasets and test different evolutionary theories.

Secondly, it can be used to build better evolutionary models and phylogenetic trees of given protein (50). The estimation of evolutionary distance is currently based on a pre-defined amino acid distance matrix among existing species. However, if we gain detailed understanding of the essentialness and mutational tolerance of each residue, then a better evolutionary model can be generated.

Moreover, it can be used to understand the genome arrangements of viruses. With the compact genome, viruses commonly use overlapping sequences to encode different proteins or multiple functions on one protein. How the DNA sequence evolves under diverse selection pressure coming from different functional requirement of proteins is not fully understood. Fitness landscape for each individual gene pin point the functions of each codon corresponding to different proteins, thus provide an essential tool for understanding the gene arrangement model and the evolution advantage of specific models for viruses (24).

1.2.2 IDENTIFICATION OF FUNCTIONAL RESIDUES

Amino acid residues in a protein have two roles: providing structural framework (structural residues) and mediating interactions with other biomolecules (functional residues). Identification and annotation of functional residues are fundamental questions in protein characterization (51–55). A number of methods have been developed to tackle these questions. Most methods utilize sequence conservation information, with the assumption that functional residues are often conserved in homolog proteins (16–18). Other methods predict functional residues based on the shapes and properties of protein 3D structures (56–62). Starting from well-known functional domains (ligand binding, catalytic, et al.), these analyses determine the similar local structures and key residues that may be related to the functions. Conservation-based methods provide valuable information on protein functional residues, but are limited by the insufficient sampling of protein sequence space in natural evolution (63).

The high-throughput platform, together with protein stability prediction and protein homologous analyses provide a systematic way to identify and annotate functional residues. We have proposed a workflow for this task (13). Firstly, the effect of mutations on protein function or viral replication can be comprehensively examined. Secondly, functional residues that are essential for protein function but do not affect protein stability are identified by protein stability prediction. And lastly, homologous structural alignment is utilized to further annotate specific biological functions (canonical versus non-canonical functions) for each functional residue. We have demonstrate the feasibility and effectiveness of this workflow with two influenza proteins:

PA and PB1(13, 63). The method showed superior sensitivity and accuracy comparing with the conservation-based method.

1.2.3 Identification of optimal drug target, drug resistant mutations and second generation drug design

The ideal drug should target a highly conserved domain on viral genome to limit the emergence of drug resistant mutations. High-throughput genetic approach enables us to systematically perturb viral genome and identify viral regions with high fitness cost upon mutations. The system can be more powerful by combining with structural analysis, where the high fitness cost region can be mapped onto protein structure to identify possible drug binding pocket. This information is valuable for drug design, docking and improvement (14).

Moreover, drug resistant mutations can be systematically identified as a direct application of this genetic platform. Two strategies have been commonly used to identify resistant mutations: isolation of resistant strains in clinic (30–32, 64), and long-term evolution in vitro under drug selection (65, 66). Through sequencing, viral mutations with resistant phenotypes can be identified. However, both strategies suffer from limited sampling of de novo mutations and stochasticity in evolution. Some important resistant mutations might not be captured due to genetic drift or clonal interference. Moreover, both strategies are highly time consuming, which makes it difficult to provide a comprehensive assessment of resistant barriers of multiple drugs. The high-throughput genetic approach provided a systematic way to evaluate the resistant profiles against drugs. Starting with a diverse library, we were able to bypass the stochastic genetic drift and measure the fitness of thousands of mutations. Then the library can be screened with and without drug selection, and drug resistant mutations can be identified by comparing the phenotypes under these two conditions.

Additionally, the information of high-throughput genetic profiling can be possibly used to enable the “next-generation drug development”. Combining the identification of drug targets and the evaluation of drug resistant mutations, genetic profiling can foresee the possibly emerging resistant mutations during the development of new drug, rather than waiting for the emergency of resistance in clinic. Then modifications can be made based on the resistant profile to further increase the resistant profile of the new drug. The process can be iterative and greatly shorten the optimization phase of current drug development, into what we called “next generation drug development”.

1.2.4 EXPLORING NEW FUNCTIONS OF VIRAL PROTEINS

Genetic profiling is also a powerful tool to reveal new functions or new mechanisms of a selected protein. Screening mutant libraries under diverse selection conditions can pin point the residues with new functions, which is usually intertwined with the fitness cost of mutations.

For example, possible antigenic sites can be identified by screening the library under antibody selection (67). Antigenic site mutations are likely to escape under antibody selection and specifically enriched in the library. Thus, the genetic profiling can be a possible method to map antibody epitopes.

Another example comes from the species specificity of influenza polymerase. Influenza virus can replicate in different hosts (avian, swine, human, et al). To achieve optimal replication in a specific host, influenza viral genome usually requires multiple mutations to facilitate the viral entry and replication (68, 69). Understanding host adaptive mutations is critical for the assessment of viral transmission and pandemic risk. By screening polymerase library in different host cells, we will be able to identify mutations that are deleterious and beneficial in each cell type. These findings can lead to better understanding of species specificity (26).

In the last chapter of my thesis, I will present a story about systematic identification and characterization of anti-interferon functions across the influenza genome. Type I interferon, including interferon α and interferon β, are dedicated to detect and signal the presence of intracellular pathogens and communicate with neighbor cells (70, 71). Signaling cascade induced by type I interferon provides the first line of defense against viral infection by inducing interferon stimulated genes (ISGs). More than 400 ISGs have now been identified, with different functions including anti-viral, immune modulatory and cell regulatory functions (72). Type I interferon system is also the linker between innate immune and adaptive immune response, which is critical for dendritic cell maturation, T cell development and antibody production(73, 74). This linker function is critical for vaccine design, since a high interferon response might lead to higher vaccine efficacy. Anti-interferon function is essential for virus to replicate efficiently in vivo. During natural evolution, viruses have developed multiple strategies to counteract the function of IFN system. For influenza virus, NS1 is the key viral protein that counter-acts IFN function. It is shown that NS1 protein can bind to RIG-I and MAVS, thus affect interferon induction. NS1 can also down-regulate interferon receptor (IFNAR) to affect signaling. Moreover, NS1 protein can interfere with the function of ISGs, such as Ser/Thr kinase PKR and the RNase L-pathway activator OAS. The anti-interferon functions from other genes are not being appreciated until recently. With the accumulation of databases and the improvement of reverse genetic technologies, researchers started to reveal the importance of other influenza proteins for their anti-interferon functions (69, 75–78). However, it is still a challenging task to separate the viral replication function of essential proteins from their anti-interferon function. Our genetic profiling system provides a unique opportunity to discover mutations in most segments across the viral genome that increase sensitivity to IFN. Furthermore, this information can be utilized to generate a safer and more effective live attenuated vaccine. By identifying and combining 8 mutations together, we generated a DAI strain that is severely attenuated in IFN competent host but able to induce strong IFN production and response. Moreover, DAI strain can induce robust antibody and T cell responses, and provide broad protection against homologous and heterologous viral challenge.

Overall, quantitative high-throughput genetic platform represent a powerful system to investigate fundamental evolution problems, identify functional residues and new functions of target protein, and facilitate drug development. With the ever increasing sequencing power, we foresee the further improvement and more broad applications of this system.

1.3 Bibliography

1. Sanjuán R, Nebot MR, Chirico N, Mansky LM, Belshaw R. 2010. Viral mutation rates. J Virol 84:9733–9748.

2. Burch CL, Chao L. 2000. Evolvability of an RNA virus is determined by its mutational neighbourhood. Nature 406:625–8.

3. Sanjuán R, Cuevas JM, Moya A, Elena SF. 2005. Epistasis and the adaptability of an RNA virus. Genetics 170:1001–1008.

4. Elena SF, Carrasco P, Daròs J-A, Sanjuán R. 2006. Mechanisms of genetic robustness in RNA viruses. EMBO Rep 7:168–173.

5. Russell C a, Kasson PM, Donis RO, Riley S, Dunbar J, Rambaut A, Asher J, Burke S, Davis CT, Garten RJ, Gnanakaran S, Hay SI, Herfst S, Lewis NS, Lloyd-Smith JO, Macken C a, Maurer-Stroh S, Neuhaus E, Parrish CR, Pepin KM, Shepard SS, Smith DL, Suarez DL, Trock SC, Widdowson M-A, George DB, Lipsitch M, Bloom JD. 2014. Improving pandemic influenza risk assessment. Elife 3:e03883.

6. Cox NJ, Trock SC, Burke SA. 2015. Development of Framework for Assessing Influenza Virus Pandemic Risk. Emerg Infect Dis 21:1372–1378.

7. Wu NC, Young AP, Dandekar S, Wijersuriya H, Al-Mawsawi LQ, Wu T-T, Sun R. 2013. Systematic identification of H274Y compensatory mutations in influenza A virus neuraminidase by high-throughput screening. J Virol 87:1193–9.

8. 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.

9. 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 influenza A virus NS segment. J Virol.

10. Olson CA, Wu NC, Sun R. 2014. A comprehensive biophysical description of pairwise epistasis throughout an entire protein domain. Curr Biol 24:2643–2651.

11. Al-Mawsawi LQ, Wu NC, Olson C, Shi V, Qi H, Zheng X, Wu T-T, Sun R. 2014. High-throughput profiling of point mutations across the HIV-1 genome. Retrovirology 11:124.

12. Wu NC, Olson CA, Du Y, Le S, Tran K, Remenyi R, Gong D, Al-Mawsawi LQ, Qi H, Wu T-T, others. 2015. Functional constraint profiling of a viral protein reveals discordance of evolutionary conservation and functionality. PLoS Genet 11:e1005310.

13. Du Y, Wu NC, Jiang L, Zhang T, Gong D, Shu S, Wu T. 2016. Annotating Protein Functional Residues by Coupling High- Throughput Fitness Profile and Homologous-Structure Analysis. MBio 7:1–13.

14. Qi H, Wu NC, Du Y, Wu T-T, Sun R. 2015. High-resolution genetic profile of viral genomes: why it matters. Curr Opin Virol 14:62–70.

15. Dai L, Du Y, Qi H, Wu NC, Lloyd-smith JO, Sun R. 2016. Quantifying the evolutionary potential and constraints of a drug-targeted viral protein. bioRxiv.

16. Glaser F, Pupko T, Paz I, Bell ER, 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.

17. Sankararaman S, Kolaczkowski B, Sjölander K. 2009. INTREPID: a web server for prediction of functionally important residues by evolutionary analysis. Nucleic Acids Res 37:W390-5.

18. Wilkins AD, Bachman BJ, Erdin S, Lichtarge O. 2012. The use of evolutionary patterns in protein annotation. Curr Opin Struct Biol 22:316–25.

19. 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 110:20248–53.

20. Doud MB, Bloom JD. 2016. Accurate Measurement of the Effects of All Amino-Acid Mutations on Influenza Hemagglutinin 1–17.

21. 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 42:1–8.

22. Thyagarajan B, Bloom JD. 2014. The inherent mutational tolerance and antigenic evolvability of influenza hemagglutinin. Elife 2014:1–26.

23. 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 b-Lactamase. Cell 160:882–892.

24. Fernandes J, Faust TB, Frankel AD. 2016. Functional Segregation of Overlapping Genes in HIV. Cell 167:1762–1766.e12.

25. Adams RM, Kinney JB, Mora T, Walczak AM. 2016. Measuring the sequence-affinity landscape of antibodies with massively parallel titration curves 1–19.

26. Taft AS, Ozawa M, Fitch A, Depasse J V, Halfmann PJ, Hill-Batorski L, Hatta M, Friedrich TC, Lopes TJ, Maher EA, Ghedin E, Macken CA, Neumann G, Kawaoka Y. 2015. Identification of mammalian-adapting mutations in the polymerase complex of an avian H5N1 influenza virus. Nat Commun 6:7491.

27. CDC. Key Facts about Influenza (Flu) & Flu Vaccine.

28. World health organization. World Health Organization. Influenza. Fact sheet no. 211.

29. Wang J, Wu Y, Ma C, Fiorin G, Wang J, Pinto LH, Lamb R a, Klein ML, Degrado WF. 2013. Structure and inhibition of the drug-resistant S31N mutant of the M2 ion channel of influenza A virus. Proc Natl Acad Sci U S A 110:1315–20.

30. Thorlund K, Awad T, Boivin G, Thabane L. 2011. Systematic review of influenza resistance to the neuraminidase inhibitors. BMC Infect Dis 11:134.

31. Dunn CJ, Goa KL. 2001. Oseltamivir: A Review of its Use in Influenza. Drugs 58:761–784.

32. Samson M, Pizzorno A, Abed Y, Boivin G. 2013. Influenza virus resistance to neuraminidase inhibitors. Antiviral Res 98:174–185.

33. Mckimm-Breschkin JL. 2013. Influenza neuraminidase inhibitors: Antiviral action and mechanisms of resistance. Influenza Other Respi Viruses 7:25–36.

34. Cox RJ, Brokstad K a, Ogra P. 2004. Influenza virus: immunity and vaccination strategies. Comparison of the immune response to inactivated and live, attenuated influenza vaccines. Scand J Immunol 59:1–15.

35. Maassab HF, Bryant ML. 1999. The development of live attenuated cold-adapted influenza virus vaccine for humans. Rev Med Virol 9:237–44.

36. Quan FS, Steinhauer D, Huang C, Ross TM, Compans RW, Kang S-M. 2008. A bivalent influenza VLP vaccine confers complete inhibition of virus replication in lungs. Vaccine 26:3352–61.

37. Manuscript A, Temperature E, Live S, Influenza A. 2013. Vaccines from Emerging Viruses 30:3691–3702.

38. Osterholm MT, Kelley NS, Sommer A, Belongia E a. 2012. Efficacy and effectiveness of influenza vaccines: a systematic review and meta-analysis. Lancet Infect Dis 12:36–44.

39. Zhang J, Nguyen D, Hu K-Q. 2016. Chronic Hepatitis C Virus Infection: A Review of Current Direct-Acting Antiviral Treatment Strategies. N Am J Med Sci (Boston) 9:47–54.

40. Barré-Sinoussi F, Ross AL, Delfraissy J-F. 2013. Past, present and future: 30 years of HIV research. Nat Rev Microbiol 11:877–883.

41. Kuritzkes DR, Lewin SR, Margolis DM, Mccune JM. 2016. A Cure for HIV Infection:“Not in My Lifetime” or “Just Around the Corner” 1:154–164.

42. Jiménez JI, Xulvi-Brunet R, Campbell GW, Turk-MacLeod R, Chen I a. 2013. Comprehensive experimental fitness landscape and evolutionary network for small RNA.Proceedings of the National Academy of Sciences of the United States of America.

43. Eyre-Walker A, Keightley PD. 2007. The distribution of fitness effects of new mutations. Nat Rev Genet 8:610–8.

44. Bataillon T, Bailey S. 2014. Effects of new mutations on fitness: insights from models and data. Ann New York Acad … 1320:76–92.

45. Desai MM, H DW, O DT and P, R RM, J WMJ and GC, E ES and LR, Buckling A MRCBMA and CN, R ESF and S, Garland J R RMR, Lenski R RMSS and TS, Gresham D DMMTCMJHTPDAWADCGBD and DMJ, B RD and RP, N C, Goddard M R GHCJ and BA, C CEC and GT, Perron G G GA and BA, Turner P E DJA and WR, Kirill S K MJIMNKAWMOH and DRN, L NA and S, A BG and G, Kryazhimskiy S K RDP and DMM, Lang G I BD and DMM, Kerr B NCBBJM and DAM, Hegreness M SNDDHD and KR, Agresti J J AEAARAKRACBJ-CMMKAMGAD and WDA, Guo M T RAHJA and WDA, Baraban L BFSMLMBNPPBJ de VJAGM and BJ, L LSF and SM, G B, A OH, Kerr B RMAFMW and BBJM, M RPB and T, Treves D S MS and AJ, Turner P E SV and LRE, E RDE and LR, Friesen M L SGTM and DM, R LB, Gallet R CTFESF and LT, Smith A M HLEMJKFTMJCMRFPGG and NC, al GG et, al HME et, al SLM et, Manna F GRMG and LT, J PC and A, C IM and S, Hegreness M SNHD and KR, G KKC and S, Zhang W SVDDMARBRCTF and AR, V ICJR and M, Barrick J E YDSYSHJHOTKSDLRE and KJF, E BJE and LR, Tenaillon O R-VAGRLMPBAFLAD and GBS, V MJ and TM, S W, Weissman D B DMMFDS and FMW, Michor F IY and NMA, Weinreich D DNDM and HD, Segre D DACGM and KR, E S, A FR, T MG and L, Martin G ESF and LT, Atwood K C SLK and RFJ, Atwood K C SLK and RFJ, L NAC and S, L KA, R LB, A OH, H M, M CJF and K, J M-S, J MS, F BW, J F, A HWG and R, de Visser J A ZCWGPJBJL and LRE, I DE and M, Sniegowski P D GPJ and LRE, Thompson D A DMM and MAW, Miralles R GPJMA and ESF, Miralles R MA and ESF, Desai M M FDS and MAW, P BJP and HJ, J BA, W JSB and HD, B HDW and JS, Perfeito L FLMC and GI, Wloch D M SKBRH and KR, R GP and L, Kessler D LHRD and TL, Tsimring L S LH and KDA, S DMM and FD, Rouzine I WJ and CJ, Rouzine I BE and WC, J PSC and K, Ridgway D LH and KD, Schiffels S SGJMV and LM, O H, Cohen E KDA and LH, Park S-C SD and KJ, J SPD and GP, Good B H RIMBDJHO and DMM, Neher R A SBI and FDS, W HM, Karasov T MPW and PDA, Sella G PDAPM and AP, Pritchard J K PJK and CG, M BNH and EA, Desai M M NLEWAM and PJB, Walczak A M NLEPJB and DMM, Seger J SWAPJJHJKZASLLPLRVJ and AFR, O'Fallon B D SJ and AFR, A OH, H GJ, Beisel C J RDRWHA and JP, T KR and B, Sanjuan R MA and ES, Cowperthwaite M C BJJ and MLA, P BAJ and BJ, R RD, D P, Barrett R D H MLK and OSP, A OH, T MG and L, A MRC and B, Barrett R D H CMR and BG, Fogle C A NJL and DMM, Blount Z D BCZ and LRE, Bridgham J T CSM and TJW, Goh C-S BAAJMWD and CFE, Schulz zur Wiesch P EJ and BS, Levin B R PV and WN, C MBA and L, G KA, DeVisser J HR and V den EH, DeVisser J HR and V den EH, DeVisser J HR and V den EH, R ES and L, F ES, D WM and B, Wloch D BR and KR, L BC and C, al TA et, R CS, Collins S R SMKNJ and WJS, al CM et, Lunzer M MSFR and DA, Bloom J D GLI and BD, Chou H-H CH-CDNFSD and MCJ, Khan A I DDMSDLRE and CTF, MacLean R C PGG and GA, Woods R J BJECTFSUKMR and LRE, L BCL and C, Barrick J E KMRSCC and LRE, McBride R OCB and TP. 2013. Statistical questions in experimental evolution. J Stat Mech Theory Exp 2013:P01003.

46. Jacquier H, Birgy A, Le Nagard H, Mechulam Y, Schmitt E, Glodt J, Bercot B, Petit E, Poulain J, Barnaud G, Gros P-A, Tenaillon O. 2013. Capturing the mutational landscape of the beta-lactamase TEM-1. Proc Natl Acad Sci U S A 110:13067–72.

47. Chevereau G, Dravecká M, Batur T, Guvenek A, Ayhan DH, Toprak E, Bollenbach T. 2015. Quantifying the Determinants of Evolutionary Dynamics Leading to Drug Resistance. PLoS Biol 13:e1002299.

48. Hietpas RT, Jensen JD, Bolon DN a. 2011. Experimental illumination of a fitness landscape. Proc Natl Acad Sci U S A 108:7896–7901.

49. Bank C, Hietpas RT, Jensen JD, Bolon DNA. 2015. A Systematic Survey of an Intragenic Epistatic Landscape. Mol Biol Evol 32:229–238.

50. Bloom JD. 2014. An Experimentally Determined Evolutionary Model Dramatically Improves Phylogenetic Fit Article Fast Track 31:1956–1978.

51. Mills CL, Beuning PJ, Ondrechen MJ. 2015. Biochemical functional predictions for protein structures of unknown or uncertain function. Comput Struct Biotechnol J 13:182–91.

52. Gonzalez-Perez A, Mustonen V, Reva B, Ritchie GRS, Creixell P, Karchin R, Vazquez M, Fink JL, Kassahn KS, Pearson J V, Bader GD, Boutros PC, Muthuswamy L, Ouellette BFF, 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 10:723–9.

53. 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.

54. Betancourt AJ, Bollback JP. 2006. Fitness effects of beneficial mutations: the mutational landscape model in experimental evolution. Curr Opin Genet Dev 16:618–23.

55. 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 30:1237–44.

56. Panchenko A, Kondrashov F, Bryant S. 2004. Prediction of functional sites by analysis of sequence and structure conservation. Protein Sci 884–892.

57. Capra J a, Laskowski R a, Thornton JM, Singh M, Funkhouser T a. 2009. Predicting protein ligand binding sites by combining evolutionary sequence conservation and 3D structure. PLoS Comput Biol 5:e1000585.

58. Tong W, Williams R, Wei Y, Murga L. 2008. Enhanced performance in prediction of protein active sites with THEMATICS and support vector machines. Protein Sci 333–341.

59. 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.

60. Skolnick J, Brylinski M. 2009. FINDSITE: a combined evolution/structure-based approach to protein function prediction. Brief Bioinform 10:378–91.

61. Pazos F, Sternberg M. 2004. Automated prediction of protein function and detection of functional sites from structure. Proc Natl Acad Sci 2004.

62. Petrova N V, Wu CH. 2006. Prediction of catalytic residues using Support Vector Machine with selected protein sequence and structural properties. BMC Bioinformatics 7:312.

63. 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.

64. Nguyen HT, Fry AM, Gubareva L V. 2012. Neuraminidase inhibitor resistance in influenza viruses and laboratory testing methods. Antivir Ther 17:159–173.

65. Molla A, Kati W, Carrick R, Steffy K, Shi Y, Montgomery D, Gusick N, Stoll VS, Stewart KD, Ng TI, Maring C, Kempf DJ, Kohlbrenner W, Laboratories A, Rd AP, Park A. 2002. In Vitro Selection and Characterization of Influenza A (A/N9) Virus Variants Resistant to a Novel Neuraminidase Inhibitor , A-315675. J Virol 76:5380–5386.

66. Hurt AC, Holien JK, Barr IG. 2009. In vitro generation of neuraminidase inhibitor resistance in A(H5N1) influenza viruses. Antimicrob Agents Chemother 53:4433–4440.

67. Doud MB, Hensley SE, Bloom JD. 2016. Complete mapping of viral escape from neutralizing antibodies. bioRxiv.

68. Travanty E, Zhou B, Zhang H, Di YP, Alcorn JF, Wentworth DE, Mason R, Wang J. 2015. Differential Susceptibilities of Human Lung Primary Cells to H1N1 Influenza Viruses. J Virol 89:11935–11944.

69. Riegger D, Hai R, Dornfeld D, Mänz B, Leyva-Grado V, Sánchez-Aparicio MT, Albrecht R a, Palese P, Haller O, Schwemmle M, García-Sastre A, Kochs G, Schmolke M. 2014. The nucleoprotein of newly emerged H7N9 influenza A virus harbors a unique motif conferring resistance to antiviral human MxA. J Virol.

70. Stetson DB, Medzhitov R. 2006. Type I interferons in host defense. Immunity 25:373–81.

71. Vilcek J. 2006. Fifty years of interferon research: aiming at a moving target. Immunity.

72. Schoggins JW, Wilson SJ, Panis M, Murphy MY, Jones CT, Bieniasz P, Rice CM. 2011. A diverse range of gene products are effectors of the type I interferon antiviral response. Nature 472:481–5.

73. Tough DF. 2004. Type I Interferon as a Link Between Innate and Adaptive Immunity through Dendritic Cell Stimulation. Leuk Lymphoma 45:257–264.

74. Bon A Le, Tough D. 2002. Links between innate and adaptive immunity via type I interferon. Curr Opin Immunol 432–436.

75. Chakrabarti AK, Pasricha G. 2013. An insight into the PB1F2 protein and its multifunctional role in enhancing the pathogenicity of the influenza A viruses. Virology 440:97–104.

76. Yoshizumi T, Ichinohe T, Sasaki O, Otera H, Kawabata S, Mihara K, Koshiba T. 2014. Influenza A virus protein PB1-F2 translocates into mitochondria via Tom40 channels and impairs innate immunity. Nat Commun 5:4713.

77. 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.

78. Liedmann S, Hrincius ER, Guy C, Anhlan D, Dierkes R, Carter R, Wu G, Staeheli P, Green DR, Wolff T, McCullers J a, Ludwig S, Ehrhardt C. 2014. Viral suppressors of the RIG-I-mediated interferon response are pre-packaged in influenza virions. Nat Commun 5:5645.

CHAPTER 2

QUANTIFYING THE EVOLUTIONARY POTENTIAL AND CONSTRAINTS OF A DRUG-TARGETED VIRAL PROTEIN

2.1 Abstract

RNA viruses are notorious for their ability to evolve rapidly under novel environments. It is known that the high mutation rate of RNA viruses can generate huge genetic diversity to facilitate viral adaptation. However, less attention has been paid to the underlying fitness landscape that represents the selection forces on viral genomes. Here we systematically quantified the distribution of fitness effects (DFE) of single amino acid substitutions (86 amino acids total) in the drug-targeted region of NS5A protein of Hepatitis C Virus (HCV). We found that the majority of non-synonymous substitutions incur large fitness costs, suggesting that NS5A protein is highly optimized in natural conditions. Furthermore, we characterized the evolutionary potential of HCV by subjecting the mutant viruses to varying concentrations of an NS5A inhibitor Daclatasvir. As the selection pressure increases, the DFE of beneficial mutations shifts from an exponential distribution to a heavy-tailed distribution with a disproportionate number of exceptionally fit mutants. The number of available beneficial mutations and the selection coefficient are both found to increase at higher levels of antiviral drug concentration, as predicted by a pharmacodynamics model describing viral fitness as a function of drug concentration. Our large-scale fitness data of mutant viruses also provide insights into the biophysical basis of evolutionary constraints and the role of the genetic code in protein evolution.

2.2 Introduction

In our evolutionary battles with microbial pathogens, RNA viruses are among the most formidable foes. HIV-1 and Hepatitis C Virus acquire drug resistance in patients under antiviral therapy. Influenza and Ebola virus cross the species barrier to infect human hosts. Understanding the evolution of RNA viruses is therefore of paramount importance for developing antivirals and vaccines and assessing the risk of future emergence events [1–3]. Comprehensive characterization of viral fitness landscapes, and the principles underpinning them, will provide us with a map of evolutionary pathways accessible to RNA viruses and guide our design of effective strategies to limit antiviral resistance, immune escape and cross-species transmission [4–6].

Although the concept of fitness landscapes has been around for a long time [7], we still know little about their properties in real biological systems. Previous empirical studies of fitness landscapes have been constrained by very limited sampling of sequence space. In a typical study, mutants are generated by site-directed mutagenesis and assayed for growth rate individually. We and others have recently developed a high-throughput technique, often referred to as “deep mutational scanning”, to profile the fitness effect of mutations by integrating deep sequencing with selection experiments [8–10]. This novel application of next generation sequencing has raised an exciting prospect of large-scale fitness measurements [11–14] and a revolution in our understanding of molecular evolution [15].

The distribution of fitness effects (DFE) of mutations is a fundamental entity in genetics and reveals the local structure of a fitness landscape [16–23]. Deleterious mutations are usually abundant and impose severe constraints on the accessibility of fitness landscapes. In contrast, beneficial mutations are rare and provide the raw materials of adaptation. Quantifying the DFE of RNA viruses is crucial for understanding how these pathogens evolve to acquire drug resistance and surmount other evolutionary challenges.

The model system used in our study is Hepatitis C Virus (HCV), a positive sense single-stranded RNA virus with a genome of ~9.6 kb. The biology of HCV has been studied extensively in the past two decades and provides an excellent model system of human RNA viruses. We applied high-throughput fitness assays to map the fitness effects of all single amino acid substitutions in domain IA (amino acid 18-103) of HCV NS5A protein (Methods). This domain is the target of several directly-acting antiviral drugs, including Daclatasvir (DCV) [24]. We profiled the DFE of HCV NS5A protein under varying levels of positive selection by tuning the concentration of the antiviral drug DCV. In addition, we studied how viral evolution is constrained by deleterious mutations that impact protein stability. Finally, we analyzed the shape of the DFE in nucleotide sequence space and analyzed how the structure of the genetic code influences protein evolution.

2.3 Results

2.3.1 PROFILING THE FITNESS LANDSCAPE OF HCV NS5A PROTEIN

To study the DFE of mutations of HCV NS5A protein, we used a previously constructed library of mutant viruses using saturation mutagenesis [11]. Briefly, each codon in the mutated region was randomized to cover all possible single amino acid substitutions. We observed 2520 non-synonymous mutations in the plasmid library, which covered 99.6% (1628 out of 1634) of all possible single amino acid substitutions, as well as 105 synonymous mutations. After transfection, we performed selection on the mutant viruses in an HCV cell culture system [25]. Mutants with frequency below a certain cutoff after transfection were assigned as lethal mutations (Methods). The relative fitness of a mutant virus to the wild-type virus was calculated based on the changes in frequency of the mutant virus and the wild-type virus after one round of selection in cell culture (**Figure S2-1**).

Our experiment provides a comprehensive profiling of the local fitness landscape of all single amino acid mutations. As expected, the fitness effects of synonymous mutations were nearly neutral, while most non-synonymous mutations were deleterious (**Figure 2-1**). After grouping together non-synonymous mutations leading to the same amino acid substitution, we found that around 90% of single amino acid mutations had fitness costs and almost half of them were found to be lethal (**Figure S2-2**). The high sensitivity to mutations in HCV NS5A, an essential protein for viral replication, is generally consistent with previous mutagenesis studies of RNA viruses [26]. Our data support the view that RNA viruses are very sensitive to the effect of deleterious mutations, possibly due to the compactness of their genomes [27,28].

Using the distribution of fitness effects of synonymous mutations as a benchmark for neutrality, we identified that only 3.4% of single amino acid mutations are beneficial (Methods). The estimated fraction of beneficial mutations is consistent with previous small-scale mutagenesis studies in viruses including bacteriophages, vesicular stomatitis virus, etc. [16,26,29,30]. Our results indicate that HCV NS5A protein is under strong purifying selection, suggesting that viral proteins are highly optimized in their natural conditions.

2.3.2 EVOLUTIONARY POTENTIAL AS A FUNCTION OF POSITIVE SELECTION

Beneficial mutations are the raw materials of protein adaptation [16]. Previous studies have found that the evolvability of proteins is a function of purifying selection [31]. In this study, we aimed to study the role of positive selection in modulating the evolutionary potential of drug-targeted viral proteins. In addition to growing viruses in the natural condition without drugs, we selected the mutant library in 10, 40 and 100 pM of a potent HCV NS5A inhibitor Daclatasvir (DCV). The drug concentrations were chosen based on in vitro IC50 of wild type HCV virus (~20 pM) to represent different levels of positive selection (mild, intermediate and strong).

By tuning the concentration of DCV, we observed a shift in the DFE of beneficial mutations (**Figure 2-2**, **Figure S2-3**). At higher drug concentrations, we observed an increase in the average selection coefficient as well as the total number of beneficial mutations (**Table 2-1**). We further tested whether the shape of this distribution changed under drug selection. Previous empirical studies supported the hypothesis that the DFE of beneficial mutations is exponential [26,29,32–39]. Following a maximum likelihood approach, we fit the DFE of beneficial mutations to the Generalized Pareto Distribution (**Figure S2-5**). The fitted distribution (**Table 2-1**) is described by two parameters: a scale parameter (τ), and a shape parameter (κ) that determines the behavior of the distribution's tail. Using a likelihood-ratio test [40], we found that the distribution was exponential ($\kappa = 0$) in the natural condition without drug selection, but shifted to a heavy-tailed distribution ($\kappa > 0$) in the presence of DCV, a condition that the wild type virus was poorly adapted to. Our observation confirms the prediction that the shape of the DFE of beneficial mutations is dependent on how well adapted the wild type is in a certain environment [16]. When individuals encounter novel environments (i.e. novel forms of selection pressure) [41,42], the fitness of the wild type is no longer top-ranking and the DFE is expected to deviate from the exponential distribution [43].

A simple pharmacodynamics model describing viral fitness as a function of drug concentration (i.e. phenotype-fitness mapping) can explain the changing spectra of beneficial mutations upon drug treatment (**Figure 2-2**). For example, mutations that reduce a protein's binding affinity to drug molecules (i.e. with a higher inhibitory concentration than wild-type) may come with a fitness cost [10]. Thus, a drug-resistant mutant that is deleterious in the absence of drug may become beneficial under drug selection, leading to an increase in the number of beneficial mutations. Moreover, the relative fitness of the drug-resistant mutant is expected to increase with stronger selection pressure (**Figure 2-2**, dashed line).

The dose response curves were previously measured for a set of mutants constructed by site-directed mutagenesis (**Figure S2-6**) [11]. Indeed, we found that the relative fitness of drug-resistant mutants increased at higher drug concentration (**Figure 2-2**); in contrast, drug-sensitive mutants became less fit under drug selection. Furthermore, based on this set of mutants with validated dose response curves, we were able to use the fitness measurements to estimate the IC50 of all mutants in our library (**Figure S2-7**). In particular, we found that a small group of mutations were highly resistant to DCV and this could explain the heavy-tailed DFE of beneficial mutations under drug selection. Overall, our results suggest that the evolutionary potential of proteins is modulated by the strength of positive selection, in addition to the previous findings on the role of purifying selection [31].

2.3.3 Deleterious mutations as evolutionary constraints

While beneficial mutations open up adaptive pathways to genotypes with higher fitness, mutations that reduce fitness impose constraints on the evolution of viruses. To understand the biophysical basis of mutational effects [44], we took advantage of the available structural information. The crystal structure of NS5A domain I is available excluding the amphipathic helix at N-terminus [45,46].

We found that the fitness effects of deleterious mutations at buried sites (i.e. with lower solvent accessibility) were more pronounced than those at surface exposed sites (**Figure 2-3**, **Figure S2-8**) [47]. Moreover, we performed simulations of protein stability for individual mutants using the PyRosetta program (Methods) [48,49]. A mutation with $\Delta\Delta G>0$, i.e. shifting the free energy difference to favor the unfolded state, is expected to destabilize the protein. We found that mutations that decreased protein stability led to reduced viral fitness (**Figure 2-3**). For example, mutations at a stretch of highly conserved residues (F88-N91) that run through the core of NS5A protein tended to destabilize the protein and significantly reduced the viral fitness (**Figure S2-9**). Mutations that increase $\Delta\Delta G$ beyond a threshold were mostly lethal. This is consistent with the threshold robustness model, which predicts that proteins become unfolded after using up the stability margin [50–52]. In contrast, mutations at some sites were highly deleterious despite having little impact on protein stability, suggesting that evolution at these sites may be under additional constraints to preserve protein function [53–55], such as RNA binding [56,57].

We further tested whether the viral replication fitness in cell culture was predictive of evolutionary landscapes of viruses in patients [58]. This is critical for the extrapolation of viral replication fitness from in vitro to in vivo [59]. We analyzed sequence diversity of HCV sequences in the database of Los Alamos National Lab (Methods). Indeed, we found that the within-patient sequence diversity at each site was highly correlated to the replication fitness measured in cell culture (Spearman's ρ=0.82, **Figure 2-3**), suggesting that fitness landscapes profiled in laboratory settings can provide insights into evolutionary pathways of viruses in nature [60].

2.3.4 The role of the genetic code in protein evolution

So far we have considered the spectrum of beneficial and deleterious mutations in the amino acid sequence space [61]. In fact, the evolution of viral proteins in the nucleotide sequence space faces additional constraints posed by the genetic code, because most mutations in RNA viruses occur as point mutations during genome replication. Our fitness data of single codon mutants (replaced by NNK) provides a unique opportunity to examine impacts of the genetic code on the evolution of proteins [62].

Due to codon degeneracy, many point mutations are synonymous and thus less likely to be deleterious than 2 or 3-nt substitutions. For non-synonymous mutations, we found that the deleterious impacts on fitness increased with the number of nucleotide substitutions (**Figure 2-4**), supporting the hypothesis that the structure of the standard genetic code can buffer the mutational load [62]. This observation is consistent with the facts that amino acids with similar biochemical properties tend to be adjacent in the genetic code [63], and that mutating to biochemically similar amino acids is less likely to decrease fitness (**Figure S2-2**).

For HCV and other RNA viruses, there is an observed transition:transversion bias in evolution [64,65]. This phenomenon has been attributed to two different, but not mutually exclusive, causes: 1) the "mutation hypothesis" argues for a transition:transversion bias in the mutation rate, which is bolstered by experimental measurement of de novo mutation rates in viruses [66]; 2) the "selection hypothesis" argues that natural selection favors amino acid replacements via transition [67]. We tested the "selection hypothesis" using the non-synonymous point mutations in our library (**Figure 2-4**). We found that the fraction of lethal mutations caused by transversions was slightly larger than transitions, but the difference was not statistically significant. Together with previous studies in other systems [67], our results suggest that the “selection hypothesis” is unlikely to be the major cause underlying the transition:transversion bias in evolution of viral proteins.

In addition, we observed a slight (though not statistically significant) enrichment of beneficial mutations in point mutations under the natural condition (**Figure 2-4**). Under drug selection, our conclusions on the shifting DFE of beneficial single amino acid mutations still held true for point mutations (**Figure S2-10**, **Table S2-1**). Furthermore, we observed that beneficial mutations were significantly depleted in 3-nt substitutions (**Figure S2-11**). However, this difference in the fraction of beneficial mutations can be confounded by the fact that 3-nt substitutions are more likely to be lethal. As pointed out in a previous study on the potential benefits of the genetic code in protein evolution [62], mutational robustness and the enrichment of beneficial mutations may actually be two sides of the same coin.

2.4 Discussion

Mutation accumulation [68] and site-directed mutagenesis [69] are traditional approaches to examine the DFE. Both methods provide pivotal insights into the shape of the DFE, yet with limitations. The site-directed mutagenesis approach requires fitness assays for each mutant and can only provide a sparse sampling of mutations. The sampling of sequence space in a mutation accumulation experiment is biased towards large-effect beneficial mutations, as they are more likely to fix in the population. In contrast, the “deep mutational scanning” approach [9,10], which utilizes high-throughput sequencing to simultaneously assay the fitness or phenotype of a library of mutants, allows for unbiased and large-scale sampling of fitness landscapes and thus is ideal for studying the characteristics of empirical DFE.

The shape of the DFE determines mutational robustness [69–71]. Our study quantified the fitness effects of single amino acid substitutions in the drug-targeted region of an essential viral protein (86 amino acids, 1628 out of 1634 possible substitutions). In general, the empirical DFE of HCV NS5A was consistent with previous findings that viral proteins were highly optimized in the natural condition and very sensitive to the effects of deleterious mutations. Moreover, given the advantages of our saturation mutagenesis, we were able to use the fitness data to test multiple hypotheses of protein evolution, including the role of the genetic code in buffering mutational load, and the cause of transition:transversion bias. In the future, profiling the DFE in a range of different systems will allow us to test the generality of our conclusions.

In our study, we have used the fitness effects of synonymous mutations to determine the threshold of neutrality. Synonymous mutations are usually expected to have no or minimal influence on phenotype or fitness, but this view is being increasingly challenged as the effect of synonymous mutations on protein expression and folding becomes elucidated, such as via mRNA secondary structures or codon usage [72,73]. One interesting observation in our selection experiments is that some synonymous mutations seemed to have phenotypic effects on drug sensitivity. Although this is not the focus of our study, understanding the mechanism of natural selection at the RNA level and its implications for molecular evolution, particularly in the context of RNA viruses, may be a fruitful area for future studies.

One often overlooked point is that DFE will vary as a function of selection pressure [31,74,75]. For example, mutations that impair function would become more deleterious with increasing pressure of purifying selection, thus leading to reduced protein evolvability [31]. In this study, we have focused on gain-of-function mutations in a novel environment. The pleiotropic effect of mutations causes the spectrum of beneficial mutations to shift between the natural condition and the condition with drug selection. Moreover, mutations enabling the new function (e.g. drug resistance) become more beneficial with increasing pressure of positive selection.

Although different systems have distinct protein-drug interactions that lead to different resistance profiles [76], the results in our study provide a general framework to study DFE of drug-targeted proteins. Future studies along this line will further our understanding of how proteins evolve new functions under the constraint of maintaining their original function [77], as exemplified in the evolution of resistance to directly-acting antiviral drugs [78]. We have also demonstrated that the fitness data could be utilized to infer drug sensitivity of mutants and inform predictive modeling of within-patient viral dynamics [4]. Quantifying the characteristics of DFE of drug-targeted proteins under different environments (e.g. varying levels of selection pressure, or conflicting selection pressures), would allow us to assess repeatability in the outcomes of viral evolution [79] and guide the design of therapies to minimize drug resistance [80].

2.5 Conclusions

Many RNA viruses adapt rapidly to novel selection pressures, such as antiviral drugs. Understanding how pathogens evolve under drug selection is critical for the success of antiviral therapy against human pathogens. By combining deep sequencing with selection experiments in cell culture, we have quantified the distribution of fitness effects of mutations in the drug-targeted domain of Hepatitis C Virus NS5A protein. Our results indicate that the majority of single amino acid substitutions in NS5A protein incur large fitness costs. Combined with stability predictions based on protein structure, our fitness data reveal the biophysical constraints underlying the evolution of viral proteins. Furthermore, by subjecting the mutant viruses to positive selection under an antiviral drug, we find that the evolutionary potential of viral proteins in a novel environment is modulated by the strength of selection pressure.

2.6 Materials and Methods

Mutagenesis

The mutant library of HCV NS5A protein domain IA (86 amino acids) was constructed using saturation mutagenesis as previously described [11]. In brief, the entire region was divided into five sub-libraries each containing 17-18 amino acids. NNK (N: A/T/C/G, K: T/G) was used to replace each amino acid. The oligos, each of which contains one random codon, were synthesized by IDT. The mutated region was ligated to the flanking constant regions, subcloned into the pFNX-HCV plasmid and then transformed into bacteria. The pFNX-HCV plasmid carrying the viral genome was synthesized in Dr. Ren Sun’s lab based on the chimeric sequence of genotype 2a HCV strains J6/JFH1.

Cell culture

The human hepatoma cell line (Huh-7.5.1) was provided by Dr. Francis Chisari from the Scripps Research Institute, La Jolla. The cells were cultured in T-75 tissue culture flasks (Genesee Scientific) at 37 oC with 5% CO2. The complete growth medium contained Dulbecco's Modified Eagle's Medium (Corning Cellgro), 10% heat inactivated Fetal Bovine Serum (Omega Scientific), 10 mM HEPES (Life Technologies), 1x MEM Non-Essential Amino Acids Solution (Life Technologies) and 1x Penicillin-Streptomycin-Glutamine (Life Technologies).

Selection

Plasmid mutant library was transcribed in vitro using T7 RiboMAX Express Large Scale RNA Production System (Promega) and purified by PureLink RNA Mini Kit (Life Technologies). 10 µg of in vitro transcribed RNA was used to transfect 4 million Huh-7.5.1 cells via electroporation by Bio-Rad Gene Pulser (246 V, 950 µF). The supernatant was collected 144 hours post transfection and virus titer was determined by immunofluorescence assay. The viruses collected after transfection were used to infect ~2 million Huh-7.5.1 cells with an MOI at around 0.1-0.2. The five mutant libraries were passaged for selection separately as previously described [81]. For the three different levels of selection pressure, the growth media was supplemented with 10 pM, 40 pM and 100 pM HCV NS5A inhibitor Daclatasvir (BMS-790052), respectively. The supernatant was collected at 144 hours post infection.

Preparation of Illumina sequencing samples

For each sample, viral RNA was extracted from 700 µl supernatant collected after transfection and after selection using QIAamp Viral RNA Mini Kit (Qiagen). Extracted RNA was reverse transcribed into cDNA by SuperScript III Reverse Transcriptase Kit (Life Technologies). The targeted region in NS5A (51-54 nt) was PCR amplified using KOD Hot Start DNA polymerase (Novagen). The Eppendorf thermocycler was set as following: 2 min at 95 °C; 25 to 35 three-step cycles of 20 s at 95 °C,15 s at 52-56 °C (sub-library #1, 52 °C; #2, 52 °C; #3, 52 °C; #4, 56 °C; #5, 54 °C) and 25s at 68 °C; 1 min at 68 °C. The number of PCR cycles are chosen based on the copy number of cDNA templates as determined by qPCR (Bio-Rad). The PCR primers are listed in **Table S2-1**. The PCR products were purified using PureLink PCR Purification Kit (Life Technologies) and prepared for Illumina Hiseq 2000 sequencing (paired-end 100 bp) following 5'-phosphorylation using T4 Polynucleotide Kinase (New England BioLabs), 3’ dA-tailing using dA-tailing module (New England BioLabs), and TA ligation of the adapter using T4 DNA ligase (Life Technologies). Each sample was tagged with a unique 3-bp customized barcodes, which were part of the adapter sequence and were sequenced as the first three nucleotides in both the forward and reverse reads [55] (**Table S2-2**).

Analysis of sequencing data

The sequencing data were parsed by SeqIO function of BioPython. The reads from different samples were de-multiplexed by the barcodes and mapped to the entire mutated region in NS5A by allowing at maximum 5 mismatches with the reference genome (**Table S2-3**) [11]. Since both forward and reverse reads cover the whole amplicon, we used paired reads to correct for sequencing errors. A mutation was called only if it was observed in both reads and the quality score at the corresponding position was at least 30. Sequencing reads containing mutations not supposed to appear in the mutant library were excluded from downstream analysis. The sequencing depth for each sub-library is at least ~105 and two orders of magnitude higher than the library complexity.

Calculation of relative fitness

For each condition of selection experiments (i.e. different concentration of Daclatasvir [DCV]), the relative fitness (RF) of a mutant virus to the wild-type virus is calculated by the relative changes in frequency after selection,

$$RF_{mut}([DCV]) = \left(\frac{f_{mut}^{T=1}}{f_{mut}^{T=0}} \right) \Big/ \left(\frac{f_{WT}^{T=1}}{f_{WT}^{T=0}} \right)$$

where $f_{mut}^{T=round}$ and $f_{WT}^{T=round}$ is the frequency of the mutant virus and the wild-type virus at round 0 (after transfection) or round 1 (after selection). The fitness of wild-type virus is normalized to 1. The fitness values estimated from one round have been shown to be highly consistent to estimates from multiple rounds of selection [11].

Mutants with less than 10 read counts in the plasmid library were filtered. A mutation was considered lethal if at least one of the two criteria was met: 1) after transfection, the mutant had less than 10 read counts; 2) after transfection, the ratio between the mutant’s frequency and the wild-type’s frequency was smaller than 10-4. The thresholds for beneficial and deleterious mutations were defined as $1+2\sigma_{silent}$ and $1-2\sigma_{silent}$, respectively. σ_{silent} is the standard deviation of the fitness effects of synonymous mutations under the natural condition (**Figure 2-1**). The fitness effects of non-synonymous mutations leading to the same amino acid substitution were averaged to estimate the fitness effect of the given single amino acid substitution.

Fitting the distribution of fitness effects of beneficial mutations

The distribution of selection coefficients of beneficial mutations were fitted to a Generalized Pareto Distribution following a maximum likelihood approach [40],

$$F(x|\kappa,\tau)=$$

$$\begin{cases} 1-(1+\frac{\kappa}{\tau}x)^{-1/\kappa}, x \geq 0, \ \ if \kappa > 0 & \text{(Frechet)} \\ 1-(1+\frac{\kappa}{\tau}x)^{-1/\kappa}, 0 \leq x < -\frac{\tau}{\kappa}, \ \ \text{i}f \kappa < 0 & \text{(Weibull)} \\ 1-e^{-x/\tau}, x \geq 0, \text{ if } \kappa = 0 & \text{(Gumbel)} \end{cases}$$

Only mutations with relative fitness higher than the beneficial threshold $1+2\sigma_{silent}$ were included in the distribution of beneficial mutations. The selection coefficients were normalized to the beneficial threshold. The shape parameter κ determines the tail behavior of the distribution, which can be divided into three domains of attraction: Gumbel domain (exponential tail, κ = 0), Weibull domain (truncated tail, κ < 0) and Fréchet domain (heavy tail, κ > 0). For each selection condition, a likelihood ratio test is performed to evaluate whether the null hypothesis κ = 0 (exponential distribution) can be rejected.

Estimation of IC50 from fitness data

We can quantify the drug resistance of each mutant in the library by computing its fold change in relative fitness,

$$W([DCV]) = \frac{RF_{mut}([DCV])}{RF_{mut}}$$

Here RF_{mut} is the relative fitness of a mutant under the natural condition (i.e. no drug). W is the fold change in relative fitness and represents the level of drug resistance relative to the wild type. W > 1 indicates drug resistance, and W < 1 indicates drug sensitivity.

This empirical measure of drug resistance can be directly linked to a simple pharmacodynamics model [78], where the viral replicative fitness is modeled as a function of drug dose,

$$W([DCV]) = \frac{RF_{mut}([DCV])}{RF_{mut}} = \left(\frac{IC_{mut}}{[DCV]+IC_{mut}} \right) \Bigg/ \left(\frac{IC_{wt}}{[DCV]+IC_{wt}} \right)$$

Here IC denotes the half-inhibitory concentration. The Hill coefficient describing the sigmoidal shape of the dose response curve is fixed to 1, as used in fitting the dose response curves of wild-type virus and validated mutant viruses (**Figure S2-6**). We can use the drug resistance score W to infer the dose response of each mutant. Because the dose response curve depends on the duration of drug treatment, we transformed Wobserved (144 hr drug treatment in selection experiments) to Wpredicted (48 hr drug treatment, **Figure S2-7**) and then calculated IC50 using the equation above.

Calculation of relative solvent accessibility

DSSP (http://www.cmbi.ru.nl/dssp.html) was used to compute the Solvent Accessible Surface Area (SASA) [82] from the HCV NS5A protein structure (PDB: 3FQM) [46]. SASA was then normalized to Relative Solvent Accessibility (RSA) using the empirical scale reported in [83].

Predictions of protein stability

ΔΔG (in Rosetta Energy Units) of HCV NS5A mutants was predicted by PyRosetta (version: “monolith.ubuntu.release-104”) as the difference in scores between the monomer structure of mutants (single amino acid mutations from site 32 to 103) and the reference (PDB: 3FQM). The score is designed to capture the change in thermodynamic stability caused by the mutation (ΔΔG) (Das and Baker 2008). The sequence of the reference protein was different from the sequence of the wild-type virus used in this study. Thus instead of directly comparing ΔΔG to fitness effects, we used the median ΔΔG caused by amino acid substitutions at each site.

The PDB file of NS5A dimer was cleaned and trimmed to a monomer (chain A). Next, all side chains were repacked (sampling from the 2010 Dunbrack rotamer library [84]) and minimized for the reference structure using the talaris2014 scoring function. After an amino acid mutation was introduced, the mutated residue was repacked, followed by quasi-Newton minimization of the backbone and all side chains (algorithm: “lbfgs_armijo_nonmonotone”). This procedure was performed 50 times, and the predicted ΔG of a mutant structure is the average of the three lowest scoring structures.

We note that predictions based on NS5A monomer structure were only meant to provide a crude profile of how mutations at each site may impact protein stability. Potential structural constraints at the dimer interface have been ignored, which is further complicated by the observations of two different NS5A dimer structures [45,46]. The reference sequence of NS5A in the PDB file (PDB: 3FQM) is different from the WT sequence used in our experiment by 20 amino acid substitutions.

Within-patient sequence diversity

Aligned nucleotide sequences of HCV NS5A protein were downloaded from Los Alamos National Lab database [85] (all HCV genotypes, ~2600 sequences total) and clipped to the region of interest (amino acid 18-103 of NS5A). Sequences that caused gaps in the alignment of H77 reference genome were manually removed. After translation to amino acid sequences, sequences with ambiguous amino acids were removed (~2300 amino acid sequences after filtering). The sequence diversity at each amino acid site was quantified by Shannon entropy.

Ethics Statement

The use of human cell lines and infectious agents in this paper is approved by Institutional Biosafety Committee at University of California, Los Angeles (IBC #40.10.2-f).

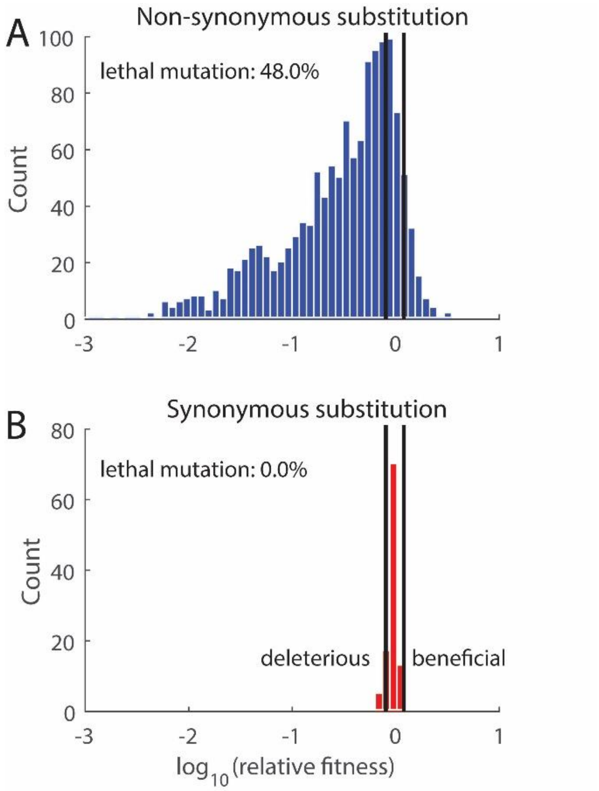


Figure 2-1. Distribution of fitness effects (DFE) of single codon substitutions of HCV NS5A protein. DFE of (A) non-synonymous substitutions and (B) synonymous substitutions. The thresholds (black lines) used for classifying beneficial, nearly neutral and deleterious mutations are determined by the variation of fitness values of synonymous substitutions (Methods). 88.7% of non-synonymous substitutions in NS5A protein are deleterious (among which 48.0% are lethal and not displayed in the histogram), 7.9% are nearly neutral and 3.4% are beneficial mutation.

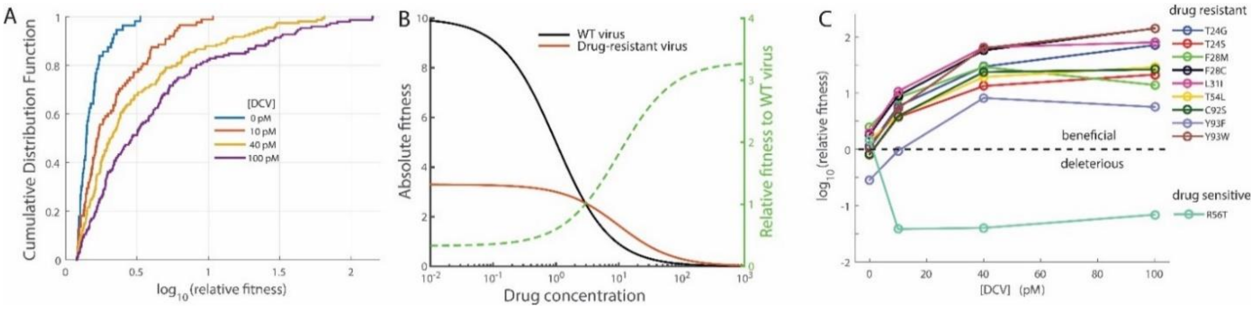


Figure 2-2. The spectrum of beneficial mutations shifts under the selection of an antiviral drug. (A) The cumulative distribution function of relative fitness of beneficial single amino acid substitutions under different selection conditions. (B) Hypothetical dose response curves of the wild-type virus and a drug-resistant mutant virus. Relative fitness of the drug-resistant mutant is expected to increase with drug concentration. (C) Relative fitness of validated drug resistant and sensitive mutations as a function of Daclatasvir concentration.

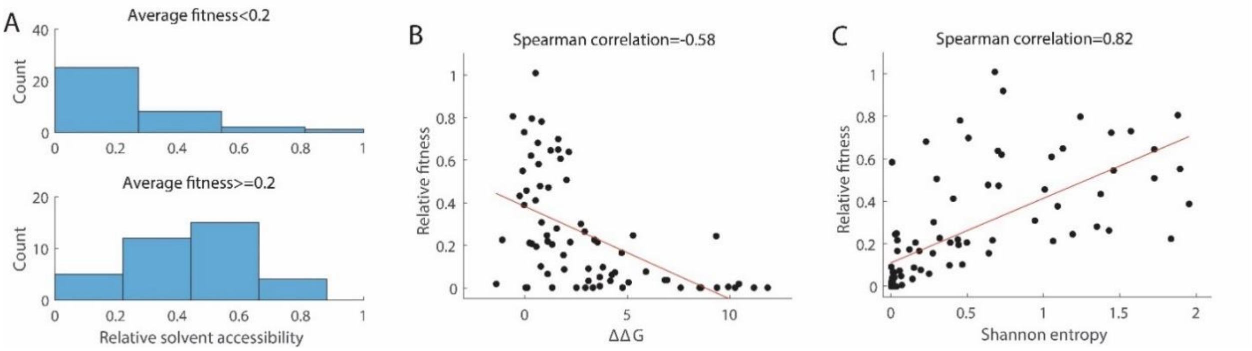


Figure 2-3. Deleterious mutations reveal constraints of protein evolution. (A) Amino acid sites that were less tolerant of mutations (average fitness of mutants <0.2) have lower relative solvent accessibility.. (B) Mutations that destabilized protein stability reduced the viral replicative fitness. Changes in folding free energy ΔΔG (Rosetta Energy Unit) of NS5A monomer were predicted by PyRosetta (Methods). The median of ΔΔG at each amino acid site is shown. (C) The within-patient sequence diversity of HCV NS5A protein at each site is highly correlated to the replicative fitness measured in cell line, suggesting that evolutionary pathways of viral proteins are indeed constrained by mutations that reduce viral replicative fitness. In (B) and (C), the average fitness of observed mutants at each amino acid site is shown. Red lines represent the fits by linear regression and are only used to guide the eye.

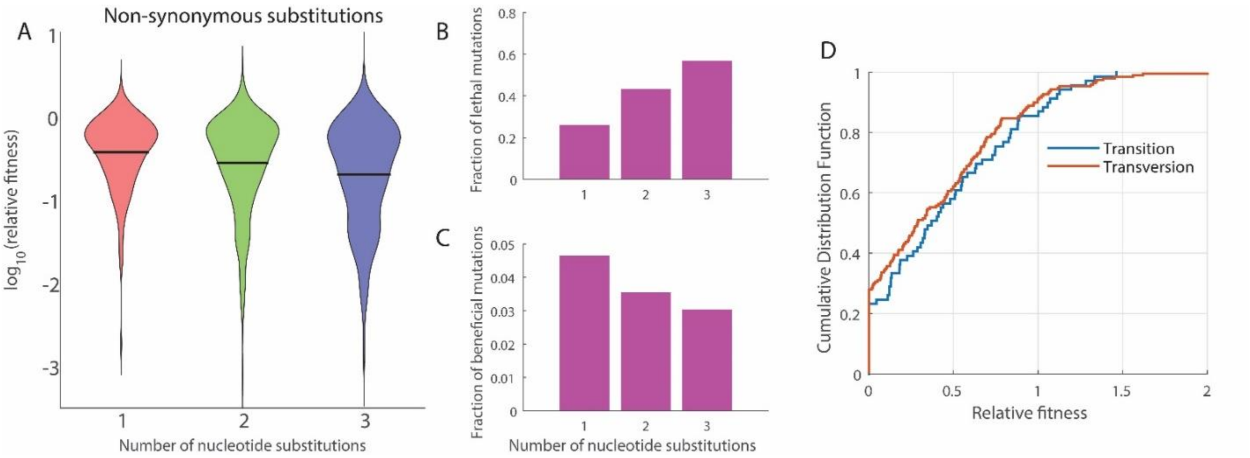


Figure 2-4. The role of the standard genetic code in viral evolution. (A) DFE of non-synonymous substitutions with 1, 2, and 3 nucleotide changes (lethal mutations are not displayed). Black lines indicates the mean. (B) The fraction of lethal mutations increases with the number of nucleotide changes, suggesting that the genetic code is optimized to buffer the mutational load (Chi-squared test, p=1.4×10-21). (C) The fraction of beneficial mutations is slightly enriched for point mutations (Chi-squared test, p=0.41). Only non-synonymous substitutions are included in the analysis. (D) For non-synonymous point mutations, the fitness effect of transitions (n=69) is slightly less deleterious than that of transversions (n=190), but the difference is not significant (two-sample Kolmogorov-Smirnov test, p=0.53).

[DCV]	*Fraction of beneficial codon substitutions*	*Scale parameter τ*	*Shape parameter κ*	*p-value*
0 pM	3.4% (56/1634)	0.26	0.27n.s.	0.09
10 pM	5.4% (88/1634)	0.63	0.62	0.0001
40 pM	9.7% (158/1634)	0.88	1.11	<0.0001
100 pM	9.3% (152/1634)	1.57	1.18	<0.0001

Table 2-1. Statistics of the distribution of fitness effects of beneficial single amino acid substitutions under varying selection pressure. n.s.: cannot reject the null hypothesis that the distribution is exponential (p>0.05).

2.7 Supplementary Materials

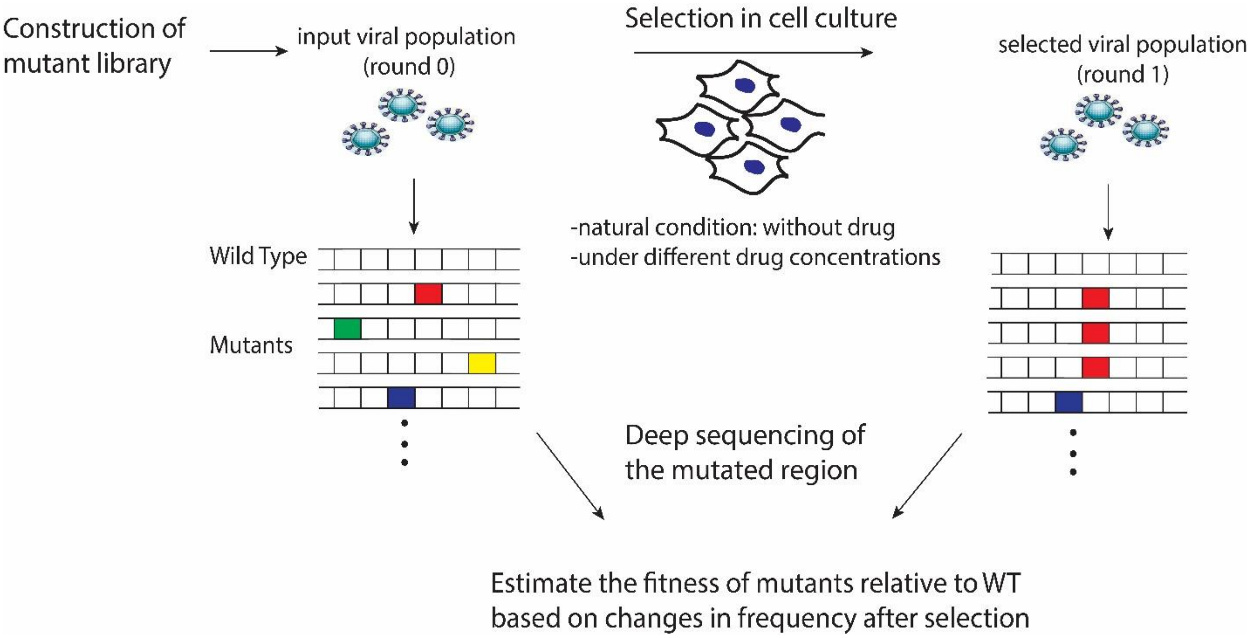

Construction of mutant library
input viral population (round 0)
Selection in cell culture
selected viral population (round 1)
Wild Type
Mutants
-natural condition: without drug
-under different drug concentrations
Deep sequencing of the mutated region
Estimate the fitness of mutants relative to WT based on changes in frequency after selection

Figure S2-1. Experimental workflow of high-throughput fitness assays. We performed the selection of the mutant virus library using HCV cell culture system. Viral RNA was extracted after transfection or after selection, and reverse transcribed into cDNA. The mutated region in NS5A protein was amplified by PCR and sequenced by Illumina HiSeq. The relative fitness of a mutant virus to the wild-type virus was calculated based on the frequency of the mutant virus and the wild-type virus at round 0 (after transfection) and round 1 (after selection). See Methods for more details.

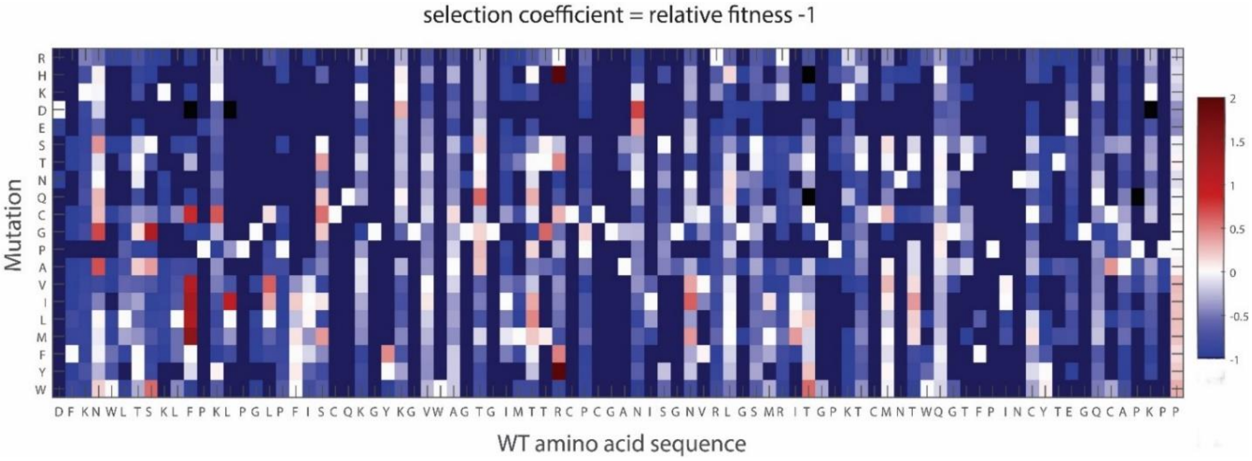


Figure S2-2. Fitness effects of single amino acid substitutions in HCV NS5A protein under native condition. Different amino acid substitutions at the same site can have different fitness effects. The missing variants are colored black; lethal variants are colored dark blue.

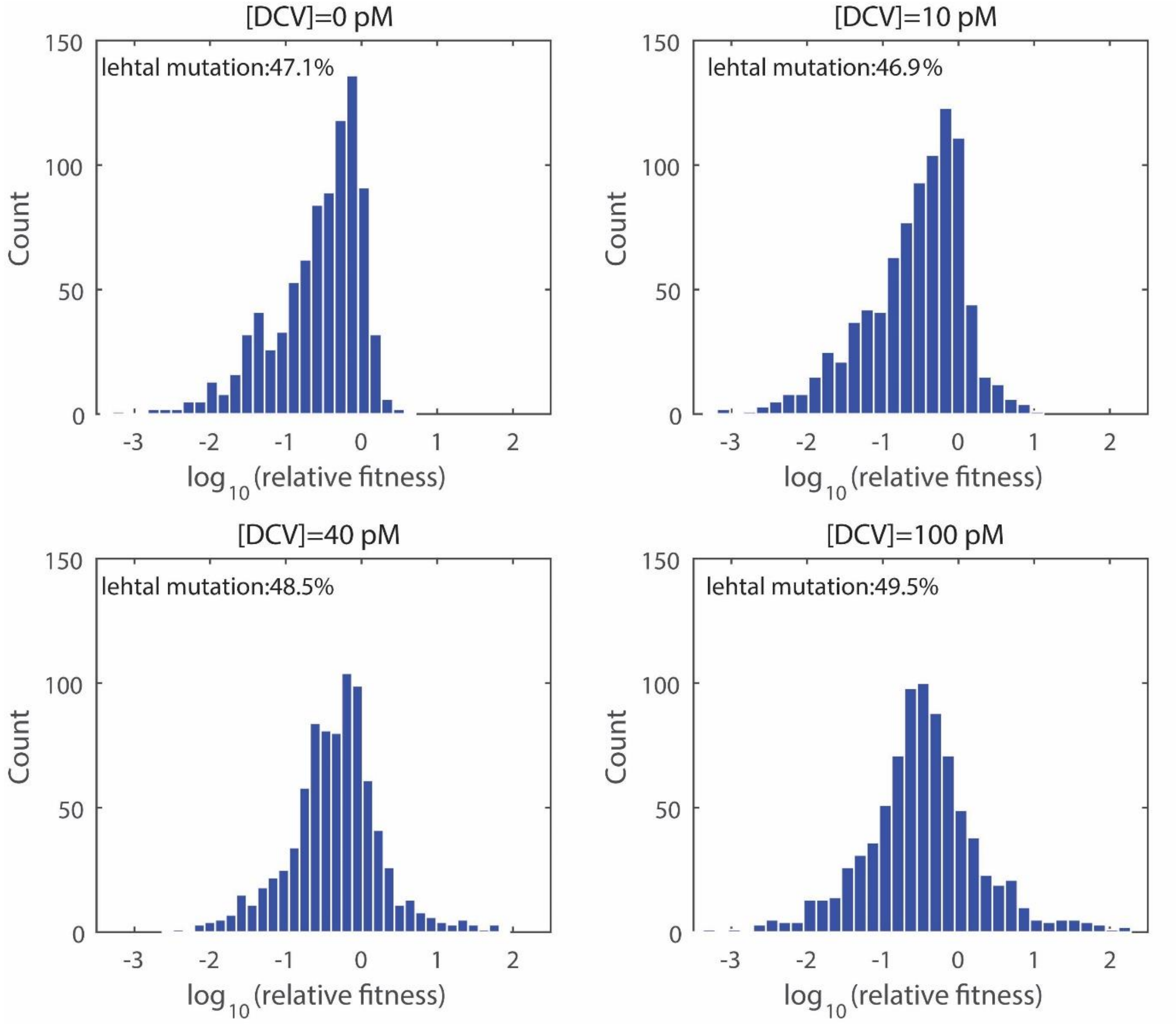


Figure S2-3. DFE of non-synonymous substitutions under drug selection.

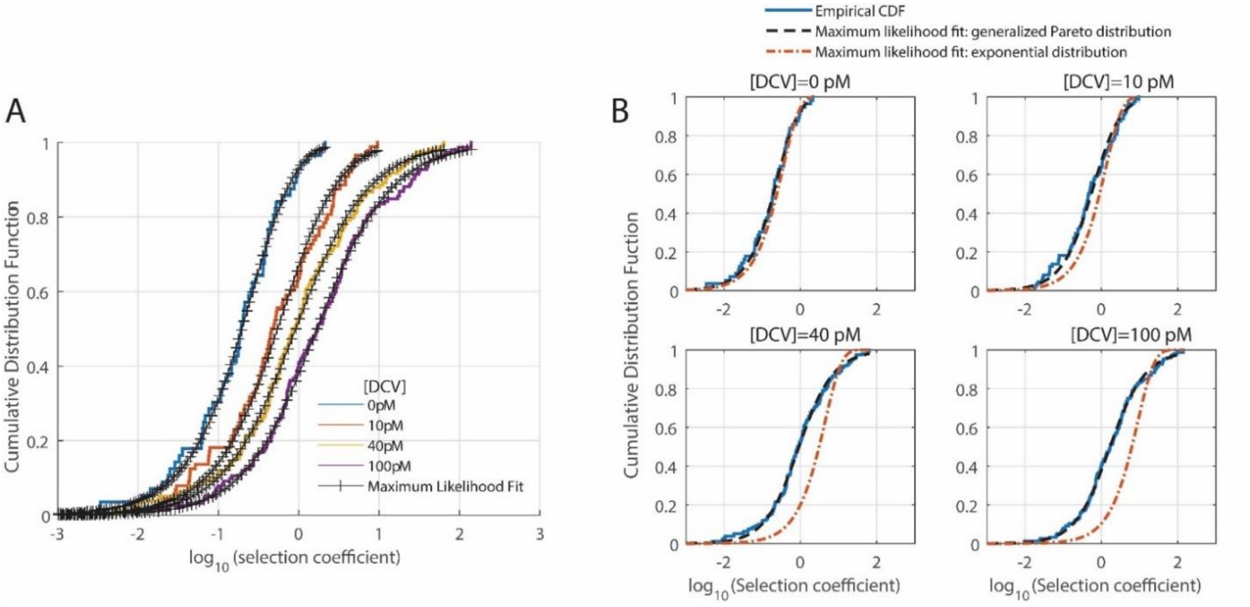


Figure S2-4. Fitted distribution of fitness effects of beneficial single amino acid mutations. (A) Comparison of the fitted distribution to data. (B)The exponential distribution fails to fit the spectrum of beneficial mutations under conditions with drug selection.

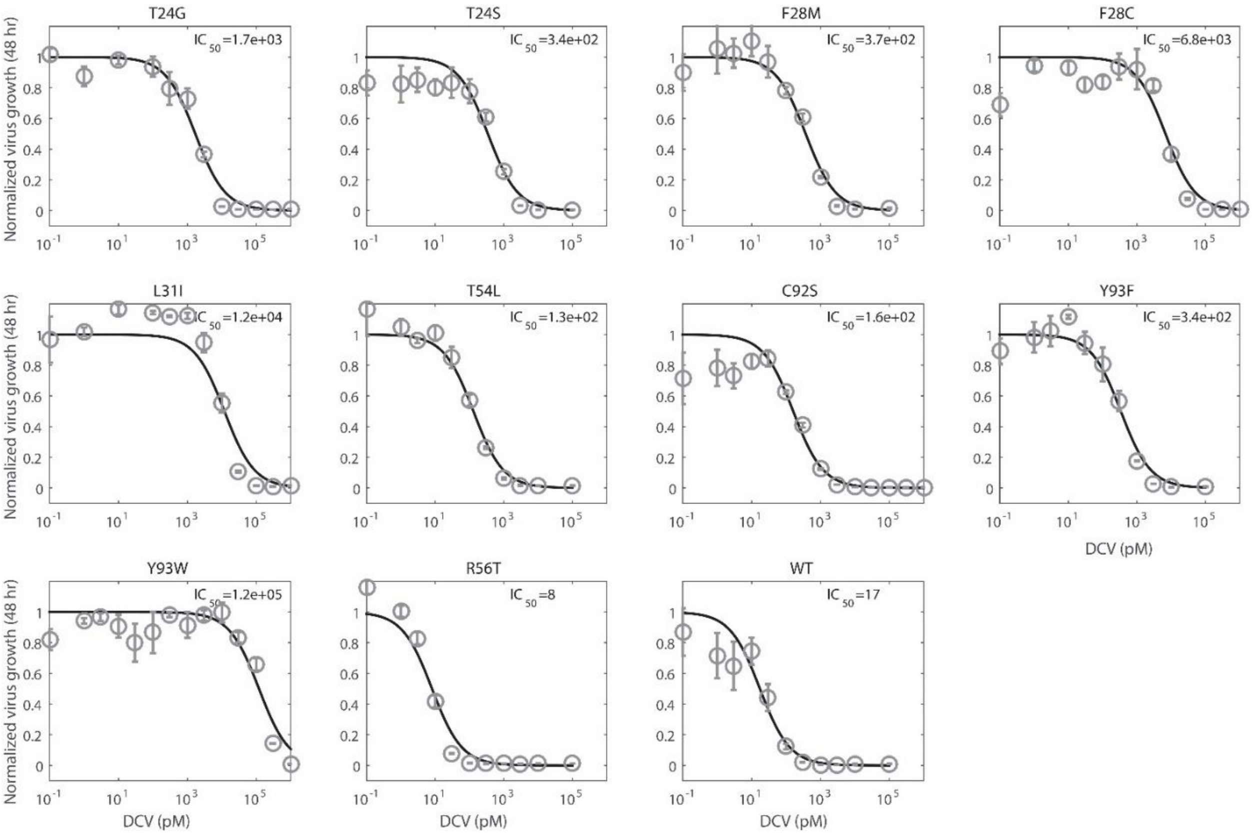


Figure S2-5. Dose response curve of validated mutants (10 drug-resistant mutant, 1 drug-sensitive mutant) and WT virus. The Hill coefficient is fixed at 1 in fitting the dose response

curves (Methods). The unit of IC50 is pM. The virus titer was measured after 48 hr of growth under drug treatment.

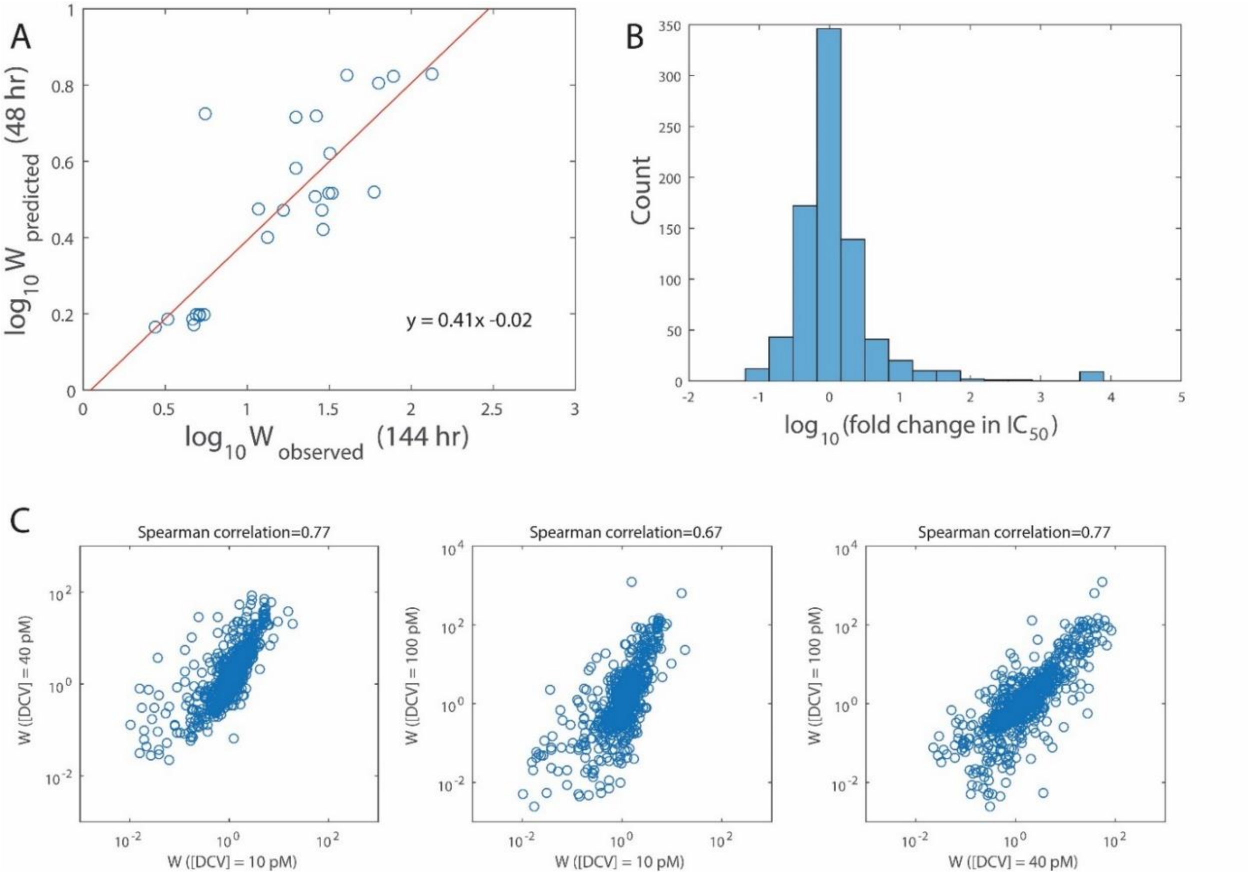


Figure S2-6. Infer IC50 from fitness data under drug selection. (A) W is the fold change in relative fitness and represents the level of drug resistance relative to the wild type (Methods). Because the dose response curve depends on the duration of drug treatment, we normalized Wobserved (144 hr drug treatment in selection experiments) to Wpredicted (48 hr drug treatment, as used in the measurement of dose response curves of validated mutants). If viral growth is always exponential, the exponent is expected to be $\frac{48}{144}=\frac{1}{3}$. The fitted exponent is larger than $\frac{1}{3}$, suggesting that virus titer starts to saturate in 144 hr. (B) The fold change in IC50 caused by a single amino acid substitution is inferred from the measured fitness profiles under native condition and under drug selection (100 pM [DCV]). The existence of a group of highly resistant mutants (>10 fold change in IC50) can explain why DFE of beneficial mutations shifts to a heavy-tailed distribution under drug selection. The resistance score of 9 single amino acid substitutions exceeds the maximum (Methods) and is manually set to 104 pM, which can be seen by the small peak in the right tail of histogram. (C) The measurement of drug resistance is consistent across different conditions of drug selection.

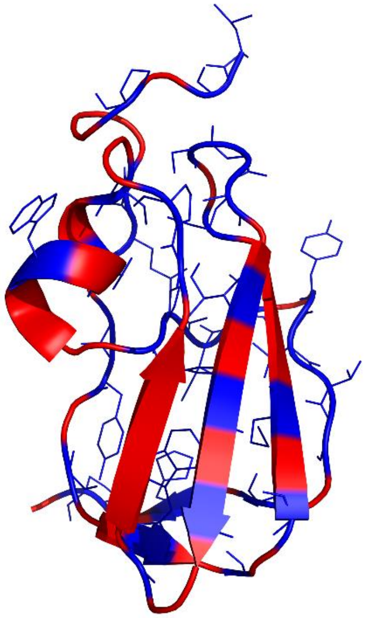


Figure S2-7. Mutations at buried sites are highly deleterious. The structure of HCV NS5A monomer is visualized by PyMOL (PDB: 3FQM, chain A). Amino acid sites with an average fitness less than 0.2 are in blue and the corresponding side chains are shown.

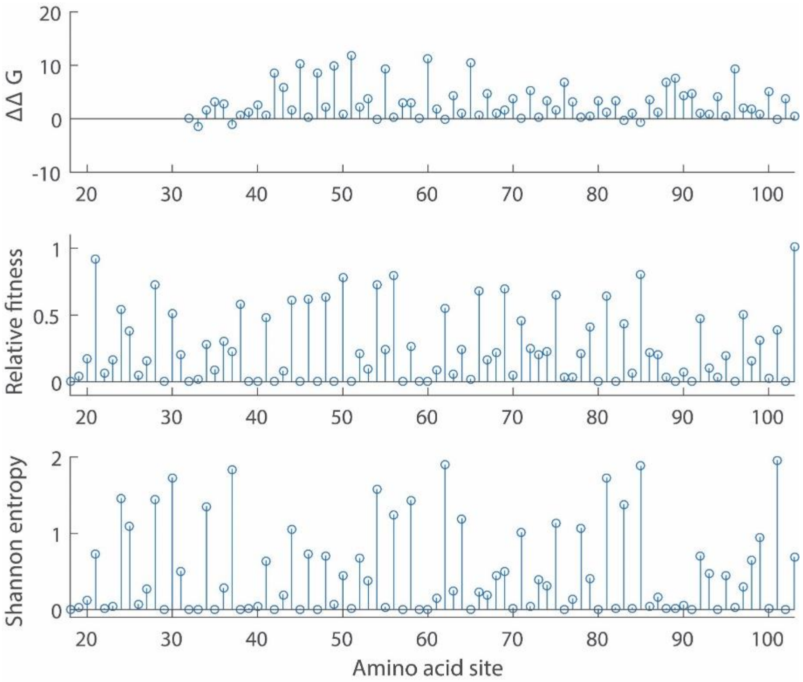


Figure S2-8. Effects of mutations on protein stability, viral fitness and sequence diversity at each amino acid site. The fitness is averaged over all observed mutants at each amino acid site. The medium of ddG at each amino acid site is shown.

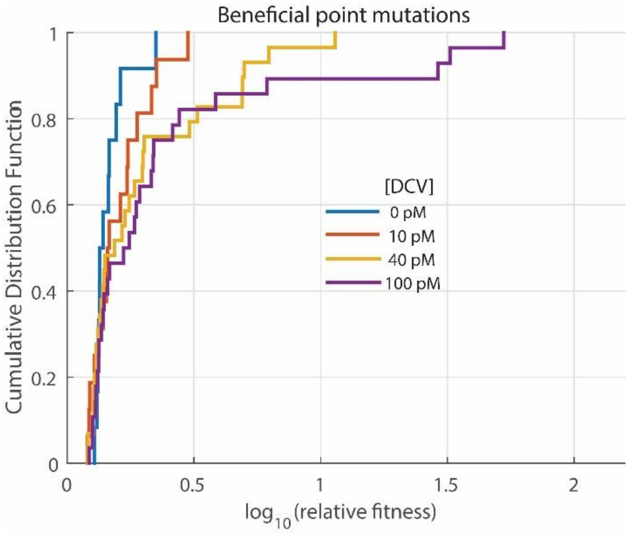


Figure S2-9. Distribution of fitness effects of beneficial point mutations. Only non-synonymous mutations are included.

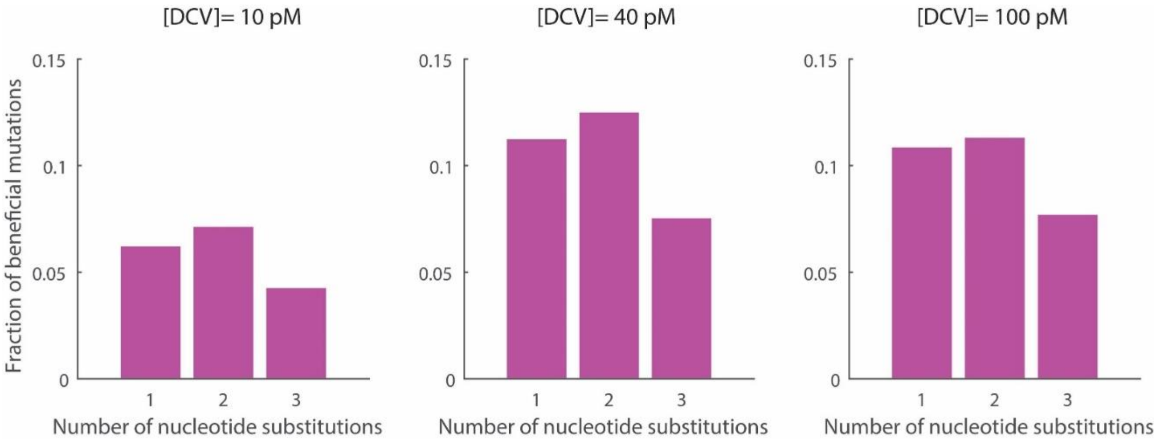


Figure S2-10. The potential role of genetic code in adaptive mutations. For non-synonymous codon substitutions, the fraction of beneficial mutations with 3-nt changes is significantly lower than that of 1-nt or 2-nt mutations under drug selection (Pearson's chi-squared test: [DCV]=10 pM, p=1.2×10-2, 40 pM: p=3.7×10-4, 100 pM: p=1.1×10-2).

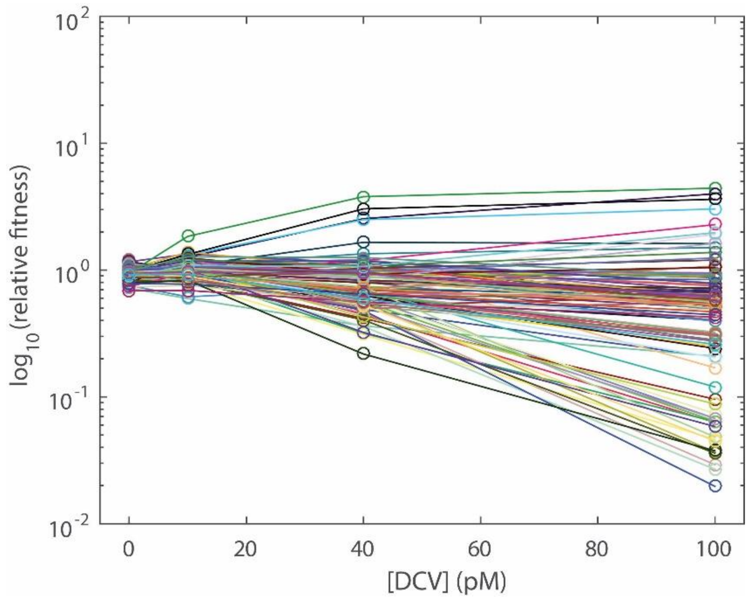


Figure S2-11. Fitness effects of synonymous mutations under drug selection. Relative fitness of some synonymous mutations increased or decreased with drug concentration, suggesting that these mutations may have phenotypic effects on drug sensitivity.

[DCV]	**Fraction of beneficial mutations**	**Scale parameter**	**Shape parameter**	**p-value**
0 pM	4.6% (12/259)	0.13	0.35n.s.	0.42
10 pM	6.2% (16/259)	0.41	0.08n.s.	0.85
40 pM	11.2% (29/259)	0.32	1.06	0.0003
100 pM	10.8% (28/259)	0.32	1.50	<0.0001

Table S2-1. Fitted parameters of the distribution of beneficial point mutations. Only non-synonymous mutations are included in the analysis. n.s.: p>0.05.

Library (amino acid site)	**Forward primer**	**Reverse primer**
1 (18-34)	*5'-GTT TGC ACC ATC TTG ACA-3'*	*5'-TTG ACA AGA GAT GAA GGG-3'*
2 (35-51)	5'-CAA GCT GCC CGG CCT C-3'	5'-GCA GCG CGT GGT CAT GAT-3'

3 (52-68)	5’-TGG GCC GGC ACT GGC-3’	5’-GAT CCT CAT AGA GCC CAG-3’
4 (69-85)	5’-CAT CTC TGG CA A TGT CCG C-3’	5’-AGC AAT TGA TAG GAA AGG CCC-3’
5 (86-103)	5’-TGC ATG AAC ACC TGG CAG-3’	5’-ATG GCG GTC TTG TAG TTC GT-3’

Table S2-2. PCR primers used in preparation of sequencing samples.

Sample	**Barcode**
Plasmid	ATG
Transfection (round 0)	CCC
No drug (round 1)	CGG
[DCV]=10 pM (round 1)	CTT
[DCV]=40 pM (round 1)	ACT
[DCV]=100 pM (round 1)	AAC

Table S2-3. Barcodes used in multiplexing Illumina sequencing samples.

2.8 Bibliography

1. Domingo E, Sheldon J, Perales C. Viral quasispecies evolution. Microbiol Mol Biol Rev. 2012;76: 159–216. doi:10.1128/MMBR.05023-11

2. Goldberg DE, Siliciano RF, Jacobs WR. Outwitting evolution: fighting drug-resistant TB, malaria, and HIV. Cell. Elsevier; 2012;148: 1271–83. doi:10.1016/j.cell.2012.02.021

3. Metcalf CJE, Birger RB, Funk S, Kouyos RD, Lloyd-Smith JO, Jansen VAA. Five challenges in evolution and infectious diseases. Epidemics. 2015;10: 40–44. doi:10.1016/j.epidem.2014.12.003

4. Ke R, Loverdo C, Qi H, Sun R, Lloyd-Smith JO. Rational Design and Adaptive Management of Combination Therapies for Hepatitis C Virus Infection. PLoS Comput Biol. Public Library of Science; 2015;11: e1004040. doi:10.1371/journal.pcbi.1004040

5. Barton JP, Goonetilleke N, Butler TC, Walker BD, McMichael AJ, Chakraborty AK. Relative rate and location of intra-host HIV evolution to evade cellular immunity are predictable. Nat Commun. Nature Publishing Group; 2016;7: 11660. doi:10.1038/ncomms11660

6. Turner PE, Elena SF. Cost of Host Radiation in an RNA Virus. Genetics. 2000;156: 1465–1470.

7. Wright S. The Roles of Mutation, Inbreeding, Crossbreeding and Selection in Evolution. 1932. pp. 356–366.

8. Thyagarajan B, Bloom JD. The inherent mutational tolerance and antigenic evolvability of influenza hemagglutinin. Elife. 2014;3. doi:10.7554/eLife.03300

9. Fowler DM, Fields S. Deep mutational scanning: a new style of protein science. Nat Methods. Nature Publishing Group, a division of Macmillan Publishers Limited. All Rights Reserved.; 2014;11: 801–807. doi:10.1038/nmeth.3027

10. Wu NC, Young AP, Dandekar S, Wijersuriya H, Al-Mawsawi LQ, Wu T-T, et al. Systematic identification of H274Y compensatory mutations in influenza A virus neuraminidase by high-throughput screening. J Virol. 2013;87: 1193–9. doi:10.1128/JVI.01658-12

11. Qi H, Olson CA, Wu NC, Ke R, Loverdo C, Chu V, et al. A quantitative high-resolution genetic profile rapidly identifies sequence determinants of hepatitis C viral fitness and drug sensitivity. PLoS Pathog. 2014;10: e1004064. doi:10.1371/journal.ppat.1004064

12. Wu NC, Young AP, Al-Mawsawi LQ, Olson CA, Feng J, Qi H, et al. High-throughput profiling of influenza A virus hemagglutinin gene at single-nucleotide resolution. Sci Rep. Nature Publishing Group; 2014;4: 4942. doi:10.1038/srep04942

13. Li C, Qian W, Maclean CJ, Zhang J. The fitness landscape of a tRNA gene. Science. 2016;352: 837–840. doi:10.1126/science.aae0568

14. Puchta O, Cseke B, Czaja H, Tollervey D, Sanguinetti G, Kudla G. Network of epistatic interactions within a yeast snoRNA. Science. 2016;352: 840–844.

15. He X, Liu L. Toward a prospective molecular evolution. Science. American Association for the Advancement of Science; 2016;352: 769–70. doi:10.1126/science.aaf7543

16. Eyre-Walker A, Keightley PD. The distribution of fitness effects of new mutations. Nat Rev Genet. 2007;8: 610–8. doi:10.1038/nrg2146

17. Bataillon T, Bailey S. Effects of new mutations on fitness: insights from models and data. Ann New York Acad …. 2014;1320: 76–92. doi:10.1111/nyas.12460

18. Burch CL, Chao L. Evolvability of an RNA virus is determined by its mutational neighbourhood. Nature. 2000;406: 625–8. doi:10.1038/35020564

19. Jacquier H, Birgy A, Le Nagard H, Mechulam Y, Schmitt E, Glodt J, et al. Capturing the mutational landscape of the beta-lactamase TEM-1. Proc Natl Acad Sci U S A. 2013;110: 13067–72. doi:10.1073/pnas.1215206110

20. Chevereau G, Dravecká M, Batur T, Guvenek A, Ayhan DH, Toprak E, et al. Quantifying the Determinants of Evolutionary Dynamics Leading to Drug Resistance. PLoS Biol. Public Library of Science; 2015;13: e1002299. doi:10.1371/journal.pbio.1002299

21. Hietpas RT, Jensen JD, Bolon DNA. Experimental illumination of a fitness landscape. Proc Natl Acad Sci U S A. 2011;108: 7896–901. doi:10.1073/pnas.1016024108

22. Bank C, Hietpas RT, Jensen JD, Bolon DNA. A systematic survey of an intragenic epistatic landscape. Mol Biol Evol. Oxford University Press; 2015;32: 229–38. doi:10.1093/molbev/msu301

23. Desai MM. Statistical questions in experimental evolution. J Stat Mech Theory Exp. 2013;2013: P01003. doi:10.1088/1742-5468/2013/01/P01003

24. Gao M, Nettles RE, Belema M, Snyder LB, Nguyen VN, Fridell R a, et al. Chemical genetics strategy identifies an HCV NS5A inhibitor with a potent clinical effect. Nature. Nature Publishing Group; 2010;465: 96–100. doi:10.1038/nature08960

25. Lindenbach BD, Evans MJ, Syder AJ, Wölk B, Tellinghuisen TL, Liu CC, et al. Complete replication of hepatitis C virus in cell culture. Science. 2005;309: 623–6. doi:10.1126/science.1114016

26. Sanjuan R, Moya A, Elena SF. The distribution of fitness effects caused by single-nucleotide substitutions in an RNA virus. Proc Natl Acad Sci. 2004;101: 8396–8401. doi:10.1073/pnas.0400146101

27. Elena SF, Carrasco P, Daròs J-A, Sanjuán R. Mechanisms of genetic robustness in RNA viruses. EMBO Rep. 2006;7: 168–73. doi:10.1038/sj.embor.7400636

28. Rihn SJ, Wilson SJ, Loman NJ, Alim M, Bakker SE, Bhella D, et al. Extreme genetic fragility of the HIV-1 capsid. PLoS Pathog. 2013;9: e1003461. doi:10.1371/journal.ppat.1003461

29. Burch C, Guyader S, Samarov D, Shen H. Experimental estimate of the abundance and effects of nearly neutral mutations in the RNA virus ϕ6. Genetics. 2007;476: 467–476. doi:10.1534/genetics.106.067199

30. Silander O, Tenaillon O, Chao L. Understanding the evolutionary fate of finite populations: the dynamics of mutational effects. PLoS Biol. 2007;5. doi:10.1371/journal.pbio.0050094

31. Stiffler MA, Hekstra DR, Ranganathan R. Evolvability as a Function of Purifying Selection in TEM-1 β-Lactamase. Cell. 2015;160: 882–892. doi:10.1016/j.cell.2015.01.035

32. MacLean RC, Buckling A. The distribution of fitness effects of beneficial mutations in Pseudomonas aeruginosa. PLoS Genet. 2009;5: e1000406. doi:10.1371/journal.pgen.1000406

33. Bataillon T, Zhang T, Kassen R. Cost of adaptation and fitness effects of beneficial mutations in Pseudomonas fluorescens. Genetics. 2011;189: 939–949. doi:10.1534/genetics.111.130468

34. Carrasco P, Iglesia F de la, Elena S. Distribution of fitness and virulence effects caused by single-nucleotide substitutions in Tobacco etch virus. J Virol. 2007;81: 12979–12984. doi:10.1128/JVI.00524-07

35. Cowperthwaite MC, Bull JJ, Meyers LA. Distributions of beneficial fitness effects in RNA. Genetics. 2005;170: 1449–57. doi:10.1534/genetics.104.039248

36. Peris JB, Davis P, Cuevas JM, Nebot MR, Sanjuán R. Distribution of fitness effects caused by single-nucleotide substitutions in bacteriophage f1. Genetics. 2010;185: 603–9. doi:10.1534/genetics.110.115162

37. Imhof M, Schlötterer C. Fitness effects of advantageous mutations in evolving Escherichia coli populations. Proc Natl Acad Sci U S A. 2001;98: 1113–1117.

38. Kassen R, Bataillon T. Distribution of fitness effects among beneficial mutations before selection in experimental populations of bacteria. Nat Genet. 2006;38: 484–8. doi:10.1038/ng1751

39. Rokyta D, Joyce P, Caudle S, Wichman H. An empirical test of the mutational landscape model of adaptation using a single-stranded DNA virus. Nat Genet. 2005;37: 441–444. doi:10.1038/ng1535

40. Beisel CJ, Rokyta DR, Wichman HA, Joyce P. Testing the extreme value domain of attraction for distributions of beneficial fitness effects. Genetics. 2007;176: 2441–9. doi:10.1534/genetics.106.068585

41. Rokyta D, Beisel C, Joyce P. Beneficial fitness effects are not exponential for two viruses. J Mol Evol. 2008;67: 368–376. doi:10.1007/s00239-008-9153-x.Beneficial

42. Schenk MF, Szendro IG, Krug J, de Visser JAGM. Quantifying the adaptive potential of an antibiotic resistance enzyme. PLoS Genet. Public Library of Science; 2012;8: e1002783. doi:10.1371/journal.pgen.1002783

43. MacLean RC, Hall AR, Perron GG, Buckling A. The population genetics of antibiotic resistance: integrating molecular mechanisms and treatment contexts. Nat Rev Genet. Nature Publishing Group; 2010;11: 405–414. doi:10.1038/nrg2778

44. Liberles DA, Teichmann SA, Bahar I, Bastolla U, Bloom J, Bornberg-Bauer E, et al. The interface of protein structure, protein biophysics, and molecular evolution. Protein Sci. Wiley Subscription Services, Inc., A Wiley Company; 2012;21: 769–785. doi:10.1002/pro.2071

45. Tellinghuisen TL, Marcotrigiano J, Rice CM. Structure of the zinc-binding domain of an essential component of the hepatitis C virus replicase. Nature. Macmillian Magazines Ltd.; 2005;435: 374–9. doi:10.1038/nature03580

46. Love RA, Brodsky O, Hickey MJ, Wells PA, Cronin CN. Crystal structure of a novel dimeric form of NS5A domain I protein from hepatitis C virus. J Virol. 2009;83: 4395–403. doi:10.1128/JVI.02352-08

47. Ramsey DC, Scherrer MP, Zhou T, Wilke CO. The Relationship Between Relative Solvent Accessibility and Evolutionary Rate in Protein Evolution. Genetics. 2011;188: 479–488. doi:10.1534/genetics.111.128025

48. Das R, Baker D. Macromolecular Modeling with Rosetta. Annu Rev Biochem. 2008;77: 363–382. doi:10.1146/annurev.biochem.77.062906.171838

49. Chaudhury S, Lyskov S, Gray JJ. PyRosetta: a script-based interface for implementing molecular modeling algorithms using Rosetta. Bioinformatics. 2010;26: 689–91. doi:10.1093/bioinformatics/btq007

50. Olson CA, Wu NC, Sun R. A comprehensive biophysical description of pairwise epistasis throughout an entire protein domain. Curr Biol. 2014;24: 2643–51. doi:10.1016/j.cub.2014.09.072

51. Wylie CS, Shakhnovich EI. A biophysical protein folding model accounts for most mutational fitness effects in viruses. Proc Natl Acad Sci. National Academy of Sciences; 2011;108: 9916–9921. doi:10.1073/pnas.1017572108

52. Bloom JD, Silberg JJ, Wilke CO, Drummond DA, Adami C, Arnold FH. Thermodynamic prediction of protein neutrality. Proc Natl Acad Sci U S A. 2005;102: 606–11. doi:10.1073/pnas.0406744102

53. Echave J, Spielman SJ, Wilke CO. Causes of evolutionary rate variation among protein sites. Nat Rev Genet. 2016;17: 109–121. doi:10.1038/nrg.2015.18

54. Jack BR, Meyer AG, Echave J, Wilke CO. Functional Sites Induce Long-Range Evolutionary Constraints in Enzymes. PLoS Biol. 2016;14: e1002452. doi:10.1371/journal.pbio.1002452

55. Wu NC, Olson CA, Du Y, Le S, Tran K, Remenyi R, et al. Functional Constraint Profiling of a Viral Protein Reveals Discordance of Evolutionary Conservation and Functionality. PLOS Genet. Public Library of Science; 2015;11: e1005310. doi:10.1371/journal.pgen.1005310

56. Hwang J, Huang L, Cordek DG, Vaughan R, Reynolds SL, Kihara G, et al. Hepatitis C virus nonstructural protein 5A: biochemical characterization of a novel structural class of RNA-binding proteins. J Virol. 2010;84: 12480–91. doi:10.1128/JVI.01319-10

57. Foster TL, Belyaeva T, Stonehouse NJ, Pearson AR, Harris M. All three domains of the hepatitis C virus nonstructural NS5A protein contribute to RNA binding. J Virol. 2010;84: 9267–77. doi:10.1128/JVI.00616-10

58. Ferguson AL, Mann JK, Omarjee S, Ndung'u T, Walker BD, Chakraborty AK. Translating HIV sequences into quantitative fitness landscapes predicts viral vulnerabilities for rational immunogen design. Immunity. 2013;38: 606–17. doi:10.1016/j.immuni.2012.11.022

59. Hart GR, Ferguson AL. Error catastrophe and phase transition in the empirical fitness landscape of HIV. Phys Rev E. 2015;91: 32705. doi:10.1103/PhysRevE.91.032705

60. Gong LI, Suchard MA, Bloom JD. Stability-mediated epistasis constrains the evolution of an influenza protein. Elife. eLife Sciences Publications Limited; 2013;2: e00631. doi:10.7554/eLife.00631

61. Wu NC, Dai L, Olson CA, Lloyd-Smith JO, Sun R. Adaptation in protein fitness landscapes is facilitated by indirect paths. Elife. 2016;5: e16965. doi:10.7554/eLife.16965

62. Firnberg E, Ostermeier M. The genetic code constrains yet facilitates Darwinian evolution. Nucleic Acids Res. 2013;41: 7420–8. doi:10.1093/nar/gkt536

63. Yampolsky LY, Stoltzfus A. The exchangeability of amino acids in proteins. Genetics. Genetics Society of America; 2005;170: 1459–72. doi:10.1534/genetics.104.039107

64. Tanaka T, Kato N, Hijikata M, Shimotohno K. Base transitions and base transversions seen in mutations among various types of the hepatitis C viral genome. FEBS Lett. 1993;315: 201–203. doi:10.1016/0014-5793(93)81163-T

65. Duchêne S, Ho SY, Holmes EC. Declining transition/transversion ratios through time reveal limitations to the accuracy of nucleotide substitution models. BMC Evol Biol. 2015;15: 312. doi:10.1186/s12862-015-0312-6

66. Acevedo A, Brodsky L, Andino R. Mutational and fitness landscapes of an RNA virus revealed through population sequencing. Nature. Nature Research; 2014;505: 686–690. doi:10.1038/nature12861

67. Stoltzfus A, Norris RW. On the Causes of Evolutionary Transition:Transversion Bias. Mol Biol Evol. 2016;33: 595–602. doi:10.1093/molbev/msv274

68. Levy SF, Blundell JR, Venkataram S, Petrov DA, Fisher DS, Sherlock G. Quantitative evolutionary dynamics using high-resolution lineage tracking. Nature. Nature Publishing Group, a division of Macmillan Publishers Limited. All Rights Reserved.; 2015;advance on. doi:10.1038/nature14279

69. Visher E, Whitefield SE, McCrone JT, Fitzsimmons W, Lauring AS. The Mutational Robustness of Influenza A Virus. PLOS Pathog. Public Library of Science; 2016;12: e1005856. doi:10.1371/journal.ppat.1005856

70. Draghi JA, Parsons TL, Wagner GP, Plotkin JB. Mutational robustness can facilitate adaptation. Nature. Nature Publishing Group; 2010;463: 353–355. doi:10.1038/nature08694

71. Visser JAGM, Hermisson J, Wagner GP, Meyers LA, Bagheri-Chaichian H, Blanchard JL, et al. Perspective: Evolution and detection of genetic robustness. Evolution. Blackwell Publishing Ltd; 2003;57: 1959–1972. doi:10.1111/j.0014-3820.2003.tb00377.x

72. Agashe D, Sane M, Phalnikar K, Diwan GD, Habibullah A, Martinez-Gomez NC, et al. Large-Effect Beneficial Synonymous Mutations Mediate Rapid and Parallel Adaptation in a Bacterium. Mol Biol Evol. 2016;33: 1542–53. doi:10.1093/molbev/msw035

73. Yang J-R, Chen X, Zhang J. Codon-by-Codon Modulation of Translational Speed and Accuracy Via mRNA Folding. PLoS Biol. Public Library of Science; 2014;12: e1001910. doi:10.1371/journal.pbio.1001910

74. Lalić J, Cuevas JM, Elena SF. Effect of host species on the distribution of mutational fitness effects for an RNA virus. PLoS Genet. 2011;7: e1002378. doi:10.1371/journal.pgen.1002378

75. Martin G, Lenormand T. The fitness effect of mutations across environments: a survey in light of fitness landscape models. Evolution. 2006;60: 2413–2427.

76. Robinson M, Tian Y, Delaney WE, Greenstein AE. Preexisting drug-resistance mutations reveal unique barriers to resistance for distinct antivirals. Proc Natl Acad Sci U S A. 2011;108: 10290–5. doi:10.1073/pnas.1101515108

77. Soskine M, Tawfik DS. Mutational effects and the evolution of new protein functions. Nat Rev Genet. Nature Publishing Group, a division of Macmillan Publishers Limited. All Rights Reserved.; 2010;11: 572–82. doi:10.1038/nrg2808

78. Rosenbloom DIS, Hill AL, Rabi SA, Siliciano RF, Nowak MA. Antiretroviral dynamics determines HIV evolution and predicts therapy outcome. Nat Med. Nature Publishing Group; 2012;18: 1378–85. doi:10.1038/nm.2892

79. de Visser JAGM, Krug J. Empirical fitness landscapes and the predictability of evolution. Nat Rev Genet. Nature Publishing Group, a division of Macmillan Publishers Limited. All Rights Reserved.; 2014;15: 480–90. doi:10.1038/nrg3744

80. Ogbunugafor CB, Wylie CS, Diakite I, Weinreich DM, Hartl DL. Adaptive Landscape by Environment Interactions Dictate Evolutionary Dynamics in Models of Drug Resistance. PLoS Comput Biol. Public Library of Science; 2016;12: e1004710. doi:10.1371/journal.pcbi.1004710

81. Qi H, Olson CA, Wu NC, Du Y, Sun R. Determining the Relative Fitness Score of Mutant Viruses in a Population Using Illumina Paired-end Sequencing and Regression Analysis. Bio-protocol. 2015;5: e1475.

82. Kabsch W, Sander C. Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features. Biopolymers. Wiley Subscription Services, Inc., A Wiley Company; 1983;22: 2577–2637. doi:10.1002/bip.360221211

83. Tien MZ, Meyer AG, Sydykova DK, Spielman SJ, Wilke CO. Maximum allowed solvent accessibilites of residues in proteins. PLoS One. 2013;8: e80635. doi:10.1371/journal.pone.0080635

84. Shapovalov M V, Dunbrack RL. A smoothed backbone-dependent rotamer library for proteins derived from adaptive kernel density estimates and regressions. Structure. 2011;19: 844–58. doi:10.1016/j.str.2011.03.019

85. Kuiken C, Yusim K, Boykin L, Richardson R. The Los Alamos hepatitis C sequence database. Bioinformatics. 2005;21: 379–384. doi:10.1093/bioinformatics/bth485

CHAPTER 3

ANNOTATING PROTEIN FUNCTIONAL RESIDUES BY COUPLING HIGH-THROUGHPUT FITNESS PROFILE AND HOMOLOGOUS STRUCTURE ANALYSIS

3.1 Abstract

Identification and annotation of functional residues are fundamental questions in protein sequence analysis. Sequence and structure conservation provide valuable information to tackle these questions. They are, however, limited by the incomplete sampling of sequence space in natural evolution. Moreover, proteins often encode multiple functions with overlapping sequences, which present challenges to accurately annotate the exact functions of individual residues with conservation-based methods. Using influenza A virus PB1 protein as an example, we presented 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 the 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 non-canonical functions of viral RNA dependent RNA polymerase (vRdRp), we performed 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 non-canonical functional residues located in the loop region of 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 with available homologous structure information.

3.2 Introduction

Amino acid residues in a protein have two roles: providing structural framework (structural residues) and mediating interactions with other biomolecules (functional residues). Identification and annotation of functional residues are fundamental questions in protein characterization (1–

5). A number of methods have been developed to tackle these questions. Most methods utilize sequence conservation information, with the assumption that functional residues are often conserved in across homolog proteins (6–8). The identified residues are then expected to perform functions similar to other homologs. Other methods predict functional residues based on the shapes and properties of protein 3D structures (9–15). Starting from well-known functional domains (ligand binding, catalytic, et al.), these analyses determine the 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 single-residue level (**Figure 3-1**)(16). Therefore, a more direct and systematic method needs to be applied for the accurate identification and annotation of functional residues.

Due to their compact genome, viruses usually encode multifunctional proteins, including viral polymerase proteins. Viral RNA-dependent RNA polymerase (vRdRp) is utilized by many RNA viruses for transcription and replication. Functions of vRdRp can be grouped into two classes: canonical vRdRp functions and non-canonical functions. 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 non-canonical functions of vRdRp, however, are specific to each virus. For example, multimerisation of HCV vRdRp is essential for viral replication. Thus, the interacting residues among HCV vRdRp are non-canonical functional residues specific to HCV virus (23, 24). Moreover, vRdRp often recruits cellular machinery for replication and plays a role in inhibiting cellular immune response (25–31). Non-canonical functional residues are usually involved in performing those functions and thus, are essential for viral replication. Non-canonical functional residues in vRdRp are difficult to be determined by commonly used methods and are not as well studied as the key residues for polymerase catalytic functions. However, the non-canonical functional residues are indispensable for a thorough protein characterization, and may act as a drug-target. As a result, it is essential to develop methods that enable identification of non-canonical functional residues.

Previously we have developed a method to systematically identify functional residues by coupling experimental fitness measurement with protein stability prediction (16). Here, we are extending this method to annotate functional residues in combination with structural comparison of homologous proteins. The method consists of three steps. Firstly, the effect of PB1 mutations on viral replication at single-nucleotide resolution is examined by saturation mutagenesis and high-throughput sequencing. Secondly, functional PB1 residues that are essential for viral growth but do not affect protein stability were identified by protein stability prediction. Lastly, homologous structural alignment was utilized to further annotate specific biological functions (canonical versus non-canonical functions) for each functional residue (**Figure 3-1**). We achieved high sensitivity in identifying and annotating the canonical polymerase functional residues. Moreover, we also identified non-canonical functional residues, which are exemplified by a cluster of residues located in the loop region of 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).

3.3 Results

3.3.1 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). Here, point mutations were randomly introduced into the 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 the entire segment 2 (**Figure S3-1**). 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 an MOI 0.05, and supernatants were collected 24 hours post-infection. The input DNA libraries, post-transfection libraries, and post-infection libraries were subjected to Illumina sequencing. To control for technical error and to assess library quality, biological duplicates were included both in transfection and subsequent infection steps (**Figure 3-2**).

Distribution of the number of mutations in the input DNA library was examined. 30-35% of input DNA library plasmids contained the desired single nucleotide mutations (**Figure S3-1**). We achieved at least 20,000x sequencing coverage for each nucleotide position (**Figure S3-1**). The library covered 94.9% of all nucleotides in segment 2, and included 98.2% of all single nucleotide mutations of observed positions (**Figure S3-2**). To further improve the accuracy of fitness quantification, we focused on the mutations that occur >0.1% in the plasmid mutant library. After this quality control, we were still able to observe 94.2% of all nucleotide positions with 63.9% of all single nucleotide mutations. More than 82% nucleotide positions were covered with two or three nucleotide mutations (**Figure S3-2**). To assess the quality and reproducibility of our mutant library, we compared the relative frequency of single mutations between biological replicates. We obtained a strong spearman correlation of 0.93 for two independent transfections, and 0.75 for infections (**Figure 3-2**). Relative fitness index (RF index) was calculated for individual mutations as the ratio of relative frequency in infection library to that in the input DNA library. The profiling data of the entire segment 2 is shown in **Figure 3-2**, where majority of mutations had a fitness cost (log10 RF index < 0).

3.3.2 Systematic identification of deleterious mutations of PB1 protein

Segment 2 of influenza A virus encoded three proteins: PB1, PB1-F2 and N40. N40 was a truncated form of PB1 protein, which 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) (**Figure S3-3**). So we focused on PB1 protein for downstream analysis. The RF indexes of silent mutations were considered as internal quality control since most, if not all, of them were expected to have comparable growth capacity with the wild type. In the fitness profile of PB1 protein, RF index of silent mutations followed a normal distribution with a mean equal to 0.9 and were significantly higher than that of nonsense mutations (two-tailed t-test, p = 4.6E-21) (**Figure S3-4**). This result confirms the presence of fitness selection and validates the data quality.

To systematically identify deleterious mutations, we chose a stringent cutoff of RF index ≤ 0.1. 2.4% silent mutations fell below the cutoff, which represented type I error. 43.1% of missense mutations that satisfied this cutoff were identified as deleterious mutations (**Figure 3-3**). We randomly selected 14 deleterious mutations and reconstructed them individually. Rescue experiments were performed and the resultant viral titers were quantified by TCID50 assay. 13 out of 14 mutant viruses had at least a 10-fold drop in viral titer as compared to WT. The other mutant also showed more than 6-fold decrease in titer (**Figure 3-3**). These results validated our approach to systematically quantify the relative fitness and identify deleterious mutations of PB1 protein.

3.3.3 Identifying functional residues by dissecting structural constraint and functional constraint

A mutation might be deleterious due to structural constraints 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 using two computational tools: I-Mutant and Rosetta ddg monomer.

I-Mutant was a supporter vector machine (SVM) based software to predict the effect of single-site mutations on protein stability (ΔΔG) (43–45). Based on 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 bat influenza A polymerase complex (PDB: 4WSB) (46, 47). Among mutations with structure information available, 64.5% were shown to be destabilizing, 33.5% were neutral and 2% were stabilizing (**Figure S3-5**). As expected, destabilizing mutations had a significantly lower value of solvent accessible surface area (SASA) (48–50) (**Figure S3-5**). To further reduce the false negative rate of identifying functional residues, we performed protein stability prediction with Rosetta for all deleterious mutations (16, 42, 44). Different from machine learning algorithm used by I-mutant, Rosetta generated structural models for single amino acid mutations based on pre-optimized wild-type structure. Using high-resolution protocol, 50 models of wild type and mutant protein structures were generated and three lowest ΔΔG were averaged based on optimized rotamers. The absolute correlation coefficient of prediction that resulted from these two methods was 0.3 (**Figure S3-5**). 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 cutoff of RF index, and non-destabilizing criteria of ΔΔG predictions from either software. A total of 297 residues were identified as functional residues.

To examine the sensitivity of our method in 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 four other methods: FireStar, Frpreq, Consurf and Concavity (6, 10, 55–58) (**Table 3-1**). 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 non-canonical polymerase functions (4 of the 8 residues), which 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).

3.3.4 Annotating functional residues with homologous structural alignment

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, A-E) (20). Since canonical vRdRp functional residues of PB1 protein are expected to be structurally conserved, they aligned well with other protein structures from vRdRp family. Therefore, homologous structural alignment might enable us to further annotate PB1 residues by distinguishing canonical and non-canonical 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 alignment (59–61). Both softwares utilize 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 features, which is suitable for alignment of distantly related protein structures (61).

Twenty representative vRdRp structures were selected from +ssRNA viruses, -ssRNA viruses, and dsRNA viruses families with criteria as previously stated (20). Briefly, representative structures were selected from each of the Baltimore classes that encoded vRdRp, including +ssRNA viruses (Caliciviriade, Flaviviridae, Picornaviridae, Cystoviridae), dsRNA viruses (Birnaviridae, Cystoviridae and Reoviridae) and –ss RNA virus (Bunyaviridae) (62–81). Structures with no mutations and with bound substrate were preferred. PDB files with the highest resolution were picked for each protein.

To ensure sufficient structural similarity, pairwise structural comparison was performed between selected protein with PB1 using TM-align. The structures with TM-score > 0.5 were kept for multiple-structural alignment, which generally indicated similar protein folding (43). **Figure S3-6** provides an example superimposition of PB1 protein with the HCV NS5B (PDB: 2XI3) with decent alignment in major protein domains (67). A total of 16 proteins were included for multiple structure alignment (MSA) with PB1 using 3DCOMB.

Root-mean-square deviation (RMSD), the measurement of 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 a RMSD value of 10 (**Figure 3-4**). Low RMSD scores represented that the residues were conserved in vRdRp family, thus more likely to carry 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 non-conserved residues (two-tailed t-test, p=0.0006, **Figure 3-4**). The RMSD of all identified functional residues of PB1 protein were plotted. The smooth curve of RMSD was fitted with loess regression. We could clearly identify the 6 conserved domains (pre-A/F, A-E) of vRdRp as valleys on the smooth curve (**Figure 3-4**). These results demonstrated the feasibility of using homologous structural alignment to identify canonical vRdRp residues.

3.3.5 Identification of non-canonical functional residues, ones involved in nuclear import of PB1 protein

43% of identified functional residues could not be aligned to other protein structures from vRdRp family. Although it could be due to poor alignment quality, it is also possible that these residues carry non-canonical functions that are essential for viral growth. Interestingly, 62% of these residues belong to the protein interface between PB1 and PB2, PA as identified by change of solvent accessible surface area (SASA) upon complex formation using Sppider (residues with at least 4% decrease in SASA and more than 5Å2 surface expose area upon complex formation) (82) (**Figure S3-6**). These interface residues also accounted for some of the peaks (residue 50-80, residues 350-400 and residues at C terminus of PB1) in the RMSD smooth curve of functional residues in **Figure 3-3**.

We then performed a detailed analysis on the non-canonical functional residues that did not locate in the heterotrimer-forming interface. When mapped onto the protein structure, some of these residues (residues 180-220) formed a noticeable cluster (**Figure 3-5**). This clustered region is unique to 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 NLS signals were reported in the β-ribbon region (Amino Acid 187-190, 207-210) to mediate PB1 nuclear import through interaction with Ran-binding protein 5 (RanBP5) (32, 83). Nonetheless, the function of this loop region is not completely clear. It is suspected to interact with viral genome in the resolved structures of influenza B and influenza C (46, 47, 84), and K198 of influenza A was suggested to be related to host adaptation (85). As the density of the loop region (residues 195-198) is missing in the influenza A polymerase crystal structure, we used kinematic loop modeling in Rosetta software to computationally re-construct the loop region (86). From the above analysis, D193 in the loop region was identified as a non-canonical functional residue. Interestingly, it was the only negatively charged residue located within a highly positive charged environment. It was 100% conserved among all the human influenza A virus PB1 sequences from the influenza research database (IRD) under purifying selection (dN/dS = 0.015) (87–89). Two positively charged residues (K197, K198) located on the opposite side of 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 more than 6 fold drop of RF index. To examine if the loop region carried possible non-canonical functions, we introduced single substitutions (D193G, K197E, K198E) and double substitutions (D193G-K197E, G193G-K198E, K197E-K198E) into PB1 protein. We also constructed mutants that carried substitutions in the NLS region (K188A-R189A, R208A-K209A) and mutants that decreased the polymerase activity (W55R, H184R, H47L and Q268L) as controls. Of note, all controls were identified as deleterious in our high-throughput fitness profile. Viral production of all mutants was measured by TCID50 assay with viral rescue experiments. D193G, D193G-K197E, G193G-K198E, K197E-K198E and the reported substitutions on NLS region (K188A-R189A) showed severe impact on viral production with no detectable viral titer post transfection (**Figure 3-5**) (32). Consistently, these mutations also showed significantly slower viral growth rate in A549 cells (**Figure 3-5**). To examine the vRdRp function of these mutations, we utilized a mini-genome replicon assay by co-transfecting a virus-inducible luciferase reporter and polymerase segments (PB2, PB1, PA, NP) in 293T cells. The reported NLS mutant (K188A-R189A), which were highly deleterious for viral replication, still had ~50% polymerase activity in the mini-genome replicon assay. Similarly, D193G and all the double substitutions (D193G-K197E, D193G-K198E, K197E-K198E) showed discordance between vRdRp function and viral growth capacity. Compared with W55R, H184R, H47L and Q268L, which remained ~0.1%- 65% polymerase activities, the fitness drop of these newly identified loop mutations was much more severe, indicating that they might carry non-canonical polymerase function of PB1 (**Figure 3-5**).

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 non-canonical functions of PB1 protein. We tested if the identified mutations in the loop region (D193G, D193G-K197E, K197E-K198E) had effects on protein nuclear import. A549 cells were infected with wild type and mutant virus at an MOI 0.1. Cells were fixed and subjected to immunofluorescence analysis (IFA) at 18h post-infection. As expected, PB1 proteins of wild type virus were mostly localized in the nucleus. However, PB1 proteins from mutant viruses were found significantly enriched in the cytoplasm, suggesting that these mutations were defective in PB1 protein nuclear import (**Figure 3-6**). More severe defects were observed for double mutations (D193G-K197E, K197E-K198E). Similar results were observed at earlier time points (8h post-infection) with an MOI 0.5 (**Figure S3-7**). Interestingly, for those PB1 mutants, the nuclear import for PA protein was also delayed, which is consistent with the notion that PA and PB1 are imported to the nucleus as a complex (32, 83, 90, 91) (**Figure S3-7**).

RanBP5 belongs to the importin-β family, which has non-classical nuclear import function (92, 93). RanBP5 has been shown to be important for influenza A PB1 nuclear import. The NLS mutations affected protein nuclear import by decreasing the 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 co-transfecting FLAG-tagged PB1 protein and HA-tagged RanBP5 protein in 293T cells. Two days later, total cell lysate was collected and subjected to IP with anti-HA antibody conjugated beads or IgG conjugated beads. As shown in **Figure 3-6**, all three mutant proteins showed decreased binding with RanBP5. Consistent with our IFA results, double mutations (D193G-K197E, K197E-K198E) showed higher reduction in protein binding. The above results indicate that the residues in the loop region are important for the nuclear import of influenza A PB1 protein through the interaction with RanBP5, which is a non-canonical function for the vRdRp family.

3.4 Discussion

For a comprehensive characterization of protein function, identification and annotation of functional residues are the fundamental tasks. Here we presented a systematic approach to achieve these tasks using influenza A PB1 as the target protein. Our approach combines high-throughput fitness profiling with mutant stability prediction and homologous structural alignment to identify and annotate canonical and non-canonical vRdRp functional residues (**Figure 3-1**). Interestingly, we identified a cluster of mutations that were highly deleterious for viral replication but with relatively intact vRdRp function. These mutations were located in the loop region of 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 for identifying and annotating functional residues, which can be applied to a wide range of proteins with homologous structural information available.

In the context of evolutionary biology, proteins from the same homolog family share a common ancestor and possess significant sequence and structural similarities (94–97). Structural similarities are postulated to be maintained by functional constraints (98, 99). Viral RNA dependent RNA polymerase (vRdRp) were likely evolved from a common ancestor (100). Although the sequence identity is ~20%, they have adopted similar structural domains and utilize similar catalytic mechanisms (20). Throughout the period of evolution, different proteins also evolved diverse functions to satisfy the need of specific organisms. Thus, the specific structural motifs that differentiate one protein from their homologous proteins may carry organism specific functions. Here we used homologous protein structure information to further annotate the diverse protein functions. Therefore, a multi-functional protein might harbor both canonical (evolutionary conserved) functions and non-canonical (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 showed that the residues in the loop region of PB1 β ribbon were 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 translocate together as a complex, while PB2 can be translocated by itself (101). RanBP5 is important for the nuclear import for PB1 and PA through the direct interaction with PB1. Besides the two reported NLSs, we showed that the mutations in the loop region also impact the interaction between PB1 and RanBP5, thus causing the defect of PB1 nuclear import. We do not have direct evidence to show whether the loop region works as a direct NLS or by affecting the nearby NLS regions. But based on the sequence of loop region, it did not fall into any of the six classes of nuclear localization signals (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 nor important for RanBP5 binding (32, 103). The detailed mechanism is 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 using cell growth rate as a read out. Applying saturated mutagenesis, we can learn which mutation is related to abnormal cell growth rate and can further use flow cytometry to differentiate cells in different phases. We can also investigate the role of mutant proteins on cancer metastasis through transwell migration assay in vitro or using mouse xenograft models in vivo. The structures of target protein or homologous protein structures 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 cryo electron 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.

3.5 Materials and Methods

Construction of influenza A segment 2 mutant library

Influenza A/WSN/33 segment 2 mutant libraries were generated using the eight-plasmid transfection system(39). In brief, the whole length influenza gene was separated into 9 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 using MegaX DH10B T1R cells (Life Technologies). As each small library was expected to have ~1000 single mutations, ~50,000 bacterial colonies were collected to cover the entirety. Plasmids from collected bacteria were midi-prepped as the input DNA library.

Transfection, infection and viral titer

To generate the mutant viral library, ~30 million 293T cells were transfected with 32ug DNA. Transfections were performed using Lipofectamine 2000 (Life Technologies). Virus was collected 72h post transfection. TCID50 were measured with A549 cells. To passage viral libraries, ~10 million A549 cells were infected with an MOI 0.05. Cells were washed with PBS three times at 2 h post-infection. Virus was collected 24 h post-infection from supernatant.

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

Sequencing library construction and data analysis

Viral RNA was extracted using QIAamp Viral RNA Mini Kit (Qiagen Sciences). DNaseI (Life Technologies) treatment was performed, followed by reverse transcription using superscript III system (Life Technologies). At least 106 viral copy numbers 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 de-multiplexed using the three-nucleotide ID. Sequencing error was corrected by filtering un-matched forward and reverse reads. Mutations were called by comparing sequencing reads with the wild-type sequence. Clones containing two or more mutations were discarded. Relative fitness index (RF index) was calculated for individual point mutations and only mutations that have frequency more than 0.1% in the DNA library were reported.

$$RF\ index_{mutant\ i} = Relative\ Frequency\ of\ Mutant\ i_{\ infection} / Relative\ Frequency\ of\ Mutant\ i_{\ plasmid}$$

Where $Relative\ Frequency\ of\ Mutant\ i = Reads\ of\ Mutant\ i\ /\ Reads\ of\ wild\ type$

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

Protein structural analysis

Chain B (PB1 protein) of PDB: 4WSB was used for protein ΔΔG prediction with single amino acid mutations(46, 47). ΔΔG prediction were performed with both I-Mutant 2.0 package and ddg_monomer in Rosetta software (43, 108). Default parameters (temperature= 25 ℃, pH =

7.0) were used in I-Mutant package. Parameters used for Rosetta were same as previous described (16, 109). $\Delta\Delta G < 0$ in I-Mutant and $\Delta\Delta G > 0$ in Rosetta mean destabilization.

DSSP was used to calculated SASA, which was then normalized to the empirical scale as described (48–50). Sppider was used to identify protein-protein interface. Residues with at least 4% reduction and more than 5 Å2 reduction in SASA upon complex formation are identified as protein-protein interface (82) .

TM-align and 3DCOMB were used for pair-wise structural alignment and multiple structural alignment (59, 61). TM-score normalized to PB1 protein were utilized.

Protein loop modeling

In the loop region of PB1 β-ribbon, electron density for residues 195-198 is missing from the x-ray crystal structure (PDB: 4WSB). Rosetta software was used to computationally re-construct the loop region, which was based on Monte Carlo sampling with exact kinematic loop closure (KIC) (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

100 ng of each of PB2, PB1 (wild-type and indicated mutations), PA, NP, 50 ng of virus-inducible luciferase reporter and 5ng PGK-renilla luciferase were transfected in 293T cells in 24 well plates (110). Cells were lysed 24 h post-transfection and luciferase assay was performed with Dual-Luciferase Assay Kit (Promega).

Immunofluorescence

Localization of wild type PB1 and PB1 mutations were determined by immunofluorescence. Infected A549 cells were fixed in 2% paraformaldehyde, permeabilized with 0.1% Triton-X100, and then blocked with 3% BSA and 10% FBS. Viral PB1 protein was detected with anti-PB1 antibody (GeneTex GTX125923). Hoechst 33342 dye was used for nucleic acid stain.

Immunoprecipitation

Immunoprecipitation 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 two days post-transfection with RIPA buffer (50 mM Tris-HCl pH 7.4, 0.5% NP-40, 150 mM KCl, 1 mM EDTA and protease inhibitor). Cell lysates were incubated with 1 μg anti-HA for 4 hours at 4℃ with constant agitation, washed with RIPA buffer for 5 times and eluted with 60 μl of SDS-PAGE sample buffer. All samples were subjected to SDS-PAGE and western blotting analysis.

Western blotting

Proteins in SDS-PAGE sample buffer were heated at 95℃, resolved by SDS-PAGE gel electrophoresis, and then transferred onto PVDF 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 using MUSCLE (88). 3000 sequences were randomly sampled for dN/dS calculation by Fubar using HyPhy (89).

Nucleotide sequence accession numbers

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

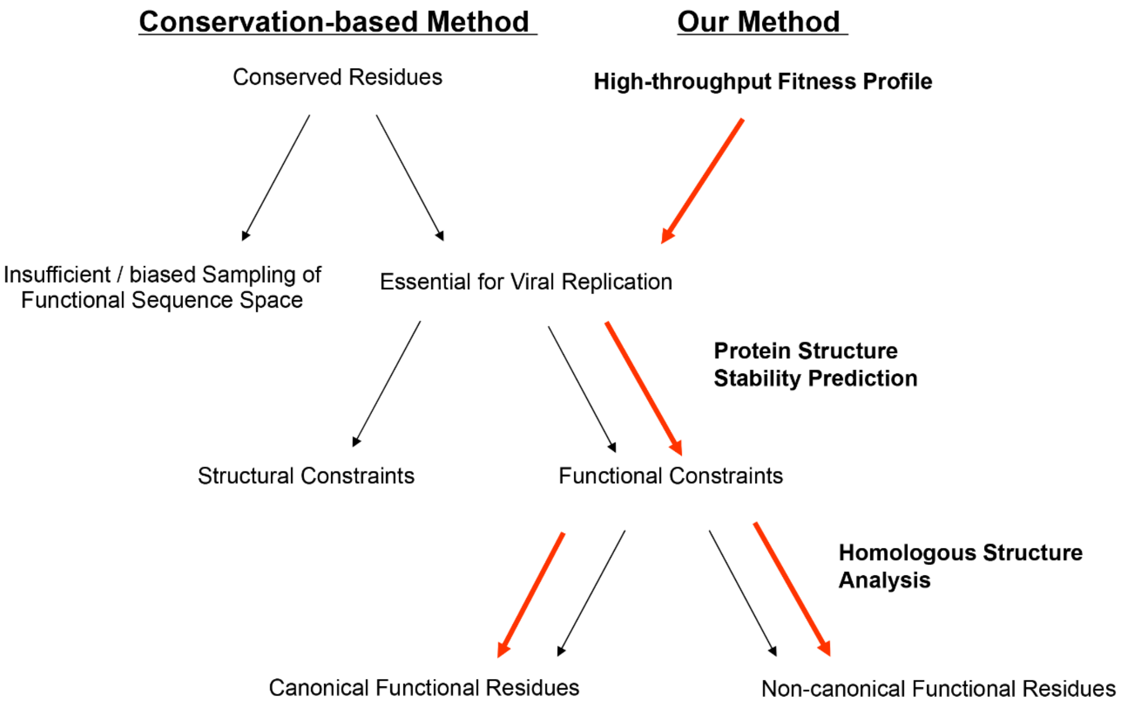

Conservation-based Method
Our Method
Conserved Residues
High-throughput Fitness Profile
Insufficient / biased Sampling of Functional Sequence Space
Essential for Viral Replication
Protein Structure Stability Prediction
Structural Constraints
Functional Constraints
Homologous Structure Analysis
Canonical Functional Residues
Non-canonical Functional Residues

Figure 3-1. Comparison of 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. Firstly, the method is limited by the insufficient sampling of protein functional space in natural evolution. Secondly, 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. Our method presented here may overcome these limitations and provide a systematic way for annotating functional residues. Using high-throughput fitness profiling, we can identify essential residues for viral replication. Through mutant stability prediction, we are able to dissect the structural and functional constraints. Homologous structural analysis is utilized to further annotate canonical and non-canonical functional residues.

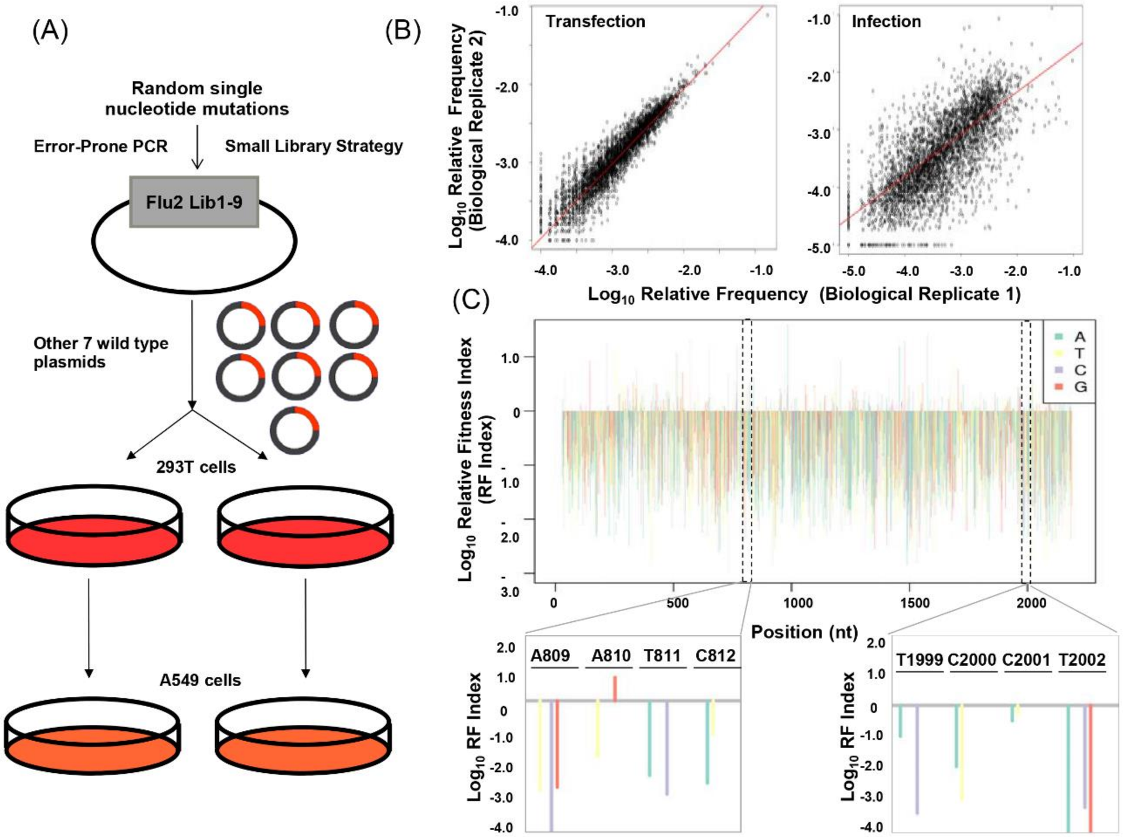


Figure 3-2. Fitness profile of influenza A virus segment 2 at single nucleotide resolution. (A) A schematic presentation of the experimental flow of high-throughput fitness profiling. Random single nucleotide mutations were introduced into influenza A/WSN/33 segment 2. Mutant viral libraries were generated by co-transfecting mutant DNA library with seven plasmids encoding the other wild-type viral fragments. Viral libraries were then passaged in A549 cells. High-throughput sequencing was performed for the plasmid mutant libraries, post transfection and post-infection viral libraries. (B) Correlation of the relative frequency of each single-nucleotide mutations between biological duplicates are shown. (C) Relative fitness scores are shown for individual mutations of influenza A/WSN/33 segment 2 in log10. Two representative regions are zoomed in to show the single nucleotide change.

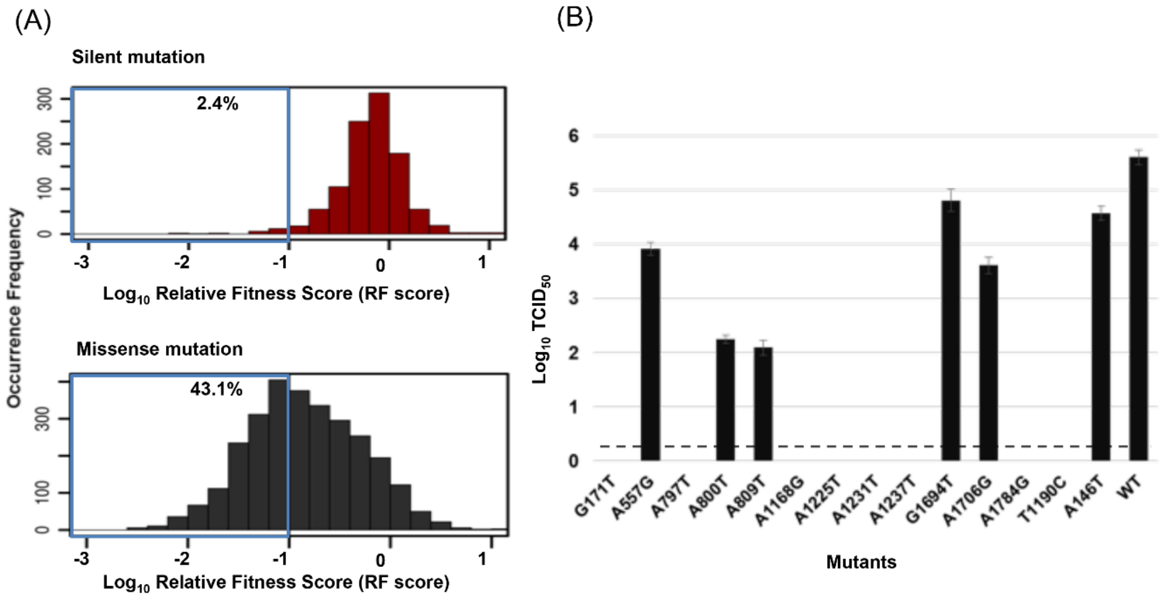


Figure 3-3. Systematic identification of deleterious mutations of PB1 protein. (A) Histogram illustrations are shown for the relative fitness distribution (RF index in log10) of silent mutations and missense mutations. Mutations of RF index ≤ 0.1 were identified as deleterious mutations. The percentage of silent mutations and missense mutations that fall below this cutoff are boxed in blue. (B) Fourteen deleterious mutations were selected and reconstructed into viral genome. TCID50 of selected single nucleotide mutations are shown. Dashed line represents the detection limit of TCID50 assay. Data is presented as means±SD from a biological duplicate.

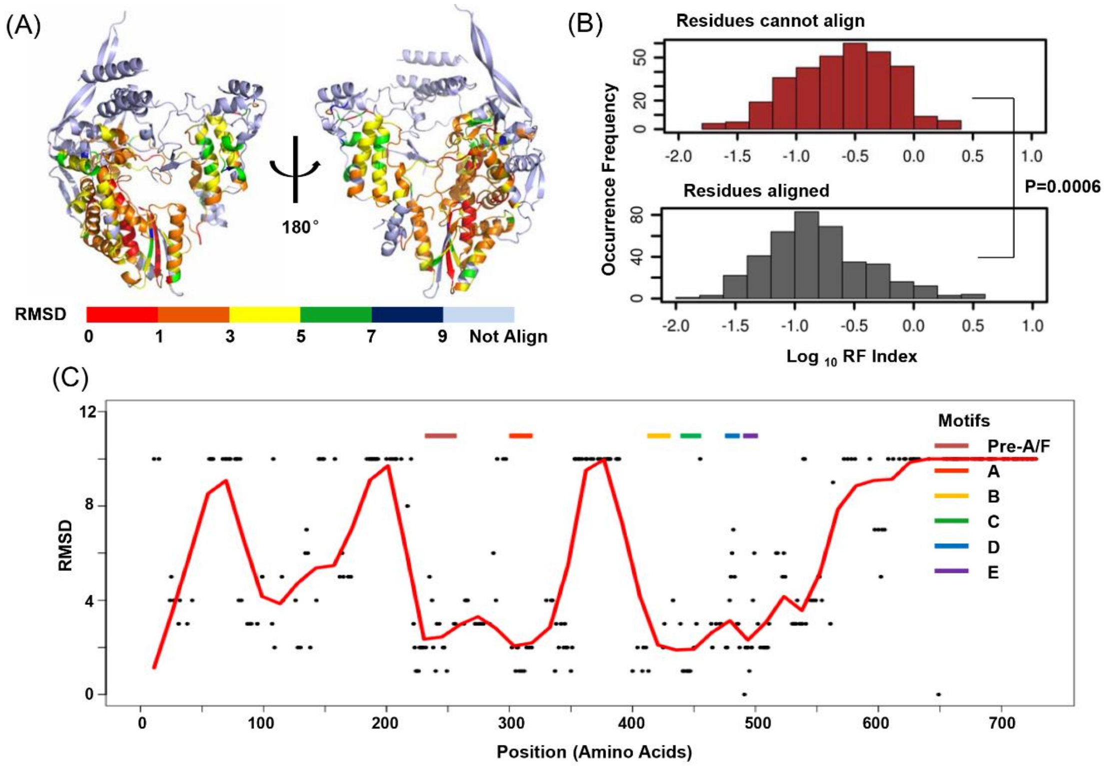


Figure 3-4. Annotation of PB1 functional residues with homologous structural alignment. (A) Multiple structure alignment was performed among PB1 and 16 other homologous structures in vRdRp family. PB1 structure is rainbow colored according to the root-mean-square deviation (RMSD) of each residue (B) Histograms of the RF indexes are shown for residues that cannot align (red) and the residues that can align to other structures in vRdRp family. RF index were significantly higher for residues that cannot be aligned (two-tailed T test, p=0.0006). (C) RMSD scores are shown for functional residues. Smooth curve was fitted by loess regression. Conserved domains (pre-A/F, A-E) of vRdRp are labeled and shown as valleys on the smooth curve of RMSD.

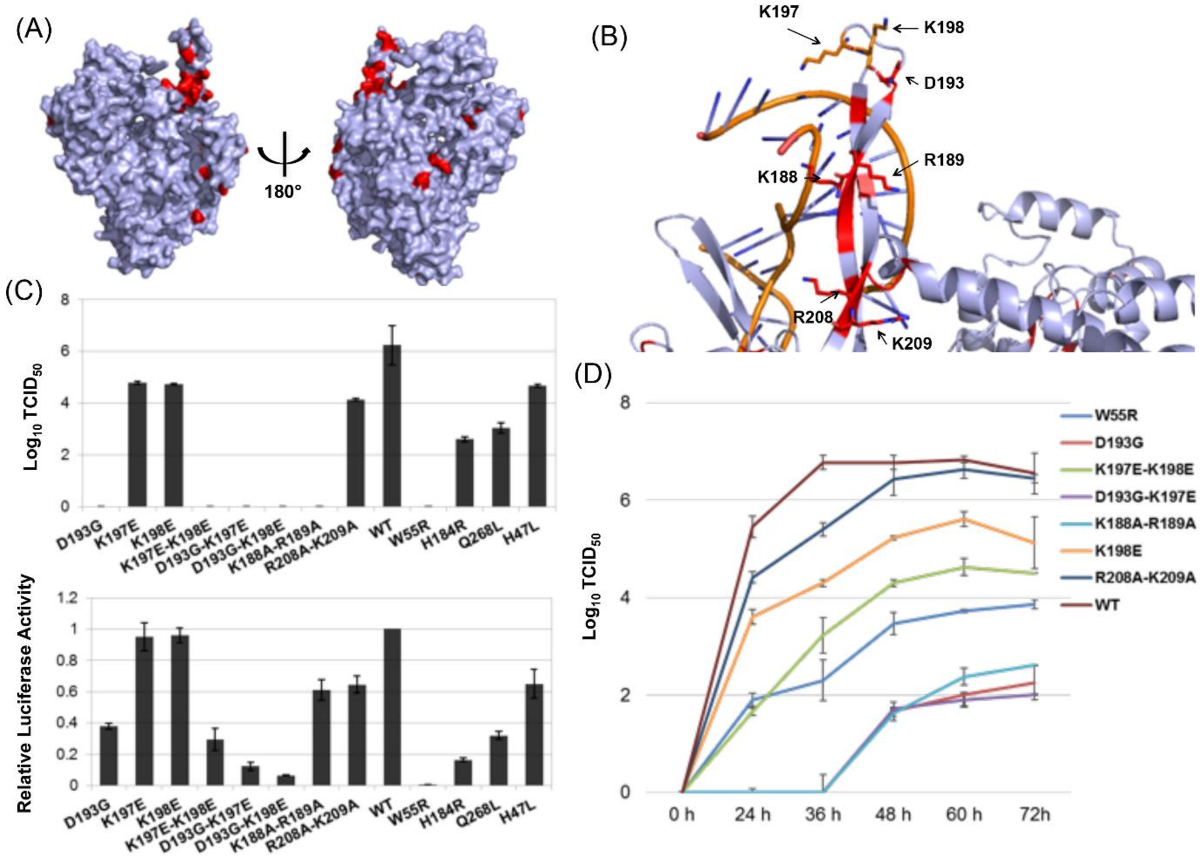


Figure 3-5. Identification of non-canonical functional residues of PB1 protein. (A,B) Non-canonical non-interface functional residues of PB1 protein are highlighted in red. A cluster of residues is located at the long twisted β-ribbon region. The non-structured loop region (amino acid 195-198) was reconstructed by Rosetta. (C) TCID50 (Upper panel) and relative polymerase activity (Lower panel) for indicated mutations are shown. The data are presented as the mean±SD from four independent biological replicates. (D) Growth curve of indicated mutations are shown. A549 cells were infected with indicated mutant virus at an MOI 0.1. Viruses were collected at indicated time points and TCID50 were measured.

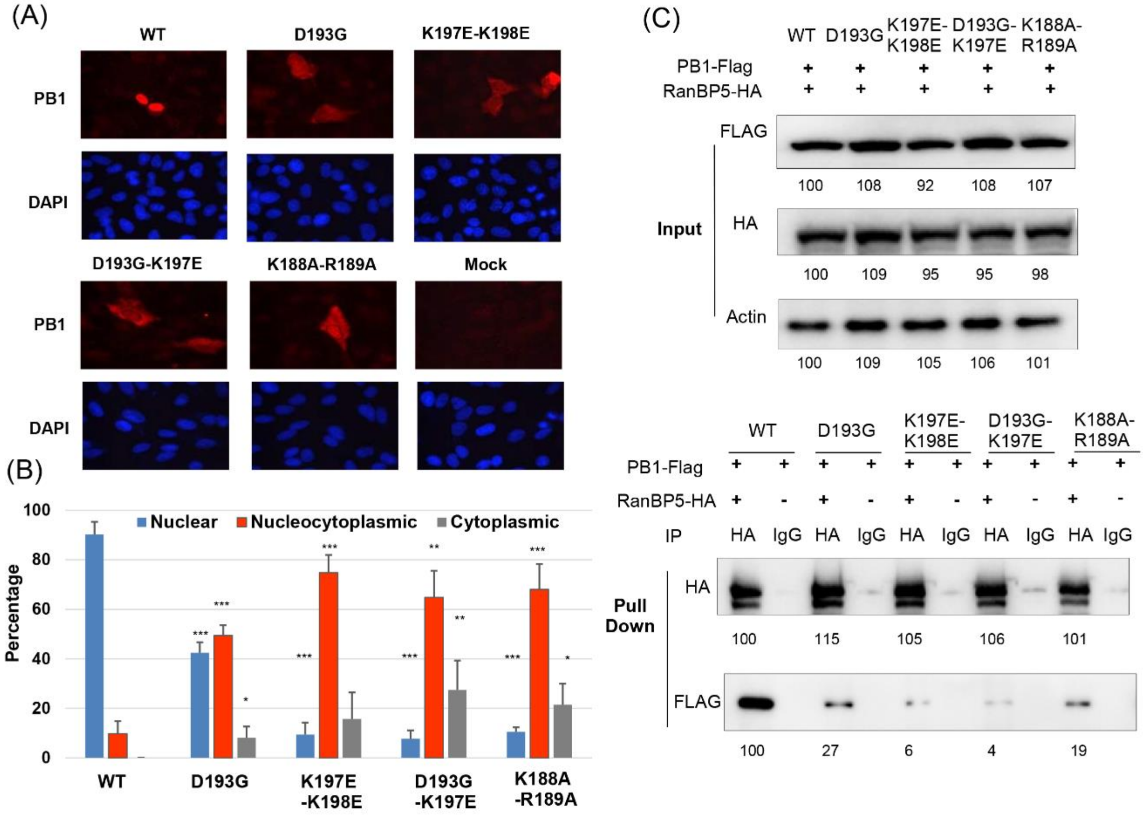


Figure 3-6. Identified non-canonical functional residues may involve in nuclear import of PB1 protein by interacting with RanBP5. (A) Cellular localizations of wild type and mutant PB1 proteins were examined by immunofluorescence. (B) Percentage of cells with different PB1 localizations. Data is presented as mean±SD from three independent biological replicates. At least 50 cells were analyzed for each replicate with ImageJ. * p<0.05; ** p< 0.01; *** p< 0.001 by two-tailed T test. (C) Interactions between PB1 proteins and RanBP5 were examined by immunoprecipitation (IP). Number below each band is the intensity quantification measured by Image Lab.

Table 3-1. Comparison of methods in 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

3.6 Supplementary Materials

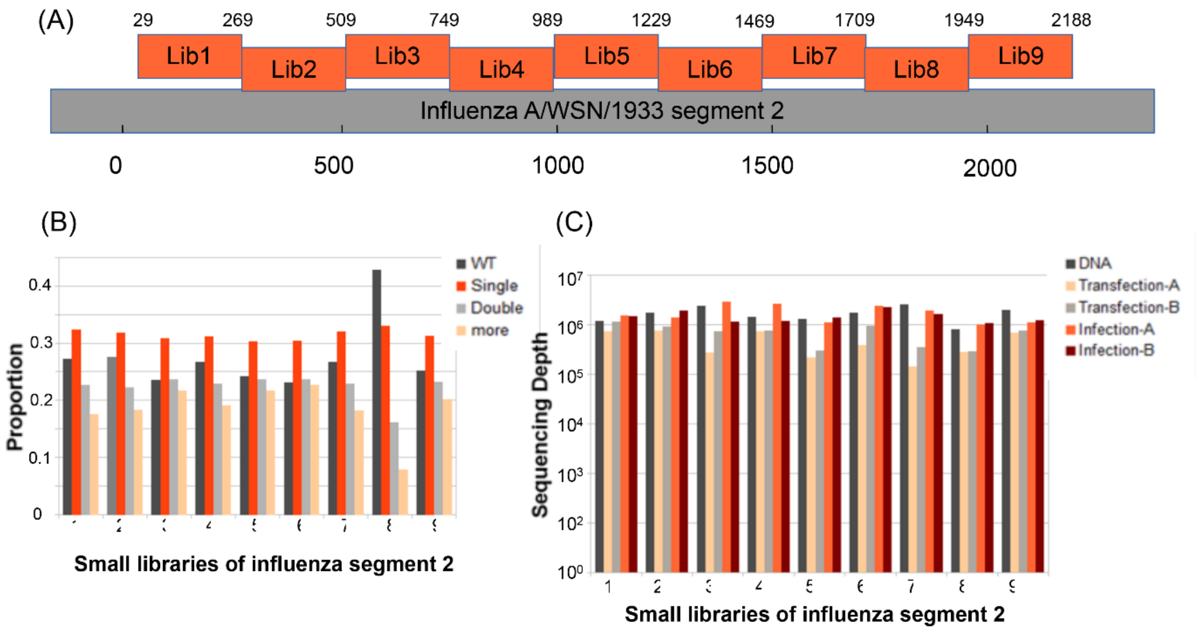


Figure S3-1. Construction of small libraries and sequencing coverage

(A) A schematic presentation of the mutagenesis library of influenza segment 2, covered by 9 small libraries of 240 bp each. The starting nucleotide position for each segment is labeled. (B) Distribution of mutations in the input DNA plasmid library is shown as a bar chart. ~30-35% of clones in the plasmid mutant library contained one mutation (single). ~25%-30% were wild type, and the rest of them contained two or multiple mutations. (C) Sequencing depth is shown for each small library.

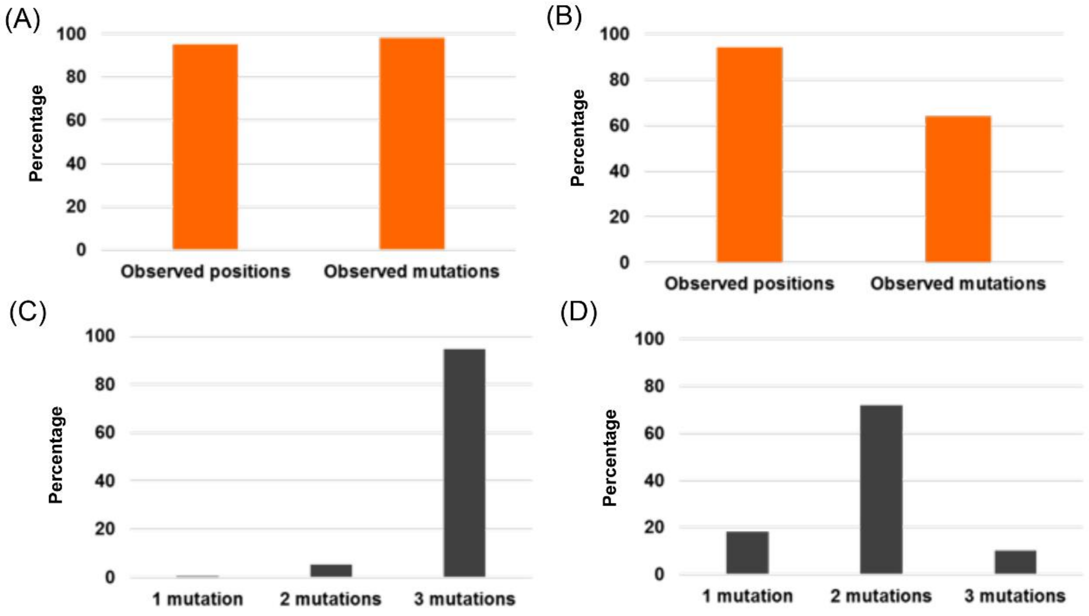


Figure S3-2. Position and mutation coverage of libraries

(A) Percentage of nucleotide positions that were observed in the fitness profile and the percentage of total single nucleotide mutations that were observed in the fitness profile prior to library quality control. (B) Percentage of nucleotide positions that were observed in the fitness profile and the percentage of total single nucleotide mutations that were observed in the fitness profile post library quality control (mutation frequency > 0.1% in the DNA library). (C) Percentage of one, two, or all three nucleotide mutations observed for all positions prior to library quality control. (D)) Percentage of one, two, or all three nucleotide mutations observed for all positions post library quality control.

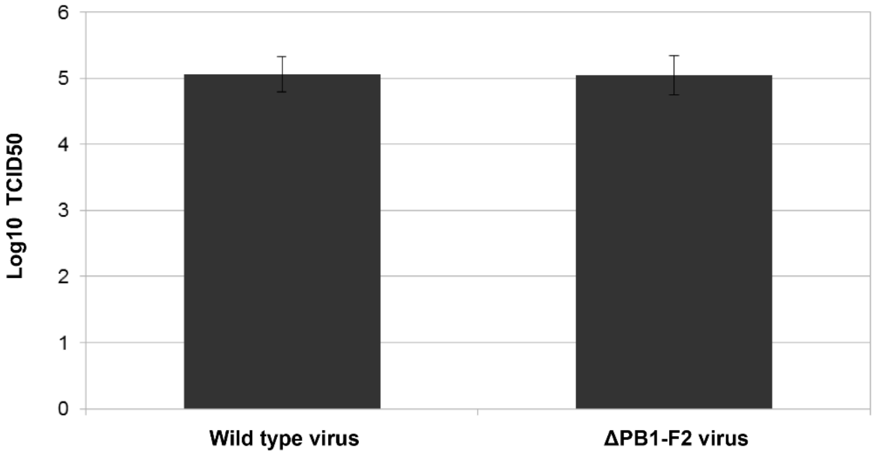


Figure S3-3. Growth Capacity of wild type and ΔPB1-F2 virus

TCID50 are shown for wild type virus and PB1-F2 knock out (ΔPB1-F2) virus in log10. No significant difference in viral growth was detected. The ΔPB1-F2 virus was generated by mutating the start codon of the gene (T120C) and introducing two stop codons at position 12 (C153G) and position 58 (G291A).

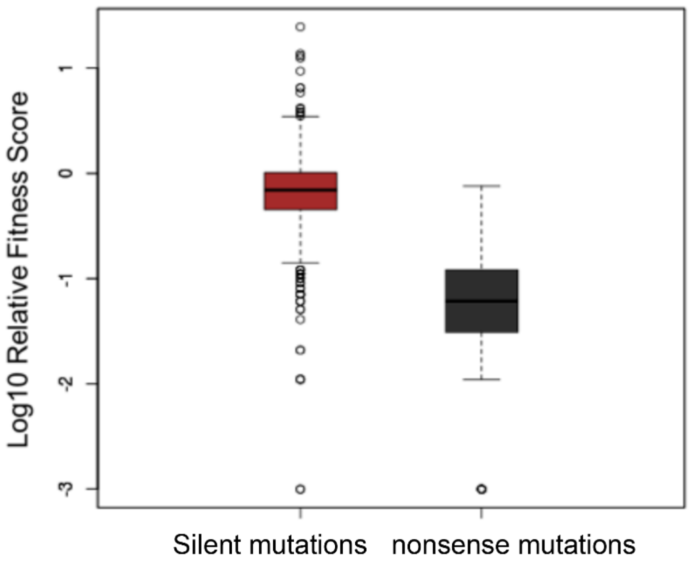


Figure S3-4. RF indexes of silent mutations and nonsense mutations

The relative fitness indexes (RF index) of silent mutations and nonsense mutations are shown with box plot. The average RF index of silent mutations is significantly larger than nonsense mutations (two-tailed T test, P=4.6E-021)

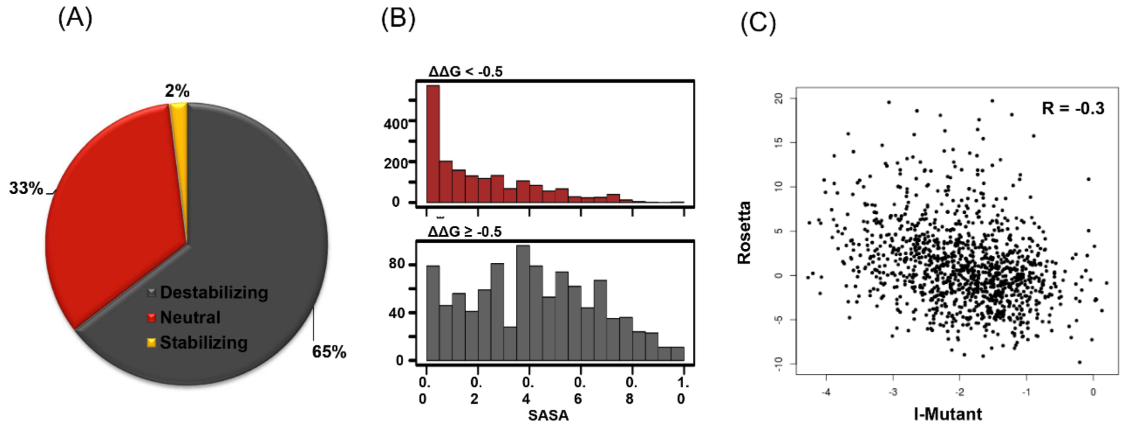


Figure S3-5. Protein stability prediction by I-Mutant and Rosetta

(A) Proportion of mutations that are destabilizing, neutral or stabilizing in PB1 protein as predicted by I-mutant 2.0 are shown as a pie chart (B) Distribution of the solvent accessible surface area (SASA) is shown for destabilizing mutations and other mutations. The destabilizing mutations showed significant lower values of SASA (two tailed T test, P=5.4E-117). (C) Correlation of ΔΔG prediction result of I-Mutant and Rosetta. Absolute Correlation coefficient of prediction that result between these two methods is 0.3. Note that the sign of ΔΔG is opposite in these two computational tools.

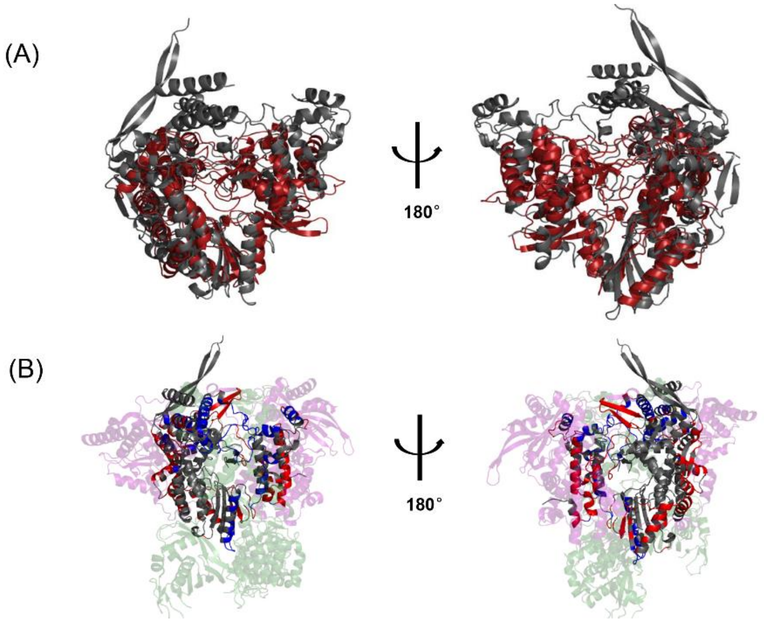


Figure S3-6. Pair-wise structural alignment of PB1 protein and interface prediction.

(A) Pair-wise alignment of PB1 protein and HCV NS5B (PDB: 2XI3) are shown. PB1 is colored in gray and NS5B is colored in crimson. (B) Possible interfaces between PA, PB1 and PB2 are shown. PA structure is shaded in purple and the possible interacting residues of PB1 are colored in red. PB2 structure is shaded in green and the possible interacting residues of PB1 are colored in blue. The interface residues were predicted by SASA changing upon complex formation using Sppider.

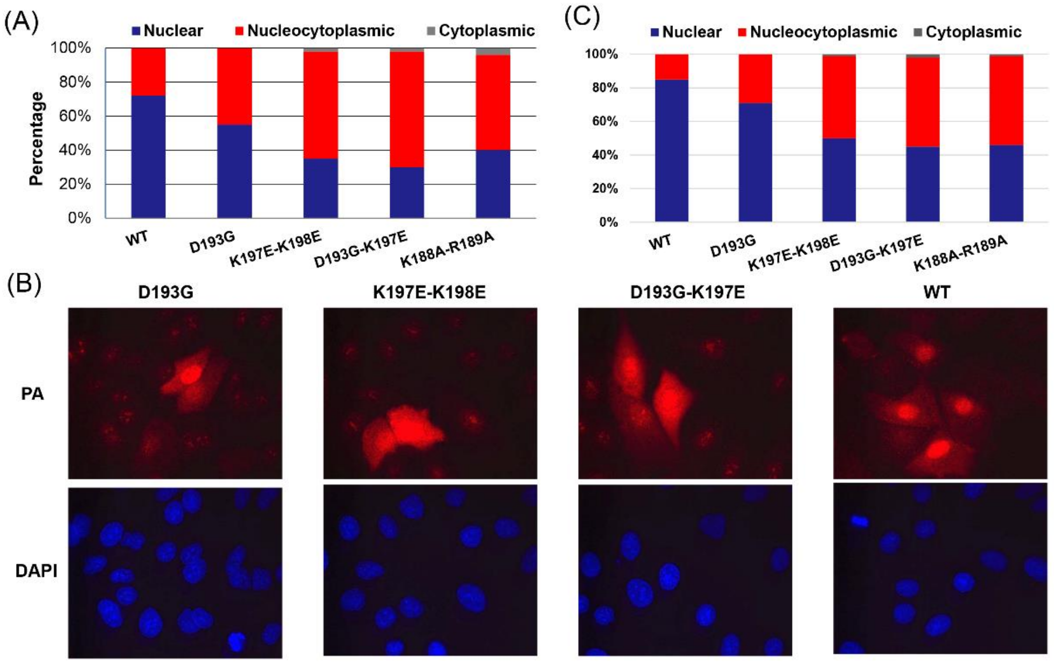


Figure S3-7. Nuclear localization of PB1 and PA protein upon PB1 mutations.

(A and C) Quantification of cellular localization (A) PB1, and (C) PA proteins of wild-type and mutant viruses. A549 cells were infected with wild type and mutated virus at MOI 0.5. Cells were fixed and subjected to immunofluorescence analysis (IFA) at 8h post-infection. At least 50 cells were analyzed for each mutant. Images were processed by ImageJ. (B) Example images are shown for the cellular localization of PA proteins of PB1 mutant viruses.

3.7 Bibliogrphy

1. Mills CL, Beuning PJ, Ondrechen MJ. 2015. Biochemical functional predictions for protein structures of unknown or uncertain function. Comput Struct Biotechnol J 13:182–91.

2. Gonzalez-Perez A, Mustonen V, Reva B, Ritchie GRS, Creixell P, Karchin R, Vazquez M, Fink JL, Kassahn KS, Pearson J V, Bader GD, Boutros PC, Muthuswamy L, Ouellette BFF, 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 10:723–9.

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.

4. Betancourt AJ, Bollback JP. 2006. Fitness effects of beneficial mutations: the mutational landscape model in experimental evolution. Curr Opin Genet Dev 16:618–23.

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 30:1237–44.

6. Glaser F, Pupko T, Paz I, Bell ER, 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.

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:W390–5.

8. Wilkins AD, Bachman BJ, Erdin S, Lichtarge O. 2012. The use of evolutionary patterns in protein annotation. Curr Opin Struct Biol 22:316–25.

9. Panchenko A, Kondrashov F, Bryant S. 2004. Prediction of functional sites by analysis of sequence and structure conservation. Protein Sci 884–892.

10. Capra J a, Laskowski R a, Thornton JM, Singh M, Funkhouser T a. 2009. Predicting protein ligand binding sites by combining evolutionary sequence conservation and 3D structure. PLoS Comput Biol 5:e1000585.

11. Tong W, Williams R, Wei Y, Murga L. 2008. Enhanced performance in prediction of protein active sites with THEMATICS and support vector machines. Protein Sci 333–341.

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.

13. Skolnick J, Brylinski M. 2009. FINDSITE: a combined evolution/structure-based approach to protein function prediction. Brief Bioinform 10:378–91.

14. Pazos F, Sternberg M. 2004. Automated prediction of protein function and detection of functional sites from structure. Proc Natl Acad Sci 2004.

15. Petrova N V, Wu CH. 2006. Prediction of catalytic residues using Support Vector Machine with selected protein sequence and structural properties. BMC Bioinformatics 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.

17. te Velthuis AJW. 2014. Common and unique features of viral RNA-dependent polymerases. Cell Mol Life Sci 71:4403–20.

18. Shatskaya GS, Dmitrieva TM. 2013. Structural Organization of Viral RNA-Dependent RNA Polymerases. Biochem Biokhimii͡a 78:231–5.

19. Ortín J, Parra F. 2006. Structure and function of RNA replication. Annu Rev Microbiol 60:305–26.

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.

21. Bruenn J a. 2003. A structural and primary sequence comparison of the viral RNA-dependent RNA polymerases. Nucleic Acids Res 31:1821–1829.

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 89:275–86.

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 J V. 2002. Oligomerization and Cooperative RNA Synthesis Activity of Hepatitis C Virus RNA-Dependent RNA Polymerase. Society 76:3865–3872.

24. Gao L, Aizaki H, He J-W, Lai MMC. 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 78:3480–8.

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.

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.

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.

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 86:8359–66.

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.

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 5:1–11.

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 M a, Hatta M, Heegel R a, 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 A a, Shukla AK, Sims A, Sova P, Tam VC, Tchitchek N, Thomas PG, Tilton SC, Totura A, Wang J, Webb-Robertson B-J, Wen J, Weiss JM, Yang F, Yount B, Zhang Q, McWeeney S, Smith RD, Waters KM, Kawaoka Y, Baric R, Aderem A, Katze MG, Scheuermann RH. 2014. A comprehensive collection of systems biology data characterizing the host response to viral infection. Sci data 1:140033.

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 92:1859–69.

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 88:10157–64.

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.

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.

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 b-Lactamase. Cell 160:882–892.

37. Fowler DM, Fields S. 2014. Deep mutational scanning: a new style of protein science. Nat Methods 11:801–7.

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 110:20248–53.

39. Hoffmann E, Neumann G. 2000. A DNA transfection system for generation of influenza A virus from eight plasmids. Proc Natl Acad Sci 97:6108–6113.

40. Chen W, Calvo P a, 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.

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.

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 33:5861–7.

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:W306–10.

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 22:553–60.

45. Thiltgen G, Goldstein R a. 2012. Assessing predictors of changes in protein stability upon mutation using self-consistency. PLoS One 7:e46084.

46. Reich S, Guilligay D, Pflug A, Malet H, Berger I, Crépin T, Hart D, Lunardi T, Nanao M, Ruigrok RWH, Cusack S. 2014. Structural insight into cap-snatching and RNA synthesis by influenza polymerase. Nature 516:361–366.

47. Pflug A, Guilligay D, Reich S, Cusack S. 2014. Structure of influenza A polymerase bound to the viral RNA promoter. Nature 516:355–60.

48. Joosten RP, Te Beek TAH, Krieger E, Hekkelman ML, Hooft RWW, Schneider R, Sander C, Vriend G. 2011. A series of PDB related databases for everyday needs. Nucleic Acids Res 39:411–419.

49. Kabsch W, Sander C. 1983. Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features. Biopolymers 22:2577–2637.

50. Tien MZ, Meyer AG, Sydykova DK, Spielman SJ, Wilke CO. 2013. Maximum allowed solvent accessibilites of residues in proteins. PLoS One 8.

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.

52. Li C, Wu A, Peng Y, Wang J, Guo Y, Chen Z, Zhang H, Wang Y, Dong J, Wang L, Qin FX-F, 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.

53. Perez DR, Donis RO. 2001. Functional Analysis of PA Binding by Influenza A Virus PB1 : Effects on Polymerase Activity and Viral Infectivity † 75:8127–8136.

54. Jung TE, Brownlee GG. 2006. A new promoter-binding site in the PB1 subunit of the influenza A virus polymerase. J Gen Virol 87:679–88.

55. López G, Valencia A, Tress ML. 2007. Firestar--Prediction of Functionally Important Residues Using Structural Templates and Alignment Reliability. Nucleic Acids Res 35:W573–7.

56. Fischer JD, Mayer CE, Söding J. 2008. Prediction of protein functional residues from sequence by probability density estimation. Bioinformatics 24:613–20.

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.

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:W235–41.

59. Zhang Y, Skolnick J. 2005. TM-align: a protein structure alignment algorithm based on the TM-score. Nucleic Acids Res 33:2302–9.

60. Zhang Y, Skolnick J. 2004. Scoring function for automated assessment of protein structure template quality. Proteins 57:702–10.

61. Wang S, Peng J, Xu J. 2011. Alignment of distantly related protein structures: algorithm, bound and implications to homology modeling. Bioinformatics 27:2537–45.

62. Collins PJ, Haire LF, Lin YP, Liu J, Russell RJ, Walker P a, Skehel JJ, Martin SR, Hay AJ, Gamblin SJ. 2008. Crystal structures of oseltamivir-resistant influenza virus neuraminidase mutants. Nature 453:1258–1261.

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 polymerases. J Mol Biol 419:198–210.

64. Zamyatkin DF, Parra F, Alonso JMM, Harki D a, Peterson BR, Grochulski P, Ng KK-S. 2008. Structural insights into mechanisms of catalysis and inhibition in Norwalk virus polymerase. J Biol Chem 283:7705–7712.

65. Fullerton SWB, Blaschke M, Coutard B, Gebhardt J, Gorbalenya A, Canard B, Tucker P a, Rohayem J. 2007. Structural and functional characterization of sapovirus RNA-dependent RNA polymerase. J Virol 81:1858–1871.

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 87:5291–5.

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.

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 101:4425–30.

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 104:9463–8.

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.

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.

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 107:22505–10.

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.

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.

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.

76. Gerlach P, Malet H, Cusack S, Reguera J. 2015. Structural Insights into Bunyavirus Replication and Its Regulation by the vRNA Promoter. Cell 161:1267–79.

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.

78. Takeshita D, Tomita K. 2012. Molecular basis for RNA polymerization by Qβ replicase. Nat Struct Mol Biol 19:229–237.

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.

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.

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.

82. Porollo A, Meller J. 2007. Prediction-Based fingerprints of protein-protein interactions. Proteins 66:630–645.

83. Deng T, Engelhardt OG, Thomas B, Akoulitchev A V, 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 80:11911–9.

84. Hengrung N, El Omari K, Martin IS, Vreede FT, Cusack S, Rambo RP, Vonrhein C, Bricogne G, Stuart DI, Grimes JM, others. 2015. Crystal structure of the RNA-dependent RNA polymerase from influenza C virus. Nature 527:114–117.

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.

86. Mandell DJ, Coutsias E a, Kortemme T. 2009. Sub-angstrom accuracy in protein loop reconstruction by robotics-inspired conformational sampling. Nat Methods 6:551–552.

87. Squires RB, Noronha J, Hunt V, Garc??a-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 Respi Viruses 6:404–416.

88. Edgar RC. 2004. MUSCLE: Multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res 32:1792–1797.

89. Kosakovsky Pond SL, Frost SDW, Muse S V. 2005. HyPhy: Hypothesis testing using phylogenies. Bioinformatics 21:676–679.

90. Broadbent AJ, Santos CP, Godbout R a, 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/. J Virol 88:12339–47.

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 89:6376–90.

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.

93. Yaseen NR, Blobel G. 1997. Cloning and characterization of human karyopherin beta3. Proc Natl Acad Sci U S A 94:4451–6.

94. Betts MJ, Guigo 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.

95. Naim HY, Niermann T, Kleinhans U, Hollenberg CP, Strasser AWM. 1991. Striking structural and functional similarities suggest that intestinal sucrase-isomaltase, human lysosomal ??-glucosidase and Schwanniomyces occidentalis glucoamylase are derived from a common ancestral gene. FEBS Lett 294:109–112.

96. Lee D, Redfern O, Orengo C. 2007. Predicting protein function from sequence and structure. Nat Rev Mol Cell Biol 8:995–1005.

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 282:903–18.

99. Russell RB. 1998. Detection of protein three-dimensional side-chain patterns: new examples of convergent evolution. J Mol Biol 279:1211–27.

100. Hansen JL, Long a M, 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 30:7353–8.

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 . J Biol Chem 284:478–485.

103. Chook YM, Suel KE. 2011. Nuclear import by karyopherin-βs: Recognition and inhibition. Biochim Biophys Acta - Mol Cell Res 1813:1593–1606.

104. Guo HH, Choe J, Loeb L a. 2004. Protein tolerance to random amino acid change. Proc Natl Acad Sci U S A 101:9205–9210.

105. Jacquier H, Birgy A. 2013. Capturing the mutational landscape of the beta-lactamase TEM-1. Proc Natl Acad Sci U S A 110:13067–13072.

106. McLaughlin RN, Poelwijk FJ, Raman A, Gosal WS, Ranganathan R. 2012. The spatial architecture of protein function and adaptation. Nature 491:138–42.

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 42:1–8.

108. Das R, Baker D. 2008. Macromolecular modeling with rosetta. Annu Rev Biochem 77:363–382.

109. Kellogg EH, Leaver-Fay A, Baker D. 2013. Role of conformational sampling in computing mutation-induced changes in protein structure and stability. Proteins 79:830–838.

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.

CHAPTER 4

SYSTEMATIC FITNESS PROFILING REVEALS DISTINCT RESISTANT BARRIERS FOR THREE DIFFERENT INFLUENZA NEURAMINIDASE INHIBITORS

4.1 Abstract

Neuraminidase inhibitors (NAIs) are widely used to as anti-viral drug for influenza virus infection. However, the low genetic barrier of NAIs facilitates the rapid development of single nucleotide drug resistant mutations. By combining saturated nucleotide mutagenesis and high-throughput sequencing, we characterized the fitness effects of single nucleotide mutations of neuraminidase (NA) and systematically identified resistant mutations for three NAIs: zanamivir, oseltamivir and AV5080. We observed that both the number and the effects of resistant mutations of AV5080 are smaller than those of zanamivir and oseltamivir, but so are their fitness costs. We used population genetic models to estimate the rate of increase in fitness under drug selection as a function of drug dosage. AV5080 showed a higher rate of increase in fitness at low drug concentrations due to the low fitness cost of resistant mutations, but also exhibited a steep drop with high drug concentrations because of lower strength of resistance. Our approach also enabled the systematic analyses of cross-resistance among different drugs, which showed to be uncommon between AV5080 and zanamivir. In summary, we performed systematic profiling of resistance mutations of three NAIs, which showed distinct resistant barriers. The generally applicable high-throughput approach may assist in evaluation of drug resistance and guild the design of personalized therapy.

4.2 Introduction

Neuraminidase inhibitors (NAIs) are the most commonly used drugs to treat influenza viral infection. As sialic acid analogs, NAIs inhibit the cleavage and spread of influenza virus by directly inhibiting neuraminidase (NA) function. Zanamivir was the first NAI introduced into clinic as a nasal spray, and was followed by the development of oseltamivir, which can be taken orally (1). Peramivir and laninamivir were also developed and approved for clinic usage in Japan (2–5).

NAIs are known to be effective for multiple isolated strains from both influenza A and influenza B. However, the high mutation rate of influenza virus and low resistant barrier of NAIs permit the rapid development of drug resistance (1, 6–8). A number of resistant mutations were reported and validated for both zanamivir and oseltamivir (7, 9–17). These drug resistant mutations, most of which contain only single nucleotide changes, can dramatically decrease the binding of NA to NAIs (18, 19). The emergence and world-wide spread of NAI resistant mutations create a major public health burden and emphasize the importance of controlling drug resistancy. New generations of NA inhibitors are actively being developed (20–22). With different binding mode or different mechanism to inhibit NA function, these new drugs might result in different mechanisms of resistance. It is important to understand the distinct resistant profiles of different NAIs in order to guide the rational use of drugs.

Two strategies have been commonly used to identify resistant mutations: isolation of resistant strains in clinic(1, 6, 7, 31), and long-term evolution in-vitro under drug selection (32, 33). Through sequencing, NA mutations with resistant phenotypes can be identified. However, both strategies suffer from limited sampling of de novo mutations and stochasticity in evolution. Some important resistant mutations might not be captured due to genetic drift or clonal interference. Importantly, neither of these methods is able to simultaneously quantify the fitness effect of resistant mutations, which is a critical component in determining the emergence of drug resistance (14, 23–30). Thus, a more systematic method needs to be established to evaluate the resistant barriers of different NAIs in a timely manner.

Cross resistance is also a major issue for the clinic usage of NAIs (31). It is reported that some mutations in functional active sites of NA (R152K, R292K) and framework residues (D198E/N/G) confer cross-resistance to both zanamivir and oseltamivir (15, 33). The traditional approach of examining cross resistance also relies on generating or isolating single mutations and test their drug resistant phenotype individually (34). A high-throughput approach to examine cross-resistance will greatly facilitate the evaluation of effective drug combination.

Here we present a systematic examination of the resistant profiles of three NA inhibitors, namely ostamivir, zanamivir and AV5080. By combining random mutagenesis and high-throughput sequencing, we identified resistance mutations for each drug. Both the resistant strength and fitness cost of resistant mutations were quantified simultaneously. We further assessed the distinct resistance barriers of three NAIs by applying a simple population genetic model. Cross resistance among the three drugs were identified and validated. The approach can systematically and rapidly evaluate the resistant barrier of NAIs. Importantly, this approach can be generally applied to other drugs with known target protein.

4.3 Results

4.3.1 SYSTEMATIC CHARACTERIZATION OF RESISTANT PROFILES OF THREE NAIS

We applied a high-throughput genetic approach to systematically profile resistant mutations of NA inhibitors. Conceptually, we sought to generate a high-dense mutant viral library, monitor the growth effect of each mutation with and without drug selection, and then quantify the fitness and drug resistance of each mutation. The mutant viral library ensured the pre-existence of mutations at relatively high frequency, minimizing the effect of genetic drift. By quantifying the enrichment of each mutation under drug selection, we bypassed the requirement of dominance or fixation of beneficial mutations to detect resistance.

Random mutagenesis was performed on the entire NA segment of influenza A/WSN/33 virus by error-prone PCR (Methods). Mutant viral libraries were reconstituted by co-transfecting the mutant NA segment together with seven other wild-type plasmids. To assess viral growth capacity under no drug condition, the mutant viral library was passaged in A549 cells for three rounds (24h per round) with MOI equal 0.05. In parallel, the mutant viral library was selected for three 24h passages under NAI drug selection. Three NAIs were included in our study including two standard NAIs: zanamivir and oseltamivir, and a newly developed NAI: AV5080 (20). Transfection and subsequent selection were performed in biological duplicates.

The mutant DNA library, reconstituted viral library and passaged virus library were subjected to next-generation sequencing with Illumina Hiseq 2000. 2746 mutations were analyzed, which covered 94.7% of total residues (429/453) and 71.2% (2746/3859) of all possible single nucleotide mutations. Selection under drug naïve condition enabled us to examine the fitness effect of each single nucleotide substitution for viral replication. The relative fitness score (RF score) for each mutant was calculated as the frequency in the selected pool (drug naïve condition) divided by the frequency in DNA library. Strong rank correlation of RF scores from biological duplicates was obtained (R=0.8) (**Figure 4-1**), indicating the reproducibility of our method. We also observed a clear separation between the distribution of RF scores of silent mutations and missense mutations, suggesting efficient selection in our passages (**Figure 4-1**). We further evaluated the essentialness of every residue by quantifying the percentage of deleterious mutations (RF score <0.2) among all profiled single nucleotide mutations in this particular residue (essentialness score). The known catalytic sites, including R118, R151, R152, D176, R224, R292, R371 and Y406 (N2 numbering) were all shown to be essential with more than 50% of mutations had deleterious effect for viral replication (**Figure 4-1**). These results validated our high-throughput approach to quantify the mutational effect on viral replication.

4.3.2 IDENTIFICATION OF RESISTANCE MUTATIONS AND DISTRIBUTION OF DRUG RESISTANCE SCORE

For a particular drug, the drug resistance score (W) of each mutation was calculated as the ratio of RF score under drug selection to that in drug naïve condition. We used synonymous mutations as internal controls, with the assumption that a synonymous mutation would have minimum, if not no effects on drug sensitivity. Log (W) of synonymous mutations followed a normal distribution centered around 0, and were comparable among three NAIs (**Figure S4-1**). We defined a mutation as drug resistant (R) if W was larger than mean+ of synonymous mutations, and as highly drug resistant (HR) if W was larger than mean+ 2

In the profile, 191 mutations were shown to be resistant for zamanivir, 70 of which we classified as highly resistant mutations (**Figure 4-2**). We detected R292K, H274Y, D199G, I117V, and E119D to be resistant, which all have been reported previously (7, 9–17). 88 mutations were resistant for osetamivir (28 highly resistant), and also included the well-known mutations: H274Y, R292K, D199G, D199N, I117V, N294S, D199E (1, 7). For the new drug AV5080, 23 mutations were shown to be drug resistant mutations. Although there is possible false positive and false negative related with the high-throughput screen, we noted that the number and resistance effects of drug resistant mutations were fewer for AV5080 compare with zamanivir and oseltamivir (**Figure S4-2**). This pattern held true across varying criteria for identifying drug resistant mutations (method). We also constructed and validated the IC50 of 7 Resistant Mutations against different drugs, which showed consistency with the profile data (**Figure 4-2**).

The distribution of W is a key parameter in determining the resistant potential of single nucleotide mutations. We compared the different distributions of W of drug resistant mutations of the three drugs using violin plots (**Figure S4-3**). There is a heavy tail in the distribution of W (i.e. highly resistant mutations) for zanamivir and oseltamivir, but not for AV5080. This difference is also evident in the cumulative distribution function of resistant mutations (**Figure S4-3**). We further fit the Generalized pareto distribution (GPD) to quantitatively examine the shape of the distributions of W (**Figure 4-2**, **Figure S4-3**) (35–45). For zanamivir and oseltamivir, the distribution of W falls in Fréchet domain, characterized by the heavier tail of a few exceptionally resistant mutants, while W of AV5080 falls in Gumbel domain with an exponential tail (46, 47).

4.3.3 Evaluating the distinct resistant barriers of drugs by combining drug resistance score and fitness cost

The emergence of resistant mutations is determined by their effects on both drug resistance and replicative fitness (48). Our fitness profiling in drug naïve condition and under drug selection provided a unique opportunity to simultaneously evaluate the resistance and fitness cost of individual mutations. We plotted the RF score and drug resistant score (W) for each of the three NA inhibitors (**Figure 4-3**). 88% of resistant mutations of zanamivir and 80% of resistant mutations of oseltamivir had fitness costs comparable to wild type. All of the top resistant mutations (W>50) of zanamivir and oseltamivir were deleterious for viral replication. However, only 50% resistant mutations of AV5080 showed a decrease in viral growth capacity. Histogram of relative fitness score of resistant mutations also clearly showed that resistant mutations of AV5080 had smaller fitness cost (**Figure S4-4**). In fact, 7 out of 8 highly resistant mutations of AV5080 are at least neutral in viral growth from our dataset. Genetic barrier is also a critical component in determining the emergence of drug resistance. As we focused on single nucleotide mutations in our profile, we analyzed the ratio of transition (Ts) and transversion (Tv) of the drug resistant mutations among the three drugs. Resistant mutations of AV5080 had a slightly higher proportion of transitions than that of zanamivir and oseltamivir (**Figure S4-5**). The frequency of Ts versus Tv is estimated to be ~10:1 for influenza virus, suggesting a lower genetic barrier for transitions than transversions (49).

To compare the distinct resistant barriers of NA inhibitors, we evaluated the rate of fitness increase at different drug concentration, which depended on both the resistant effect and fitness cost (50–52). We developed a simplified population genetic model in the regime of strong-selection-weak-mutation considering only single nucleotide mutations (methods). We noted that the shape of the curve are different for different drugs, which is mostly determined by the spectrum of beneficial mutations at a particular drug concentration (**Figure 4-3**). At low drug concentration, AV5080 showed higher rate of fitness increase attributed to the low fitness cost of resistant mutations. However, due to less number of resistant mutations and smaller resistant effect, the rate of fitness increase also drops fast at higher drug concentrations. At concentrations ~5-6 fold of wild type IC50, the rate drop to almost zero for AV5080.

4.3.4 Cross resistance is rare between AV5080 and zanamivir, oseltamivir

Importantly, our system provides a rapid and comprehensive evaluation of cross resistance among different drugs, which is informative for clinical drug usage. We observed that 40% highly resistant mutations of ostamivir are also resistant for zanamivir (**Figure 4-4**). All the top-ranked resistant mutations are shared between ostamivir and zanamivir (**Figure S4-6**). Interestingly, there seems to be limited cross-resistance between AV5080 and the other two drugs, especially with Zanamivir (**Figure S4-4**). We further validated the AV5080 drug sensitivity of 4 mutations (I222L, N294S, I222R, R292K) that were shown to be resistant to either zanamivir or oseltamivir (**Figure 4-2**). Using NA activity assay, we observed that the resistance of these mutations to AV5080 is greatly reduced (**Figure 4-4**). All the mutations have IC50<1nM for AV5080. Out of the four mutations, I222R has the highest IC50 against AV5080, which is 4.1 fold higher than wild type. However, the IC50 of I222R is almost 18 fold higher compared with wild type for osletamivir. The low cross-resistance is consistent with previous reports when examining AV5080 among a diverse panel of influenza strains.

Mapping the resistant mutations of all three drugs onto the NA protein structure (pdb:3b7e), we noticed that a cluster of resistant residues centered around the catalytic region for ostamivir and zanamivir, differed from AV5080 (**Figure S4-7**). As the catalytic region is essential for NA function, the resistant mutations on this region would have more impact on viral replication. This may explain why the fitness cost of ostamivir and zanamivir resistant mutations are higher than

AV5080. Future studies on the crystal structure of NA with AV5080 may provide better explanation regarding the different resistant profiles.

4.4 Discussion

The emerging mutations that are resistant to current NA inhibitors makes it imperative to develop new anti-virals (20–22). It is essential to understand the resistant profiles of different drugs for their rational usage. Here we presented a systematic approach to examine the resistant profiles of three NA inhibitors, namely ostamivir, zanamivir and AV5080. Focusing on single nucleotide mutations, we identified mutations that are resistant to each of the drug. Both the resistant strength and fitness cost of resistant mutations were quantified simultaneously. By calculating the rate of fitness increase as a function of drug dose, we assessed the distinct resistance barriers of three NAIs. Furthermore, we evaluated the cross-resistance among different drugs.

This high-throughput genetic approach provided a systematic way to evaluate the resistant profiles of drugs (53–56). Here we quantified viral growth by tracking the changing of viral copy number through high-throughput sequencing. This allowed us to not have to rely on the dominance of beneficial mutations, which is a non-sensitive and time-consuming requirement of long-term evolution experiments. In this experiment, we quantified viral fitness using an in-vitro system. A similar assay can be established in vivo by infecting mutant libraries in mice with and without drug treatment. Moreover, this approach can be extended to study other drugs and allow us to quickly identify resistant mutants on the target gene.

We developed a simplified model to evaluate the distinct resistant barriers of each drug. Under the regime of strong-selection-weak-mutation, we quantified the rate of fitness increase under different drug concentrations. The shape of the curve are mostly determined by the spectrum of beneficial mutations at a particular drug concentration, which reflect the resistant effect as well as the fitness cost and genetic barrier of these beneficial mutations. Although this regime not always hold true for viral populations (large population size, large mutation rate), the ‘rate of fitness increase” would be a useful measurement to compare the emergence of drug resistant mutations among different drugs.

We also want to mention some limitations and future directions of current study. Firstly, we used random mutagenesis to introduced mutation for the entire NA segment and used 11 Illumia Hiseq reads to cover the entire library. Although we tried to limit the mutation rate, but there is still chance that multiple mutations occur at the same viral clone. The short read of Hiseq cannot enable us to tell if there is multiple-mutations occur at long distance. This may create noise in assessing the relative fitness score of particular mutation. Secondly, we generated the NA library using H1N1/WSN/33 virus as background, and only focusing on single nucleotide mutations. However, different genetic backgrounds might have different profiles of resistant mutations, especially with the existence of compensatory mutations. The resistant profile will also differ considering multiple mutations and epistasis. Nevertheless, our system can be quickly applied to other genetic backgrounds, and enable us to find the compensatory mutations of drug resistant mutations (58).

The understanding of distinct resistant barriers and cross-resistance can improve the clinic usage of drugs. Based on their profiles, we can rationally choose the combination of drugs that minimize the emergence of possible resistance. Moreover, it can also facilitate the improvement of drug development. An ideal drug would have both high genetic barrier and high fitness barrier. In other words, there needs to be several nucleotide mutations in order to be resistant and inflict a high fitness cost to viral replication. However, it is very rare that first generation drugs can have both high genetic and high fitness barriers. Thus, current drug development is usually iterative. After resistant mutations are observed through clinical usage of first generation drug, modifications or derivatives are made to overcome the resistance and improve the efficacy. This clinical accumulation of drug resistant mutations can be time consuming. High throughput resistant profiling provides an efficient way to evaluate the resistant barriers and thus facilitates the “next generation” of drug design. Furthermore, it can help in the modification of drugs to avoid the highly replicable single resistant mutations (57).

4.5 Materials and Methods:

Construction of influenza NA mutant library

Influenza A/WSN/33 NA mutant libraries were generated using the eight-plasmid transfection system (59). In brief, random mutagenesis was performed for the entire NA segment with error-prone polymerase Mutazyme II (Stratagene). We used NA segment with R194G mutation as our template for error-prone mutagenesis, as R194G mutation was reported to stabilize NA structure and increase viral replication capacity (58, 60). The amplified segment was gel purified, BsaI digested, ligated to the vector and transformed using MegaX DH10B T1R cells (Life Technologies). As the library was expected to have ~3000 single mutations, ~100,000 bacterial colonies were collected to cover the entirety. Plasmids from collected bacteria were midi-prepped as the input DNA library. On average, there are ~3 mutations per plasmid.

Transfection, infection and viral titer

To generate the mutant viral library, ~75 million 293T cells were transfected with 80ug DNA. Transfections were performed using Lipofectamine 2000 (Life Technologies). Virus was collected 72h post transfection. TCID50 were measured with A549 cells. For one round of viral passage, ~10 million A549 cells were infected with an MOI 0.05. Cells were washed with PBS three times at 2 h post-infection. Virus was collected 24 h post-infection from supernatant. Three rounds of passaging were performed with and without NA inhibitor selection. Biological duplicates were performed for transfection and following passage steps.

Individual mutant viral plasmids were generated by quick-change system. To generate single mutant virus, ~2 million 293T cells were transfected with 10 μg DNA. Viral titer were measured by TCID50 assay using A549 cells.

Sequencing library construction and data analysis

Viral RNA was extracted using QIAamp Viral RNA Mini Kit (Qiagen Sciences). DNaseI (Life Technologies) treatment was performed, followed by reverse transcription using Superscript III system (Life Technologies). At least 106 viral copy numbers were used to amplify the mutated segment.

The entire NA segment was separated into 11 small amplicons for sequencing, each around 94bp. Three nucleotides population IDs were introduced into both the 5' and 3' end of amplicons to separate between different samples, and six nucleotides molecular index tags were used for sequencing error correction.

Deep sequencing was performed with Illumina sequencing Hiseq PE100. Raw sequencing reads were de-multiplexed using the six-nucleotide ID. Sequencing error was corrected by filtering un-matched forward and reverse reads. Mutations were called by comparing sequencing reads with the wild-type sequence. Relative fitness score (RF scores) was calculated for individual point mutations and only mutations that have frequency more than 0.01% in the DNA library were reported.

$$\begin{aligned} &RF\ Score_{mutant\ i} \\ &\qquad = Relative\ Frequency\ of\ Mutant\ i_{\ infection} \\ &\qquad /Relative\ Frequency\ of\ Mutant\ i_{\ plasmid} \end{aligned}$$

Where $Relative\ Frequency\ of\ Mutant\ i = Reads\ of\ Mutant\ i\ /\ Total\ Reads$

We noted that we did not normalized to the relative frequency of wild type, but using total reads instead. As we used random mutagenesis for library construction and more than one mutation might occur in one plasmid. The short reads of Hiseq did not allow us to read through the entire NA segment, thus for a particular mutation occur in the sequenced region, there are potentially other mutations in the genetic background. Similarly, reads without mutations are not necessarily WT, so we cannot accurately calculate the relative frequency of WT. We are using “RF scores” here as a proxy for fitness.

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

Identification of drug resistant mutations

Synonymous mutations were used as benchmark for the identification of resistant mutations. For a fair comparison, we used the same cutoff for all three drugs in the main text. The cutoff is based on the mean and standard deviation of synonymous mutations across all drugs (mean=1.40,   = 2.89). We also examined the effect of different cutoffs, including using the distribution of synonymous mutation according to each drug, or based on 95% and 99% confidential interval curve. The result is consistant across varying criterias, that the number and resistance effects of drug resistant mutations were fewer for AV5080 compare with zamanivir and oseltamivir.

Calculation of resistant barrier

To compare the likelihood of emergence of drug resistance mutations under different drugs, we estimated the rate of increase in fitness of a virus population ($\frac{dF}{dt}$) under the simplified assumption of strong-selection-weak-mutation:

$$\frac{dF}{dt} \sim \mu \times P(s) \times s$$

$$s = f_{mut} - f_{wt}$$

P(s): probability that one mutant with selection coefficient s get fixed in a population of N. P(s)= 0 for s<=0.

μ: beneficial mutation rate, which is proportionally to the number of mutations

s: selection coefficient

fmut([drug=x]): fitness of one mutant at drug concentration x

fwt([drug=x]): fitness of wild type virus at drug concentration x

Where

$$P(s)_{\ muti} = \frac{1 - e^{-s}}{1 - e^{-Ns}} \approx 1 - e^{-s}\ when\ N\ is\ big\ enough\ and\ s > 0$$

$$s\ is\ selection\ coefficient, N\ is\ population\ size$$

μ (transition): μ (transversion) = 10:1

For each drug, the inhibition of viral fitness is modeled by a Hill function

$$f_{WT,[drug]} = f_{WT,[drug=0]} \times \frac{1}{1 - (\frac{[drug]}{IC50_{WT}})^{h}}$$

$$f_{muti,[drug]} = f_{muti,[drug=0]} \times \frac{1}{1 - (\frac{[drug]}{IC50_{muti}})^{h}}$$

$IC50\ is\ the\ half\ maximal\ inhibitory\ concentration, h\ is\ hill\ coefficient\ of\ drug$

$f_{WT,[drug=0]}/f_{muti,[drug=0]}$ is the RF scores for mutations under drug naïve condition.

The Hill coefficient (h), was fitted for each drug with data collected from wild type virus. For each mutation, IC50 were converted from drug resistant score using hill function, under the assumption that h is consistent between wild type and mutations. Selection coefficient (relative fitness) was normalized to each round of replication of influenza virus, using 6h as average time for one round of replication (61).

f_normalized=f^1/4

Neuraminidase activity assay.

106 TCID50 virus were used to measure the NA activity. NA activity was assayed using the fluorogenic substrate methylumbelliferyl-α-d-N-acetylneuraminic acid (MUNANA) (purchased

from Sigma) following a previously described protocol (http://www.nisn.org/documents/A.Hurt_Protocol_for_NA_fluorescence.pdf). Cell culture media were used as background control.

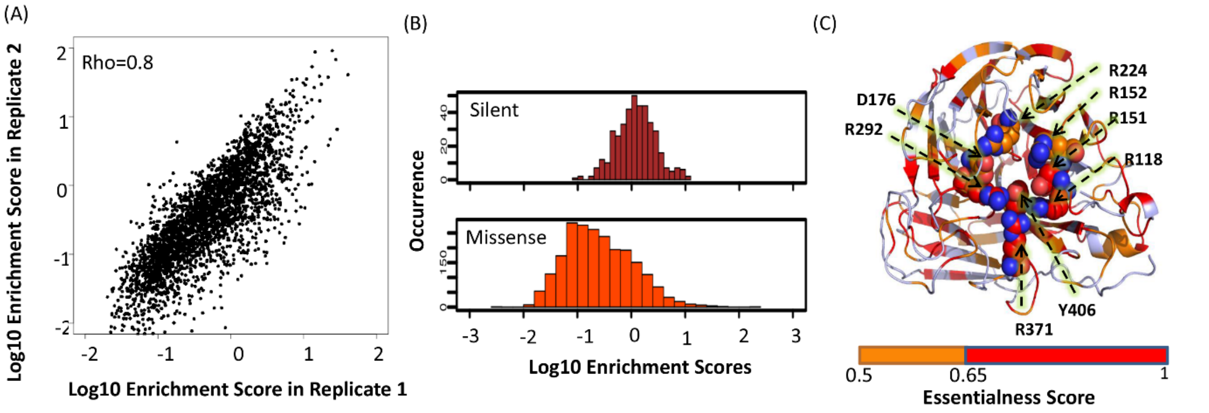
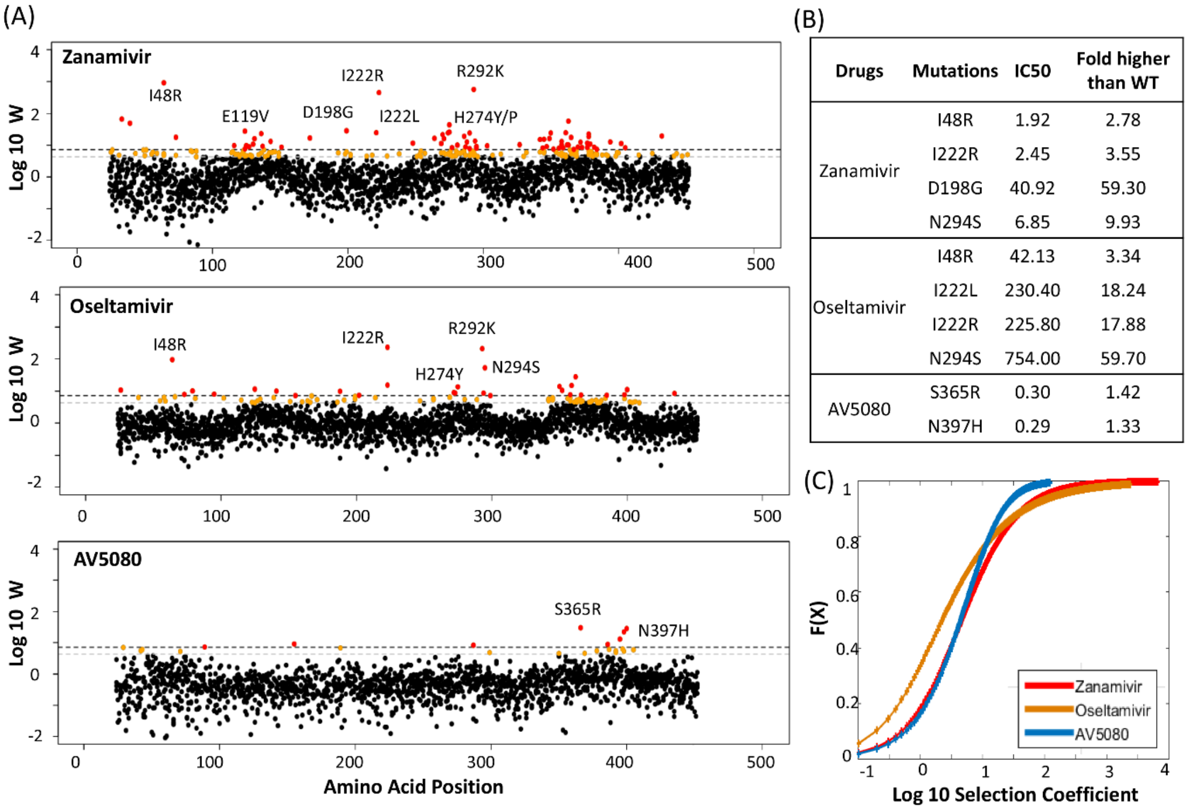


Figure 4-1. Systematic characterization of resistant profiles of three NAIs

(A) Correlation of relative fitness scores (RF scores) of NA mutations obtained from biological duplicates is shown. (B) Distribution of the RF scores for silent and missense mutations is shown. The clear separation between silent and missense mutations suggested sufficient selection pressure during passages. (C) Essentialness scores were calculated for each residue and mapped onto protein structure (PDB:3b7e). The catalytic core residues are all shown to be essential for viral replication.



Drugs	Mutations	IC50	Fold higher than WT
Zanamivir	I48R	1.92	2.78
	I222R	2.45	3.55
	D198G	40.92	59.30
	N294S	6.85	9.93
Oseltamivir	I48R	42.13	3.34
	I222L	230.40	18.24
	I222R	225.80	17.88
	N294S	754.00	59.70
AV5080	S365R	0.30	1.42
	N397H	0.29	1.33

Figure 4-2. Identification of resistance mutations and distribution of drug resistance score

(A) Scatter plots are shown for the drug resistant score (W) of each NA mutations by the amino acid positions. Resistant mutations were colored in orange, highly resistant mutations were colored in red. The top resistant mutations were labeled. (B) Validation of drug resistant mutations by NA activity assay. (C) The distribution of drug resistant scores (W) is fitted into Generalized pareto distribution (GPD). Selection coefficient of each resistant mutant is rescale from the minimum W within one drug (S resistnat mut = W resistant mut- W min resistant mut).

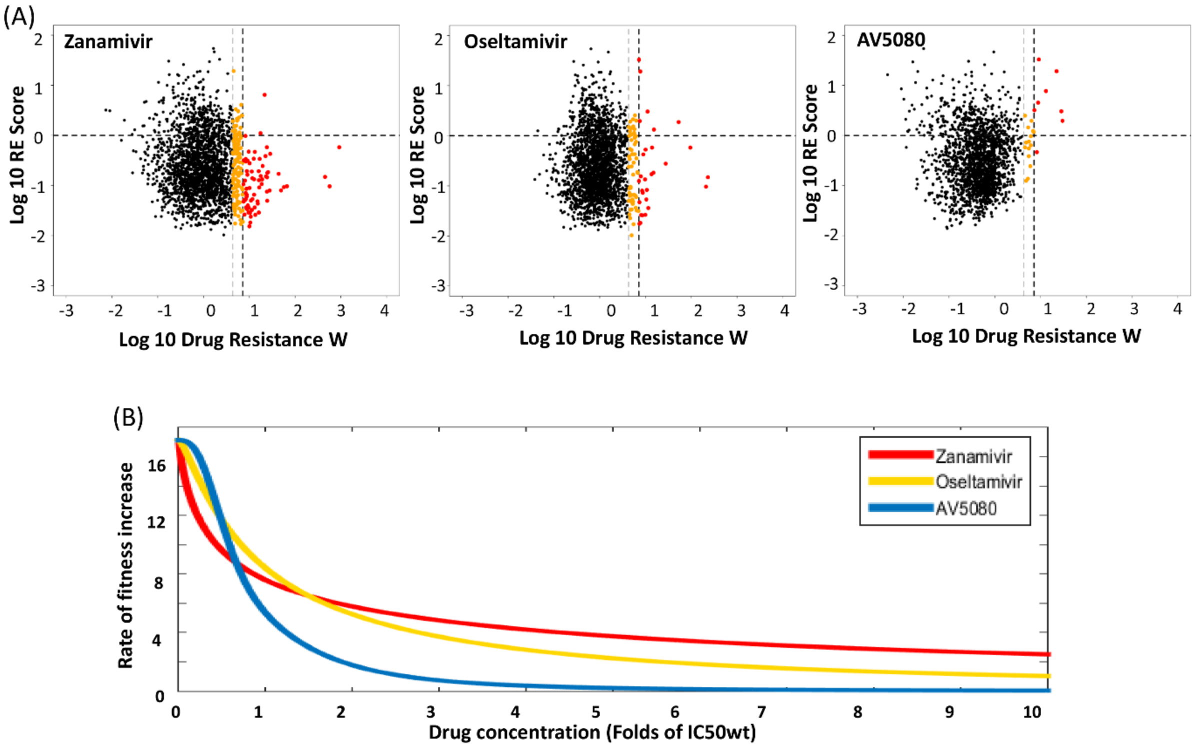


Figure 4-3. Evaluating the unique resistant barriers of drugs by combining drug resistance score, fitness cost and genetic barrier

(A)Scatter plots are shown the drug resistance scores (W) and relative fitness scores (RF score) for each NA mutations for zanamivir, oseltamivir and AV5080. (B) Rate of fitness increase is calculated as the function of drug concentration as fold of IC50 for each drug.

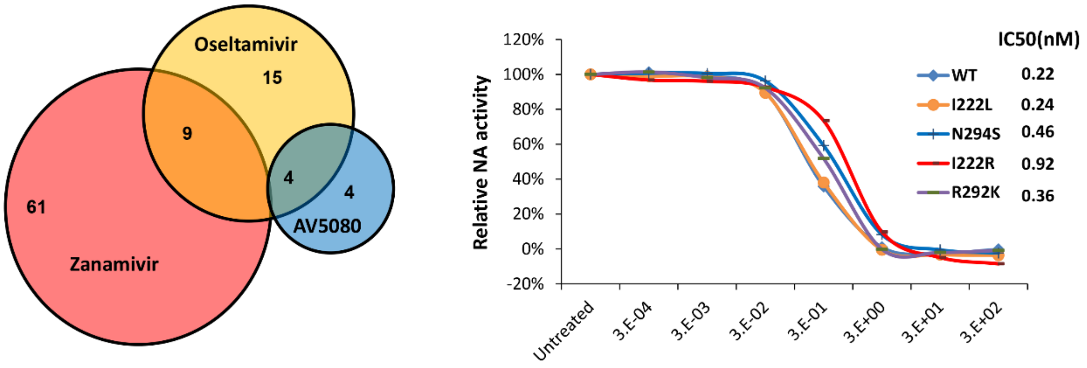


Figure 4-4. Cross resistance is rare between AV5080 and other NAIs

(A) Venn plot is shown the number of highly resistant mutations for each NAI that cross-resistance to other drugs. (B) Drug sensitivity of 4 zanamivir or oseltamivir resistant mutations against AV5080 were examined by NA activity assay. The IC50 of AV5080 for each mutations were calculated.

4.6 Supplementary Materials

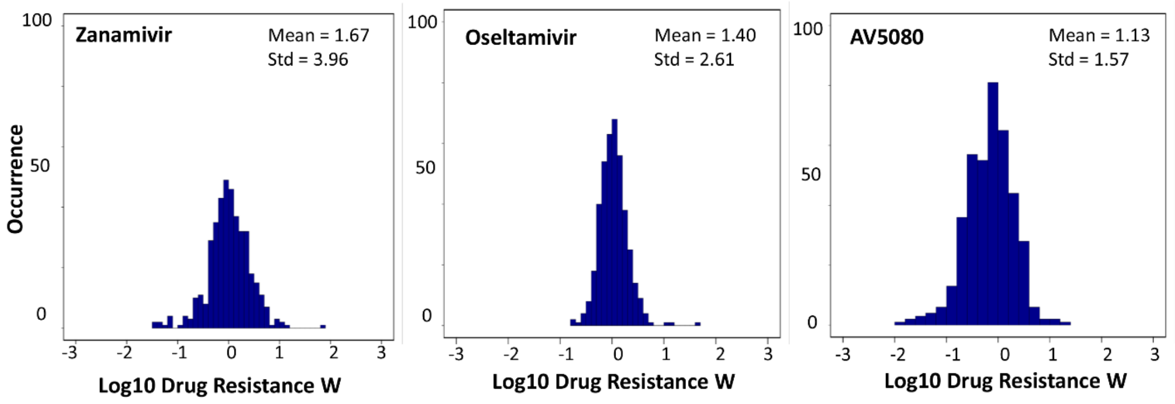


Figure S4-1. Distribution of drug resistance scores of silent mutations

Histograms are shown the distribution of drug resistant score of silent mutations for zanamivir (left), oseltamivir (middle) and AV5080 (right). Silent mutations were expected to have little, if not no effects on drug resistance, thus were used here as bench mark. Three drugs showed similar distribution centered around 0.

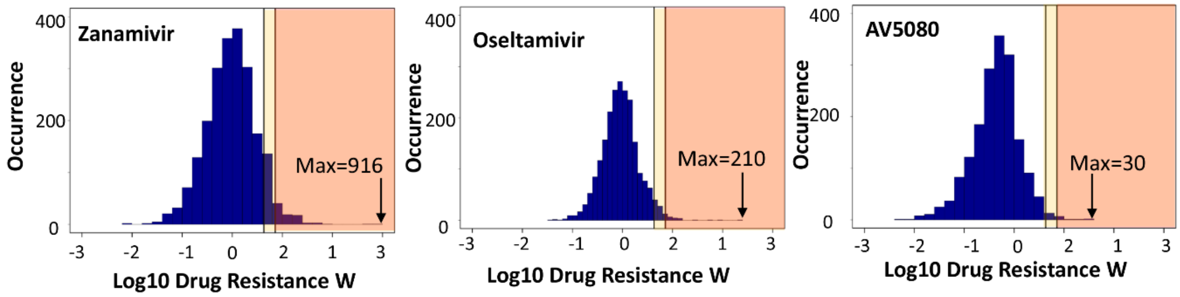


Figure S4-2. Distribution of drug resistant score (W) of NA mutations

Distribution of drug resistant score (W) of each NA mutations were shown for zanamivir, oseltamivir and AV5080. The fractions of resistant mutations were shaded in orange, and the fraction of highly resistant mutations were shaded in red. The maximum drug resistant score for each NA inhibitors were labeled.

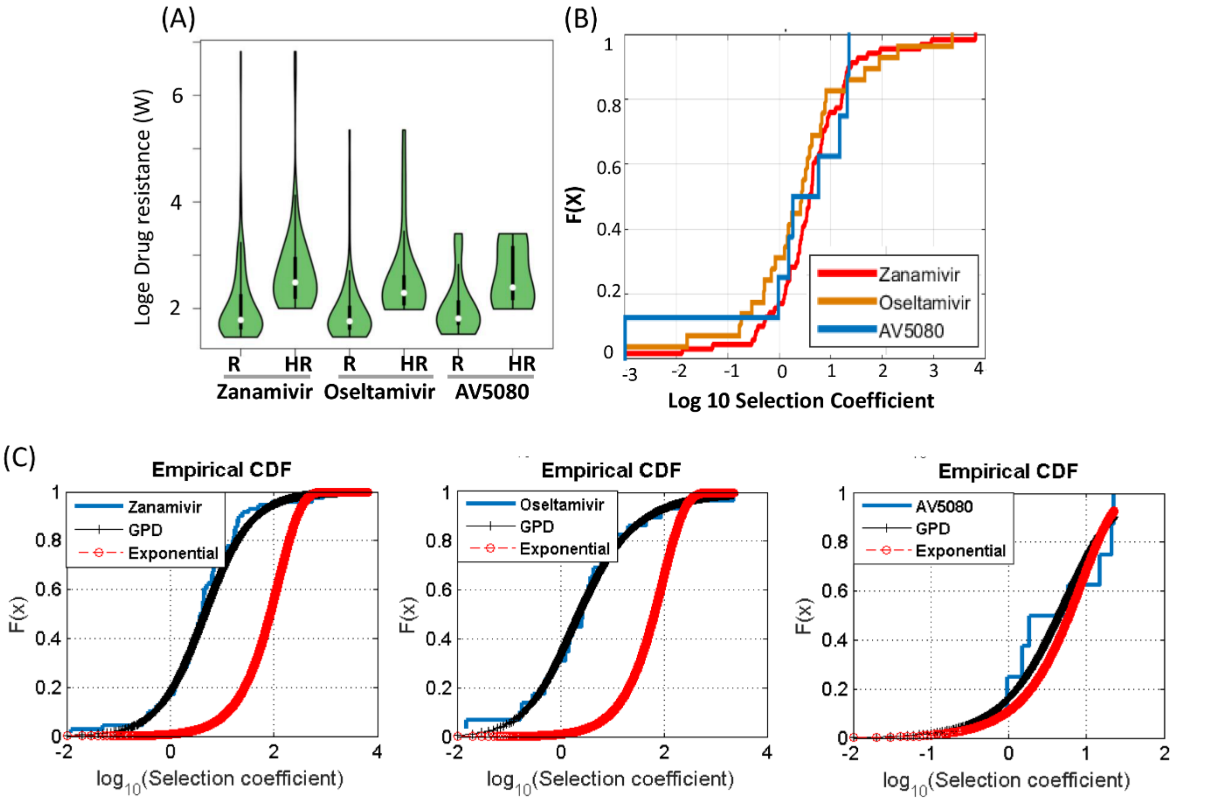


Figure S4-3. Identification of drug resistant mutations

(A) Distribution of drug resistant scores are shown as violin plot for resistant (R) and highly resistant mutations (HR) of three NAIs. (B) Cumulative distribution function (CDF) of resistant mutations were plotted for each drug. Selection coefficient of each resistant mutant is rescale from the minimum W within one drug (S resistnat mut = W resistant mut- W min resistant mut). (C) The distribution of W is fitted into Generalized pareto distribution (GPD), and the fitted curve for each drug is shown.

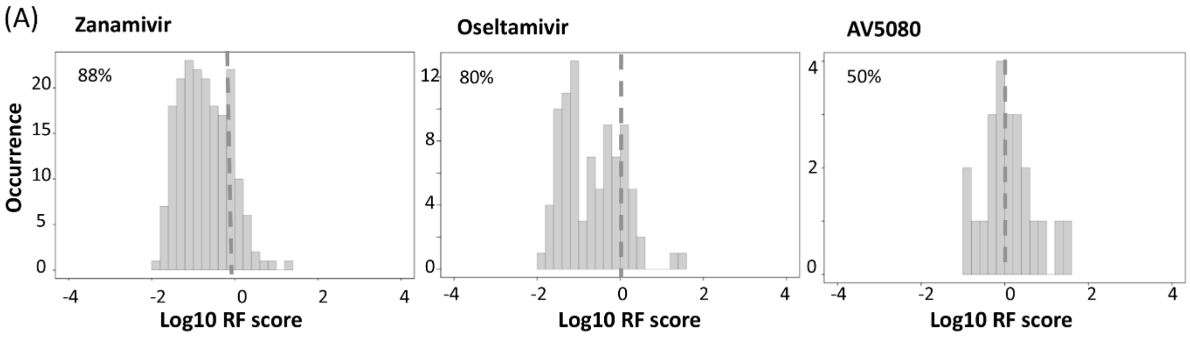


Figure S4-4. Distribution of relative fitness scores of drug resistant mutations.

Histograms are shown the distribution of relative fitness score of drug resistant mutations for zanamivir (left), oseltamivir (middle) and AV5080 (right). Drug resistant mutations for zanamivir and oseltamivir showed higher fitness cost with AV5080. 88% of resistant mutations of zanamivir and 81% of resistant mutations of oseltamivir had fitness costs comparable to wild type. However, only 50% resistant mutations of AV5080 showed a decrease in viral growth capacity.

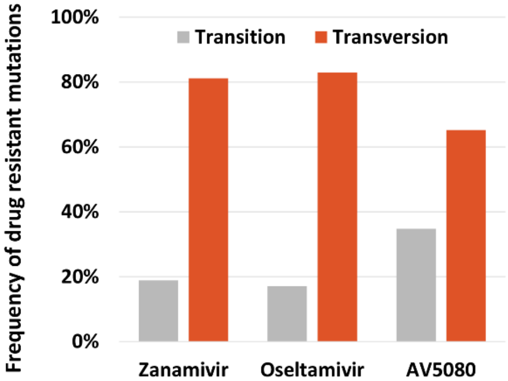


Figure S4-5. Frequency of resistant mutations caused by transition and transversion of each drug.

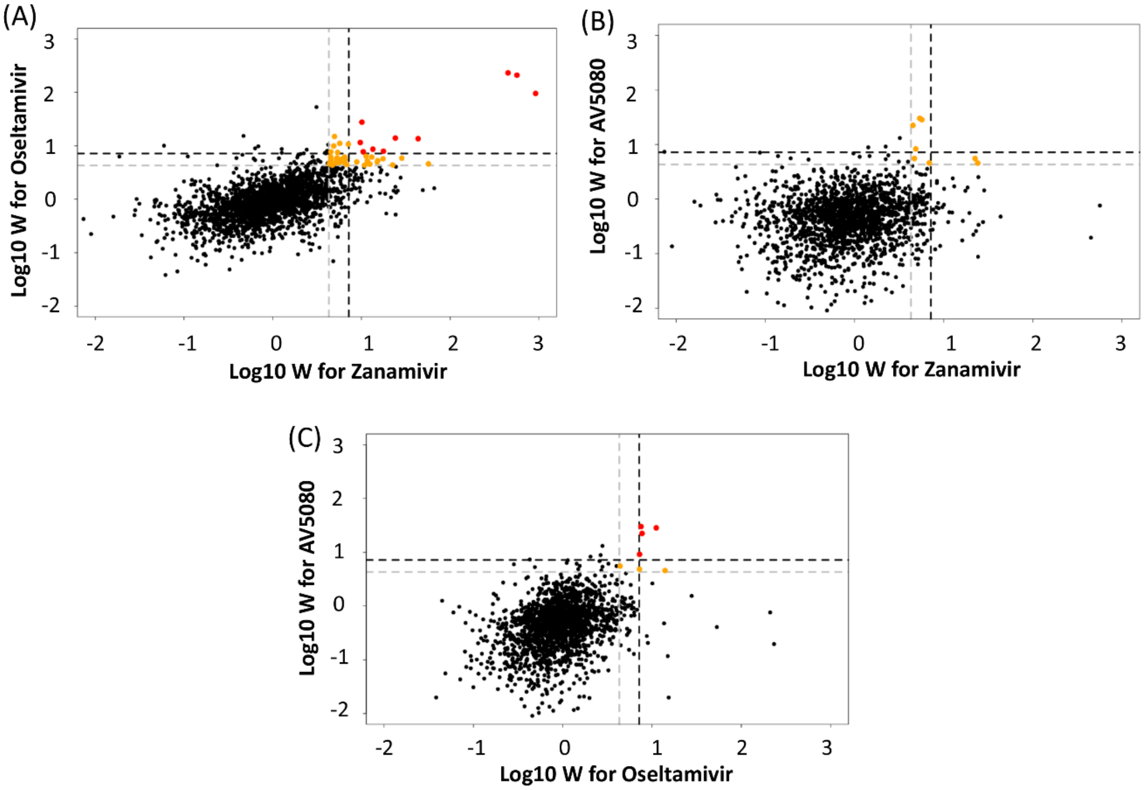


Figure S4-6. Scatter plots were shown the drug resistance score of each NA mutations, between each pair of NAIs.

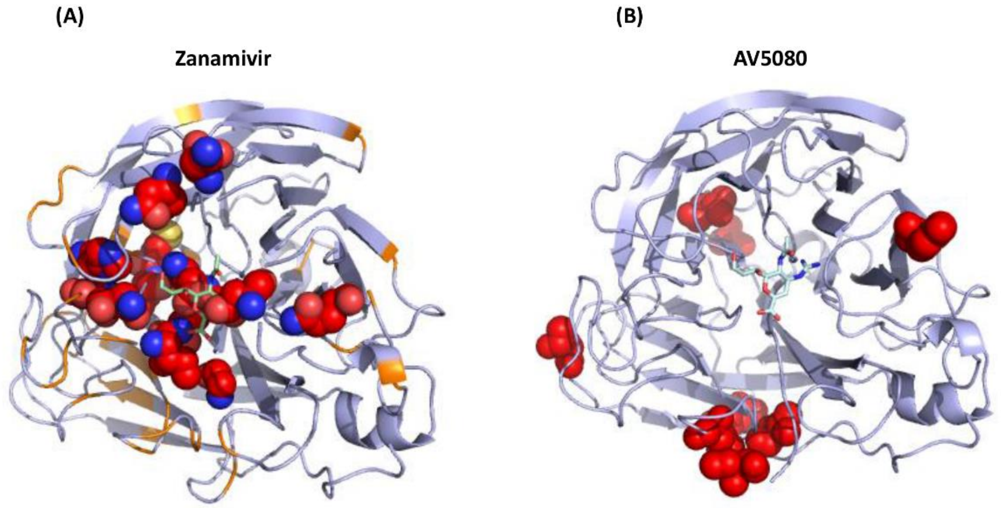


Figure S4-7. Structural mapping of drug resistant mutations.

(A) Locations of zanamivir resistant mutations were shown with NA structures (pdb: 7b3e). Residues with resistant mutations were labeled as orange and the ones located around catalytic sites were marked as red. Zanamivir molecule were co-crystal with NA and shown in green. (B): Resistant mutations of AV5080 were shown in red. They were located outside of the catalytic sites.

4.7 Bibliography

1. Dunn CJ, Goa KL. 2001. Oseltamivir: A Review of its Use in Influenza. Drugs 58:761–784.

2. Milena M McLaughlin PharmD MSc EWSB& MGIMM. 2015. Peramivir: an intravenous neuraminidase inhibitor. Expert Opin Pharmacother 16:1889–1900.

3. Leang S, Kwok S, Sullivan SG, Maurer-stroh S, Kelso A, Ian G. 2013. Peramivir and laninamivir susceptibility of circulating influenza A and B viruses 135–139.

4. Donck L Ver, Cox E, Jonge HR De, Schuurkes JAJ. 2016. Long-acting neuraminidase inhibitor laninamivir octanoate as post-exposure prophylaxis for influenza. Clin Infect Dis 1–26.

5. Kashiwagi S, Watanabe A, Ikematsu H, Awamura S, Okamoto T, Uemori M, Ishida K. 2013. Laninamivir octanoate for post-exposure prophylaxis of influenza in household contacts: a randomized double blind placebo controlled trial. J Infect Chemother 19:740–9.

6. Thorlund K, Awad T, Boivin G, Thabane L. 2011. Systematic review of influenza resistance to the neuraminidase inhibitors. BMC Infect Dis 11:134.

7. Samson M, Pizzorno A, Abed Y, Boivin G. 2013. Influenza virus resistance to neuraminidase inhibitors. Antiviral Res 98:174–185.

8. Mckimm-Breschkin JL. 2013. Influenza neuraminidase inhibitors: Antiviral action and mechanisms of resistance. Influenza Other Respi Viruses 7:25–36.

9. Abed Y, Baz M, Boivin G. 2006. Impact of neuraminidase mutations conferring influenza resistance to neuraminidase inhibitors in the N1 and N2 genetic backgrounds. Antivir Ther 11:971–976.

10. Aoki FY, Boivin G, Roberts N. 2007. Influenza virus susceptibility and resistance to oseltamivir. Antivir Ther 12:603–616.

11. Baz M, Abed Y, McDonald J, Boivin G. 2006. Characterization of multidrug-resistant influenza A/H3N2 viruses shed during 1 year by an immunocompromised child. Clin Infect Dis 43:1555–1561.

12. Chidlow GR, Harnett GB, Williams SH, Tempone SS, Speers DJ, Hurt AC, Deng YM, Smith DW. 2010. The detection of oseltamivir-resistant pandemic influenza A/H1N1 2009 viruses using a real-time RT-PCR assay. J Virol Methods 169:47–51.

13. Duan S, Govorkova EA, Bahl J, Zaraket H, Baranovich T, Seiler P, Prevost K, Webster RG, Webby RJ. 2014. Epistatic interactions between neuraminidase mutations facilitated the emergence of the oseltamivir-resistant H1N1 influenza viruses. Nat Commun 5:5029.

14. Duan S, Boltz DA, Seiler P, Li J, Bragstad K, Nielsen LP, Webby RJ, Webster RG, Govorkova EA. 2010. Oseltamivir-resistant pandemic H1N1/2009 influenza virus possesses lower transmissibility and fitness in ferrets. PLoS Pathog 6:1–10.

15. Ferraris O, Lina B. 2008. Mutations of neuraminidase implicated in neuraminidase inhibitors resistance. J Clin Virol 41:13–19.

16. Hurt AC, Holien JK, Parker M, Kelso A, Barr IG. 2009. Zanamivir-resistant influenza viruses with a novel neuraminidase mutation. J Virol 83:10366–73.

17. Gubareva L V, Kaiser L, Matrosovich MN, Soo-Hoo Y, Hayden FG. 2001. Selection of influenza virus mutants in experimentally infected volunteers treated with oseltamivir. J Infect Dis 183:523–531.

18. Collins PJ, Haire LF, Lin YP, Liu J, Russell RJ, Walker P a, Skehel JJ, Martin SR, Hay AJ, Gamblin SJ. 2008. Crystal structures of oseltamivir-resistant influenza virus neuraminidase mutants. Nature 453:1258–1261.

19. Gubareva L V. 2004. Molecular mechanisms of influenza virus resistance to neuraminidase inhibitors. Virus Res 103:199–203.

20. Ivachtchenko A V., Ivanenkov YA, Mitkin OD, Yamanushkin PM, Bichko V V., Shevkun NA, Karapetian RN, Leneva IA, Borisova O V., Veselov MS. 2014. Novel oral anti-influenza drug candidate AV5080. J Antimicrob Chemother 69:1892–1902.

21. Fu L, Bi Y, Wu Y, Zhang S, Qi J, Li Y, Lu X, Zhang Z, Lv X, Yan J, Gao GF, Li X. 2016. Structure-Based Tetravalent Zanamivir with Potent Inhibitory Activity against Drug-Resistant Influenza Viruses. J Med Chem 59:6303–6312.

22. Jie Y, Shuwen L, Langying D, Shibo J. 2016. A new role of neuraminidase (NA) in the influenza virus life cycle: implication for developing NA inhibitors with novel mechanism of action. Rev Med Virol 26:242–250.

23. Baz M, Abed Y, Simon P, Hamelin ME, Boivin G. 2010. Effect of the neuraminidase mutation H274Y conferring resistance to oseltamivir on the replicative capacity and virulence of old and recent human influenza A(H1N1) viruses. J Infect Dis 201:740–745.

24. Burnham AJ, Baranovich T, Marathe BM, Armstrong J, Webster RG, Govorkova EA. 2014. Fitness costs for influenza b viruses carrying neuraminidase inhibitor-resistant substitutions: Underscoring the importance of e119a and h274y. Antimicrob Agents Chemother 58:2718–2730.

25. Govorkova EA. 2013. Consequences of resistance: In vitro fitness, in vivo infectivity, and transmissibility of oseltamivir-resistant influenza A viruses. Influenza Other Respi Viruses 7:50–57.

26. Seibert CW, Rahmat S, Krammer F, Palese P, Bouvier NM. 2012. Efficient transmission of pandemic H1N1 influenza viruses with high-level oseltamivir resistance. J Virol 86:5386–9.

27. Hamelin ME, Baz M, Abed Y, Couture C, Joubert P, Beaulieu ?? Dith, Bellerose N, Plante M, Mallett C, Schumer G, Kobinger GP, Boivin G. 2010. Oseltamivir-resistant pandemic A/H1N1 virus is as virulent as its wild-type counterpart in mice and ferrets. PLoS Pathog 6:1–10.

28. Hurt AC, Nor'e SS, McCaw JM, Fryer HR, Mosse J, McLean AR, Barr IG. 2010. Assessing the viral fitness of oseltamivir-resistant influenza viruses in ferrets, using a competitive-mixtures model. J Virol 84:9427–9438.

29. Rameix-Welti MA, Enouf V, Cuvelier F, Jeannin P, Van Der Werf S. 2008. Enzymatic properties of the neuraminidase of seasonal H1N1 influenza viruses provide insights for the emergence of natural resistance to oseltamivir. PLoS Pathog 4:1–5.

30. Zurcher T, Yates PJ, Daly J, Sahasrabudhe A, Walters M, Dash L, Tisdale M, McKimm-Breschkin JL. 2006. Mutations conferring zanamivir resistance in human influenza virus N2 neuraminidases compromise virus fitness and are not stably maintained in vitro. J Antimicrob Chemother 58:723–732.

31. Nguyen HT, Fry AM, Gubareva L V. 2012. Neuraminidase inhibitor resistance in influenza viruses and laboratory testing methods. Antivir Ther 17:159–173.

32. Molla A, Kati W, Carrick R, Steffy K, Shi Y, Montgomery D, Gusick N, Stoll VS, Stewart KD, Ng TI, Maring C, Kempf DJ, Kohlbrenner W, Laboratories A, Rd AP, Park A. 2002. In Vitro Selection and Characterization of Influenza A (A/N9) Virus Variants Resistant to a Novel Neuraminidase Inhibitor , A-315675. J Virol 76:5380–5386.

33. Hurt AC, Holien JK, Barr IG. 2009. In vitro generation of neuraminidase inhibitor resistance in A(H5N1) influenza viruses. Antimicrob Agents Chemother 53:4433–4440.

34. Mishin VP, Hayden FG, Gubareva L V, Hemother ANAGC. 2005. Susceptibilities of Antiviral-Resistant Influenza Viruses to Novel Neuraminidase Inhibitors. Antimicrob Agents Chemother 49:4515–4520.

35. MacLean RC, Buckling A. 2009. The distribution of fitness effects of beneficial mutations in Pseudomonas aeruginosa. PLoS Genet 5:e1000406.

36. Gutierrez a, Laureti L, Crussard S, Abida H, Rodríguez-Rojas a, Blázquez J, Baharoglu Z, Mazel D, Darfeuille F, Vogel J, Matic I. 2013. β-Lactam antibiotics promote bacterial mutagenesis via an RpoS-mediated reduction in replication fidelity. Nat Commun 4:1610.

37. Bataillon T, Zhang T, Kassen R. 2011. Cost of adaptation and fitness effects of beneficial mutations in Pseudomonas fluorescens. Genetics 189:939–949.

38. Sanjuán R, Moya A, Elena S. 2004. The distribution of fitness effects caused by single-nucleotide substitutions in an RNA virus. Proc Natl ….

39. Carrasco P, Iglesia F de la, Elena S. 2007. Distribution of fitness and virulence effects caused by single-nucleotide substitutions in Tobacco etch virus. J Virol 81:12979–12984.

40. Cowperthwaite MC, Bull JJ, Meyers LA. 2005. Distributions of beneficial fitness effects in RNA. Genetics 170:1449–57.

41. Peris JB, Davis P, Cuevas JM, Nebot MR, Sanjuán R. 2010. Distribution of fitness effects caused by single-nucleotide substitutions in bacteriophage f1. Genetics 185:603–9.

42. Imhof M, Schlötterer C. 2001. Fitness effects of advantageous mutations in evolving Escherichia coli populations. Proc Natl Acad Sci U S A 98:1113–1117.

43. Kassen R, Bataillon T. 2006. Distribution of fitness effects among beneficial mutations before selection in experimental populations of bacteria. Nat Genet 38:484–8.

44. Rokyta D, Joyce P, Caudle S, Wichman H. 2005. An empirical test of the mutational landscape model of adaptation using a single-stranded DNA virus. Nat Genet 37:441–444.

45. Burch C, Guyader S, Samarov D, Shen H. 2007. Experimental estimate of the abundance and effects of nearly neutral mutations in the RNA virus ϕ6. Genetics 476:467–476.

46. Eyre-Walker A, Keightley PD. 2007. The distribution of fitness effects of new mutations. Nat Rev Genet 8:610–8.

47. MacLean RC, Hall AR, Perron GG, Buckling A. 2010. The population genetics of antibiotic resistance: integrating molecular mechanisms and treatment contexts. Nat Rev Genet 11:405–414.

48. 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 b-Lactamase. Cell 160:882–892.

49. Li X, Zhang Z, Yu A, Ho SYW, Carr MJ, Zheng W, Zhang Y, Zhu C, Lei F, Shi W. 2014. Global and local persistence of influenza A(H5N1) Virus. Emerg Infect Dis 20:1287–1295.

50. Robinson M, Tian Y, Delaney WE, Greenstein AE. 2011. Preexisting drug-resistance mutations reveal unique barriers to resistance for distinct antivirals. Proc Natl Acad Sci USA 108:10290–10295.

51. Taft AS, Ozawa M, Fitch A, Depasse J V, Halfmann PJ, Hill-Batorski L, Hatta M, Friedrich TC, Lopes TJ, Maher EA, Ghedin E, Macken CA, Neumann G, Kawaoka Y. 2015. Identification of mammalian-adapting mutations in the polymerase complex of an avian H5N1 influenza virus. Nat Commun 6:7491.

52. Weltman JK, Skowron G, Loriot GB. 2007. Influenza A H5N1 hemagglutinin cleavable signal sequence substitutions. Biochem Biophys Res Commun 352:177–180.

53. 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 influenza A virus NS segment. J Virol.

54. 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.

55. Du Y, Wu NC, Jiang L, Zhang T, Gong D, Shu S, Wu T. 2016. Annotating Protein Functional Residues by Coupling High- Throughput Fitness Profile and Homologous-Structure Analysis. MBio 7:1–13.

56. Jiang L, Liu P, Bank C, Renzette N, Prachanronarong K, Yilmaz LS, Caffrey DR, Zeldovich KB, Schiffer CA, Kowalik TF, Jensen JD, Finberg RW, Wang JP, Bolon DNA. 2016. A Balance between Inhibitor Binding and Substrate Processing Confers Influenza Drug Resistance. J Mol Biol 428:538–553.

57. Qi H, Wu NC, Du Y, Wu TT, Sun R. 2015. High-resolution genetic profile of viral genomes: Why it matters. Curr Opin Virol 14:6270.

58. Wu NC, Young AP, Dandekar S, Wijersuriya H, Al-Mawsawi LQ, Wu T-T, Sun R. 2013. Systematic identification of H274Y compensatory mutations in influenza A virus neuraminidase by high-throughput screening. J Virol 87:1193–9.

59. Hoffmann E, Neumann G. 2000. A DNA transfection system for generation of influenza A virus from eight plasmids. Proc Natl Acad Sci 97:6108–6113.

60. Bloom JD, Gong LI, Baltimore D. 2010. Permissive secondary mutations enable the evolution of influenza oseltamivir resistance. Science (80-) 328:1272–1275.

61. Samji T. 2009. Influenza A: Understanding the viral life cycle. Yale J Biol Med 82:153–159.

CHAPTER 5

CTL ESCAPE OF HIV-1 GAG EPITOPES IS DETERMINED BY MUTATIONAL EFFECTS ON REPLICATIVE FITNESS AND MHC-I BINDING

5.1 Abstract

Certain “protective” HLA alleles, such as B*57, B*27, are associated with superior control of HIV replication mediated by the CD8+ Cytotoxic T-Lymphocyte (CTL) response. However, the mechanisms of this superior protection is still not fully understood. Here we used a combination of quantitative high-throughput fitness profiling, in silico binding prediction, and analysis of patient samples to systematically compare the differences between protective and non-protective HLA alleles in HIV Gag region. We observed that CTL escape of HIV-1 Gag epitopes is determined by both mutational effects on viral replication and MHC-I binding affinity. Mutations in epitopes presented by protective HLA alleles come with significantly higher fitness cost and less reduction in binding to MHC-I. A linear model combining both properties strongly correlated with HLA allele’s protectiveness ranking. Finally, mutations observed in vivo in HIV-1 controllers and progressors followed a similar pattern, with a higher fitness cost and smaller effect on HLA-binding observed among epitope mutants in controllers. Overall, our results suggest that protective effect of certain HLA alleles derives from their ability to target conserved epitopes where escape mutations come at higher fitness costs and less abrogation of HLA binding.

5.2 Introduction

CD8+ Cytotoxic T-lymphocytes (CTLs) is one of the most important immune pressures that limit HIV replication in vivo (1–3). The importance of CTL in HIV control has been demonstrated in laboratory experiments and clinical observations (1, 4–6). Isolation of CTLs from patients shows robust killing of HIV in vitro (7). In macaque models, depletion of CTLs resulting in consistent relapse of viraemia (8). Moreover, HIV escape mutations in CTL epitopes were observed through individual and population level analysis (9–11). CTLs eradicate HIV infected cells through the recognition of short, viral-derived peptides that are presented on the surface of infected cells via major histocompatibility complex (MHC) class I molecules (encoded by Human Leukocyte Antigen (HLA) genes in human). The genotype of HLA alleles determines the presentation of HIV epitopes on MHC-I and the activation of CD8 T cells (12, 13), thus are the major contributor to CTL protection. Although CTL plays essential role in mediating HIV control, it failed in most of patients with disease progression and viral rebound. HIV can escape from CTL killing in multiple ways, including interference with intracellular epitope processing, reduced binding with MHC-I and undermining the recognition of CTL (12, 14–16). It was shown that reduced binding between viral derived peptides and MHC-I complex via mutations at binding sites is the major reason of CTL escape (14), which take account of more than 30% of escape mutations.

Protective HLA alleles, enriched in long term non-progressors (LTNPs), provides superior viral control (13, 17–21). Understanding the mechanisms related to LTNPs and protective HLA alleles may shed lights on potential functional cures of HIV (17, 22, 23). One of the appealing hypothesis is that the fitness cost of CTL escape mutations determines long-term control of HIV. For LTNPs with protective HLA alleles, the escape mutations either occur very late during viral infection together with compensatory mutations, or occur early without viral rebound (13, 22, 23). These observations suggest that there may be significant fitness cost associated with CTL escape mutations in long-term non-progessors (12, 24–26). The fitness cost of some CTL escape mutations has also been investigated (24, 25, 27–29). However, previous empirical studies usually require the identification and characterization of CTL escape mutation from patient samples, which are greatly limited by sample size. A systematic examination of fitness cost of epitope mutations is necessary to obtain a detailed comparison of the effect of mutations in epitopes restricted by protective and nonprotective HLA alleles (30–34). Moreover, as abrogation of HLA binding is the major reason of CTL escape (14), protective HLA alleles may correlate with less loss-of-binding epitope mutations. A comprehensive MHC binding analysis with all epitope mutations will greatly facilitate the understanding of this issue.

In this study, we used a combination of quantitative high-throughput fitness profiling, in silico binding prediction and analysis of patient samples to study CTL escape mutations in HIV-1 Gag. In particular, we examined how mutations in epitopes influence viral replicative fitness and MHC-I binding, and compared the differences between epitopes targeted by protective and non-protective HLA alleles. In comparison to mutations in epitopes targeted by non-protective HLA alleles, we observed that mutations in epitopes corresponding to protective HLA alleles have higher fitness cost and less reduction in binding with MHC-I. The same difference was observed for mutations in HIV-1 Gag sequences from clinical samples of long-term non-progressors and progressors.

5.3 Result

5.3.1 QUANTITATIVE FITNESS PROFILING OF MUTATIONS IN HIV-1 GAG

We have previously demonstrated the feasibility of high-throughput genetics to systematically evaluate the fitness effect of single mutations in HIV, HCV and Influenza Virus (30, 32, 35, 36). In this study, we generated a single nucleotide mutant library by dividing the entire HIV-1 (NL4-3) Gag region into 3 fragments (Methods). Diverse viral libraries were reconstituted from 293T cells followed by two successive passages in CEM cells. Relative Fitness scores (RF scores) were calculated for individual mutations as the ratio of relative frequency in round 2 passage library to that in the input viral library (**Figure 5-1**), as quantified by high-throughput sequencing. The mutant library covered 74% (3340/4509) all possible single nucleotide mutation and 27% single amino acid mutations (2788/10200) in Gag. The clear separation of the RF scores between silent mutations and missense mutations suggested significant selection in the passage process (**Figure S5-1**). We further quantified the average fitness effect of missense mutations as well as the fraction of lethal mutations for the 4 Gag proteins. Different Gag proteins showed diverse levels of mutation tolerability (**Figure 5-1**, **Figure S5-1**). Missense mutations of capsid protein showed the strongest deleterious effect on viral fitness, with around 20% lethal to viral replication. The fitness effects of missense mutations in individual Gag proteins also correlated well with natural sequence conservation, validating the accuracy of our fitness profiling (37) (**Figure 5-1**, **Figure S5-1**).

As Gag-specific CTLs are important for viral control (1, 24, 38), we examined if mutations in previously characterized CTL epitopes carry a higher fitness cost than other regions in general. CTL epitopes were defined according to the 2013 update of best-characterized epitopes from the Los Alamos Database (37). We obtained 1314 missense mutations within epitopes and 968 mutations in other regions from our fitness profile. The mutations in epitope region showed significantly lower relative fitness (two-tailed T test, $p<0.001$), suggesting that CTLs target mutation intolerant regions of the protein. Moreover, 70% epitopes (50/72) were located on capsid protein, which was the most conserved one on Gag.

5.3.2 Computational predictions of CTL escape mutations in HIV-1 Gag

Several algorithms have been developed to predict peptide binding to MHC-I (39–41), which greatly facilitate the identification and investigation of MHC escape mutations. NetMHC is the state-of-art predictor based on artificial neural network (40, 41), with accuracy of prediction achieving up to an 80% correlation with experimental data (42). Using NL4-3 as wild type sequence, we systematically calculated the binding affinity (Kd) of all single amino acid mutations for Gag epitopes with netMHC-4.0. A total of 58 epitopes were included in the analysis after filtering out the ones with Kd > 10,000 nM.

Peptides bind to MHC primarily through anchor residues, which are usually located at position 2 and C terminus (9th and 11th residue according to the length of peptide) of the peptide, and the MHC pocket. Since mutations of anchor residues are more likely to abrogate MHC-peptide binding, we first examined the effect of mutations at different positions within an epitope. As expected, the greatest binding drop was observed at anchor residues (2nd, 9th and 11th residue) (**Figure 5-2**), validating the accuracy of using netMHC4.0 to predict binding affinity.

The binding properties of different Gag epitopes were then examined. We plotted the distribution of Kd increase for each epitope and the histogram of average Kd increase across all the epitopes (**Figure 5-2**). Notably, different epitopes showed different property of binding drop upon mutations. HLA binding by some of the epitopes (for example those presented by: B5101, A2402) is robust, with few mutations leading to an increase in Kd, while others (for example those presented by: B0801, A2601) are more sensitive to mutations. By comparing the in silico binding affinity prediction with fitness profile of each epitope, we further evaluated the relationship between MHC binding with viral fitness. A weak but significant negative correlation were observed between the fitness epitope variants and its relative Kd increase (**Figure 5-2**, rho=-0.093, p=0.003). Consistent with previous work, this negative correlation may reflect intrinsic properties of escape and non-escape mutations (25).

5.3.3 GAG EPITOPES TARGETED BY PROTECTIVE HLA ALLELES ARE DIFFICULT TO ESCAPE CTL RECOGNITION

Systematic fitness profiling and binding prediction of mutations throughout all of gag enabled us to quantitatively examine the differences between protective or non-protective HLA alleles. Previous work has suggested that protective HLA types present highly conserved epitopes, which may contribute to overall control of viremia. We first examined several well characterized epitopes presented by HLA B*57 and B*27 (KK10, KF11, TW10) and observed that mutations in these epitopes lead to a greater loss in fitness than well characterized epitopes presented by A*02 (SL9) (**Figure 5-3**). To systematically determine whether pattern of presenting conserved epitopes holds true across protective HLA types, we ranked HLA alleles by the ratio of their prevalence in HIV controllers to their prevalence in progressors based on TIHIVC study (2). The top 5 HLAs were grouped and defined as protective HLA alleles, while bottom 5 were defined as non-protective HLA alleles. We observed that mutations on protective HLA alleles related Gag epitopes showed a significantly lower relative fitness score than the non-protective ones (One tail t-test, p= 0.005, **Figure 5-3**). These results suggested that protective HLA alleles may target more conserved regions that are less tolerant of mutations.

Next, we examined if protective and non-protective HLA presented epitopes harbor different mutational effect on binding affinity to corresponding MHC-I. Interestingly, we observed that epitopes presented by protective HLA types showed significantly less HLA binding affinity drop upon single amino acid mutations (One tail t-test, p= 0.05, **Figure 5-3**). These analysis suggested that the epitopes targeted by protective HLA alleles have at least two important properties. Firstly, it is more conserved with mutations leading to more fitness cost. Secondly, it is harder to escape from MHC binding, as single mutations resulted in less binding loss.

We further examined the ability of these two properties (fitness cost of and HLA-binding loss with mutations) to explain the “protectiveness” of certain HLAs. Each individual property did not correlate significantly with the ranking of HLAs (**Figure S5-2**), however, a linear combination of both properties significantly correlated with the ranking of 35 HLAs (p=0.01) (**Figure 5-3**). The coefficients of these two variants were 0.612 (fitness cost upon mutations) and 0.388 (binding loss with mutations), respectively. This linear model was trained 1000 times with 20 randomly picked HLAs and was then applied to 10 other HLAs for prediction. Average spearman correlation coefficient of predicted and actual HLA ranking was 0.342, and was significantly higher than zero (p=0.01).

5.3.4 CTL ESCAPE MUTATIONS IN HIV-1 GAG DURING INTRAPATIENT EVOLUTION

We extend our observation from systematic profiling to patient sample analysis. PBMC samples were collected from 4 paired rapid progressors and long term non-pregressors (LTNP) from Multicenter AIDS cohort study (MACS). Rapid progressors proceeded to AIDS related death at year four, while LTNPs maintained a stable CD4 cell count (less than 10% drop, **Figure S5-3**). All subjects were ART naïve, so that the major selective pressure for viral evolution was restricted to the host immune system. For each patient, we obtained PBMC samples at two time points; a first sample collected at time of enrollment in the MACS cohort (time point 1), and a second sample collected four years later (time point 2). DNA was extracted from 10,000,000 PBMC and the entire gag region (1500bp) was amplified and deep sequenced (**Figure S5-3**). Mutations were determined for each nucleotide according to the consensus sequence of corresponding patient. A phylogenetic tree analysis indicated clear distinction of consensus sequences from different patients (**Figure S5-3**). We also reconstructed full-length viral haplotypes by PredictHaplo. As an indicator of reconstruction accuracy, mutation frequency calculated from reconstructed haplotypes were highly correlated with raw data (**Figure S5-3**). For both groups of patients, we observed an accumulation of mutations in specific HLA related epitopes (example shown in **Figure 5-4**), and an increase of Shannon entropy of epitope regions at second time point (**Figure S5-3**).

We further examined the effect that observed CTL epitope mutations had on the viral fitness and HLA binding efficiency. Mutations were defined as variants that arose at time point 2 relative to the consensus sequence at time point 1. Relative fitness values of corresponding mutations were extracted from our profile data, and compared between LTNPs and progressors. Consistent with previous studies, mutations in CTL epitope regions observed in the controllers reduced HIV fitness significantly more than those in progressors (17, 22, 23). Mutation that occurred outside of epitope region were also compared as control, which showed no difference in fitness cost between groups (**Figure 5-4**). Additionally, mutations observed in the rapid progressors resulted in a substantial drop in MHC-I binding, while those observed in LTNPs had a much smaller effect on HLA-binding (**Figure 5-4**). Although the differences for fitness cost and

MHC binding are not significant between two groups of patients due to small sample size, we observed the same trend as our previous analysis.

5.4 Discussion

As the MHC-I restricted CTL response is the major immune pressure that restricts HIV replication in vivo, escape mutations also occurred frequently in epitope region based on host HLA (9, 13–15, 18, 43–45). The occurrence of escape mutations is also highly predictable. It is observed that protective HLA alleles often provoke mutations that have high fitness cost, or require the pre-existence of compensatory mutations (12, 25, 43, 46). Our results presented here also showed that protective HLA alleles target conserved regions where single nucleotide mutations would result in more severe fitness cost. Even minor fitness cost would be beneficial for patients than fitness-neutral mutations (12, 17, 22). Lower HIV replication capacity is significantly associated with reduced rate of CD4 decline, which is the marker of long term non-progressor.

Viral growth under CTL pressure depended on two properties: the viral replication capacity (fitness) and the ability to escape CTL selection. Although escape of CTL pressure can occur upstream, during or downstream of MHC-epitope binding (13, 14, 47), mutations that compromise MHC-peptide is shown to be the most common kind, which is estimated to take account of over 20-30% of observed CTL escape mutations (14). Here using netMHC4.0, we predicted the binding affinity of all epitope mutations with corresponding MHC. We observed a weak but significant negative correlation between the fitness cost and binding affinity of mutations, representing the possible tradeoff between these two properties for HIV escape. The similar tradeoff was observed previously with drug resistant mutations for their fitness and drug resistance (48).

Long term non-progressors provide a model for HIV functional cure (17, 22, 23). T cell mediated vaccine was under active development, in the hope of achieving robust T cell mediated response to control viral replication (49–53). Current strategy focused on using naturally conserved region in HIV genome as immunogens, on the assumption that escape mutations will have higher fitness cost. However, natural conservation have limited and biased sampling. Thus, a residue that is conserved in nature does not mean it cannot tolerant mutation. Furthermore, optimal MHC binding affinity is also an important aspect of immunogen design. With the systematic profile of fitness cost and binding affinity of all single nucleotide mutations, we may have a better idea which region is suitable as immunogen for vaccine development.

Here by combining high throughput fitness profile and in silicon MHC binding prediction, we observed that protective HLA alleless preferentially targeted conserved sites. Moreover, protective HLA alleless also recognize the regions that is harder to escape. The combination of fitness cost and MHC binding affinity can together explain the protectiveness of HLAs. Finally, we want to point out some limitations and future directions of this current study. Firstly, we used random mutagenesis to introduced mutation for the Gag region. Although we tried to limit the mutation rate, but there is still chance that multiple mutations occur at the same viral clone. This may create noise in assessing the fitness score of particular mutation. Secondly, we only focused on single nucleotide mutations and did not consider secondary mutations in the current study. A library with mutations covering multiple sites will be ideal to assess the effects of compensatory mutations on viral fitness and CTL escape.

5.5 Materials and methods

Construction of mutant libraries for HIV Pol polyprotein

HIV genomic DNA covering the whole gag genes in the replication competent proviral plasmid NL43 were divided into 3 fragments (**Figure 5-1**). Mutations were randomly introduced into each fragment by performing error-prone PCR using the Mutazyme II DNA polymerase (Stratagene, La Jolla, CA) and the fragment-specific primers. Mutated segment were then ligated back into proviral backbone. The ligation products with desired mutation rate were further electroporated into high-efficiency MegaX DH10B T1R electrocompetent cells (Invitrogen). About 50000 bacteria colonies were collected for each small libraries.

Transfection, viral titering and passage of HIV mutant libraries

Human embryonic kidney 293T cell line was used for transfection to rescue each viral mutant library Transfections were performed using Lipofectamine 2000 (Invitrogen, Life Technologies). For rescue of mutant libraries, 16 µg of library plasmid were applied to transfect 15 million of 293T cells with 50~70% confluence, then cells were rinsed with PBS at 12-14 hours post transfection, and maintained in fresh DMEM growth media supplemented with 10% FBS and 1X penicillin and streptomycin. Transfection was performed in duplicate in T75 flasks. The supernatant was harvested at 72 hours post transfection, filtered through a 0.45 µm disposable syringe filter (Olympus), and stored at -80oC in multiple small aliquots. The TCID50 of viral supernatants were measured using the Ghost 3-X4/R5 indicator cells (gift of Dr. Matthew Marsden) which derived from human osteosarcoma cells and stably transfected with the HIV LTR driving hGFP construct (24). To passage each viral mutant library, ~30 million of CEM T lymphocyte cell line was used for infection at low multiplicity of infection (MOI=0.05) supplemented with 2 µg/ml polybrene (Sigma). At 14~16 hours post infection, cells were centrifuged at 1000 rpm for 5min and washed with PBS followed by the addition of fresh RPMI 1640 growth medium. Extracellular viruses were harvested at ~5 days post-infection when syncytia formation can be observed in 60~80% of cells. Two passages were performed for each library.

Library preparation for deep sequencing

Viral RNAs were isolated from the viral supernatant using QIAamp Viral RNA Mini Kit (Qiagen Sciences). Virion RNAs were treated with DNaseI-amp (Life Technologies). Viral RNAs were reverse transcribed to cDNA using Superscript III reverse transcriptase (Life Technologies).

The plasmid mutant libraries or cDNA from the viral mutant libraries (transfection or infection) were amplified using KOD hot start DNA polymerase generating ~120bp amplicons. Illumina Hiseq 2000 PE100 were used for sequencing.

Sequencing data analysis

Burrows-Wheeler Aligner (BWA) (19) were used to map sequencing reads to reference sequences. Pair end reads were used for error correction. Relative frequency of each mutant were calculated for each condition, and the relative fitness score were calculated as the changing of relative frequency in the passaged library compare with in transfected library. To further improve data quality, mutations with frequency < 0.01% of transfection library were filtered, and G-A hyper mutation were filtered (3).

Natural sequence conservation analysis

6097 HIV-1 subtype B complete Gag sequences were downloaded from Los Alamos database, no filters were applied for sampling time, country, or patient information. All patient samples were aligned. Diversity of every position is defined as Shannon Entropy.

netMHC4.0 for binding prediction

The change of binding affinity (i.e. Kd increase) was calculated by netMHC4.0. Epitope sequences and MHC subtype were paired according to the best-characterized HIV-1 CTL epitopes from Los Alamos Database (54).For general analysis, the binding of all possible single mutations on epitope regions was simulated. The binding affinity drop is the ratio of Kd between mutated epitope and the epitope sequence in Los Alamos Database. For patient sequence analysis, the epitopes were also extracted from the same database. And the binding affinity change was calculated by the ratio of Kd of observed epitope sequences.

Protective and non-protective HLA alleles subtypes were defined according to TIHIVC study. The ratio of prevalence of HLA in HIV-1 non-progressors and progressors were ranked. And top 5 non-progressor-prevalent HLAs were defined as protective HLA alleless, and vice versa.

Linear regression

The fitness of an HLA was defined as the mean log10 fitness of all screened missense mutations on its targeting epitopes. Then all HLA was ordered by this value and assigned as HLA fitness ranking. The binding affinity drop (I.e. Kd increase) of all possible amino acid substitution on the HLA's targeting epitopes was logged and averaged. Then al HLA was ordered by this value and assigned as HLA binding loss rank. These two HLA parameters were used as input for a linear model to predict the protectiveness rank of HLA. The protectiveness was defined as the ratio of HLA subtype prevalence in non-progressor versus progressors. For cross validation, 20 HLAs are random picked as training dataset to calculate the coefficients for HLA fitness rank and HLA binding loss rank in linear model, the other 10 HLAs are used as testing datasets. The correlation coefficient of predicted HLA protectiveness rank and the reported HLA protectiveness rank was calculated by spearman correlation test. 1000 cross validation tests were carried out. The null hypothesis that correlation coefficients are zero was rejected by student 's t-test.

Rank= 0.612*fitness rank + 0.388* binding loss rank.

Patient sample sequencing and analysis

PBMC samples from 4 paired rapid progressors and non-pregressors were kindly provided by Multicenter AIDs cohort study (MACS). All of these patients are treatment naïve. For each patient, we obtained two PBMC samples from 4 years apart, where the first sample was collected at the earliest time point in MACS cohort for corresponding patient. Rapid progressor proceed to AIDS phase and died at year four, while non-progressor remain stable CD4 cells count. DNA were extracted from 10,000,000 PBMC from each patient. Entire gag region (1500bp) were amplified by nested PCR. The gel purified PCR product were then random fragmented by sonication to 200-700bp. The fragmented library were prepared for high throughput sequencing with Illumina Hiseq 2000.

Patient sequences were mapped onto NL4-3 strain of HIV-1 genome. The haplotypes of Gag gene were constructed by PredictHaplo1.0 (55). Consensus sequences were called according to the frequencies of haplotypes. Mutations were identified if a haplotype is different from the consensus sequence of the population. The tree of haplotypes is contructed using Phylip. Escape mutations were called from haplotypes that are different from time point 1 consensus sequence. If a mutation caused more than 2 fold of Kd increase, then it was defined as escape mutation. All the custom codes are deposited at https://github.com/Tian-hao/HIV-clinical/.

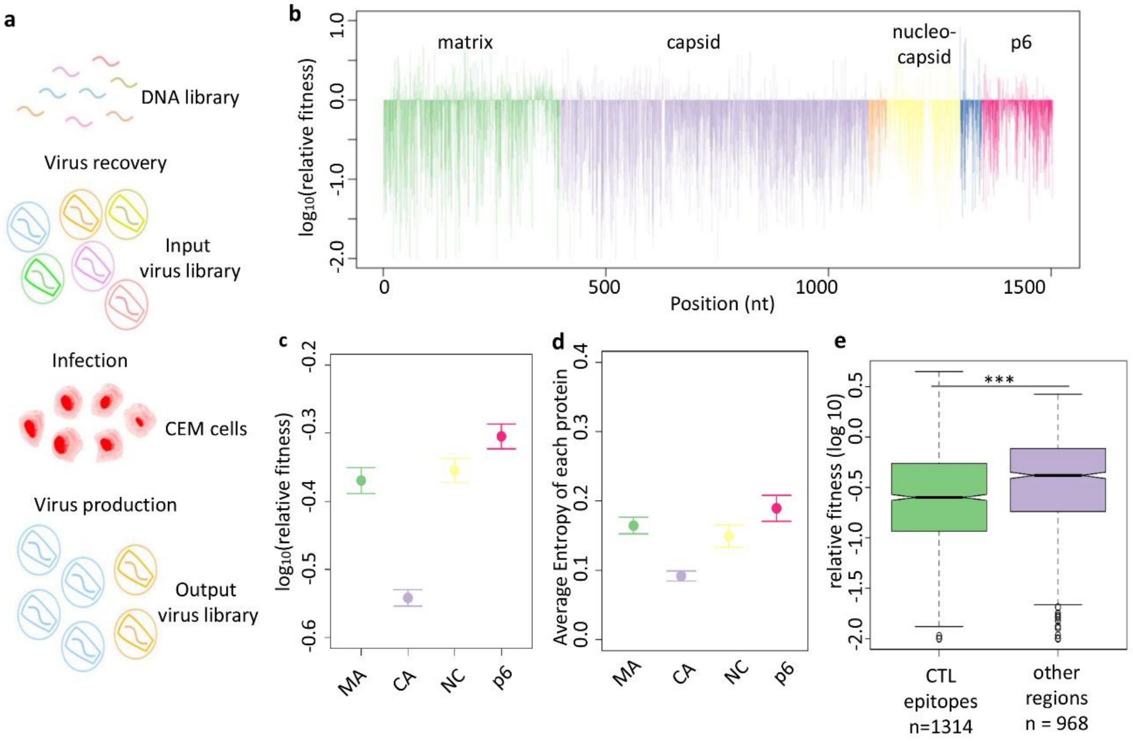


Figure 5-1. Quantitative high-throughput fitness profiling of HIV-I Gag region

a) Workflow of fitness profiling. We introduced single nucleotide substitutions to NL4-3 plasmid Gag region by error prone PCR. Mutated plasmids were transfected into 15 million HEK-293T cells to recover virus library. We infected 30 million CEM cells with virus library at an MOI of 0.05. We then harvested virus in supernatant at 7 days post infection. The fitness of a certain mutation was defined as relative frequency in the output virus pool versus input virus pool. b) Fitness profile of Gag. The relative fitness of each mutation was calculated as the ratio of fitness to wild type. Different color represented different subunit of Gag as labeled. Orange and blue represents spacer peptide 1 and spacer peptide 2 respectively. c) Average relative fitness of missense mutations. Error bar represents standard error. d) Average entropy of each protein were calculated based on naturally occurred variants in HIV database e) Relative fitness of mutations within best characterized CTL epitope regions. CTL epitopes were defined according to the 2013 update of best-characterized epitopes from the Los Alamos Database. 1314 mutations within CTL epitopes and 968 mutations outside of CTL epitopes are calculated. Student t-test has a p value smaller than 0.001. MA, matrix; CA, capsid; SP1, spacer peptide 1; NC, nucleocapsid; SP2, spacer peptide 2.

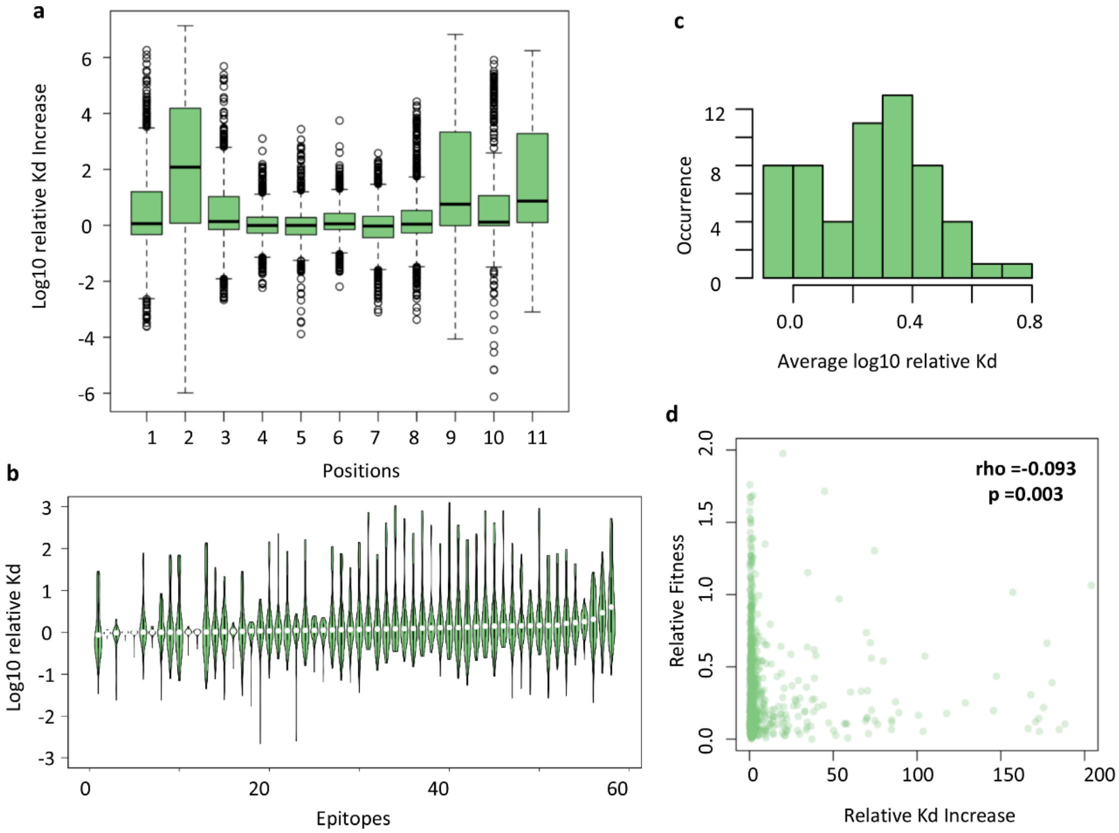


Figure 5-2. Systematic evaluate mutant binding efficiency through netMHC-4.0

a) Affinity drop of single mutations in different epitope positions. The binding affinity of mutated epitopes was predicted by netMHC-4.0. All best-characterized epitopes from Los Alamos database were calculated by netMHC-4.0. The epitopes with binding Kd more than 10,000 nM were filtered out. In the violin plot, every violin stands for an epitope, which contains 171~209 mutations, depending on the length of the epitope. b) Affinity drop of single mutations in epitope regions. c) Distribution of average affinity drop of single mutations in epitope regions. The geometric mean of affinity drop was calculated for every epitope. d) The correlation between relative fitness and binding affinity drop for all missense mutations. The relative fitness of amino acid is defined as the ratio of the mutant to NL4-3 wild type in figure 1 settings. Different codons were averaged for a mean fitness. The binding affinity drop caused by a mutation was defined as the binding affinity of mutated epitopes to NL4-3 epitopes. All possible epitopes containing this position and corresponding best-characterized MHCs were averaged for a mean binding affinity drop. 989 mutations were calculated for their relative fitness and binding affinity drop. Spearman correlation test returned a p value of 0.003 and a rho value of -0.093.

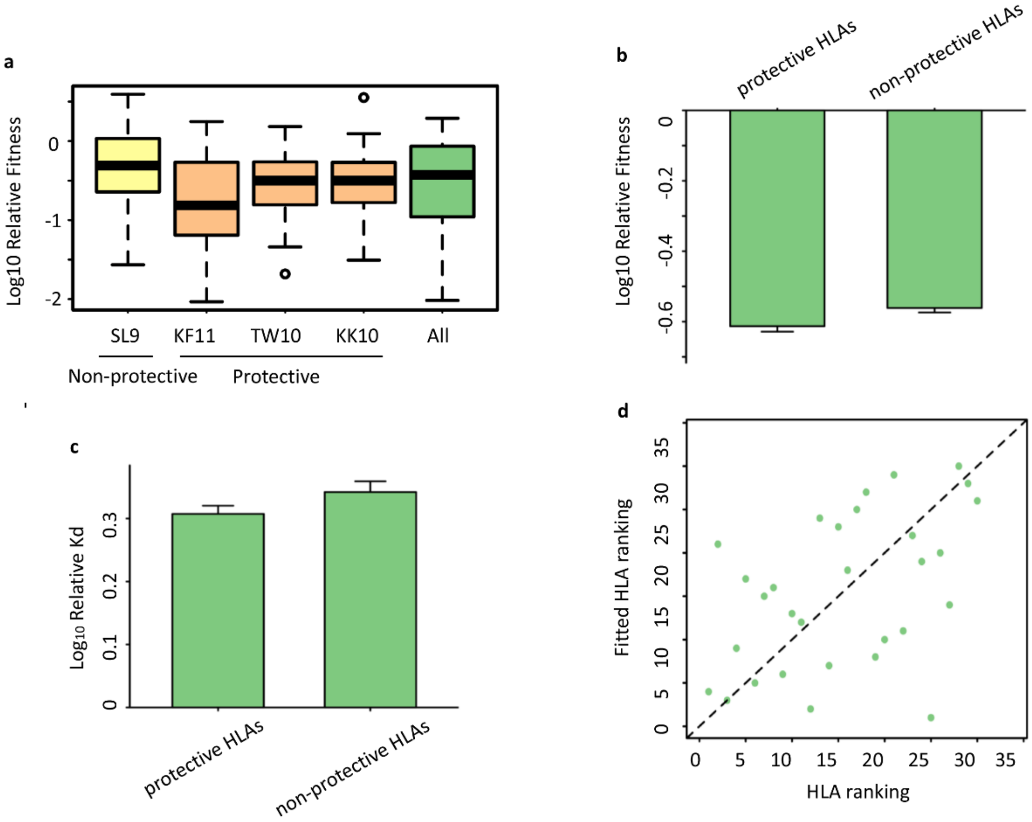


Figure 5-3. Systematic comparison of differences between protective or non-protective HLA alleles

a) Bar plots showing the relative fitness of selected CTL epitope mutations. Red bars represented protective HLA alleles, yellow bars represent non-protective HLA alleless , and green bar is the average of all epitopes on Gag. b) The relative fitness of mutations at protective HLA alleles target epitopes and non-protective HLA alleles target epitopes. HLA protectiveness was defined as the ratio of their prevalence in HIV controllers to the prevalence in progressors. Top 5 HLAs were grouped and defined as protective HLA alleless. Bottom 5 HLAs were called non-protective HLA alleless. Mutations on their best characterized CTL epitopes were analyzed. 1042 mutations on protective HLA alleles's epitopes and 792 mutations on non-protective HLA alleles's epitopes were analyzed. One tail t-test returned a p value of 0.005. c) The binding affinity drop of mutations at protective HLA alleles target epitopes and non-protective HLA alleles target epitopes. The binding affinity drop is an average drop of all possible mutations in the best characterized epitope regions. Protective and non-protective HLA alleless were defined as panel b. 2299 possible mutations of 121 residues on protective HLA alleles's epitopes were analyzed. Correspondingly, non-protective HLA alleless had 2850 possible mutations on 150 residues. One tail t-test returned a p value of 0.05. d) The correlation between fitted HLA protectiveness and reported HLA protectiveness. HLAs were ranked according to their protectiveness defined in panel b. The fitted protectiveness was ranked by linearly adding up average relative fitness and average binding affinity drop. The coefficients of these two parameters were 0.612 and 0.388. The spearman correlation test has a p value of 0.010. This linear model was trained 1000 times with 20 randomly picked HLAs and was then applied to 10 other HLAs for prediction. Average spearman correlation coefficient of 1000 tests was 0.342, and was significantly higher than zero.

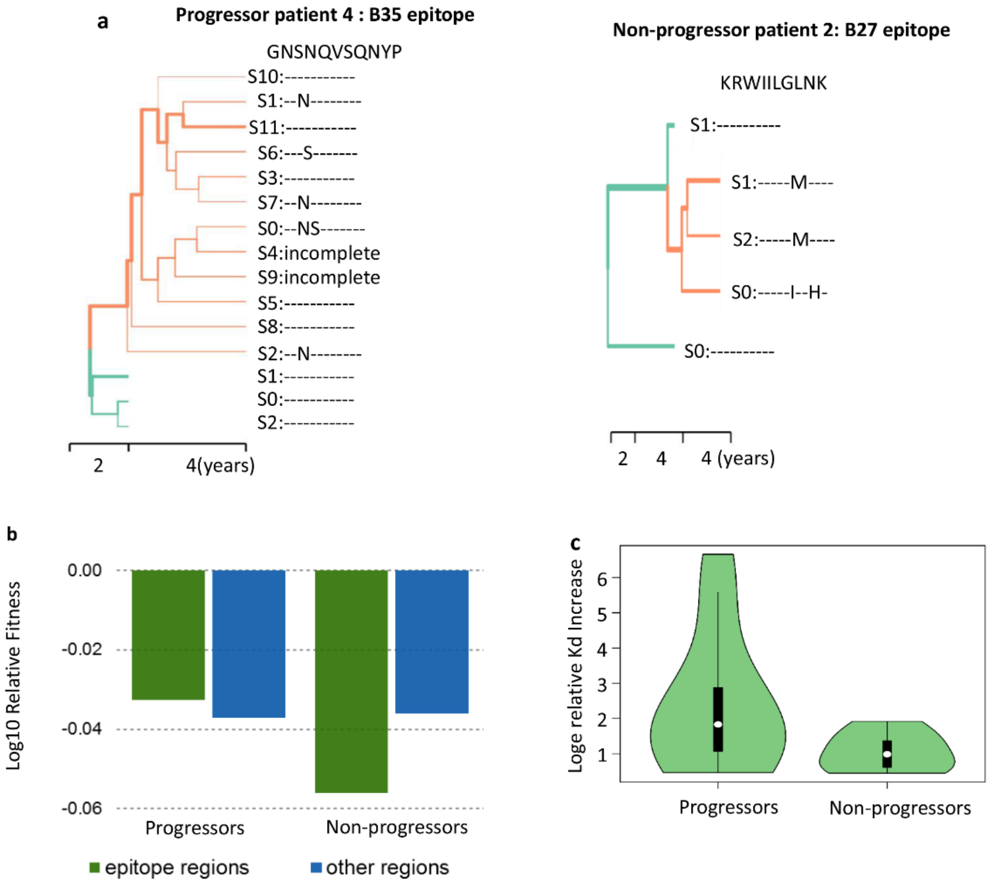


Figure 5-4. HIV intrapatient evolution of 4 paired progressors and non-progressors

a) Phylogenetic tree of virus haplotypes in patients. HIV populations in progressors and non-progressors were sampled at the beginning and the end of a four years' study. HIV Gag region in patients' PBMCs was deep sequenced. Haplotypes were assembled by PredictHaplo. MCC trees were constructed by BEAST. A representative tree in either elite controller or progressor was shown. The escape mutations on a representative HLA epitope were detailed. The width of branch was proportional to the abundance of the haplotype in the population. The color represented sampling time of the haplotype. Green was first batch of the samples. Orange was second batch of the samples. Incomplete haplotypes were provirus that had stop codon in coding region. The amino acids directly before and after the epitope were also shown. CTL epitopes were defined according to the 2013 update of best-characterized epitopes from the Los Alamos Database. c&d) Relative fitness (c) and MHC binding drop (d) of mutations in different group of patients. Mutations are defined as mutations occur in time point 2 compare with concensus sequence from time point 1 of individual patient. Binding affinity is predicted by netMHC-4.0. For every patient, binding affinity is calculated for corresponding HLA and virus haplotypes. The fitness of observed escape mutations are calculated.

5.6 Supplementary Materials

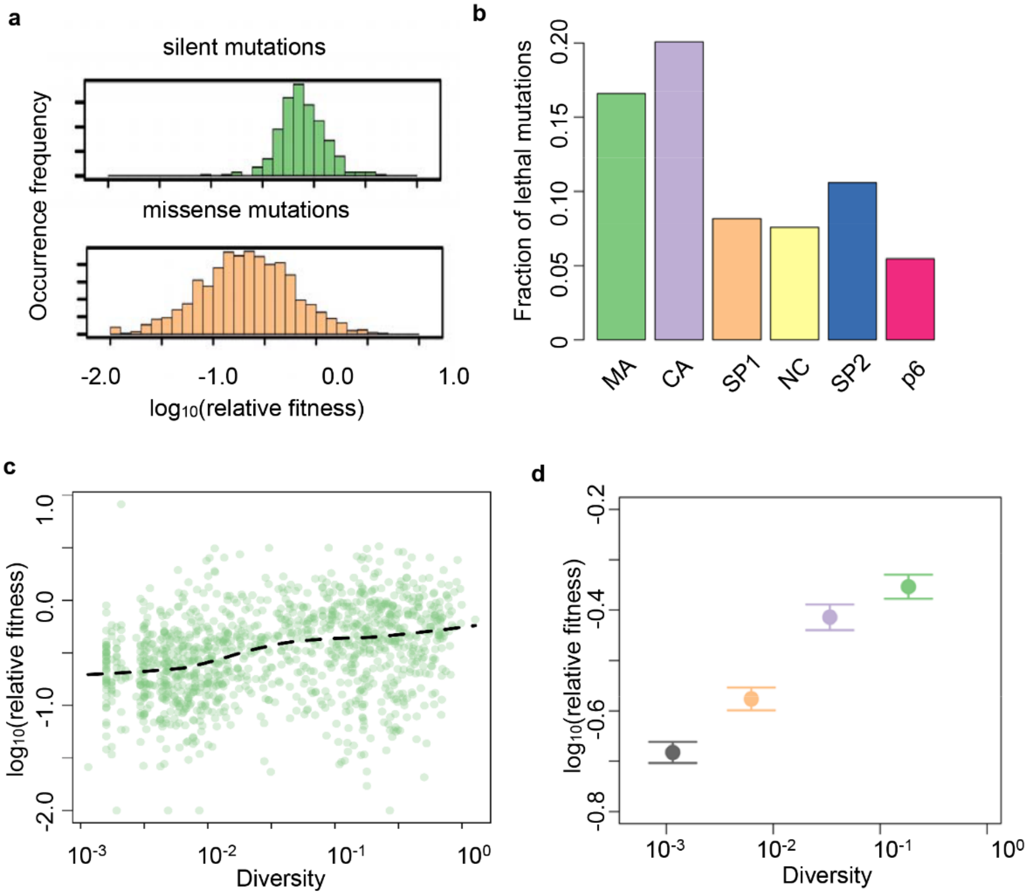


Figure S5-1. Fitness profiling and diversity.

a) Distribution of silent, nonsense and missense mutation for capsid protein. Similar distributions were obtained for other protein samples. b) Fraction of lethal mutations in missense mutations. Lethal mutations were defined as mutations with relative fitness lower than -0.95, which was relative fitness of bottom 2.5% of silent mutations. Each subunit had 807, 1499, 98, 462, 104, 366 missense mutations respectively. c) Correlation of relative fitness and natural variance. All HIV-1 subtype B sequences in Los Alamos database were extracted to calculate natural variance. For each position on Gag, Shannon diversity was calculated as an indicator of natural variance. The relative fitness of a position was the average relative fitness of all its missense mutations, which gave us a robust estimation of nucleotide fitness. The dashed line is local regression using LOESS. Spearman's correlation coefficient is 0.3379, p-value of correlation test is lower than 10-10. d) Average relative fitness of nucleotides in different natural variance groups. Nucleotides were grouped evenly by their Shannon diversity. Group1 had diversity 0.001~0.006. Group 2 had diversity 0.006~0.033. Group 3 had diversity 0.033~0.184. Group 4 had diversity 0.184 and more. Each groups had 293, 287, 311 and 287 nucleotides positions respectively. Error bar represents standard error.

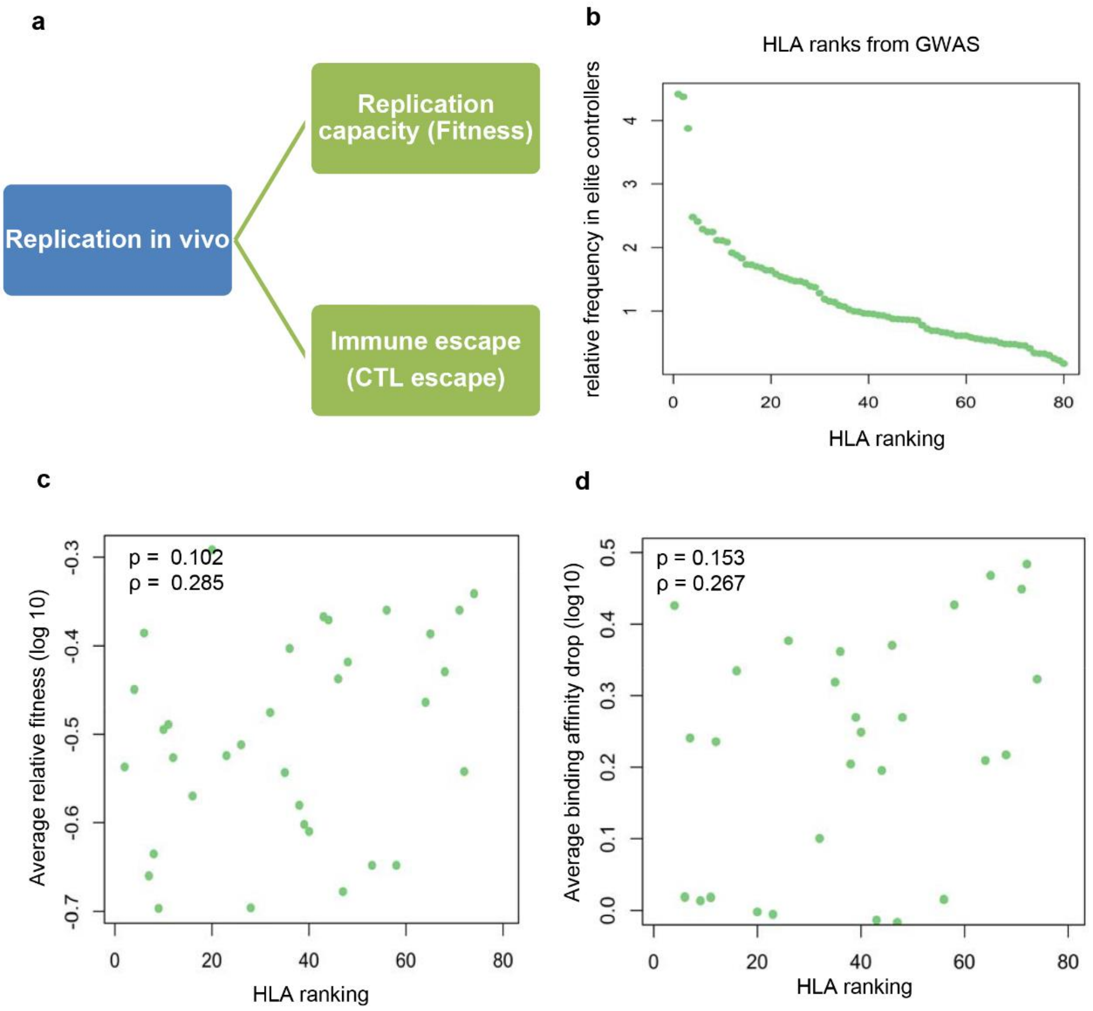


Figure S5-2. Fitness profiling and diversity.

a) Schemetic plot shown the repliacation of HIV in patient depend both on the replication capacity and the ability to escape from human immune selection pressure, especially the CTL pressure. b) HLA ranking of 58 epitopes of Gag region is shown. The ranking is calculated based on the relative frequency of specific HLA in non-progressors versus in progressors. Data were obtained from the TIHIVC study. c) The correlation between HLA protectiveness and fitness constraints on CTL epitopes. HLA protectiveness was ranked as in panel b. The fitness constraint of an HLA was defined as the average fitness of all mutations on its best-characterized CTL epitopes. The spearman correlation test of HLA protectiveness and fitness constraints return a p value of 0.102. d) The correlation between HLA protectiveness and virus escape potential. Virus escape potential was defined as the average binding affinity drop of mutations on CTL epitopes. Binding affinity of mutated epitopes and wild type epitopes were predicted by netMHC-4.0. The spearman correlation test has a p value of 0.153.

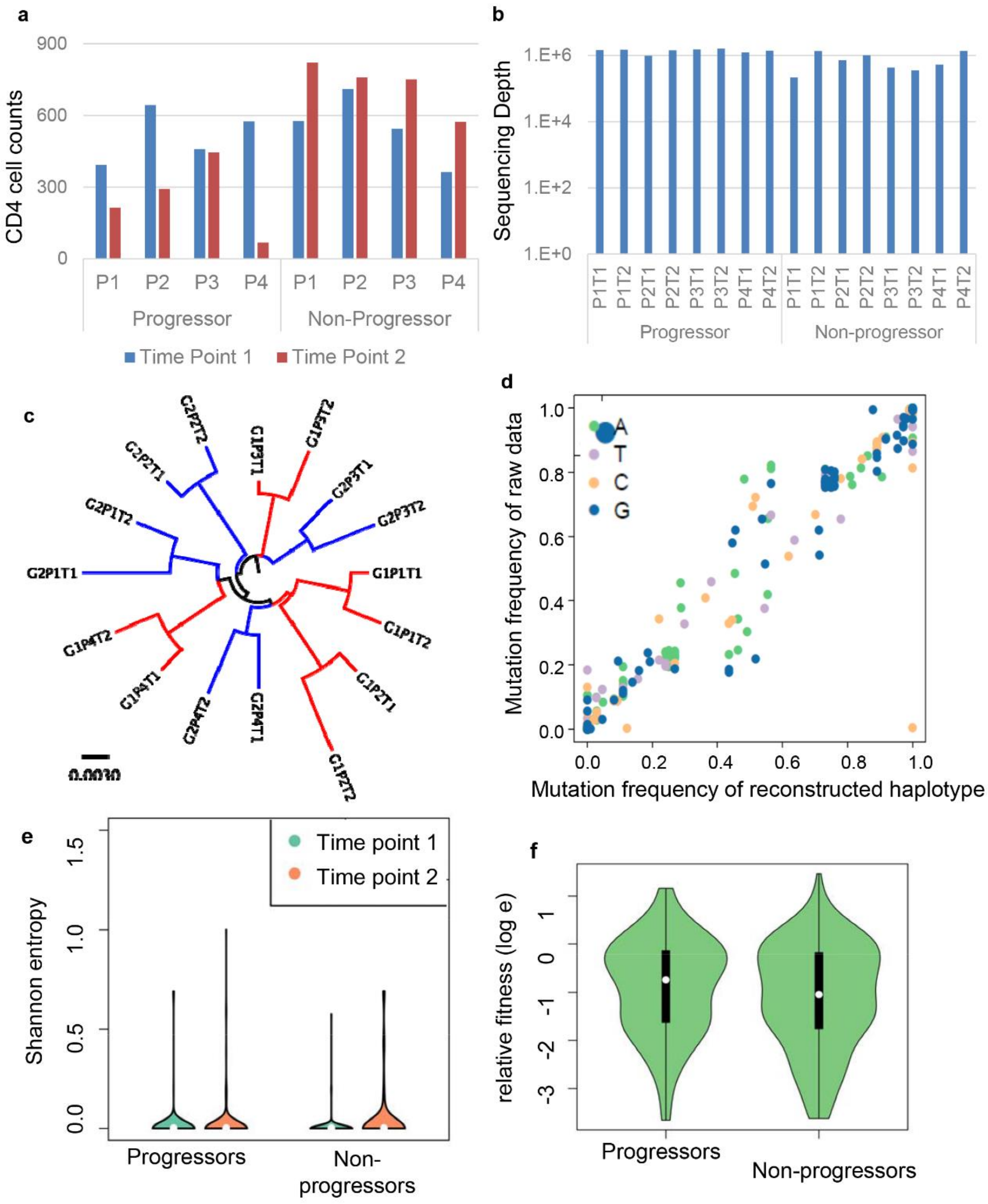


Figure S5-3. Patient sample sequencing

a) The absolute number of CD4 cells for each patient at time point 1 and time point 2. b) The sequencing depth of each patient sample. c) Phylogenetic tree constructed by the consensus sequence of each patient at each time point. d) Correlation of mutation frequency in the raw data and the reconstructed viral haplotypes.

e) Shannon entropy of epitope region were calculated for progressors and non-progressors for two time points. f) Relative fitness of mutations in different group of patients. Mutations are defined as mutations occur in time point 2 compare with concensus sequence from time point 1 of individual patient.

5.7 Bibliography

1. Borrow P, Lewicki H, Hahn BH, Shaw GM, Oldstone MB. 1994. Virus-specific CD8+ cytotoxic T-lymphocyte activity associated with control of viremia in primary human immunodeficiency virus type 1 infection. J Virol 68:6103–6110.

2. Pereyra F, Jia X, Mclaren PJ, Kadie CM, Carlson JM, Heckerman D, De PIW, Sullivan J, Gonzalez E, Davies L, Camargo A, Moore JM, Gupta S, Crenshaw A, Burtt NP, Guiducci C, Cutrell E, Rosenberg R, Moss KL, Lemay P, Leary JO, Schaefer T, Verma P, Toth I, Block B, Rothchild A, Lian J, Proudfoot J, Alvino DML, Vine S, Addo MM, Allen TM, Altfeld M, Henn MR, Gall S Le, Streeck H, Walker BD, Clinical A, Group T. 2011. The Major Genetic Determinants of HIV-1 Control Affect HLA Class I Peptide Presentation. Science 330:1551–1557.

3. Schmitz E, Kuroda MJ, Santra S, Sasseville VG, Simon MA, Lifton MA, Racz P, Tenner-racz K, Dalesandro M, Scallon BJ, Ghrayeb J, Forman MA, Montefiori DC, Rieber EP, Letvin NL, Reimann KA. 1999. Control of Viremia in Simian Immunodeficiency Virus Infection by CD8 ⁒ Lymphocytes. Science (80-) 283:857–860.

4. McMichael A, Rowland-Jones S. 2001. Cellular immune responses to HIV. Nature 410:980–987.

5. Ogg GS, Jin X, Bonhoeffer S, Dunbar PR, Nowak MA, Monard S, Segal JP, Cao Y, Rowland-jones SL, Cerundolo V, Hurley A, Markowitz M, Ho DD, Nixon DF, Mcmichael AJ. 1998. Quantitation of HIV-1 – Specific Cytotoxic T Lymphocytes and Plasma Load of Viral RNA. Science (80-) 279:2103–2106.

6. Streeck H, Nixon DF. 2010. T cell immunity in acute HIV-1 infection. J Infect Dis 202:1–11.

7. Yang OO, Kalams SA, Rosenzweig M, Trocha A, Jones N, Koziel M, Walker BD, Johnson RP. 1996. Efficient lysis of human immunodeficiency virus type 1-infected cells by cytotoxic T lymphocytes. J Virol 70:5799–5806.

8. Schmitz JE, Kuroda MJ, Santra S, Sasseville VG, Simon MA, Lifton MA, Racz P, Tenner-Racz K, Dalesandro M, Scallon BJ, Ghrayeb J, Forman MA, Montefiori DC, Rieber EP, Letvin NL, Reimann KA. 1999. Control of viremia in simian immunodeficiency virus infection by CD8+ lymphocytes. Science 283:857–60.

9. Goulder PJ, Brander C, Tang Y, Tremblay C, Colbert R a, Addo MM, Rosenberg ES, Nguyen T, Allen R, Trocha a, Altfeld M, He S, Bunce M, Funkhouser R, Pelton SI, Burchett SK, McIntosh K, Korber BT, Walker BD. 2001. Evolution and transmission of stable CTL escape mutations in HIV infection. Nature 412:334–338.

10. Cao J, McNevin J, Malhotra U, McElrath MJ. 2003. Evolution of CD8+ T cell immunity and viral escape following acute HIV-1 infection. J Immunol 171:3837–3846.

11. Allen TM, Altfeld M, Geer SC, Kalife ET, Moore C, Desouza I, Feeney ME, Eldridge RL, Maier EL, Kaufmann DE, Lahaie MP, Reyor L, Tanzi G, Johnston MN, Brander C, Draenert R, Rockstroh JK, Jessen H, Rosenberg ES, Mallal S a, Walker BD. 2005. Selective Escape from CD8. Society 79:13239–13249.

12. Kløverpris HN, Leslie A, Goulder P. 2016. Role of HLA adaptation in HIV evolution. Front Immunol 6.

13. Carlson JM, Le AQ, Shahid A, Brumme ZL. 2015. HIV-1 adaptation to HLA: A window into virus-host immune interactions. Trends Microbiol 23:212–224.

14. Lin GG, Scott JG. 2012. HIV escape mutations occur preferentially at HLA binding sites of CD8 T cell epitopes. AIDS 100:130–134.

15. Martin E, Carlson JM, Le AQ, Chopera DR, McGovern R, Rahman MA, Ng C, Jessen H, Kelleher AD, Markowitz M, Allen TM, Milloy M-J, Carrington M, Wainberg MA, Brumme ZL. 2014. Early immune adaptation in HIV-1 revealed by population-level approaches. Retrovirology 11:64.

16. Moore CB, John M, James IR, Christiansen FT, Witt CS, Mallal S a. 2002. Evidence of HIV-1 adaptation to HLA-restricted immune responses at a population level. Science 296:1439–1443.

17. Walker BD. 2006. Elite control of HIV Infection: implications for vaccines and treatment. Top HIV Med 15:134–6.

18. Pereyra F, Heckerman D, Carlson JM, Kadie C, Soghoian DZ, Karel D, Goldenthal A, Davis OB, DeZiel CE, Lin T, Peng J, Piechocka A, Carrington M, Walker BD. 2014. HIV control is mediated in part by CD8+ T-cell targeting of specific epitopes. J Virol 88:12937–48.

19. McLaren PJ, Coulonges C, Ripke S, van den Berg L, Buchbinder S, Carrington M, Cossarizza A, Dalmau J, Deeks SG, Delaneau O, De Luca A, Goedert JJ, Haas D, Herbeck JT, Kathiresan S, Kirk GD, Lambotte O, Luo M, Mallal S, van Manen D, Martinez-Picado J, Meyer L, Miro JM, Mullins JI, Obel N, O'Brien SJ, Pereyra F, Plummer FA, Poli G, Qi Y, Rucart P, Sandhu MS, Shea PR, Schuitemaker H, Theodorou I, Vannberg F, Veldink J, Walker BD, Weintrob A, Winkler CA, Wolinsky S, Telenti A, Goldstein DB, de Bakker PIW, Zagury JF, Fellay J. 2013. Association Study of Common Genetic Variants and HIV-1 Acquisition in 6,300 Infected Cases and 7,200 Controls. PLoS Pathog 9.

20. Apps R, Qi Y, Carlson JM, Chen H, Gao X, Thomas R, Yuki Y, Prete GQ Del, Goulder P, Brumme ZL, Brumme CJ, John M, Mallal S, Nelson G, Bosch R, Goedert JJ, Buchbinder S, Kirk GD, Fellay J, Mclaren P. 2012. Influence of HLA-C expression level on HIV control. Science (80-) 340:87–90.

21. van Manen D, van 't Wout AB, Schuitemaker H. 2012. Genome-wide association studies on HIV susceptibility, pathogenesis and pharmacogenomics. Retrovirology 9:70.

22. Autran B, Descours B, Avettand-Fenoel V, Rouzioux C. 2011. Elite controllers as a model of functional cure. Curr Opin HIV AIDS 6:181–187.

23. Okulicz JF. 2012. Elite controllers and long-term nonprogressors: Models for HIV vaccine development? J AIDS Clin Res 3:1–4.

24. Sunshine JE, Larsen BB, Maust B, Casey E, Deng W, Chen L, Westfall DH, Kim M, Zhao H, Ghorai S, Lanxon-Cookson E, Rolland M, Collier AC, Maenza J, Mullins JI, Frahm N. 2015. Fitness-Balanced Escape Determines Resolution of Dynamic Founder Virus Escape Processes in HIV-1 Infection. J Virol 89:10303–10318.

25. Mostowy R, Kouyos RD, Hoof I, Hinkley T, Haddad M, Whitcomb JM, Petropoulos CJ, Keşmir C, Bonhoeffer S. 2012. Estimating the fitness cost of escape from hla presentation in HIV-1 protease and reverse transcriptase. PLoS Comput Biol 8.

26. Cabello CM, Bair WB, Lamore SD, Ley S, Alexandra S, Azimian S, Wondrak GT. 2010. Translating HIV sequences into quantitative fitness landscapes predicts viral vulnerabilities for rational immunogen design. Immunity 46:220–231.

27. Boutwell CL, Carlson JM, Lin TH, Seese a, Power K a, Peng J, Tang Y, Brumme ZL, Heckerman D, Schneidewind a, Allen TM. 2013. Frequent and variable cytotoxic-T-lymphocyte escape-associated fitness costs in the human immunodeficiency virus type 1 subtype B Gag proteins. J Virol 87:3952–3965.

28. Schneidewind A, Brockman MA, Yang R, Adam RI, Li B, Le Gall S, Rinaldo CR, Craggs SL, Allgaier RL, Power KA, Kuntzen T, Tung C-S, LaBute MX, Mueller SM, Harrer T, McMichael AJ, Goulder PJR, Aiken C, Brander C, Kelleher AD, Allen TM. 2007. Escape from the dominant HLA-B27-restricted cytotoxic T-lymphocyte response in Gag is associated with a dramatic reduction in human immunodeficiency virus type 1 replication. J Virol 81:12382–93.

29. Martinez-Picado J, Martínez MA. 2008. HIV-1 reverse transcriptase inhibitor resistance mutations and fitness: A view from the clinic and ex vivo. Virus Res 134:104–123.

30. Al-Mawsawi LQ, Wu NC, Olson C, Shi V, Qi H, Zheng X, Wu T-T, Sun R. 2014. High-throughput profiling of point mutations across the HIV-1 genome. Retrovirology 11:124.

31. Qi H, Olson CA, Wu NC, Ke R, Loverdo C, Chu V, Truong S, Remenyi R, Chen Z, Du Y, others. 2014. A quantitative high-resolution genetic profile rapidly identifies sequence determinants of hepatitis C viral fitness and drug sensitivity. PLoS Pathog 10:e1004064.

32. Du Y, Wu NC, Jiang L, Zhang T, Gong D, Shu S, Wu T. 2016. Annotating Protein Functional Residues by Coupling High- Throughput Fitness Profile and Homologous-Structure Analysis. MBio 7:1–13.

33. Haddox HK, Dingens AS, Bloom JD, Zanini F, Brodin J, Thebo L, Lanz C, Bratt G, Albert J, Chen F, Vallender E, Wang H, Tzeng C, Li W, Mikkelsen T, Hillier L, Eichler E, Zody M, Jaffe D, Yang S, Langergraber K, Prüfer K, Rowney C, Boesch C, Crockford C, Fawcett K, Albert J, Abrahamsson B, Nagy K, Aurelius E, Gaines H, Nyström G, Wei X, Decker J, Wang S, Hui H, Kappes J, Wu X, Richman D, Wrin T, Little S, Petropoulos C, Olshevsky U, Helseth E, Furman C, Li J, Haseltine W, Sodroski J, Cordonnier A, Montagnier L, Emerman M, Basmaciogullari S, Babcock G, Ryk D Van, Wojtowicz W, Sodroski J, Freed E, Myers D, Risser R, Lu M, Stoller M, Wang S, Liu J, Fagan M, Nunberg J, Jacobs A, Sen J, Rong L, Caffrey M, Zwick M, Jensen R, Church S, Wang M, Stiegler G, Kunert R, Pantophlet R, Saphire E, Poignard P, Parren P, Wilson I, Burton D, Li Y, O'Dell S, Walker L, Wu X, Guenaga J, Feng Y, Lynch R, Wong P, Tran

L, O’Dell S, Nason M, Li Y, Dahirel V, Shekhar K, Pereyra F, Miura T, Artyomov M, Talsania S, Ferguson A, Mann J, Omarjee S, Ndung’u T, Walker B, Chakraborty A, Zanini F, Puller V, Brodin J, Albert J, Neher R, Zanini F, Neher R, Fowler D, Araya C, Fleishman S, Kellogg E, Stephany J, Baker D, Fowler D, Fields S, Boucher J, Cote P, Flynn J, Jiang L, Laban A, Mishra P, Jr RM, Poelwijk F, Raman A, Gosal W, Ranganathan R, Roscoe B, Thayer K, Zeldovich K, Fushman D, Bolon D, Firnberg E, Labonte J, Gray J, Ostermeier M, Olson C, Wu N, Sun R, Melnikov A, Rogov P, Wang L, Gnirke A, Mikkelsen T, Bloom J, Qi H, Olson C, Wu N, Ke R, Loverdo C, Chu V, Thyagarajan B, Bloom J, Stiffler M, Hekstra D, Ranganathan R, Doud M, Ashenberg O, Bloom J, Kitzman J, Starita L, Lo R, Fields S, Shendure J, Mishra P, Flynn J, Starr T, Bolon D, Doud M, Bloom J, Mavor D, Fraser J, Wu N, Young A, Al-Mawsawi L, Olson C, Feng J, Qi H, Starita L, Young D, Islam M, Kitzman J, Gullingsrud J, Hause R, Wu N, Olson C, Du Y, Le S, Tran K, Remenyi R, Steichen J, Kulp D, Tokatlian T, Escolano A, Dosenovic P, Stanfield R, Jardine J, Kulp D, Havenar-Daughton C, Sarkar A, Briney B, Sok D, Al-Mawsawi L, Wu N, Olson C, Shi V, Qi H, Zheng X, Peden K, Emerman M, Montagnier L, Kwong P, Wyatt R, Robinson J, Sweet R, Sodroski J, Hendrickson W, Pancera M, Majeed S, Ban Y, Chen L, Huang C, Kong L, Anken E van, Sanders R, Liscaljet I, Land A, Bontjer I, Tillemans S, Korber B, Foley B, Kuiken C, Pillai S, Sodroski J, Chakrabarti L, Emerman M, Tiollais P, Sonigo P, Yuste E, Reeves J, Doms R, Desrosiers R, Li Y, Luo L, Thomas D, Kang O, Sheehy A, Gaddis N, Choi J, Malim M, Refsland E, Stenglein M, Shindo K, Albin J, Brown W, Harris R, Ho Y, Shan L, Hosmane N, Wang J, Laskey S, Rosenbloom D, Cuevas J, Geller R, Garijo R, López-Aldeguer J, Sanjuán R, Hiatt J, Patwardhan R, Turner E, Lee C, Shendure J, Jabara C, Jones C, Roach J, Anderson J, Swanstrom R, Kinde I, Wu J, Papadopoulos N, Kinzler K, Vogelstein B, Zhang T, Wu N, Sun R, Sloan R, Wainberg M, Guo H, Choe J, Loeb L, Shafikhani S, Siegel R, Ferrari E, Schellenberger V, Bloom J, Silberg J, Wilke C, Drummond D, Adami C, Arnold F, Johnson W, Desrosiers R, Ohgimoto S, Shioda T, Mori K, Nakayama E, Hu H, Nagai Y, Pugach P, Kuhmann S, Taylor J, Marozsan A, Snyder A, Ketas T, Wang W, Nie J, Prochnow C, Truong C, Jia Z, Wang S, Moore J, Cao Y, Qing L, Sattentau Q, Pyati J, Koduri R, Sullivan N, Sun Y, Li J, Hofmann W, Sodroski J, Guttman M, Cupo A, Julien J, Sanders R, Wilson I, Moore J, Stewart-Jones G, Soto C, Lemmin T, Chuang G, Druz A, Kong R, Sanders R, Vesanen M, Schuelke N, Master A, Schiffner L, Kalyanaraman R, Taeye S de, Ozorowski G, Peña A de la, Guttman M, Julien J, Kerkhof T van den, Rizzuto C, Wyatt R, Hernández-Ramos N, Sun Y, Kwong P, Hendrickson W, Wain-Hobson S, Vartanian J, Henry M, Chenciner N, Cheynier R, Delassus S, Shaw G, Arya B, Chang S, Bowman B, Weiss J, Garcia R, White T, Parmley J, Chamary J, Hurst L, Cuevas J, Domingo-Calap P, Sanjuán R, Subramaniam A, DeLoughery A, Bradshaw N, Chen Y, O’Shea E, Losick R, Haas J, Park E, Seed B, Watts J, Dang K, Gorelick R, Leonard C, Jr JB, Swanstrom R, Fernandes J, Jayaraman B, Frankel A, Malim M, Hauber J, Le S, Maizel J, Cullen B, Emerman M, Vazeux R, Peden K, Bloom J, Bloom J, Kyte J, Doolittle R, Bloom J, Weinreich D, Delaney N, DePristo M, Hartl D, Ortlund E, Bridgham J, Redinbo M, Thornton J, Freed E, Martin M, Wang W, Essex M, Lee T, Silva J da, Coetzer M, Nedellec R, Pastore C, Mosier D, Gasser R, Hamoudi M, Pellicciotta M, Zhou Z, Visdeloup C, Colin P, Abram M, Ferris A, Shao W, Alvord W, Hughes S, Spielman S, Wilke C, Nielsen R, Yang Z, Shankarappa R, Margolick J, Gange S, Rodrigo A, Upchurch D, Farzadegan H, Overbaugh J, Morris L, Lee W, Syu W, Du B, Matsuda M, Tan S, Wolf A, Meyer A, Wilke C, Starcich B, Hahn B, Shaw G, McNeely P, Modrow S, Wolf H, Modrow S, Hahn B, Shaw G, Gallo R, Wong-Staal F, Wolf H, Moore P, Gray E, Morris L, Sok D, Pauthner M, Briney B, Lee J, Saye-Francisco K, Hsueh J, Zwick M, Labrijn A, Wang M, Spenlehauer C, Saphire E, Binley J, Blattner C, Lee J, Sliepen K, Derking R, Falkowska E, Peña A de la, Falkowska E, Le K, Ramos A, Doores K, Lee J, Blattner C, Huang J, Kang B, Pancera M, Lee J, Tong T, Feng Y, Scharf L, Scheid J, Lee J, West A, Chen C, Gao H, Kwong P, Mascola J, Nabel G, Ramsey D, Scherrer M, Zhou T, Wilke C, Zhou T, Lynch R, Chen L, Acharya P, Wu X, Doria-Rose N, Zhou T, Georgiev I, Wu X, Yang Z, Dai K, Finzi A, Klein F, Diskin R, Scheid J, Gaebler C, Mouquet H, Georgiev I, Zolla-Pazner S, Cardozo T, Yusim K, Kesmir C, Gaschen B, Addo M, Altfeld M, Brunak S, Kouyos R, Leventhal G, Hinkley T, Haddad M, Whitcomb J, Petropoulos C, Lee B, Sharron M, Montaner L, Weissman D, Doms R, Kabat D, Kozak S, Wehrly K, Chesebro B, Levy J, Berger E, Murphy P, Farber J, Ablashi D, Berneman Z, Kramarsky B, Whitman J, Asano Y, Pearson G, Wei X, Decker J, Liu H, Zhang Z, Arani R, Kilby J, Reed L, Muench H, Stamatakis A, Touw W, Baakman C, Black J, Beek T te, Krieger E, Joosten R, Kabsch W, Sander C, Tien M, Meyer A, Sydykova D, Spielman S, Wilke C, Pancera M, Zhou T, Druz A, Georgiev I, Soto C, Gorman J, Zhang M, Gaschen B, Blay W, Foley B, Haigwood N, Kuiken C, Leonard C, Spellman M, Riddle L, Harris R, Thomas J, Gregory T. 2016. Experimental Estimation of the Effects of All Amino-Acid Mutations to HIV's Envelope Protein on Viral Replication in Cell Culture. PLOS Pathog 12:e1006114.

34. Wu NC, Olson CA, Du Y, Le S, Tran K, Remenyi R, Gong D, Al-Mawsawi LQ, Qi H, Wu T-T, others. 2015. Functional constraint profiling of a viral protein reveals discordance of evolutionary conservation and functionality. PLoS Genet 11:e1005310.

35. 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.

36. 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.

37. Http://www.hiv.lanl.gov/. Los Alamos HIV databases.

38. Mothe B, Llano A, Ibarrondo J, Daniels M, Miranda C, Zamarreño J, Bach V, Zuniga R, Pérez-Álvarez S, Berger CT, Puertas MC, Martinez-Picado J, Rolland M, Farfan M, Szinger JJ, Hildebrand WH, Yang OO, Sanchez-Merino V, Brumme CJ, Brumme ZL, Heckerman D, Allen TM, Mullins JI, Gómez G, Goulder PJ, Walker BD, Gatell JM, Clotet B, Korber BT, Sanchez J, Brander C. 2011. Definition of the viral targets of protective HIV-1-specific T cell responses. J Transl Med 9:208.

39. Gutteridge A, Bartlett GJ, Thornton JM. 2003. Using A Neural Network and Spatial Clustering to Predict the Location of Active Sites in Enzymes. J Mol Biol 330:719–734.

40. Nielsen M, Lundegaard C, Blicher T, Lamberth K, Harndahl M, Justesen S, Røder G, Peters B, Sette A, Lund O, Buus S. 2007. NetMHCpan, a method for quantitative predictions of peptide binding to any HLA-A and -B locus protein of known sequence. PLoS One 2.

41. Andreatta M, Nielsen M. 2015. Gapped sequence alignment using artificial neural networks: Application to the MHC class i system. Bioinformatics 32:511–517.

42. Lundegaard C, Lamberth K, Harndahl M, Buus S, Lund O, Nielsen M. 2008. NetMHC-3.0: accurate web accessible predictions of human, mouse and monkey MHC class I affinities for peptides of length 8-11. Nucleic Acids Res 36:509–512.

43. Wang YE, Li B, Carlson JM, Streeck H, Gladden AD, Goodman R, Schneidewind A, Power KA, Toth I, Frahm N, Alter G, Brander C, Carrington M, Walker BD, Altfeld M,

Heckerman D, Allen TM. 2009. Protective HLA class I alleles that restrict acute-phase CD8+ T-cell responses are associated with viral escape mutations located in highly conserved regions of human immunodeficiency virus type 1. J Virol 83:1845–55.

44. Schellens IMM, Navis M, van Deutekom HWM, Boeser-Nunnink B, Berkhout B, Kootstra N, Miedema F, Keşmir C, Schuitemaker H, van Baarle D, Borghans JAM. 2011. Loss of HIV-1-derived cytotoxic T lymphocyte epitopes restricted by protective HLA-B alleles during the HIV-1 epidemic. Aids 25:1691–700.

45. Mothe B, Llano A, Ibarrondo J, Daniels M, Miranda C, Zamarreño J, Bach V, Zuniga R, Pérez-Álvarez S, Berger CT, Puertas MC, Martinez-Picado J, Rolland M, Farfan M, Szinger JJ, Hildebrand WH, Yang OO, Sanchez-Merino V, Brumme CJ, Brumme ZL, Heckerman D, Allen TM, Mullins JI, Gómez G, Goulder PJ, Walker BD, Gatell JM, Clotet B, Korber BT, Sanchez J, Brander C. 2011. Definition of the viral targets of protective HIV-1-specific T cell responses. J Transl Med 9:208.

46. Carlson JM, Brumme CJ, Martin E, Listgarten J, Brockman MA, Le AQ, Chui CKS, Cotton LA, Knapp DJHF, Riddler SA, Haubrich R, Nelson G, Pfeifer N, Deziel CE, Heckerman D, Apps R, Carrington M, Mallal S, Harrigan PR, John M, Brumme ZL. 2012. Correlates of protective cellular immunity revealed by analysis of population-level immune escape pathways in HIV-1. J Virol 86:13202–16.

47. Carrington M, O'Brien SJ. 2003. The influence of HLA genotype on AIDS. Annu Rev Med 54:535–551.

48. 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 b-Lactamase. Cell 160:882–892.

49. Gray GE, Laher F, Lazarus E, Ensoli B, Corey L. 2016. Approaches to preventative and therapeutic HIV vaccines. Curr Opin Virol 17:104–109.

50. Streeck H. 2016. Designing optimal HIV-vaccine T-cell responses. Curr Opin HIV AIDS 1.

51. Rolland M, Manocheewa S, Swain JV, Lanxon-Cookson EC, Kim M, Westfall DH, Larsen BB, Gilbert PB, Mullins JI. 2013. HIV-1 conserved-element vaccines: relationship between sequence conservation and replicative capacity. J Virol 87:5461–7.

52. Rolland M, Nickle DC, Mullins JI. 2007. HIV-1 group M conserved elements vaccine. PLoS Pathog 3:1551–1555.

53. Allen TM, Altfeld M. 2008. Crippling HIV one mutation at a time. J Exp Med 205:1003–1007.

54. Llano A, Williams A, Olvera A, Silva-Arrieta S, Brander C. 2013. Best-Characterized HIV-1 CTL Epitopes: The 2013 Update. HIV Mol Immunol 2013 3–25.

55. Prabhakaran S, Rey M, Zagordi O, Beerenwinkel N, Roth V. 2014. HIV haplotype inference using a propagating dirichlet process mixture model. IEEE/ACM Trans Comput Biol Bioinforma 11:182–191.

CHAPTER 6

GENOME-WIDE IDENTIFICATION OF ANTI-INTERFERON FUNCTIONS IN INFLUENZA A VIRUS ENABLING RATIONAL VACCINE DESIGN

6.1 Abstract

Generation of a safe yet highly immunogenic live attenuated vaccine represents a challenge in vaccinology. Using influenza A virus as an example, we present a novel approach for vaccine development: systematically identifying and eliminating immune evasion functions on the virus genome, while maintaining the replication fitness in vitro. Applying single-nucleotide resolution high-throughput genomics approach, we simultaneously measured the fitness and IFN sensitivity of mutations across the entire viral gnome. We have identified novel IFN-sensitive mutations on PB2, PA, PB1 and M1, in addition to NS1, suggesting influenza virus possesses multiple anti-IFN mechanism in different viral segments to evade host defense. By combining 8 mutations into one viral genome, we successfully generated a deficient in anti-interferon (DAI) strain that are highly attenuated in IFN competent host. DAI was able to induce transient IFN response and persistent adaptive immune responses in vivo. This live attenuated influenza vaccine showed superior property as a safe, effective and broad protectingng live attenuated influenza vaccine. This approach is applicable to vaccine development against other pathogens.

6.2 Introduction

Type I interferon (IFN) system, which is one of the most critical components of innate immune response, plays essential roles in limiting viral replication (1–4). The signaling cascade induced by type I IFN provides the first line of defense against viral infection by inducing the expression of hundreds of interferon-simulated genes (ISGs), many of which carry antiviral activities5. IFN system is also the bridge between the innate and adaptive immune responses , which is critical for dendritic cell maturation, T cell development and antibody production (6–11). During natural evolution, many viruses have evolved to encode multiple anti-IFN functions. Influenza A virus, for example, is able to inhibit IFN production and abrogate the functions of multiple ISGs12. NS1 is the key influenza viral protein that has been studied extensively for its anti-IFN mechanism, partly due to the un-essentialness of NS1 for viral replication (13). With the accumulation of databases and the improvement of reverse genetic technologies, researches started to appreciate the importance of other influenza proteins for their anti-interferon functions (14–18). However, as viral replication function and immune evasion function are often intertwined, it is difficult to dissect the anti-IFN functions encoded by viral proteins that are essential for viral replication.

Influenza viruses constantly evolve and adapt to diverse selection pressures, which pose a challenge for safe and efficient vaccine design. Current annual vaccine for influenza virus contain two types: trivalent inactivated vaccine (TIV) and live attenuated influenza vaccine (FluMist)19,20. Protection elicited by TIV is mostly through antibody response, which lacks cross recognition nor broad protection for strains from different HA classes. Attenuation of FluMist comes from5 mutations on polymerase complex, which renders the cold adapted (ca),temperature sensitive (ts) and live attenuated (att) phenotype. However, as no immune boosting mechanism is encoded in FluMist, the attenuation causes reduced immunogenicity, therefore lower efficacy (21–23). A safe vaccines with higher efficiency and broader protection are thus the pursued goal (24–27). As IFN plays a key role in immune response, systematic elimination of anti-IFN functions on the viral genome is a potential approach for vaccine development (28–30). With the abruption of anti-IFN functions, the vaccine will be highly attenuated in IFN competent host. Moreover, it will likely to generate superior immunogenicity due to the induction of intensive IFN response. Thus, safety and efficacy requirement can be achieved simultaneously. However, the key challenge in this approach is the comprehensive identification of anti-IFN functions of the whole viral genome in order to generate a highly interferon sensitive/inducing strain while maintaining the viral replication fitness in vitro for vaccine production.

We tackled this challenge by applying single-nucleotide resolution high-throughput genomics (**Figure 6-1**) (31–34). A diverse mutant library was generated with mutations distributed across the entire genome. The system enable us to quantitatively examine the growth capacity and IFN sensitivity of all single mutations in a massively parallel process. We identified novel anti-IFN sites on multiple vial segments, which can serve as specific starting points for further mechanism studies. By combining 8 IFN sensitive mutations across the viral genome, we successfully generated a DAI strain that are highly attenuated in IFN competent host. DAI strain showed superior properties as a safe and effective live attenuated influenza vaccine, which can induce strong antibody and T cell response, and provide broad protection against homologous and heterologous viral challenge. Importantly, this approach is broadly applicable to vaccine development against other pathogens.

6.3 Result

6.3.1 Fitness profile of influenza A virus genome at single nucleotide resolution

The mutant plasmid library was constructed on influenza A/WSN/1933 (H1N1) virus through error-prone PCR. We have improved the method by dividing the entire genome into 52 small sub-libraries for optimal mutant frequency and sequencing accuracy, as shown previously with PB1, PA and M1 (**Figure S6-1**) (35–37). Viral libraries were reconstituted in 293T cells by co-transfecting the library of mutant plasmids with the other 7 plasmids encoding wild type viral proteins. For identifying anti-IFN function across the entire viral genome, selection of all viral libraries were performed in A549 cells with or without exogenous IFN treatment (IFNa2, IC80, methods) (**Figure 6-1**). In each condition, the relative fitness score (RF score) of a mutant virus was calculated as the ratio of relative frequency of the corresponding mutation in the infection library to that in the plasmid library. 90% nucleotide positions across the genome were covered, and 95% of all possible single mutations at those positions were detected in the plasmid library.

Biological duplicates in transfection as well as infection showed strong correlations (**Figure S6-2**). Relative growth capacity of 26 randomly chosen missense mutations was measured individually and correlated well with the RF scores measured in the screen (**Figure S6-3**). Using synonymous mutations as control, 49% missense mutations across the whole genome were deleterious, suggesting the general essentialness of the RNA viral genome38,39. Additionally, there were explicit differences in relative essentialness among different influenza viral proteins. All polymerase components (PB2, PB1, PA and NP) were less tolerant of mutation, in contrast to the relative flexibility of M2 and NS1 (**Figure S6-4**).

6.3.2 SYSTEMATICALLY IDENTIFICATION OF IFN-SENSITIVE MUTATIONS

Although the RF scores of mutations are generally well correlated with and without IFN selection, we observed IFN-sensitive mutations widely spread on the other segments, in addition to NS1 (40,41). These mutations were nearly neutral in A549 cells, but highly deleterious under IFN selection (**Figure 6-1**, **Figure S6-5**). These mutations are potential candidates for vaccine design because they have strong interferon sensitive phenotype while maintaining replication capacity. As a confirmation, mutants harboring substitutions in the RNA binding domain of NS1 protein, especially in the positions of R37, R38, K41 and R46, were observed to be highly sensitive to IFN treatment in our system (42,43,12). New mutations (such as G45S) in this domain were also observed and validated to be IFN sensitive (**Figure S6-5**). We focused on the surface exposed and structurally clustered residues on internal proteins for further validation (**Figure 6-1**, **Figure S6-5**). 24 single mutations were constructed individually. All the mutations were neutral or nearly neutral for viral replication. The polymerase activities were also nearly intact for the mutations on polymerase proteins (**Figure S6-7**). Compared with wild type WSN (WT) virus, all mutants increased in IFN sensitivity, while 8 mutations on internal proteins reached statistical significance (Two-tailed T test, p<0.05, **Figure 6-1**). We further examined if these mutations could induce higher IFN production. Among the constructed mutants, the three mutations on PB2 (N9D, Q75H, T76A) and three mutations on M1 (N36Y, R72Q, S225T) induced significantly higher IFN-β (**Figure 6-1**, **Figure S6-8**) and ISG54 (**Figure S6-8**) expression, both of which are IRF-3 target genes44. Consistent results were obtained in the nuclear translocation of IRF3 post-infection (**Figure S6-9**). Moreover, these 6 IFN-β inducible mutants were no longer sensitive to IFN treatment in IFN-deficient Vero cells (**Figure S6-8**). We also observed that the higher induction of IFN is MAVS dependent and STING independent: knocking out MAVS abolished the induction but no effect was observed in STING knockout THP1 cells (**Figure S6-10**). These results suggested that the three mutations on PB2 (N9D, Q75H, T76A) and three mutations on M1 (N36Y, R72Q, S225T) might lose their functions of inhibiting IFN production at an early stage post-infection, as the basis for their IFN sensitive phenotype. The other two mutations (PB1 L155H and PA E181D), however, did not induce higher IFN production. They were still IFN-sensitive in Vero cells, suggesting that they might lose the inhibition of downstream IFN pathway (**Figure S6-8**). The identification of novel IFN-sensitive mutations will facilitate the understanding of the mechanisms for pathogenesis associated with influenza-induced inflammation. They also set up the stage for differential screenings using overexpression or knock-out with members of IFN-production or response pathways to further identify the cellular factors interacting with these mutants45.

6.3.3 DAI VIRUS HARBORING 8 IFN SENSITIVE MUTATIONS DISPLAYS STRONG IFN SENSITIVE AND IFN INDUCTION PHENOTYPE IN VITRO

To maximize the sensitivity to IFN and maintain viral growth capacity, we examined 122 different combinations of mutations and came up with a strain containing 8 mutations (PB2: N9D, Q75H, and T76A; M1: N36Y, R72Q and S225T; NS1: R38A and K41A on NS1), which we named as DAI (Deficient of Anti-Interferon) virus. The IFN-sensitivity of DAI is significantly higher than any individual mutation, indicating the accumulative effect of mutations on PB2 and M1, in addition to NS1 (**Figure 6-2**). Gene expression data at early and late time points post-infection showed that DAI virus induced higher IFN production and responses in A549 cells (**Figure 6-2**, **Figure S6-11**). These results have been validated in human primary alveolar macrophages, the important target for influenza infection (Ref), in which DAI induced almost 50 fold of IFN-β expression than WT virus (**Figure 6-2**). The phenotype of DAI virus is not limited to WSN background, as introducing the same 8 mutations into PR8 background also led to high IFN sensitivity and IFN production (**Figure S6-11**). As expected, the growth of DAI virus showed ~2 log attenuation in A549 cells, but restored in IFN-deficient Vero cells (**Figure 6-2**). To further understand the global gene expression change, we performed mRNA sequencing with WT, NS1 R38A/K41A, DAI and mock virus-infected A549 cells at 6 hours post-infection. Compared with mock, the expression of 120 genes was significantly upregulated (fold change > 2 and p<0.001) in DAI infected-cells, among which 24% belong to IFN response genes (**Figure 6-2**, **Figure S6-12**). Despite the similar replication ability, DAI virus exhibited stronger ability in induction of IFN response than NS1 R38A/K41A virus. Further analysis of the upregulated genes by Gene Ontology (GO) enrichment revealed that the type I IFN production and response related pathways were the dominant pathways activated by DAI virus comparing with both mock infected or WT infected cells (**Figure 6-2**, **Figure S6-12**). More importantly, there was no difference in expression of genes related to inflammation pathways between DAI and NS1 mutant-infected cells, suggesting the specificity of DAI virus in perturbing IFN pathway (**Figure S6-12**) (46).

6.3.4 DAI VIRUS IS HIGHLY ATTENUATED IN IFN COMPETENT HOST

As DAI virus showed strong IFN sensitive phenotype in vitro, we next investigated the replication and pathogenesis of DAI virus in a mouse model. Balb/c mice were intranasally inoculated with WT virus or DAI virus at dosages of 105 or 106 TCID50. While WT virus (106

TCID50) induced significant weight loss in all tested animals, the same amount of DAI did not cause weight loss nor clinical symptoms (**Figure 6-3**). We then compared viral replication in mice lung tissues infected with indicated strains (**Figure 6-3**, **Figure S6-13**). The five amino acid changes on the cold adapted live attenuated virus (FluMist) were constructed into WSN background (WSN-CA) and used as a comparison47,48. WSN-CA is known to be replication capable in lower temperature (33℃,nasal cavity) but highly attenuated in high temperature (39℃,lung). Replication of DAI virus was significantly lower than WT and NS1 R38A-K41A mutant, and comparable with WSN-CA virus. Further comparison of viral replication kinetics (**Figure 6-3**) indicates a robust viral replication with WT infection which peaks at 48 hours. In contrast, no increase in viral copy number was detected with DAI infected mice. Although highly attenuated in replication, DAI virus showed transient yet significant upregulation of certain IFN-related genes at 6 and 24h, and the response was rapidly reduced (**Figure 6-3**). However, WT virus induced a robust proinflammatory response through the course of investigation, exampled with the high induction of CXCl10 at 48 and 120h post infection (**Figure 6-3**). These data correlate well with the results from histology analysis of infected lungs and BAL cytospins (**Figure 6-3**, **Figure S6-13**). Sustained inflammatory response resulted in the massive infiltration of neutrophils and lymphocytes in WT-infected lungs, but not with DAI infection (**Figure 6-3**). We also examined the cytokine response with bronchoalvelar lavage (BAL) samples at 48h post-infection by luminex multiplex assay (BioRad) and ELISA (**Figure 6-3**, **Figure S6-13**). WT virus showed significantly higher induction of IL-6 and CXCL1, representing a more severe inflammation response and pathological state. However, DAI virus induce higher amount IL-12 and G-CSF, which is important for lymphocyte stimulation and T cell development. Furthermore, the replication of DAI virus was fully restored to WT level in IFNAR-/- mice (**Figure 6-3**), indicating that DAI virus maintain its intrinsic replication competence and IFN response is the key determinant to attenuate DAI virus in WT mice (**Figure 6-3**). The above results documented that DAI virus is highly attenuated in IFN competent mice, while able to stimulate prompt and specific IFN response.

6.3.5 DAI VIRUS INDUCES STRONG AND BOARD ADAPTIVE IMMUNE RESPONSES

We then examined whether DAI virus could induce robust adaptive immune response for protection since IFN plays critical roles in stimulating T and B cells. Mice sera and BAL samples were collected at 28 days post single dose vaccination with WT, DAI and WSN-CA viruses. DAI virus induced robust HA antibody responses, as measured by ELISA, hemagglutination inhibition and neutralization antibody assay (**Figure 6-4**, **Figure S6-14**). The level of HA antibody response is lower than WT due to the attenuation in viral replication, but significantly higher than WSN-CA virus. Surprisingly, we detected HA specific antibody against other strains (such as PR8: H1 and Viet04: H5) within HA group (**Figure S6-14**), which was unable to induce by inactivated influenza vaccines as reported previously 49. In addition, specific antibodies (IgG) against NP, NA and M1 proteins were also detected for DAI vaccinated sera with concentration comparable with WT, which were demonstrated to play important role in limiting viral replication50,51 (**Figure 6-4**, **Figure S6-14**). Importantly, secretory IgA response was also elicited by DAI vaccination against both HA and NP proteins, suggesting the induction of mucosal immune response (**Figure 6-4**, **Figure S6-15**). We then took one step further to examine the epitope coverage of the neutralizing antibody generated by DAI virus. Further utilizing the high-throughput genetic approach, we screened the single nucleotide mutant library on HA protein in the presence or absence of serum antibody generated from WT or DAI vaccinated mice (52). Possible mutations that were not neutralized by sera were observed in both head (Ca2 and Sa sites) and stem regions, while no significant difference was detected between the two groups (WT and DAI) in terms of the number and distribution of mutations (**Figure 6-4**, **Figure S6-15**). This suggested that although DAI virus induces slightly lower amount of HA antibody than WT, the breath and complexity, however, are comparable. In addition, we determined the ability of DAI virus to induce T cell response. Impressively, compared with WT, DAI virus elicited similar levels of NP-specific and PB1-specific CTL response examined by tetramer staining of CD8 T cells (**Figure 6-4**, **Figure S6-16**). Similar levels of influenza specific CD4 cell and memory T cell responses were also observed (**Figure 6-4**, **Figure S6-16**). T cell repertoire complexity were further explored by deep sequencing of TCR loci. V usage and clonality of NP specific CD8 T cells, as well as the numbers of expanding CD8 T clones were all comparable between WT and DAI vaccinated mice, suggesting the strength and breadth of DAI induced T cell response (**Figure S6-17**) (53). Collectively, these data indicate that despite the highly attenuated ability in replication, DAI virus is comparable in inducing adaptive immune response, comparing to WT virus.

6.3.6 DAI VIRUS PROTECTS VACCINATED MICE AGAINST HOMOLOGOUS AND HETEROLOGOUS VIRAL CHALLENGE

As DAI virus was able to induce robust humoral and cellular immune response, we next sought to determine whether it offers real protection against homologous and heterologous viral challenges. Firstly, we infected mice with 105 TCID50 of WT virus 28 days post vaccination with WT, DAI and WSN-CA strains and quantified viral titers in infected lung tissues by RT-qPCR and TCID50 assay. A titer drop of ~3 log in DAI vaccinated mice was observed in comparison with mock vaccination, more significant than the WSN-CA vaccinated group (**Figure 6-4**, **Figure S6-18**, **Figure S6-20**). Stronger inhibition of WT challenge can be achieved through higher dose vaccination or two doses of vaccination (S18C,D). More importantly, DAI vaccination provided a full protection against a lethal dose of WSN challenge (**Figure 6-4**), with no sign of weight loss, indicating a superior protection from DAI compare with recently reported PTC-4A virus (**Figure 6-4**) (54). To decide whether DAI vaccination can provide protection against heterologous strains, we then challenged mice with PR8, ACal/04/09 and A/X-31) at lethal dose. Strong protection was observed across the board in terms of survival rate, percentage of body weight loss, and clinic scores (**Figure 6-4**, **Figure S6-18**). Also, strong secondary T cell response was observed for all challenge groups, suggesting an essential role of T cell response for the broad protection (**Figure S6-19**). These data demonstrate that DAI virus is an ideal candidate for influenza vaccine development.

6.4 Discussion

Here we demonstrate the feasibility of a novel approach for rational vaccine design: systematically identifying and eliminating immune evasion functions on the virus genome by high throughput genetic screening and using the attenuated immune evasion deficient virus as master donor virus in vaccine development. We chose the IFN system as the target for its function as first line of innate immunity and the critical linkage to adaptive immune response. By studying the whole genome anti-IFN function of influenza virus, we have identified several novel anti-IFN sites in addition to the NS1 gene, suggesting influenza virus possesses multiple anti-IFN function mechanisms to evade host anti-viral defense. The underlying mechanisms can be a future study of interest, and the responsible interacting host factors can be identified through cellular effector screening using the mutant viruses or mutant proteins in comparison with the wild types. By incorporating 8 mutations, we have generated a novel candidate vaccine virus, DAI, which displays superior IFN sensitivity and induction, while significantly attenuated viral growth in vivo. We have also demonstrated that the DAI virus has superior immunogenicity, is able to elicit strong humoral and cellular immunity in mice, and provides efficient and broad protection against different strains of influenza virus. These results suggest identifying and eliminating whole genome anti-IFN function of influenza leads to a safer and more effective vaccine design.

Usually, the intensity and breath of immune responses are parallel to the level of viral replication. Many attenuated vaccines exhibit a reduced immunogenicity when the replication capacity of the vaccines is reduced to obtain the safety. Through defining the impact of every amino acid in the viral genome in different conditions, the single-nucleotide resolution high-throughput genetic system will enable the screenings, rationally selection and re-engineering the virus towards the desired properties. In this study, by quantitatively measuring the replication fitness of each mutant, with mutations distributed across the entire genome, we obtained the opportunities to uncouple the anti-IFN functions from the intrinsic replication capacity of many essential genes of the virus. This novel information has enabled us to successfully increase the immunogenicity (at least per viral genome) of the DAI virus and significantly attenuate its replication in vivo, while maintaining intrinsic replication capacity in IFN-deficient hosts. Similar screenings can be applied to other immune responses, such as, other cytokines, NK cells or CD8 T cells. Moreover, screens can be set up in mice, which will enable the determination of viral immune evasion functions in vivo and further increase immunogenicity at multiple steps of immune responses. This method has the potential to be broadly applicable to designing vaccines against other pathogens, and be used to improve the properties of viral agents to modulate immune responses against cancers, by selecting mutants that preferentially replicate in tumor cells and stimulate local immune responses to neoantigens.

6.5 Materials and Methods

Viruses and Cells

Influenza A/WSN/33 virus (WSN) and Influenza A/Puerto Rico/8/1934 (PR8) were used as the model systems. Eight-plasmid transfection system was utilized to reconstitute WT virus 55,56. A live attenuated cold adapted WSN strain (WSN-CA) was generated by introducing 5 mutations in 3 polymerase genes: PB1 (391E, 581G, and 661T), PB2 (265S), and NP (34G). These 5 mutations were derived from A/Ann Arbor/6/60, which was used as the current vaccine strain for FluMist 47,57,58. Influenza A/X-31(H3N2) and A/California/04/2009 were used for challenge experiments.

293T cells were cultured in DMEM with 10% FBS. MDCK cells were cultured in DMEM with 10% FBS, but changed to Optimum with 0.8mg/ml TPCK trypsin for viral infection with PR8 backgrounds. Human epithelial A549 cells and THP1 cells were cultured in RPMI1640 with 10% FBS. Knock out THP1 cells were generated by CRISPR-Cas9 using a lentivirus expressing a gRNA (targets STING or MAVS) and Cas9-T2A cassette59. THP1 cells were transduced with lentivirus and selected with 5µg/ml puromycin (Life Technologies). Human macrophages were isolated from de-identified donor lungs and cultured as previously described 60. The Committee for Oversight of Research and Clinical Training Involving Decedents and University of Pittsburgh Institutional Review Board approved use of the human tissues.

Construction of influenza mutant library

Influenza A/WSN/33 mutant libraries were generated using the eight-plasmid transfection system as previously described 35–37. In brief, the whole length influenza genome was separated into 52 small 240bp 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 or BsmbI digested, ligated to the vector and transformed using MegaX DH10B T1R cells (Life Technologies). As each small library was expected to have ~1000 single mutations, ~50,000 bacterial colonies were collected to cover the complexity. Plasmids from collected bacteria were midi-prepped as the input DNA library.

Transfection, infection and viral titer of reconstituted viral libraries

To generate the mutant viral library, ~30 million 293T cells were transfected with 32ug DNA. Transfections were performed using Lipofectamine 2000 (Life Technologies). Virus was collected 72h post transfection. TCID50 were measured with A549 cells by observing CPE. To passage viral libraries under natural condition, ~10 million A549 cells were infected with an MOI 0.05. Cells were washed with PBS three times at 2 h post-infection. Virus was collected 24 h post-infection from supernatant. To passage viral libraries under type I interferon treatment, ~10 million A549 cells were pre-treated with 1000U/ml IFNa2 (PBL Assay Science) for 20h and then infected with viral libraries at an MOI 0.05. Cells were washed with PBS three times at 2 h post-infection, and 1000U/ml IFNa2 were added back into culture media. Virus was collected 24 h post-infection from supernatant, clarified of debris and stored at -80 oC in aliquot.

Sequencing library construction and data analysis

Viral RNA was extracted using QIAamp Viral RNA Mini Kit (Qiagen Sciences). DNaseI (Life Technologies) treatment was performed, followed by reverse transcription using superscript III system (Life Technologies). At least 106 viral copy numbers were used to amplify the mutated segment. The amplified segment was then digested with BpuEI or BpmI 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 de-multiplexed using the three-nucleotide ID. Sequencing error was corrected by filtering un-matched forward and reverse reads. Mutations were called by comparing sequencing reads with the wild-type sequence. Clones containing two or more mutations were discarded. Relative fitness index (RF index) in each condition was calculated for individual point mutations and only mutations that have frequency more than 0.05% in the DNA library were reported.

$$RF\ index_{mutant\ i} = Relative\ Frequency\ of\ Mutant\ i_{\ infection} / Relative\ Frequency\ of\ Mutant\ i_{\ plasmid}$$

Where $Relative\ Frequency\ of\ Mutant\ i = Reads\ of\ Mutant\ i\ /\ Reads\ of\ wild\ type$

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

Construction of single mutant viruses

Single mutations were generated using PCR-based site-directed mutagenesis strategy. For mutant on WSN background, 1.5 million 293T cells were transfected with 4ug DNA. Transfections were performed using Lipofectamine 2000 (Life Technologies). Virus was collected 72h post transfection. TCID50 were measured with A549 cells. Mutant viruses were further amplified in A549 cells, then supernatants were collected, clarified of debris and stored at -80 oC in aliquot. Virion RNA was also extracted, and reverse transcribed to cDNA. Coding sequences of all mutants were PCR amplified, and subjected to Sanger DNA sequencing to confirm correct sequence. To measure the growth curve, ~1 million A549 cells were infected with MOI 0.1 and supernatant were collected at the indicated time.

To generate mutant on PR8 background, 1.5 million 293T/MDCK cocultured cells were transfected with 4ug DNA in DMEM + 10% FBS. Media were changed to Optimum + 0.8mg/ml TPCK Trypsin 24h post transfection. Virus was collected 72h post transfection. TCID50 were measured with MDCK cells. Mutant viruses were further amplified in MDCK cells with Optimum + 0.8mg/ml TPCK Trypsin, then supernatants were collected, clarified of debris and stored at -80 oC in aliquot.

Selection of possible IFN sensitive mutations

As previous studies focused on the anti-IFN function of NS1 protein, structural proteins were focused (PB2, PB1, PA, NP, M1) for further validation of novel anti-IFN functional residues .The basic criteria includes: 1) RF scores of fitness > 0.7 2) RF scores of fitness under interferon selection <0.2. 3) Residues that when mutated into different amino acids gives similar interferon sensitive phenotypes are preferred. Moreover, we mapped the list of potential interferon sensitive residues onto protein structures (PDB: 4WSB, 1RUZ, 2IQH, 1EA3) 61–66 and preferably picked the residues that clustered on the protein surface for validation. Up to 7 mutations per segment were selected.

Validation of IFN sensitivity of single mutations

1 million A549 cells were pre-treated with 1000U/ml IFNa2 (PBL Assay Science) for 20h or leave untreated. Then cells were infected with indicated virus at an MOI 0.1. Cells were washed with PBS three times at 2 h post-infection, and 1000U/ml IFNa2 were added back into culture media. Supernatant was collected 24 h post-infection and viral copy number were quantified by RT-qPCR. Relative IFN sensitivity for each mutant were calculated as the fold of copy number drop with IFN selection, as compare with wild type.

Polymerase activity assay

100 ng of each of PB2, PB1, PA, NP, 50 ng of virus-inducible luciferase reporter and 5ng PGK-renilla luciferase were transfected in 293T cells in 24 well plates 67. Cells were lysed 24 h post-transfection and luciferase assay was performed with Dual-Luciferase Assay Kit (Promega).

Immunofluorescence

Localization of IRF3 protein was determined by immunofluorescence. Infected A549 cells were fixed in 2% paraformaldehyde, permeabilized with 0.1% Triton-X100, and then blocked with 3% BSA and 10% FBS. Viral NP protein was detected with anti-NP monoclonal antibody (GeneTex). IRF3 were detected with anti-IRF3 rabbit polyclonal antibody (Cell signaling). Hoechst 33342 dye was used for nucleic acid stain.

RNA extraction, reverse transcription and real-time PCR

Viral RNA was extracted using QIAamp Viral RNA Mini Kit (Qiagen Sciences). Total cellular RNA was extracted from infected cells with Purelink RNA Mini Kit (Ambion) or Trizol

(Thermo Fisher Scientific) following product manuals, reverse transcribed by Superscript III Reverse Transcriptase (Thermo Fisher) or qscript (Quanta Biosciences). Quantitative real time PCR was performed using Taq polymerase and SYBG. For viral copy number, standard curve ranging from 103 to 108 was used for calculation and normalized to mock or WT infected cells. For cellular RNA, results were calculated using the 2-ΔΔCT method (citation). Sequences of primers used to quantify RNA are as follows:

	Gene	Forward	Reverse
Human	ISG15	CAT GGG CTG GGA CCT GAC G	CGC CAA TCT TCT GGG TGA TCT G
	ISG54	CAG CTG AGA ATT GCA CTG CAA	GTA GGC TGC TCT CCA AGG AA
	RIGI	TGC GAA TCA GAT CCC AGT GTA	TGC CTG TAA CTC TAT ACC CAT GT
	OAS	CAA GCT CAA GAG CCT CAT CC	TGG GCT GTG TTG AAA TGT GT
	IFNb	AGG ACA GGA TGA ACT TTG AC	TGA TAG ACA TTA GCC AGG AG
Mice	Ifnb	CCC TAT GGA GAT GAC GGA GA	CCC AGT GCT GGA GAA ATT GT
	Ifng	ACA GCA AGG CGA AAA AGG AT	TGA GCT CAT TGA ATG CTT GG
	Mx1	TCT GAG GAG AGC CAG ACG AT	CTC AGG GTG TCG ATG AGG TC
	Rnaseh	GAT TCC AAG AAA GCT GTC CG	TAG AGA ACC AGC CGT CCA AG
Influenza	NP	GAC GAT GCA ACG GCT GGT CTG	ACC ATT GTT CCA ACT CCT TT

Examination of mutation combinations

To maximize the sensitivity to IFN and maintain viral growth capacity, we examined 122 different combinations of mutations for their viral growth. Combinations were generated by randomly choosing one genotype from the following 5 categories:

PB2	M1	PB1	PA	NS1
WT	WT	WT	WT	WT

N9D	R72Q	L155H	E181E	R38A+K41A
N9D+T76A	R72Q+S225T			
N9D+Q75H+T76A	N36Y+R72Q+S225T			

1.5 million 293T cells were transfected with 4ug DNA containing corresponding mutations. Virus was collected 72h post transfection and used to infect 293T cells overexpressing virus-inducible mCherry reporter in 48 well format. WT virus were used as control. mCherry intensity were examined under microscope at 24h and 48h post infection, which proportional to the growth capacity of corresponding virus. Viable mutant combination that contain the largest amount of mutations were considered as possible vaccine candidate.

RNA-Seq library preparation and sequencing

RNA-seq libraries were prepared using a modified method based on ScriptSeq mRNA-Seq library preparation kit (Epicentre)68. Multiplex sequencing was performed by 50bp single-end read with Illumina HiSeq 2000 machine at UCLA Clinical Microarray Core. Raw reads were aligned to human genome assembly (hg19) or mouse genome assembly (mm10) using TopHat under default parameter69,70. Results were quantified by reads per million total reads (RPM). Differential expression analysis were performed with edgeR 71. Gene ontology enrichment analysis was done through metascape 72. Genes related to certain cellular pathways were extracted from MsigDB46.

Mouse studies

6-8 weeks old female BALB/c mice were used for safety and antibody studies. 6-8 weeks old female C57BL/6J mice were used for T cell studies. IFNAR-/- mice were used for viral growth studies.

Firstly, 6-8 weeks Balb/c mice were anesthetized with isoflurane (IsoFlo, Henry Schein) and intranasal inoculated with WT virus or DAI virus at a dose of 105 or 106 TCID50 in a volume of 30ul. Body weight loss were monitored daily for 14 days. To quantify viral growth in mice lung tissues, mice were intranasal inoculated with indicated virus at a does of 105 TCID50 and sacrificed at 2 days post infection. DMEM media were used as control for infection. To quantify viral titer by TCID50, lung tissues were harvested, homogenized, and freeze-thaw three times to release the virus. To quantify viral copy number, RNA was extracted from mice lung tissue using Trizon (Thermo Fisher Science). Similarly, to quantify the viral growth curve and gene expression in mice lung tissue, mice lung samples were collected at 2, 6, 24, 48 and 120 hours post infection. Viral copy number and gene expression were quantified by qRT-PCR. All animal experiments were performed in accordance with the guidelines of the animal protocol approved by UCLA.

For pathology and cytokine studies, 6-8 weeks old female BALB/c mice were intranasal inoculated with indicated virus at a does of 105 TCID50. Bronchoalveolar lavage (BAL) samples were collected at Day 2 and Day 9 post infection. Albumin concentration was determined using mouse albumin ELISA Quantitation kit (Bethyl Alboratories Inc) and cytokine response was analyzed by Lincoplex according to manufacturer’s guide (BioRad). BAL cell cytospin slides were stained with a HEMA-3 stain kit (Fisher Scientific) for inflammatory cell differential counts. In addition, lung tissues were fixed with 10% neutral buffered formalin (EMD Millipore, Billerica, MA) and paraffin embedded for histology. Hematoxylin and Eosin (H&E) stained lung tissue slides were scored for their pathology (Score 1-5) :

1 = no observable pathology.

2 = perivascular/peribronchus or lung parenchyma inflammatory infiltration < 25% of the lobe section.

3 = perivascular/peribronchus or lung parenchyma inflammatory infiltration 25%-50% of the lobe section.

4 = perivascular/peribronchus or lung parenchyma inflammatory infiltration 50%-75% of the lobe section.

5 = perivascular/peribronchus or lung parenchyma inflammatory infiltration >75% of the lobe section.

For antibody studies, 6-8 weeks old female BALB/c mice were intranasal inoculated with indicated virus at a does of 105 TCID50. DMEM media was used as control. Sera samples were collected at 14, 21 and 28 days post infection (N=3). Sera and BAL samples were collected at 28 days post infection (N=4). Sera samples were used for immunoglobulin G (IgG) antibody detection, hemagglutination inhibition assays and neutralization assay. They are also used for a screen of antibody escape mutations on HA protein. BAL samples were used for immunoglobulin A (IgA) antibody detection.

For T cell studies, 6-8 weeks old female C57BL/6J mice were intranasal inoculated with indicated virus at a does of 105 TCID50. To examine primary T cell response, lung and spleen were harvested at 10 days post infection. Fresh cells were used for tetramer staining with flow cytometry. For peptide stimulation, spleens were harvested at 28 days post infection. To examine secondary T cell response, lung and spleen were harvested at 14 days post challenge of indicated virus after 28 days vaccination of DAI virus.

For protection studies, 6-8 weeks old female BALB/c mice were intranasal inoculated with indicated virus at a does of 105 TCID50. DMEM media were used as control. Mice were then challenged with 105 TCID50 WT. Viruses were quantified by both TCID50 and RT-qPCR assay at day 2 post challenge.

To examine homologous and heterologous protection, 6-8 weeks old female C57BL/6J mice were intranassally vaccinated with virus at a does of 105 TCID50 or PBS as control. At day 28 post vaccination, PR8 (H1N1), Cal/04 (H1N1) and WSN (H1N1) were used for intranasal challenge at a dose of LD90 which was 6000 EID50, 8000 EID50 and 14,000 EID50 respectively. X-31(H3N2) was given at 45,000 EID50 (LD50). Mice body weight loss was monitored twice daily for 14 days. Clinical score were used to quantify the clinic symptoms with:

0 = no visible signs of disease
1 = slight ruffling of fur
2 = ruffled fur, reduced mobility

3 = ruffled fur, reduced mobility, rapid breathing
4 = ruffled fur, minimal mobility, huddled appearance, rapid and/or labored breathing
5 = found dead

Enzyme-linked immunosorbent assay (ELISA)

Viral protein specific IgG and IgA were detected using Enzyme-linked immunosorbent assay (ELISA). 96 well ELISA plate (Costar, Corning Inc, USA) were coated with 1ug/ml recombinant viral proteins (HA-WSN, HA-PR8, HA-HK68, HA-Viet04, NP, NA and M1) in bicarbonate/carbonate buffer at pH 9.5 at concentration of 1ug/ml at 4°C over night. Wells were washed with PBST between each steps for 3-5 times. Wells were then blocked with 10% FBS in PBS for 1h at room temp. Serum samples for viral protein-specific IgG detection, or BAL samples for virus-specific IgA detection, were diluted in blocking buffer and added to wells for 4°C over night. HRP-conjugated anti-mouse IgG antibody (Cell Signaling) or HRP-conjugated anti-mouse IgA antibody (Thermo Fisher) were diluted in blocking buffer and added into wells for 1h at room temp. Then SIGMA FAST OPD (Sigma) was used to detect the plate at OD 450nm.

Hemagglutination Inhibition (HAI) assay

Mice sera were pre-treated at 56°C for 30 min. 4 HA units of WT virus were incubated with 2 fold serially diluted sera in room temp for 1h in V shaped 96 well plate. The starting concentration of sera is 1:4. Washed turkey red blood cells (Lampire) at a concentration of 0.5% were added into each well and incubated at room temp for 30 min. The HA titer is then read as the highest dilution of serum that prevents hemagglutination.

Antibody Neutralization assay

Sera were pre-treated at 56°C for 30 min. WT virus were incubated with serial diluted sera in 37°C for 1h. Control virus was incubated with PBS. Then TCID50 were measured using A549 cells.

Detection of viral-specific CD8 T cells

Tetramer staining and flow cytometry were used to detect viral specific CD8 T cells. Lung tissues were minced and digested with 5mg/ml collagenase A for 30min at 37°C. Then cells were transferred through a single cell strainer and digestion was stopped with HBSS/2mM EDTA for 10min at room temp. Spleen samples were directly transferred through the single cell strainer to make single cell suspension. Red blood cells were lysed using ACK lysis buffer and reaction was stopped with 10% FBS in DMEM media. Cells were washed with FACS buffer (2% FBS in PBS) and stained with tetramer against NP epitope (H-2Db + influenza A virus NP366–374 , ASNENMETM) and PB1 epitope (H-2Kb influenza A virus PB1703-711, SSYRRPVGI), conjugated with Phycoerythrin (PE). Then cells were washed with FACS buffer and stained with anti-mouse CD3-eflurofore450 and anti-mouse CD8a-FITC antibody (eBioscience). Samples were analyzed with Forsetta flow cytometer.

Screen of antibody escape mutations

Inhibition of mice serum for WT replication was measured by antibody neutralizing assay. Reconstituted viral HA single nucleotide library (estimated MOI=0.1) were incubated with serum at concentration of IC80 at 37°C for 2h. Unvaccinated mice serum was used as mock control. 30 million A549 cells were infected by each library after serum incubation. PBS wash was performed twice at 2h post infection. Then serum was added into culture media that matched IC80. Supernatants were collected 48h post infection, viral RNA was extracted and reverse transcribed. Relative frequencies of each mutant were determined as described above.

Relative enrichment score (RE score) of each mutant were calculated by comparing the relative frequency of mutant under serum selection (WT or DAI) compare with mock serum selection. Mutant with RE score >5 were selected for each selection condition as possible serum escape mutations. Mutations occur in at least two mice serum in the WT or DAI group were selected and compared.

Peptide stimulation of spleenocytes

Single cell suspension is prepared from spleen tissues and stimulated with indicated peptide for 6h at the concentration of 1uM, with the presence of 1ug/ml brefeldin A. Cells were stained with CD3-eflurofore 450 and CD8-FITC for 15min at 4°C. Then cells were washed with FACS wash buffer twice and fixed with IC fixation buffer (eBioscience) for 20min at room temperature. Without washing, 2ml of 1X permeabilization buffer (eBioscience) were added into each sample and centrifuge at 1600rpm at room temperature for 5min. Samples were washed with FACS wash buffer twice. Then IFNg-PE antibody was used to stain for 1h at room temperature. Samples were analyzed with Forsetta flow cytometer.

Three peptides (synthesized by Thermo Fisher Scientific) were used for stimulation:

NP 366-374: ASNENMETM.
PB1703-711: SSYRRPVGI.
PA224-235: SSLEKFRAYV.

Detection of viral-specific CD4 T cells

Spleenocytes were stimulated with WT WSN virus for 16h at 37° C with the presence of 1ug/ml brefeldinA. Cells were stained with CD3-eflurofore 450, CD4-FITC and intracellularly stained with IFNg-PE. Percentage of IFNg positive CD4 cells was quantified by flow cytometry.

T cell repertoire sequencing

TCRB gene of influenza NP366–374 specific CD8 cells and lung infiltrating CD8 T cells were sequenced.

To isolate NP-specific CD8 cells, mice lung and spleen tissue were harvested and single cell suspension were generated. T cell population were enriched using EasySep Mice T cell isolation kit (Stemcell) and stained with CD3-eflurofore450, CD8a-FITC and NP366–374 tetramer conjugated with PE. NP positive CD8 T cells were sorted out. For lung CD8 cells, lung tissues were harvested at 10 days post infection from WT, DAI and mock infected mice. CD8 T cells were enriched using EasySep Mice CD8 T cell isolation kit (Stemcell). DNA from isolated T cells was extracted using QLAamp DNA blood mini kit (Qiagen). TCRB genes were amplified and deep sequenced with immunoSEQ assies from Adaptive Biotechology.

VDJ recombination was analyzed for each sample. Clonality was calculated for NP-specific T cells as the inverse of the normalized version of Shannon's entropy. Expanding T cells were defined as the ones with TCRB rearrangement detected in NP specific T cells and occure > 0.1% in the total population, but did not detected in the mock control.

Statistical analysis

All numerical data were calculated and plotted with mean+/- SD. Results were analyzed by unpaired Student's t test. Differences were considered statistically significant when $p<0.05$ (*) or $p<0.01$ (**).

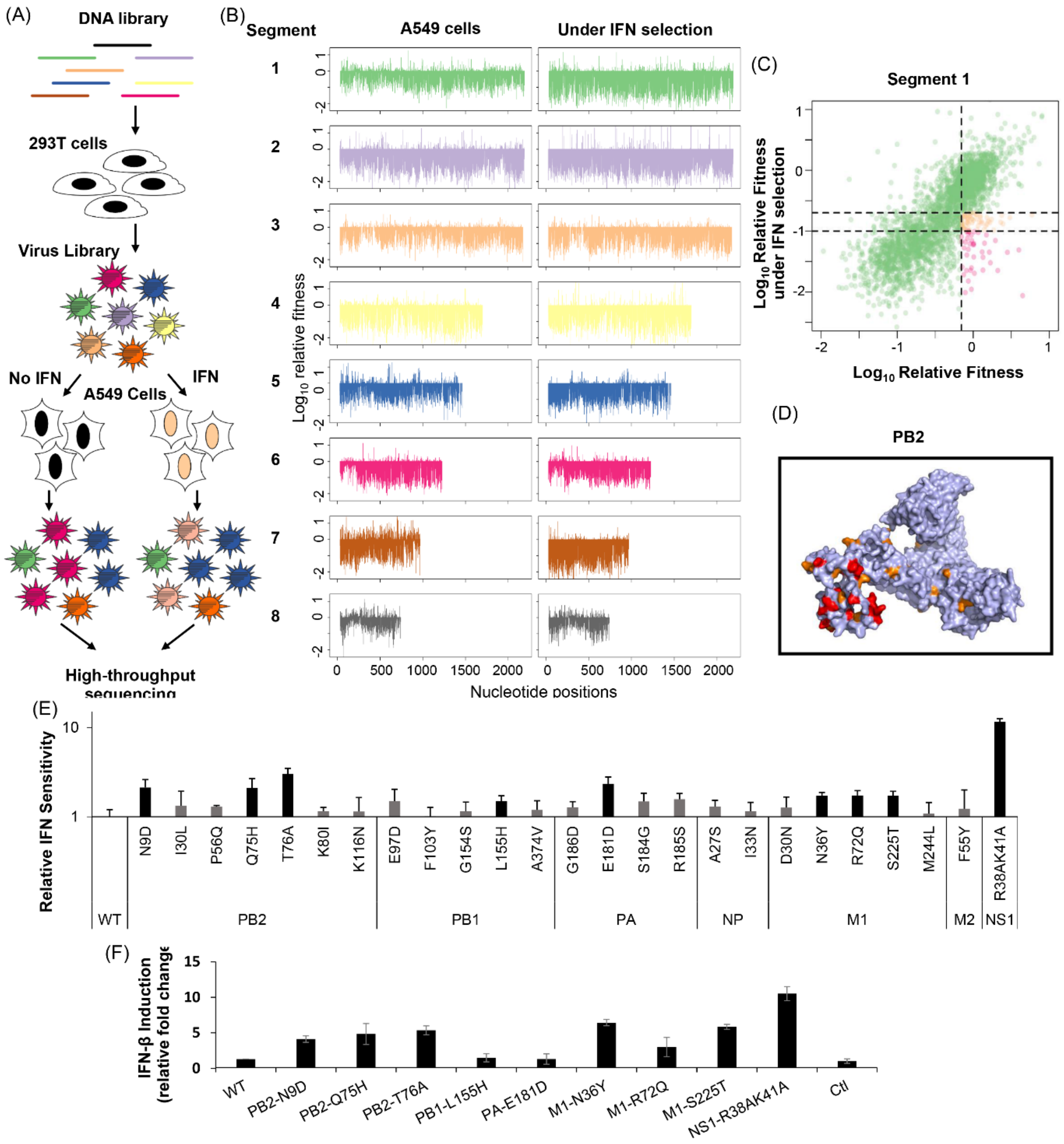


Figure 6-1. Systematically identification of IFN-sensitive mutations using high-throughput genetic system

(A) A schematic presentation of the experimental flow of high-throughput genetic system and its utilization to systematically identify IFN sensitive mutations. (B) Relative fitness scores (RF scores) were shown for individual mutations in A549 cells without (left panel) and withIFN selection (right panel). (C&D) Selection of possible IFN-sensitive mutations across A/WSN/33 genome. RF scores with and without IFN for each mutation were plotted as scatter plots. Mutations with RF scores > 0.7 without IFN and < 0.3 (orange) or <0.2 (red) with IFN treatment were highlighted (C) and mapped onto protein structures (D). PB2 protein was shown as example (PDB: 4WSB). (E) Validation of IFN sensitivity with twenty-four single mutations (N=3).

8 mutations shown in black bars were significantly more sensitive to IFN compared with WT (Two tail T test, p<0.05). (F) Induction of IFN-β gene expression was shown for these 8 mutations as indicated (N=3).

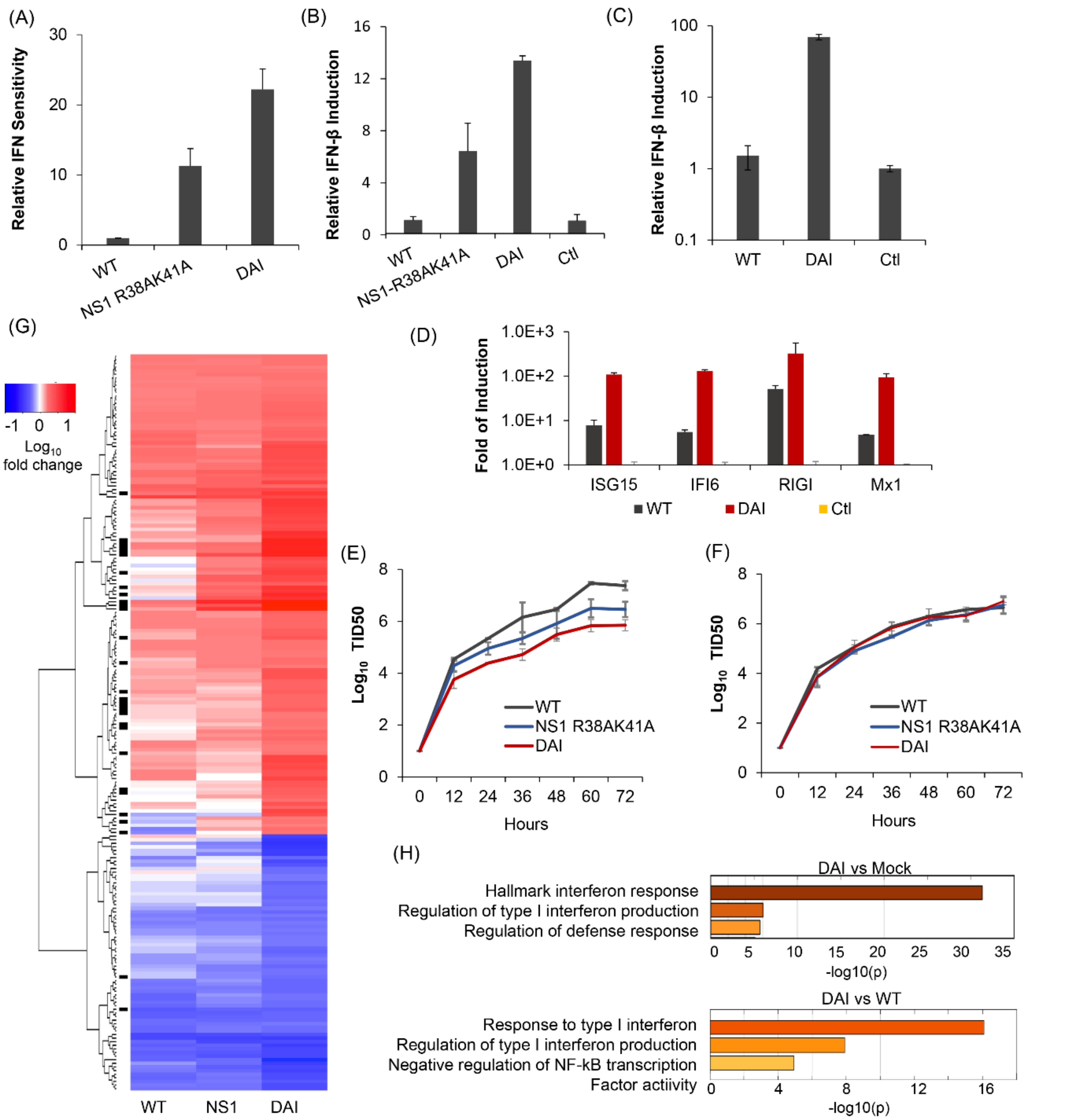


Figure 6-2. Combination of mutations (DAI) generates stronger IFN sensitive and IFN induction phenotype

(A) IFN sensitivity of DAI virus compared with WT and NS1 R38A-K41A mutation (N=4). (B&C) Induction of IFN-β gene expression by DAI virus in (B) A549 cells and (C) primary human alveolar macrophages (N=3).Cells were infected with indicated virus for 6h with MOI 1. (D) Induction of indicated ISGs in THP-1 cells infected with indicated virus for 24h (with MOI 0.1

(N=2). (E&F) Replication of WT, NS1 R38A-K41A, and DAI virus in (E) A549 and (F) Vero (F) cells by TCID50 assay. (G) Global gene expression change in A549 cells infected with indicated virus at MOI 1 for 6h by mRNA sequencing (N=2). Heat map of genes that were significantly upregulated or downregulated in DAI infected A549 cells (fold change >2 and p<0.001). IFN pathway related genes are marked at the left side. (H) GO analysis were shown for the upregulated genes of DAI infected A549 cells compare with mock infected cells (up panel) and WT infected cells (bottom panel).

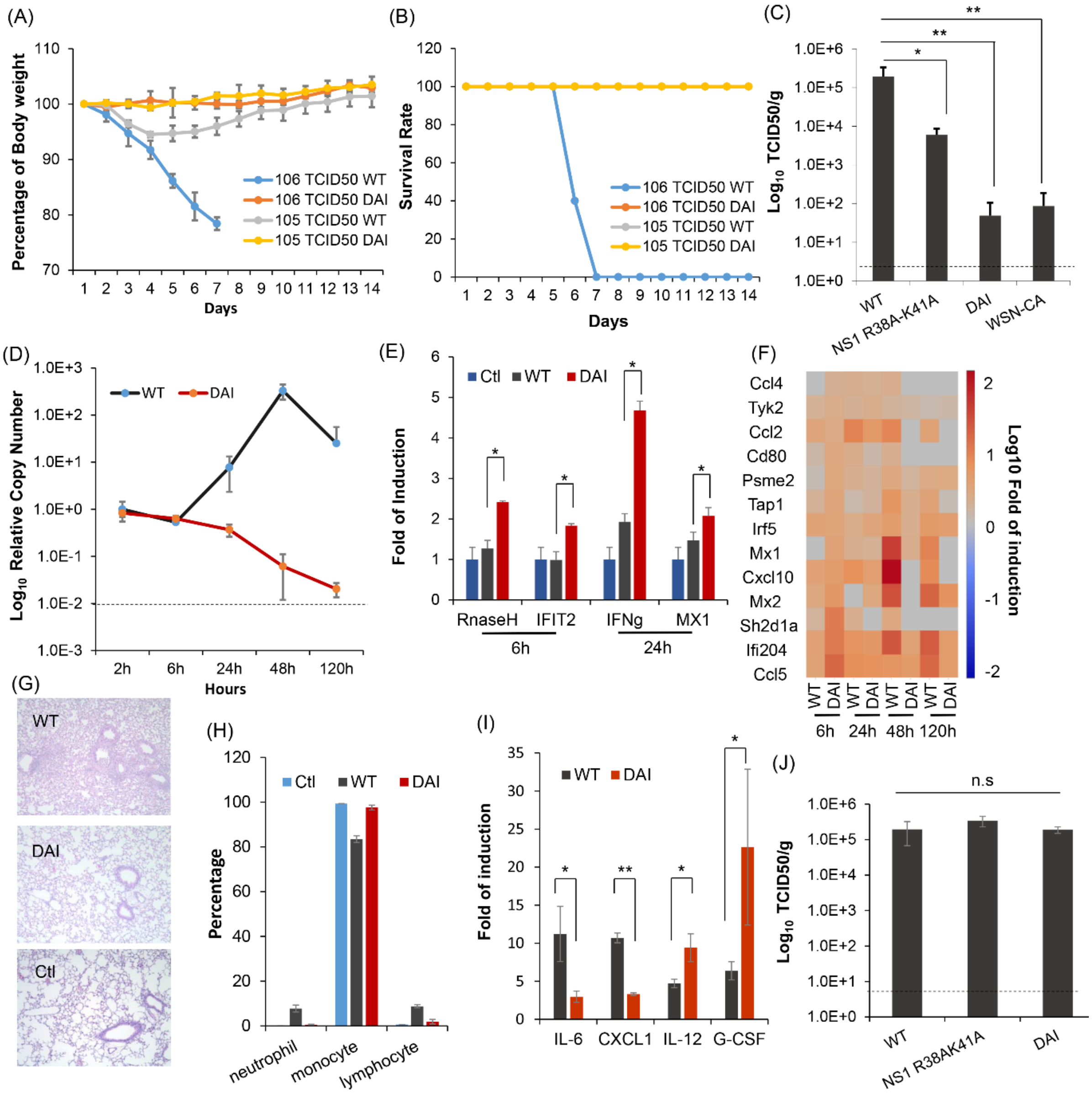


Figure 6-3. DAI virus is replication deficient in vivo

(A&B) Pathogenesis of intranasal infection of indicated virus was shown with (A) percentage of body weight loss and (B) survival rate (N=5). 20% or more body weight loss was considered lethal. (C) Viral titers in mice lung tissue at 2 days post infection (N=3). (D) Replication kinetics of of WT and DAI virus in mouse lung tissues (N=3). Viral copies were quantified by RT-qPCR and normalized to the WT infected lung tissue at 2 h post infection. Dashed line represents mock infected mice. (E) Transcripts of indicated ISGs were examined for mouse lung samples infected with WT or DAI virus at 6h or 24h (N=2). (F) Induction of IFN inducible genes for WT and DAI infected mouse lung samples were examined by mRNA sequencing. Data is shown with fold of induction compared with mock-infected lung samples as heatmap (N=2). (G&H) HE staining of lung tissues (G) and BAL cytospin (H) of WT and DAI infected mice at day 9 post infection. (I) Levels of indicated cytokines in BAL samples collected at day 2 post infection by luminex multiplex assay (N=3). (J) Replication of indicated virus in lung tissues of IFNAR-/- mice at day 2 post infection (N=3).

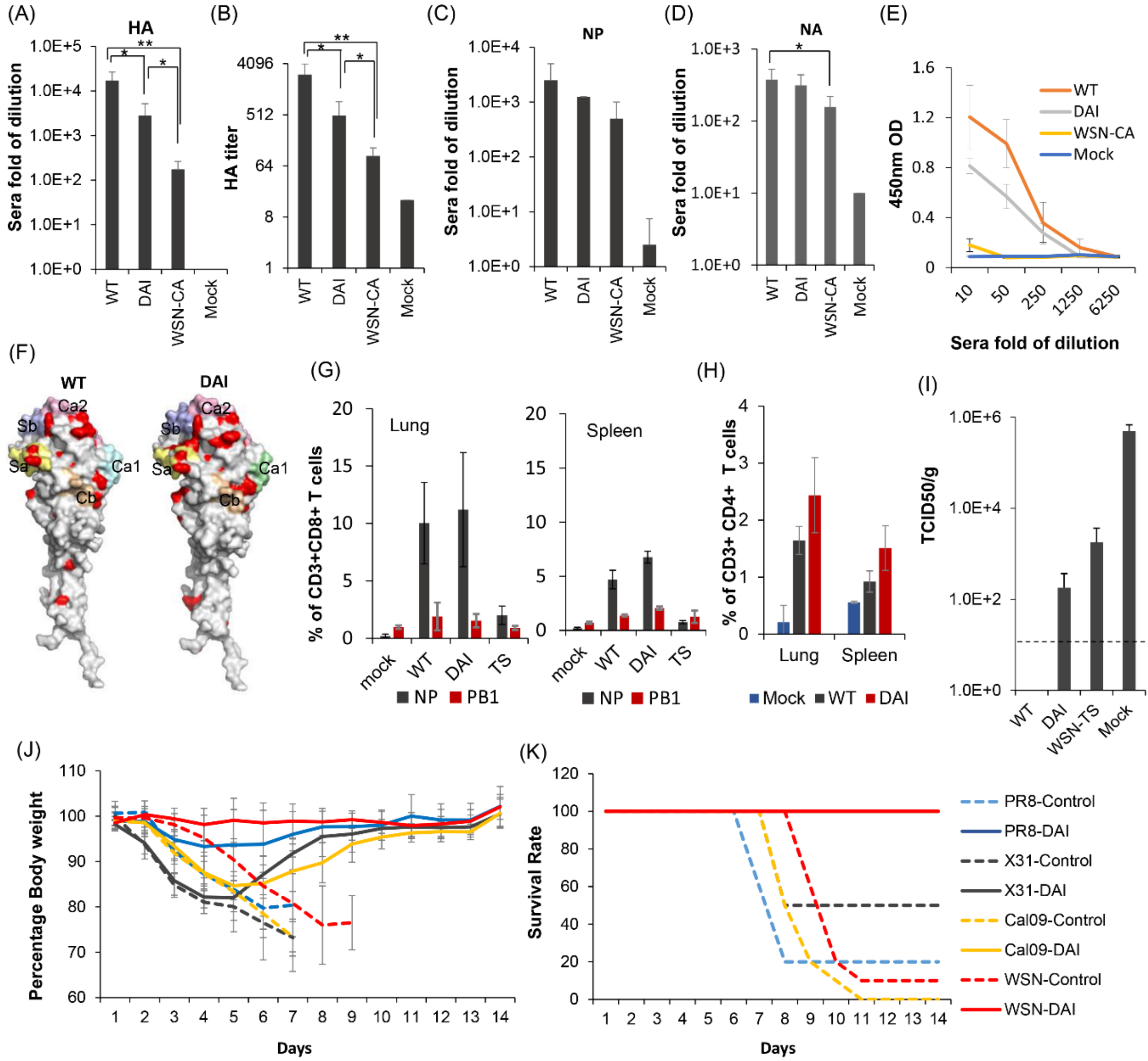


Figure 6-4. DAI virus induces strong and board adaptive immune response and protects mice from challenge

(A&B) Specific HA binding IgG and neutralizing antibodies from infected mice sera were examined by (A) ELISA and (B) hemagglutinin inhibition (HAI) assay (N=4). (C&D) Specific (C) NP (D) NA binding IgG from infected mouse sera were examined by ELISA (N=4). (E) HA binding IgA from BAL were examined by ELISA (N=4). (F) Possible neutralizing antibody escape mutations selected under mouse sera were mapped onto HA structure (N=5). (G) Flow analysis of Primary T cell response by influenza A virus NP366–374 (NPP, ASNENMETM) and PB1703-711 (SSYRRPVGI) tetramer staining (N=4). (H) CD4 T cell response were examined by IFNg intracellular staining (N=3). (I) Protection of WT infection were quantified by the viral load lung samples of vaccinated mice at day 2 post challenge (N=4). (J&K) Protection of DAI vaccinated mice from challenge of homologous and heterologous strains is shown with percentage of body weight loss and (J) survival rate (K) daily for 14 days (N=10).

6.6 Supplementary Materials

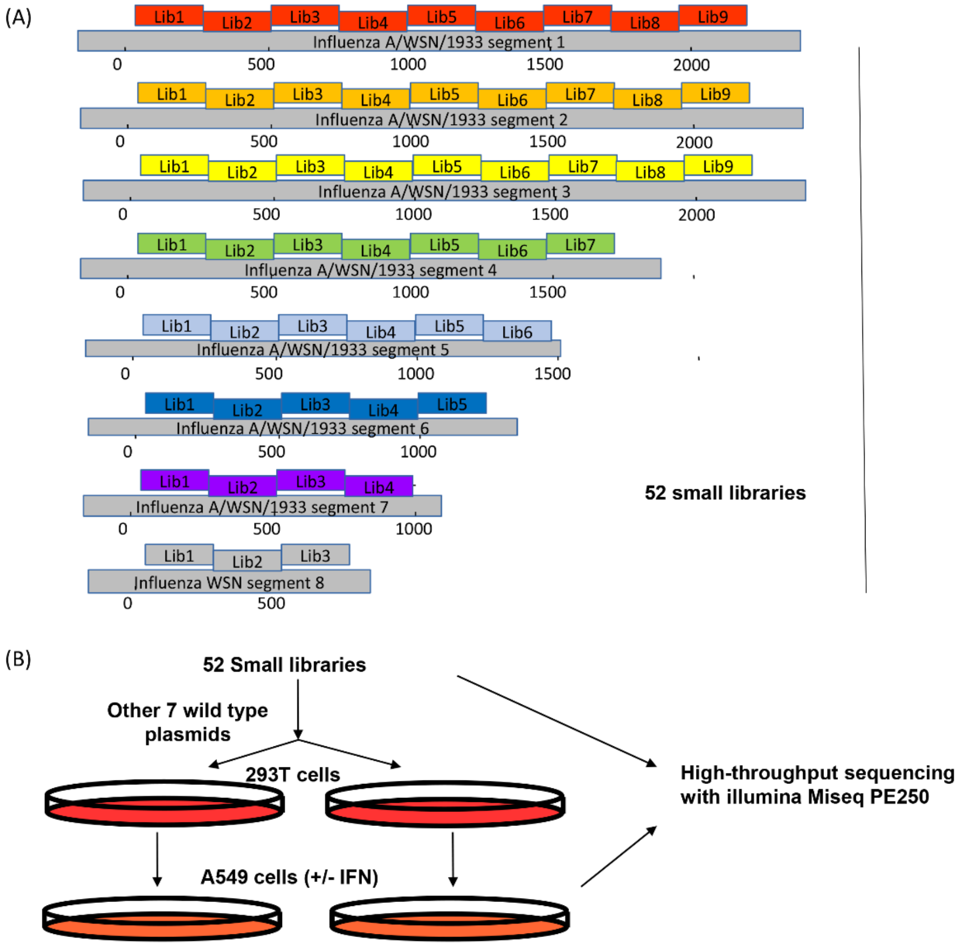

(A)
Lib1 Lib2 Lib3 Lib4 Lib5 Lib6 Lib7 Lib8 Lib9
Influenza A/WSN/1933 segment 1
0 500 1000 1500 2000
Lib1 Lib2 Lib3 Lib4 Lib5 Lib6 Lib7 Lib8 Lib9
Influenza A/WSN/1933 segment 2
0 500 1000 1500 2000
Lib1 Lib2 Lib3 Lib4 Lib5 Lib6 Lib7 Lib8 Lib9
Influenza A/WSN/1933 segment 3
0 500 1000 1500 2000
Lib1 Lib2 Lib3 Lib4 Lib5 Lib6 Lib7
Influenza A/WSN/1933 segment 4
0 500 1000 1500
Lib1 Lib2 Lib3 Lib4 Lib5 Lib6
Influenza A/WSN/1933 segment 5
0 500 1000 1500
Lib1 Lib2 Lib3 Lib4 Lib5
Influenza A/WSN/1933 segment 6
0 500 1000
Lib1 Lib2 Lib3 Lib4
Influenza A/WSN/1933 segment 7
0 500 1000
Lib1 Lib2 Lib3
Influenza WSN segment 8
0 500
52 small libraries
(B)
52 Small libraries
Other 7 wild type plasmids
293T cells
A549 cells (+/- IFN)
High-throughput sequencing with illumina Miseq PE250

Figure S6-1. Construction of single nucleotide mutant libraries across the entire influenza genome. (A) The arrangement of small mutant libraries is shown. Mutant plasmid libraries were constructed on the backbone of influenza A/WSN/1933 (H1N1) strain. In order to control mutant library size, we divided the entire viral genome into 240bp small fragments. For each small segment, error-prone polymerase were used to introduce random single nucleotide mutations to generate a mixed mutant population. Each resulting plasmid population is considered a mutant plasmid library, with ~1000 different mutations. A total of 52 small plasmid libraries are established to cover the whole genome.

(B) Schematic plot was shown for the experimental procedures. 30 million 293T cells are transfected with the each plasmid library together with seven other wild type plasmids to reconstitute the mutant virus library. 15 million A549 cells were then used to passage the viral library with MOI 0.05 for 24h. Biological duplications were conducted for both transfection and infection steps. Mutant plasmid library, mutant viral library after transfection and infection were prepared for Illumina Miseq with 250bp paired-end reads. At least 20000 reads depth were obtained for each nucleotide position. Relative fitness score (RF score) of each single nucleotide mutation were calculated as the ratio of relative frequency in the infection library compared with DNA library. To further increase the measurement of viral fitness, we filtered the mutations that only occur < 0.05% in the input library. After all the quality control, we obtained the high quality fitness data for over 75% across the genome under natural condition or under IFN selection.

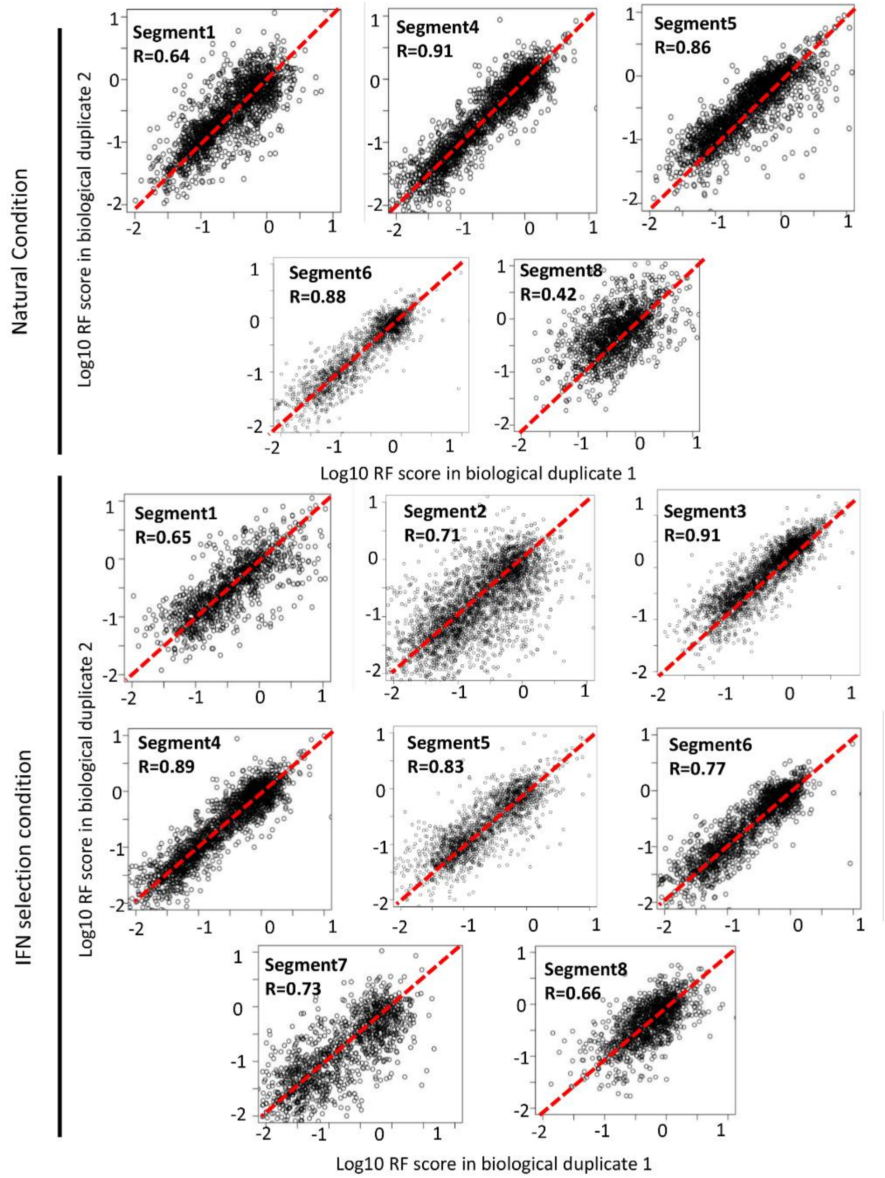


Figure S6-2. Correlation of RF scores from independent biological replicates. Scatter plots shown the correlation between RF scores in biological replicates. Reconstitution and selection of mutant libraries were performed for each segment separately. Independent biological replicate were introduced for both transfection and infection steps. Strong correlation were obtained for the RF scores between replicates across all segments. Mutant RF scores without IFN selection were reported before for segment 2, 3 and 7, thus were not shown here. The final RF scores for each mutants were calculated as the average score between two replicates.

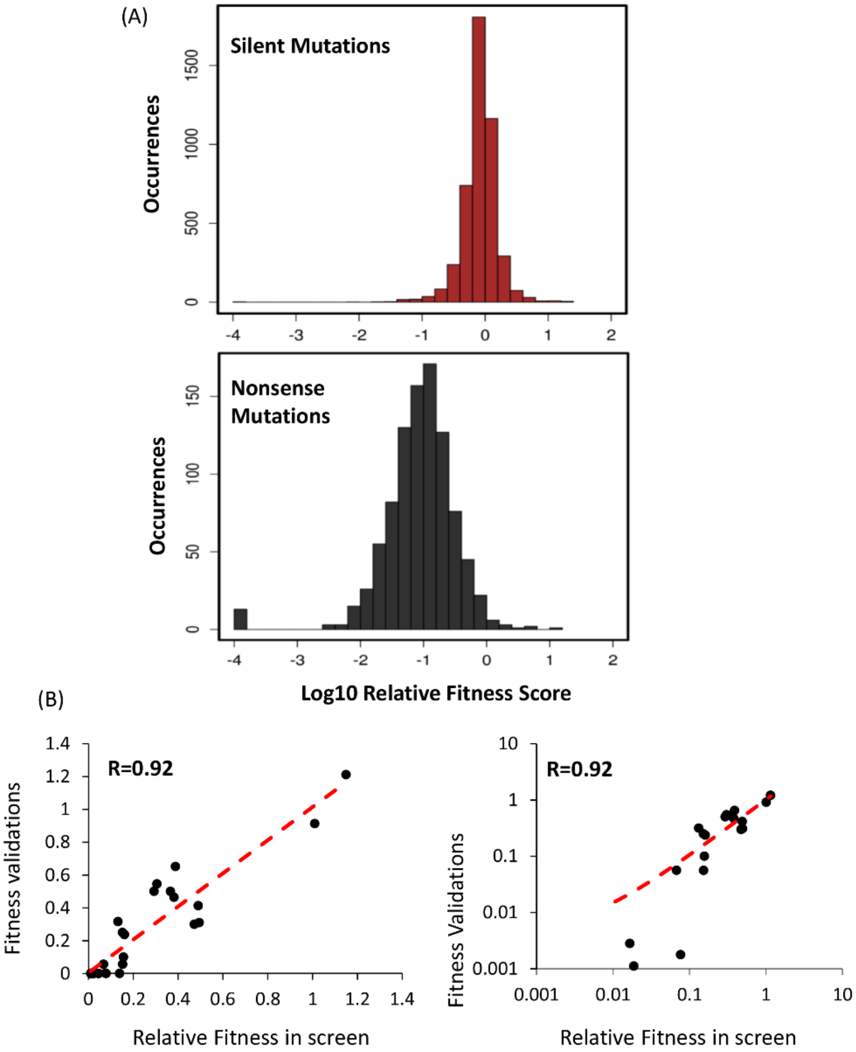


Figure S6-3. Validations of high throughput fitness profile. (A) Histograms of the RF scores of silent and nonsense mutations were shown. RF score of silent mutations were centered around 1 (log10 RF scores centered around 0, mean=0.08, s=0.28), suggesting the majority of silent mutations are neutral for viral replication. A clear separation of distributions were observed between silent and nonsense mutations, suggesting sufficient selection pressure in our system. (B) 26 single mutations were randomly selected across the genome and reconstructed individually in the content of the whole virus. Relative growth capacity of these single mutations were examined by TCID50 assay and compared with RF scores in the screen. As 7 of the single mutants have a TCID50 below detection limit, scatter plot were shown both in

linear (left) and log (right) scale. The relative growth validated as single mutant correlated well with screen data.

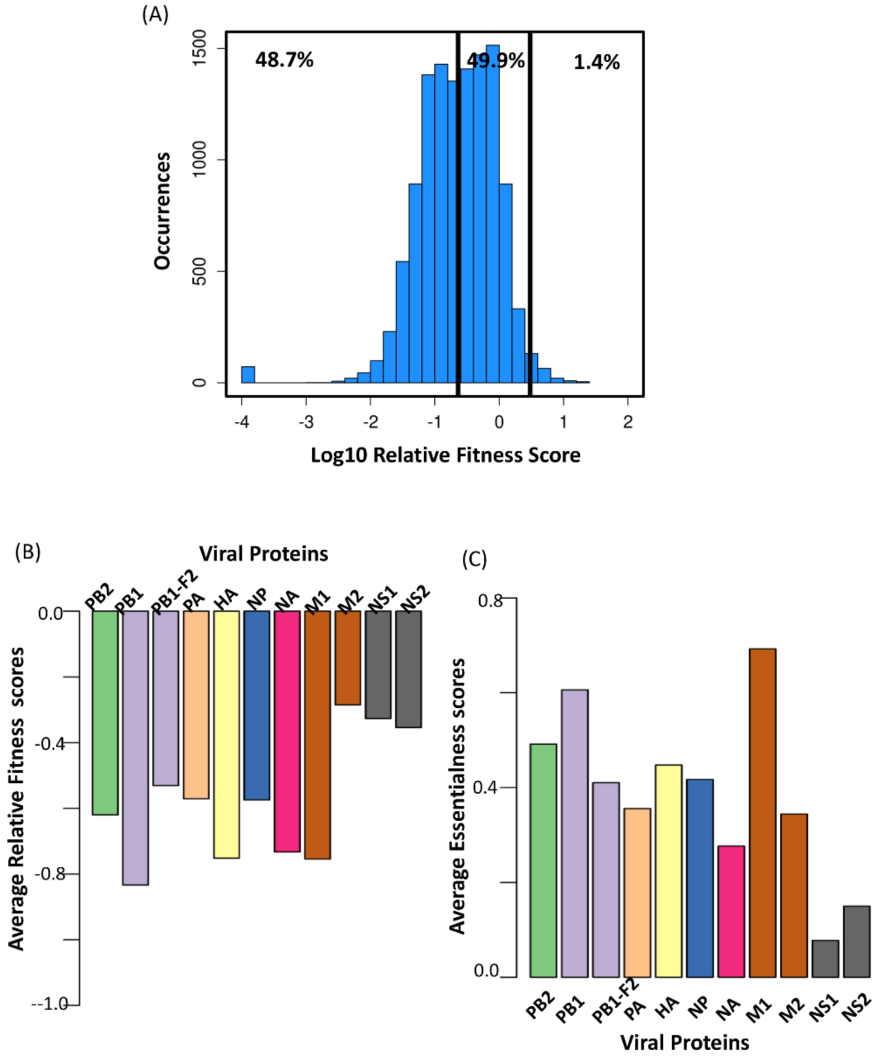


Figure S6-4. Distribution of missense mutations and relative conservation of different viral proteins. (A) Distribution of fitness effect for all single nucleotide missense mutations were shown. Using silent mutations as control, we define a mutation to be deleterious if the RF score < mean-2s, to be beneficial if the RF > mean+2s, and to be nearly neutral if within. Across the entire genome, around 48.7% single nucleotide missense mutations are lethal and only 1.4% are beneficial. (B&C) Different influenza viral proteins showed diverse level of mutation toleration, shown here with two indexes: average RF scores and essentialness score. Average RF score for each protein were calculated across all the missense mutation. The essentialness score is the percentage of missense mutations that are highly deleterious or lethal (RF scores <0.1). The results correlates well between these two indexes, with the polymerase complex related proteins (PB2, PB1, PA, NP) shows high conservation while NS1 and M2 are relative tolerable of single nucleotide mutations.

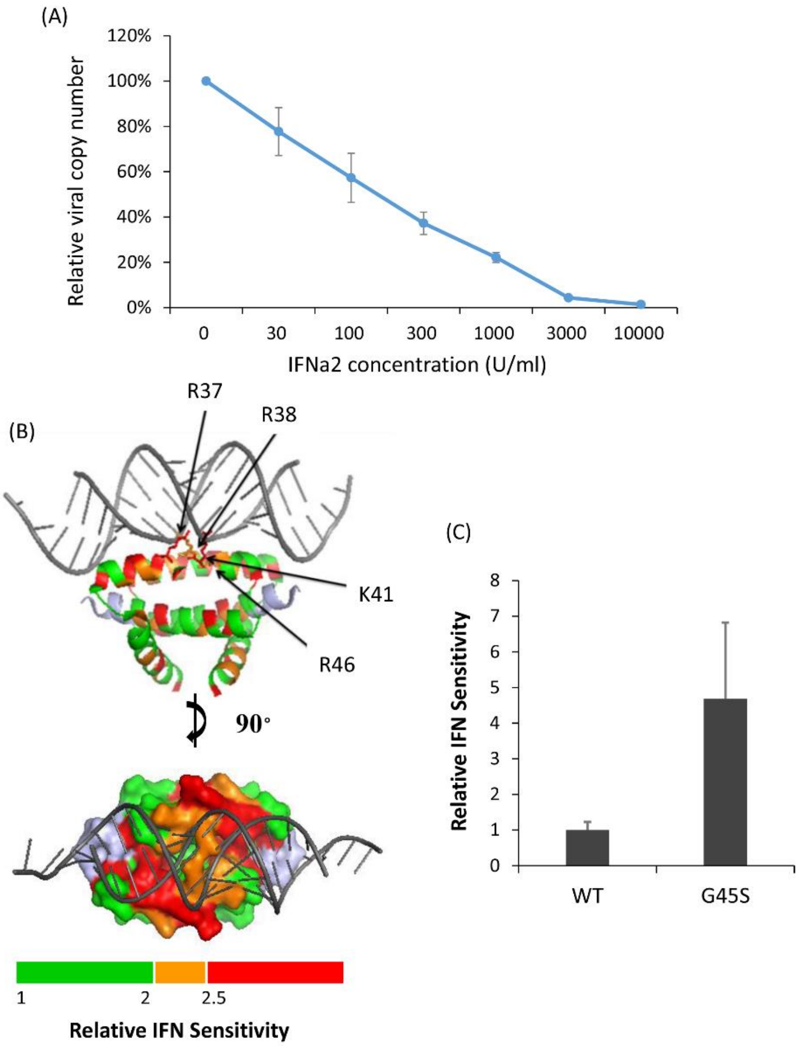


Figure S6-5. Interferon selection of mutant library. (A) Does response curve of Type I IFN (IFNa2) on wild type WSN (WT) viral replication were measured. A549 cells were pre-treated with different doses of IFNa2 for 20h, and infected with WT virus at MOI equal 0.1. Cells were washed twice with PBS at 2 hours post infection and equal concentration of IFNa2 were added back into the cells. Supernatant were collected 24h post infection and viral copy number were measured by RT-qPCR. 1000U/ml were selected as the concentration for screening, which is at IC80. (B &C) IFN sensitive mutations on NS1 RNA binding domain were mapped onto protein structure. Consistent with previous reports, were identified R37, R38, K41, R46 in the RNA binding domain as key residues interfering with IFN function (B). These amino acids cannot tolerant mutations under interferon selection, but did not impede viral replication without interferon treatment. Other unreported mutations on NS1 RNA binding domain were also identified to be highly interferon sensitive, such as G45S (C). G45S were reconstructed individually in the content of the whole virus. Relative IFN sensitivity were validated and compared to wild type WSN virus. Data is presented as means ± s from a biological duplicate.

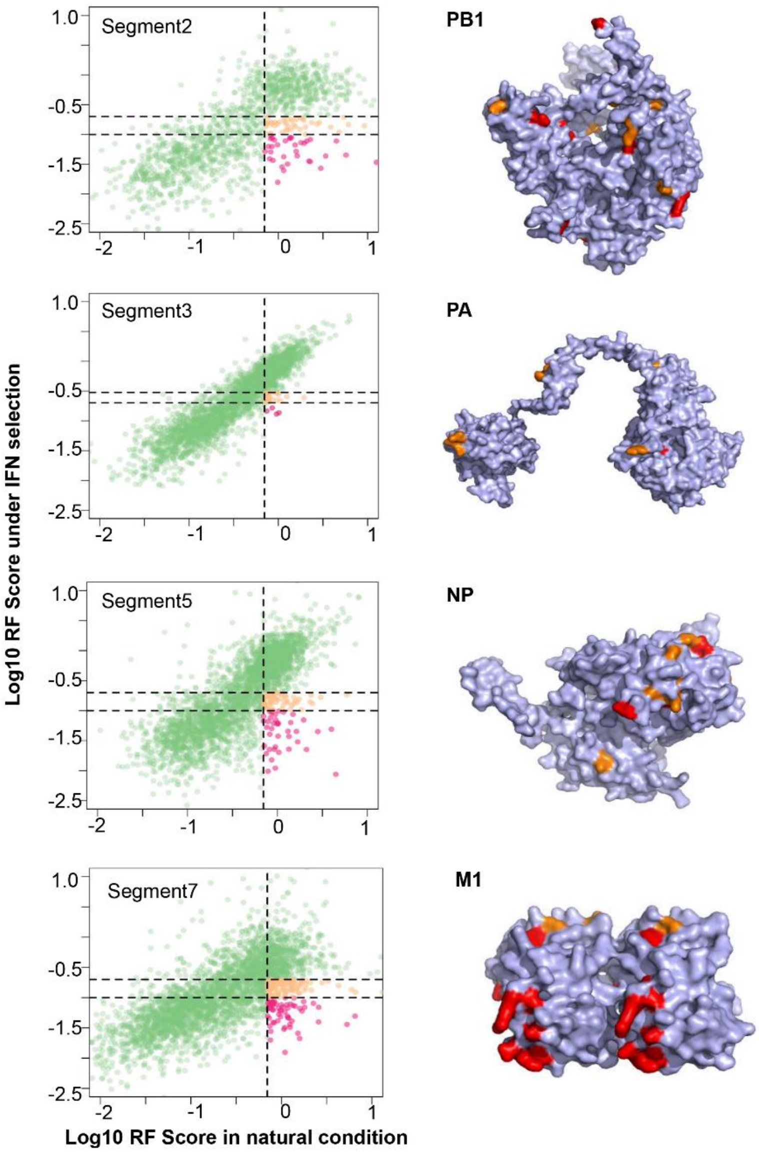


Figure S6-6. Selection of possible IFN sensitive mutations for validation. Selection of possible IFN sensitive mutations for validation were shown. Structural proteins were focused (PB2, PB1, PA, NP, M1).The basic criteria includes:1) RF scores of fitness > 0.7 2) RF scores of fitness under interferon selection <0.2. 3) Residues that when mutated into different amino acids gives similar interferon sensitive phenotypes are preferred. The scatter plots (left side) shows mutations with RF scores > 0.7 and RF scores with IFN < 0.3 (orange) and <0.2 (red). Moreover, we mapped the list of potential interferon sensitive residues onto protein structures (right side). We reasoned that if a residue is interacting with interferon pathway, then it is more likely to be located on the exposable surface. Moreover, it is likely that multiple mutations around the same domain or same pocket showed a similar phenotype. Thus, we preferably picked the residues that clustered on the protein surface for validation. Up to 7 mutations per segment were selected.

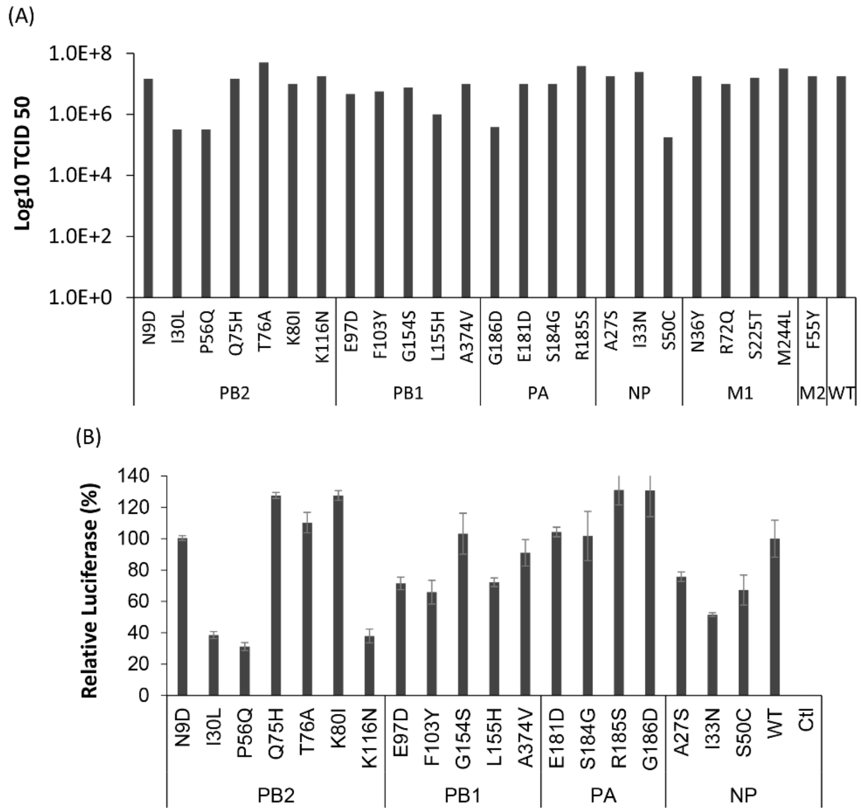


Figure S6-7. Validation of interferon sensitive single mutations. (A)Viral growth capacity of IFN sensitive mutants were shown. 28 possible IFN sensitive mutant were selected across the genome and reconstructed individually in the content of the whole virus. Virus were reconstituted by cotransfecting plasmid containing the single mutation together with 7 plasmids encoding other wild type proteins. A549 cells were infected with each mutant at an MOI = 0.1. Supernatant were collected 24h post infection and viral titer were measured by TCID50.

(B) Relative polymerase activity for single mutations in the polymerase complex are shown. The data are presented as the mean±sfrom three independent biological replicates.

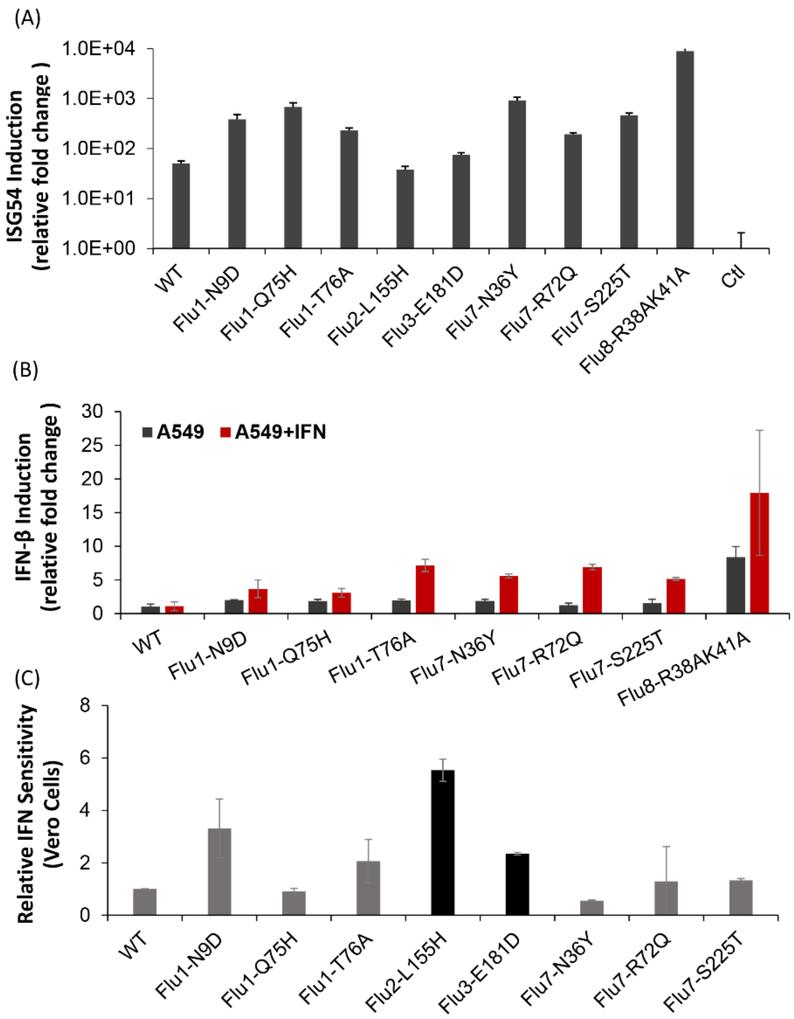


Figure S6-8. Induction of IFN production of IFN sensitive mutations

(A) Induction of ISG54 gene expression were shown for 8 mutations as indicated. A549 cells were infected with each mutation at MOI 1 and cellular RNA were extracted 6h post infection. ISG54 gene induction were quantified by RT-qPCR and calculated as fold of induction compare with mock infected cells. The data are presented as the mean±s from three independent biological replicates.

(B) The induction of IFNb is more significant for cells pre-treated with IFN. Induction of IFNb gene expression were shown for 8 mutations as indicated. A549 cells were pretreated with 1000U/ml IFNa2 for 20h or mock-treated. Cells were then infected with each mutation at MOI 1 and cellular RNA were extracted 6h post infection. ISG54 gene induction were quantified by RT-qPCR and calculated as fold of induction compare with WT infected cells.

(C) IFN sensitivity of selected mutant were examined in vero cells. Vero cells were infected with each mutation at MOI 0.1 with and without 1000U/ml IFN selection. Relative IFN sensitivity were shown for each mutation normalized to wild type WSN virus. In contrast to A549 cells, only 2 mutations showed in black bars were significantly more sensitive to IFN compare with WT (Two-tailed T test, $p<0.05$). These two mutations did not show to induce higher IFNb gene production.

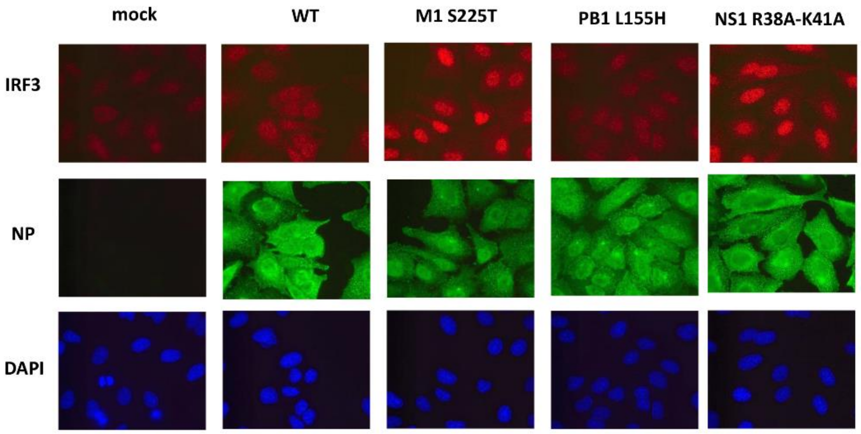


Figure S6-9. Nuclear translocation of IRF3

Nuclear translocation of IRF3 upon infection with different mutations were examined by immunofluorescence. A549 cells were infected with wild type and mutated virus at MOI 1. Cells were fixed and subjected to immunofluorescence analysis (IFA) at 8h post infection.

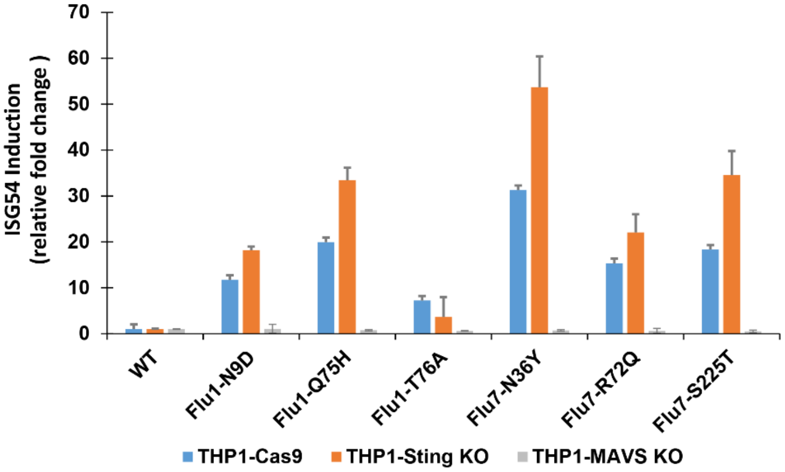


Figure S6-10. MAVS are required for the high induction of IFNb of PB2 and M1 mutants

Induction of ISG54 gene expression were examined in STING or MAVS knock out THP1 cells. THP1 cells stable overexpressing Cas9 protein were used as control. Cells were infected with each mutation at MOI 1 and cellular RNA were extracted 6h post infection. ISG54 gene induction were quantified by RT-qPCR and calculated as fold of induction compare with WT infected cells. The data are presented as the mean±s from three independent biological replicates.

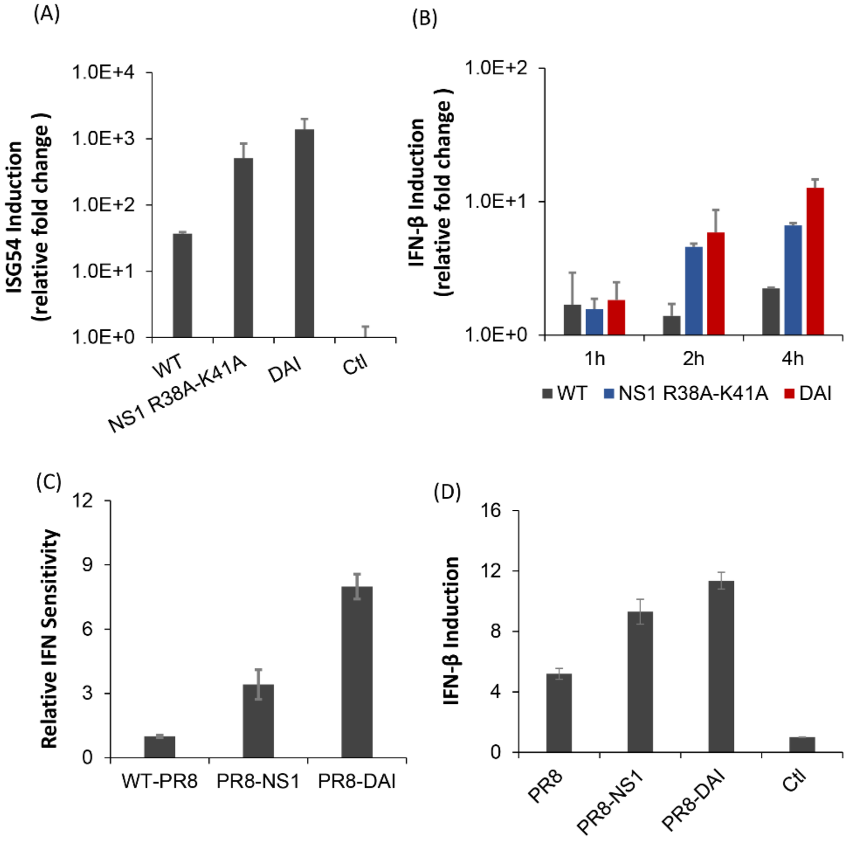


Figure S6-11. Induction of IFN production of DAI virus

(A) Induction of ISG54 gene expression were shown for mutations as indicated. A549 cells were infected with indicated mutant at MOI 1 and cellular RNA were extracted 6h post infection. ISG54 gene induction were quantified by RT-qPCR and calculated as fold of induction compare with mock infected cells. The data are presented as the mean±s from three independent biological replicates.

(B) Induction of IFNb gene expression were examined at 1h, 2h and 4h post infection with indicated virus. The data are presented as the mean±s from biological duplicates.

(C&D) Relative IFN sensitivity and IFNb induction were examined for DAI virus in PR8 background. Same 8 mutations (PB2: N9D, Q75H, T76A, M1: N36Y, R72Q, S225T, NS1: R38A, K41A) were introduced in PR8 reverse genetic system. The data are presented as the mean±s from biological duplicates.

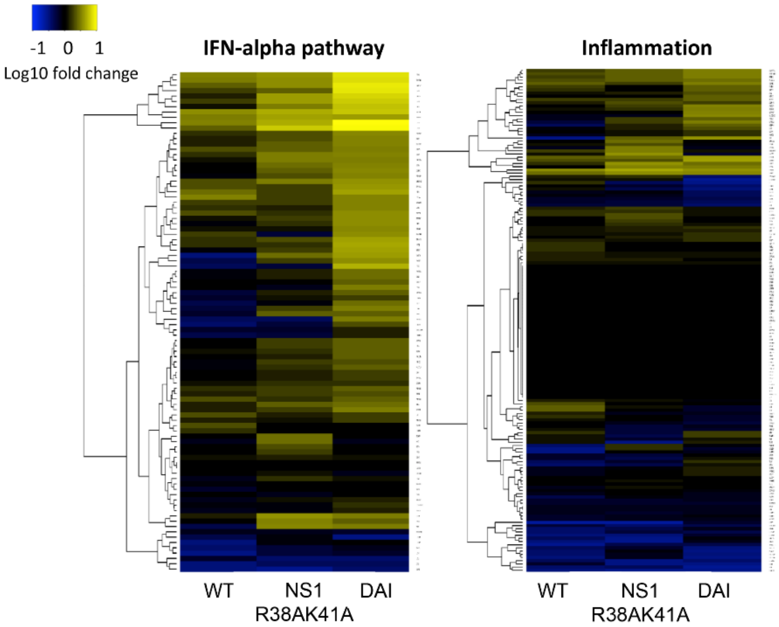


Figure S6-12. DAI has highly induction of IFN but not inflammation response

Heatmap were shown for genes expression pattern of interferon a (left panel) and inflammation (right panel) response in WT, NS1 R39AK41A and DAI infected A549 cells. The gene expression level is quantified by mRNA sequencing. A549 cells was infected with indicated virus at MOI 1 for 6h. Cellular mRNA were extracted and subjected to mRNA sequencing. Gene expression change were compared between each infected cells with mock infected A549 cells and showed as fold change in log10. Biological duplicate were included for each condition. Genes belong to certain pathway were collected from MsigDB. DAI infection significantly upregulated genes in interferon a pathway, but not inflammation.

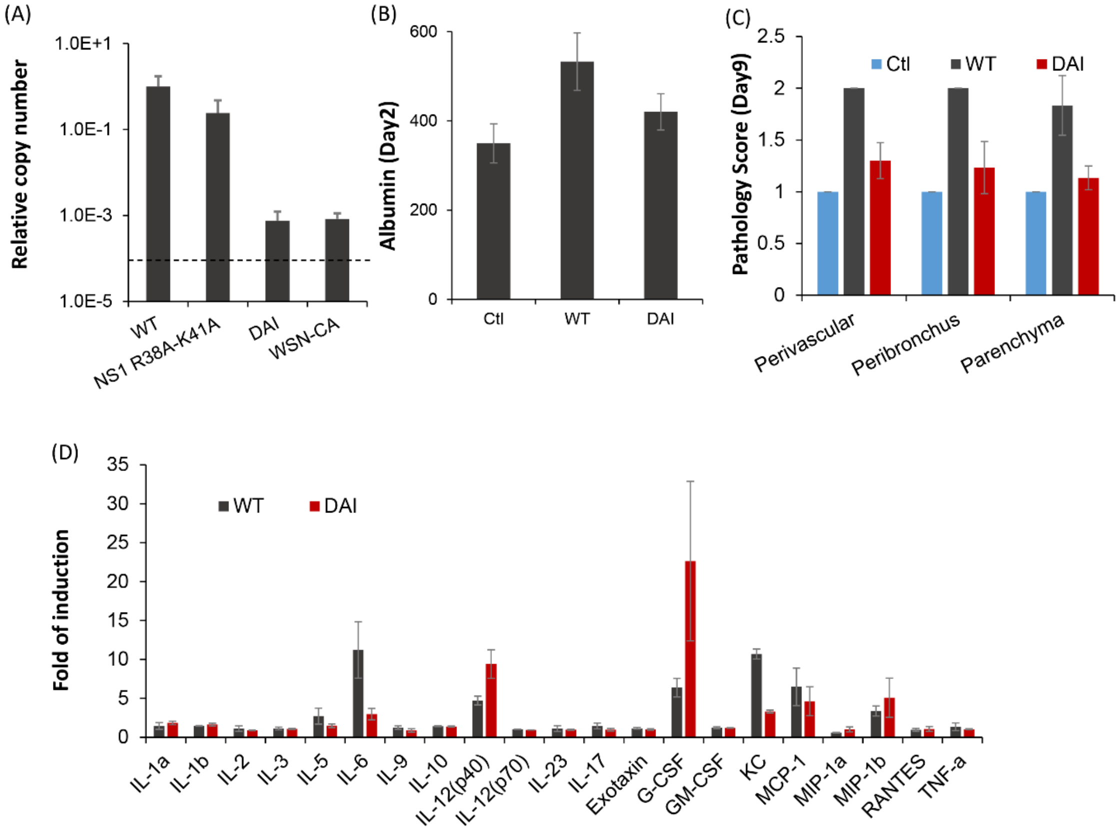


Figure S6-13. Replication and induction of cytokine of DAI virus in IFN competent mice

(A)Viral titer in lung tissue are shown for indicated virus as relative copy number. Female Balb/c mice age 6-8 weeks were intranasally infected with 105 TCID50 of indicated virus, mice lung tissue were harvested at 2 days post infection (N=3). Viral copy number were examined by RT-qPCR and normalized to WT infected mice. Dashed line represent mock infected mice.

(B) Epithelial integrity is examined by albumin concentration in BAL sample collected at day 2 post-infection (N=3). WT infected mice showed significantly higher albumin concentration, suggesting the more severe loss of lung integrity.

(C) Pathology score were shown for mice lung HE staining slides at day 9 post infection. 4-5 areas were averaged for each mice. The data are presented as the mean±s from three mice per group.

(D) Induction of indicated cytokines were examined with luminex multiplex assay with BAL samples collected at day 2 post infection (N=3). Fold of induction for each cytokine were calculated for WT and DAI infected mice compare to control mice with PBS infection.

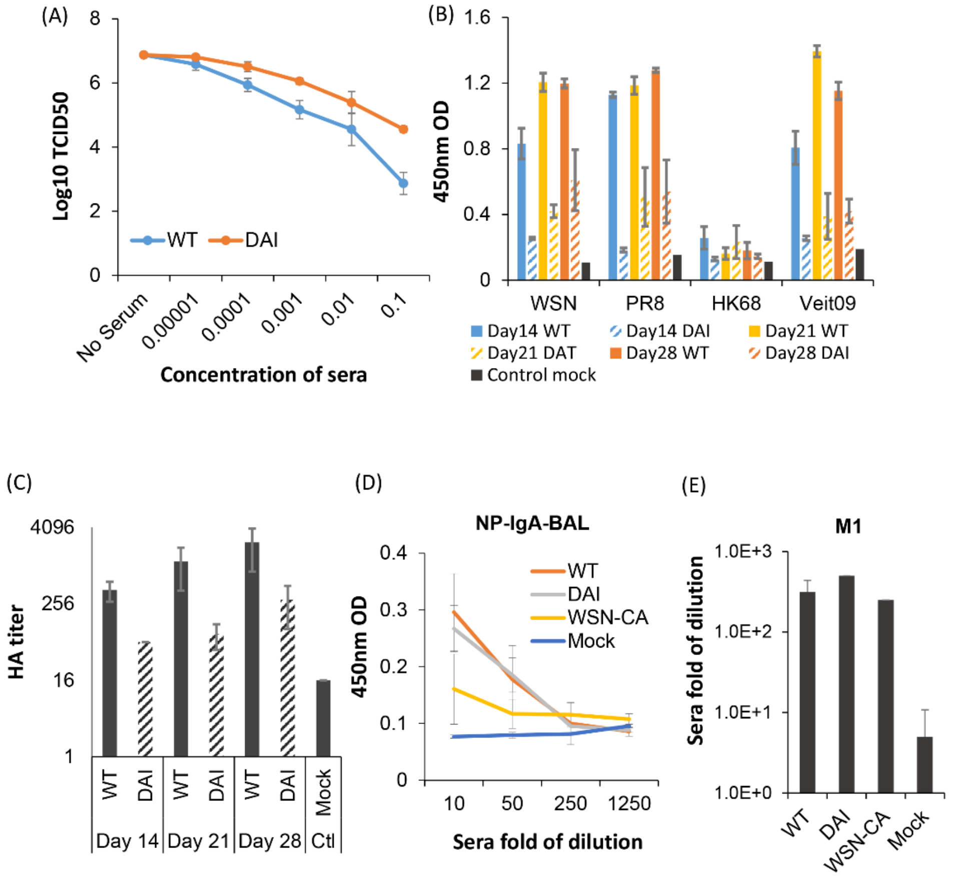


Figure S6-14. Antibody response induced by DAI vaccination.

(A) Neutralizing antibody in vaccinated mice sera were examined. Female Balb/c mice age 6-8 weeks were intranasally infected with 105 TCID 50 of indicated virus (N=3). Serum were obtained at day 28 post infection and heat inactivated for viral neutralization assay.

(B&C) Diverse HA binding IgG and HA neutralizing antibodies were examined in vaccinated mice sera at different time points by ELISA (B) and hemagglutinin inhibition (HAI) assay (C). Female Balb/c mice were intranasal infected with 105 TCID50 WT and DAI-1 virus (N=3). At 14, 21 and 28 days post infection, serum samples were collected and HA-specific antibody were assessed through ELISA. 4 types of HA protein were purified from different strains of virus and used as target for IgG binding: WSN (H1), PR8 (H1), HK68 (H3) and Viet04 (H5). WT and DAI infected mice elicit antibody response against other strains (PR8: H1, Viet04: H5) within HA group (**Figure S6-14**C), but not across different HA group (HK68: H3). Similar to the WT infected group, the antibody titer in DAI immunized mice showed steady increase from Day 14, 21 to 28 days post vaccination

(D) NP binding IgA were examined in BAL samples by ELISA. Female Balb/c mice age 6-8 weeks were intranasally infected with 105 TCID50 of indicated virus (N=3), BAL samples were obtained at day 28 post infection.

(E) M1 binding IgG antibody were examined in mice sera by ELSA. The data are presented as the mean±s from three mice.

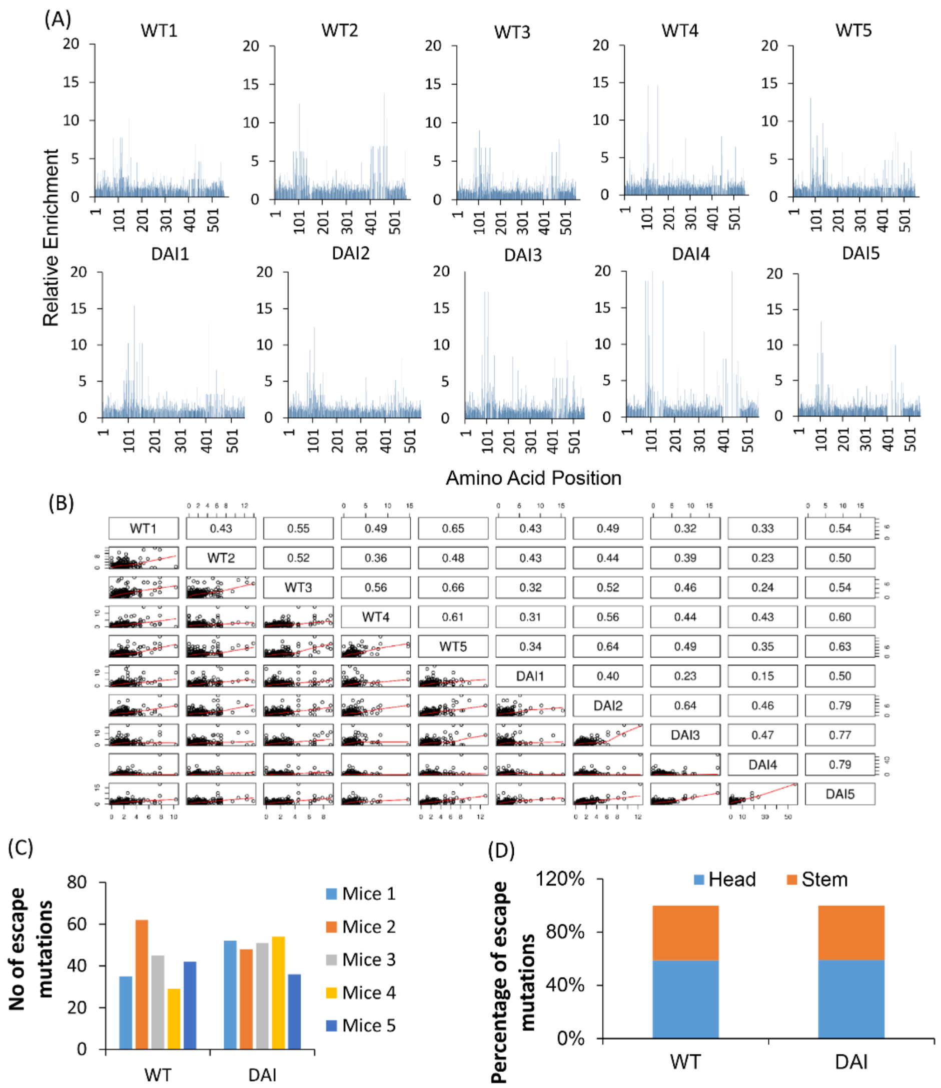


Figure S6-15. Antibody escape mutations screen with HA single mutant library

(A) Relative enrichment of each mutations on HA protein were shown. HA single nucleotide mutant libraries were selected under WT or DAI infected mice sera (N=5) at the concentration of IC80. Mock infected mice sera were used as control. Relative enrichment of each mutant were calculated as the relative fitness under sera selection compared with control.

(B) Correlation were shown for the relative enrichment scores of each single nucleotide mutations on HA protein under different mice sera selection conditions. Modest correlation were obtained among all the conditions.

(C) Numbers of escape mutations were shown. Escape mutations for mice sera were defined as the ones with relative enrichment > 5 for each condition. Similar numbers of escape mutations were detected from WT or DAI vaccinated mice sera.

(D) Percentage of escape mutations located in the head and stem region of HA proteins are shown. 51 and 61 possible escape mutations were identified to occur in more than one mice in WT and DAI infected group (N=5). ~60% of these mutations were located in the head domain of

HA protein, while ~40% of them are in the stem region. The percentage is similar between two groups.

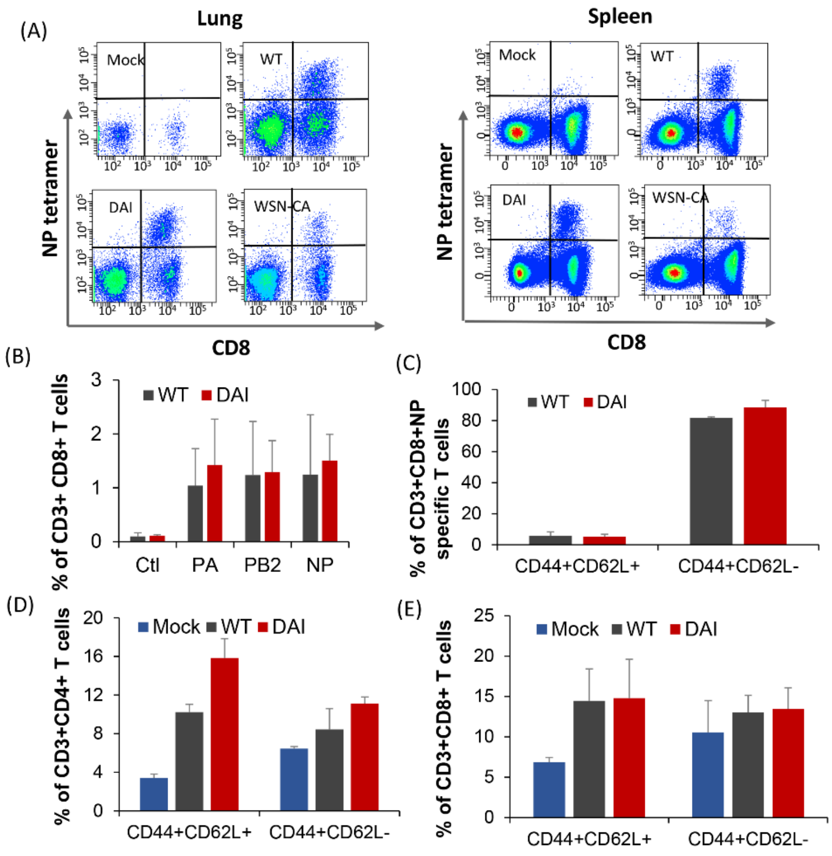


Figure S6-16. Robust T cell response induced by DAI virus

(A) CD8 T cell response were examined by tetramer staining and low cytometry. Representative flow cytometry dot plots were shown for lung and spleen samples. Female C57/B6 mice were intranasally infected with 105 TCID 50 of indicated virus (N=5). Single cell suspension were made from lung and spleen tissues harvested at 10 days post infection. 1 million cells were subjected to flow cytometry analysis with CD3, CD8 and tetramer complexes with H-2Db + influenza A virus NP366–374 (NPP). Black circles indicated portion of NP epitope specific CD8 cells, which is double positive for NP tetramer and CD8 staining.

(B) Long term CD8 T cell response were examined by peptide stimulation in spleen tissues. Female C57/B6 mice were intranasally infected with 105 TCID50 of WT and DAI virus (N=4). Spleen tissue were collected 28 days post infection and directly transferred through the single cell strainer to make single cell suspension. Spleenocytes were then stimulated with indicated peptide for 6h at the concentration of 1uM, with the presence of 1ug/ml brefeldinA. Cells were stained with CD3, CD8 and intracellularly stained with IFNg. Percentage of IFNg positive CD8 cells were quantified for each peptide.

(C-E) Percentage of effector/effector memory T cells and central memory T cells were examined for (C) NP specific CD8 T cells, (D) CD4 and (E) CD8 T cells. Female C57/B6 mice were intranasally infected with 105 TCID50 of WT and DAI virus (N=4). Spleen tissue were harvested 10 days post infection. Spleenocytes were stained with CD3-eflurofore450, CD8-FITC, CD4-APC, CD44-APC eflurofore 710, CD62L- Percp cy5.5 and NP tetramer conjugated with PE.

Percentage of effector/effector memory T cells (CD44+CD62L-) and central memory T cells (CD44+CD62L+) were quantified. DAI virus vaccination elicit strong CD8, CD4 T response with robust central memory T cells which is essential for long-term and cross-class protection.

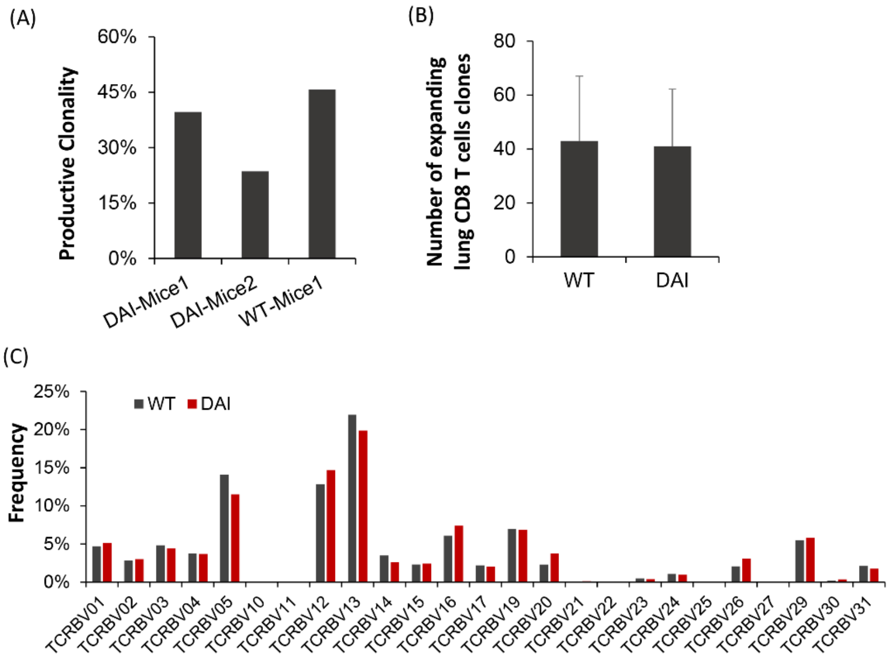


Figure S6-17. T cell repertoire analysis

(A) Deep sequencing were performed for TCRB gene of influenza NP366–374 specific CD8 cells. Lung and spleen tissue were harvested at day 10 post infection. NP specific T cells were sorted out and genomic DNA were extracted and prepared for sequencing. Clonality of NP specific T cells is shown with two DAI infected mice and one WT infected mice.

(B) Numbers of expanding CD8 T cells clones were shown for WT and DAI virus infected mice as compare with mock (N=2). Lung CD8 cells were isolated at day 10 post infection and TCRB gene were deep sequenced. Expanding T cells were defined as the ones with TCRB rearrangement detected in NP specific T cells or occure > 0.1% in the total population, but did not detected in the mock control.

(C) TCR Vb gene usage were analyzed for both NP specific CD8 T cells for DAI and WT infected mice.

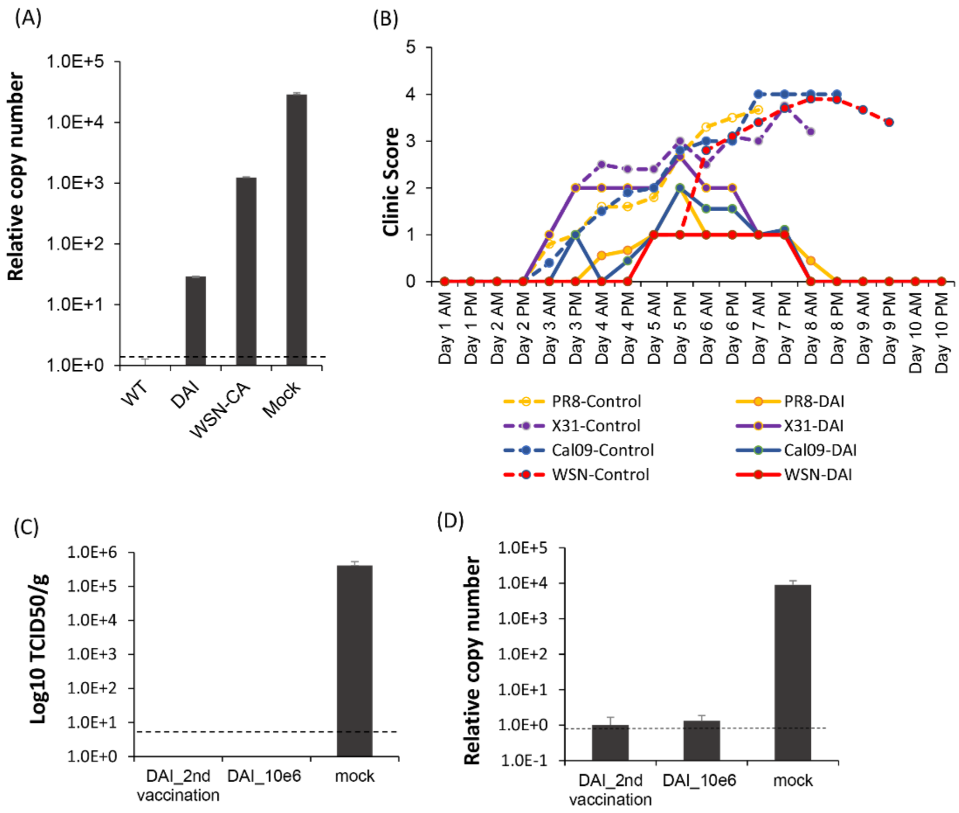


Figure S6-18. Protection of DAI vaccinated mice from challenge

(A) Protection of WT infection were quantified by the viral load and relative viral copy number in lung samples of vaccinated mice. Female Balb/c mice age 6-8 weeks were intranasally vaccinated with 105 TCID 50 of indicated virus (N=4). Mice were challenged with 105 TCID 50 of WT virus at 28 days post vaccination. Lung tissue were extracted at 2 days post challenge. Viral copy number were quantified by RT-qPCR and normalized to WT vaccinated mice. Dashed line represented mock infected mice, which is the detection limit of RT-qPCR assay.

(B) Clinic scores of mice challenged with homologous and heterologous viral strains. DAI vaccinated or mock (PBS) vaccinated C57/B6 mice were challenged with WSN, PR8, and ACal/04/09 at a dose of LD90 and X-31 at LD50 (N=10). Clinic score were obtained twice daily for 10 days.

(C&D) Protection of WT infection were examined with high does or two vaccinations with DAI strain. 106 TCID 50 DAI virus were used as high does vaccination for 28 days. Two vaccination were performed with 105 TCID 50 at 28 days apart. Mice were challenged with 105 TCID 50 of WT virus and viral growth were quantified by both RT-qPCR and TCID50 in lung samples at day 2 post infection.

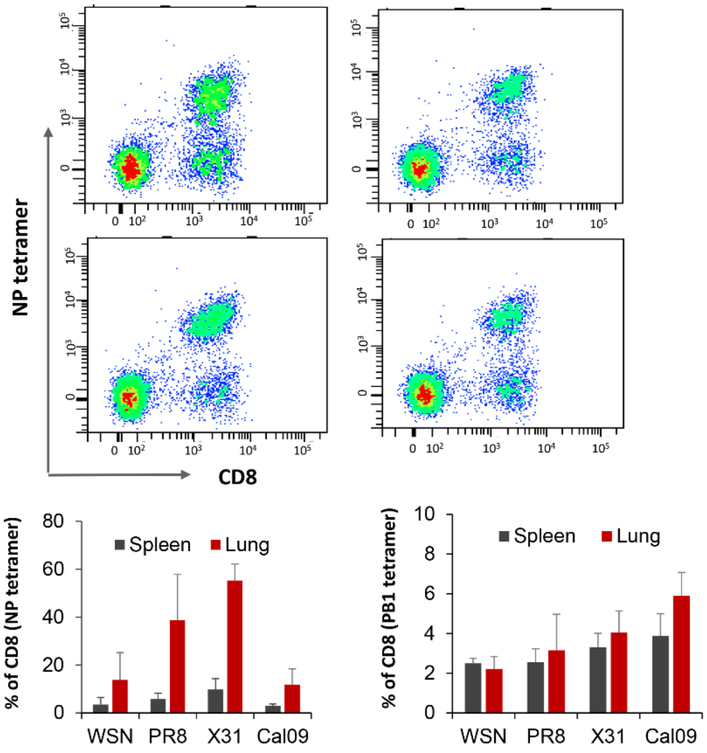


Figure S6-19. Secondary T cell response post challenge

Secondary T cell response were examined by tetramer staining and flow cytometry. DAI vaccinated mice were challenged with WSN, PR8, ACal/04/09 and X-31 after 28 days of vaccination. 14 days after challenge, lung and spleen samples were collected from vaccinated mice from each challenge group (N-5). Specific CD8 T cells agains5 H-2Db + influenza A virus NP366–374 (NPP, ASNENMETM) and H-2Kb influenza A virus PB1703-711 (SSYRRPVGI) were examined. Robust CD8 T cell rebound response toward NP epitope were observed for all the challenge groups.

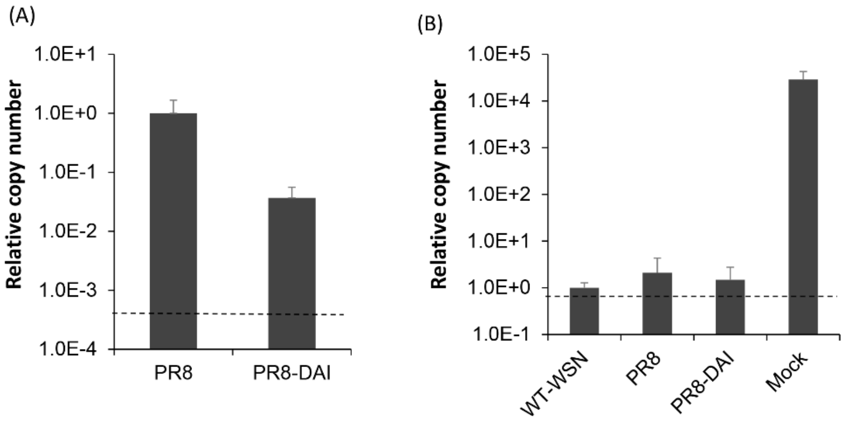


Figure S6-20. Replication and protection of DAI virus in PR8 background.

(A) Viral titer in lung tissue are shown for indicated virus as relative copy number. Female Balb/c mice age 6-8 weeks were intranasally infected with 104 TCID50 of indicated virus, mice lung tissue were extracted at 2 days post infection (N=3). Viral copy number were examined by RT-qPCR and normalized to WT infected mice. Dashed line represent mock infected mice.

(B) Protection of WT infection were quantified by the viral load in lung samples of vaccinated mice at day 2 post challenge (N=4). DAI virus in PR8 background also showed attenuation in vivo and protection against viral challenge.

6.7 Bibliography

1. Randall, R. E. & Goodbourn, S. Interferons and viruses: an interplay between induction, signalling, antiviral responses and virus countermeasures. *J. Gen. Virol.* **89,** 1–47 (2008).

2. García-Sastre, A. Induction and evasion of type I interferon responses by influenza viruses. *Virus Res.* **162,** 12–8 (2011).

3. González-Navajas, J. M., Lee, J., David, M. & Raz, E. Immunomodulatory functions of type I interferons. *Nat. Rev. Immunol.* **12,** 125–35 (2012).

4. Iwasaki, A. & Pillai, P. S. Innate immunity to influenza virus infection. *Nat. Rev. Immunol.* **14,** 315–28 (2014).

5. Schoggins, J. W. *et al.* A diverse range of gene products are effectors of the type I interferon antiviral response. *Nature* **472,** 481–5 (2011).

6. Huber, J. P. & Farrar, J. D. Regulation of effector and memory T-cell functions by type I interferon. *Immunology* **132,** 466–74 (2011).

7. Bon, A. Le, Schiavoni, G. & D'Agostino, G. Type I interferons potently enhance humoral immunity and can promote isotype switching by stimulating dendritic cells in vivo. *Immunity* **14,** 461–470 (2001).

8. Tough, D. F. Type I Interferon as a Link Between Innate and Adaptive Immunity through Dendritic Cell Stimulation. *Leuk. Lymphoma* **45,** 257–264 (2004).

9. Welsh, R. M., Bahl, K., Marshall, H. D. & Urban, S. L. Type 1 interferons and antiviral CD8 T-cell responses. *PLoS Pathog.* **8,** e1002352 (2012).

10. Bon, A. Le & Tough, D. Links between innate and adaptive immunity via type I interferon. *Curr. Opin. Immunol.* 432–436 (2002).

11. Crouse, J., Kalinke, U. & Oxenius, A. Regulation of antiviral T cell responses by type I interferons. *Nat. Rev. Immunol.* **15,** 231–242 (2015).

12. Hale, B., Albrecht, R. & García-Sastre, A. Innate immune evasion strategies of influenza viruses. *Future Microbiol.* **62623,** 1–29 (2010).

13. Hale, B. G., Randall, R. E., Ortin, J. & Jackson, D. The multifunctional NS1 protein of influenza A viruses. *J. Gen. Virol.* **89,** 2359–2376 (2008).

14. Graef, K. M. *et al.* The PB2 subunit of the influenza virus RNA polymerase affects virulence by interacting with the mitochondrial antiviral signaling protein and inhibiting expression of beta interferon. *J. Virol.* **84,** 8433–45 (2010).

15. Yoshizumi, T. *et al.* Influenza A virus protein PB1-F2 translocates into mitochondria via Tom40 channels and impairs innate immunity. *Nat. Commun.* **5,** 4713 (2014).

16. Varga, Z. T. *et al.* 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 (2011).

17. Liedmann, S. *et al.* Viral suppressors of the RIG-I-mediated interferon response are pre-packaged in influenza virions. *Nat. Commun.* **5,** 5645 (2014).

18. Riegger, D. *et al.* The nucleoprotein of newly emerged H7N9 influenza A virus harbors a unique motif conferring resistance to antiviral human MxA. *J. Virol.* (2014). doi:10.1128/JVI.02406-14

19. Finch, C., Li, W. & Perez, D. Design of Alternative Live Attenuated Influenza Virus Vaccines. *Influ. Pathog. Control. II* 205–235 (2015). doi:10.1007/82

20. Cox, R. J., Brokstad, K. a & Ogra, P. Influenza virus: immunity and vaccination strategies. Comparison of the immune response to inactivated and live, attenuated influenza vaccines. *Scand. J. Immunol.* **59,** 1–15 (2004).

21. Osterholm, M. T., Kelley, N. S., Sommer, A. & Belongia, E. a. Efficacy and effectiveness of influenza vaccines: a systematic review and meta-analysis. *Lancet Infect. Dis.* **12,** 36–44 (2012).

22. Tricco, A. C. *et al.* Comparing influenza vaccine efficacy against mismatched and matched strains: a systematic review and meta-analysis. *BMC Med.* **11,** 153 (2013).

23. Darvishian, M., Bijlsma, M. J., Hak, E. & van den Heuvel, E. R. Effectiveness of seasonal influenza vaccine in community-dwelling elderly people: a meta-analysis of test-negative design case-control studies. *Lancet. Infect. Dis.* **14,** 1228–39 (2014).

24. Burton, D. R., Poignard, P., Stanfield, R. L. & Wilson, I. A. Broadly neutralizing antibodies present new prospects to counter highly antigenically diverse viruses. *Science (80-.).* **337,** 183–186 (2012).

25. Wei, C.-J. Induction of Broadly Neutralizing H1N1. *Science (80-.).* **329,** 1060–1064 (2010).

26. Impagliazzo, A. *et al.* A stable trimeric influenza hemagglutinin stem as a broadly protective immunogen. *Science (80-.).* **349,** 1301–1306 (2015).

27. Nabel, G. J. & Fauci, A. S. Induction of unnatural immunity: prospects for a broadly protective universal influenza vaccine. *Nat. Med.* **16,** 1389–1391 (2010).

28. Krammer, F. & Palese, P. Advances in the development of influenza virus vaccines. *Nat. Rev. Drug Discov.* **14,** 167–82 (2015).

29. Talon, J. & Salvatore, M. Influenza A and B viruses expressing altered NS1 proteins: a vaccine approach. *Proc. Natl. Acad. Sci.* **97,** 4309–4314 (2000).

30. Wu, T. T., Qian, J., Ang, J. & Sun, R. Vaccine prospect of Kaposi sarcoma-associated herpesvirus. *Curr. Opin. Virol.* **2,** 482–488 (2012).

31. Qi, H. *et al.* A quantitative high-resolution genetic profile rapidly identifies sequence determinants of hepatitis C viral fitness and drug sensitivity. *PLoS Pathog* **10,** e1004064 (2014).

32. Doud, M. B. & Bloom, J. D. Accurate Measurement of the Effects of All Amino-Acid Mutations on Influenza Hemagglutinin. 1–17 (2016). doi:10.3390/v8060155

33. Heaton, N. S., Sachs, D., Chen, C.-J., Hai, R. & Palese, P. Genome-wide mutagenesis of influenza virus reveals unique plasticity of the hemagglutinin and NS1 proteins. *Proc. Natl. Acad. Sci. U. S. A.* **110,** 20248–53 (2013).

34. Taft, A. S. *et al.* Identification of mammalian-adapting mutations in the polymerase complex of an avian H5N1 influenza virus. *Nat. Commun.* **6,** 7491 (2015).

35. Hoffmann, E., Krauss, S., Perez, D., Webby, R. & Webster, R. Eight-plasmid system for rapid generation of influenza virus vaccines. *Vaccine* **20,** 3165–3170 (2002).

36. Hoffmann, E. & Neumann, G. A DNA transfection system for generation of influenza A virus from eight plasmids. *Proc. Natl. Acad. Sci.* **97,** 6108–6113 (2000).

37. Wu, N. C. *et al.* High-throughput profiling of influenza A virus hemagglutinin gene at single-nucleotide resolution. *Sci. Rep.* **4,** 4942 (2014).

38. Wu, N. C. *et al.* High-throughput identification of loss-of-function mutations for anti-interferon activity in the influenza A virus NS segment. *J. Virol.* **88,** 10157–64 (2014).

39. Wu, N. C. *et al.* Functional Constraint Profiling of a Viral Protein Reveals Discordance of Evolutionary Conservation and Functionality. *PLoS Genet.* **11,** e1005310 (2015).

40. Du, Y. *et al.* Annotating Protein Functional Residues by Coupling High- Throughput Fitness Profile and Homologous-Structure Analysis. *MBio* **7,** 1–13 (2016).

41. Wu, N. C. *et al.* Coupling high-throughput genetics with phylogenetic information reveals an epistatic interaction on the influenza A virus M segment. *BMC Genomics* **17,** 1 (2016).

42. Elena, S. F., Carrasco, P., Daròs, J.-A. & Sanjuán, R. Mechanisms of genetic robustness in RNA viruses. *EMBO Rep.* **7,** 168–173 (2006).

43. Dai, L. *et al.* Quantifying the evolutionary potential and constraints of a drug-targeted viral protein. *bioRxiv* (2016). doi:10.1101/078428

44. Pérez-Cidoncha, M. *et al.* An unbiased genetic screen reveals the polygenic nature of the influenza virus anti-interferon response. *J. Virol.* **88,** 4632–46 (2014).

45. Hale, B. G., Randall, R. E., Ortín, J. & Jackson, D. The multifunctional NS1 protein of influenza A viruses. *J. Gen. Virol.* **89,** 2359–76 (2008).

46. Gack, M. U. *et al.* Influenza A virus NS1 targets the ubiquitin ligase TRIM25 to evade recognition by RIG-I. *Cell Host Microbe* **5,** 439–449 (2010).

47. Yoh, S. M. *et al.* PQBP1 is a proximal sensor of the cGAS-dependent innate response to HIV-1. *Cell* **161,** 1293–1305 (2015).

48. Qi, H. *et al.* Systematic identification of anti-interferon function on hepatitis C virus genome reveals p7 as an immune evasion protein. **114,** 3–8 (2018).

49. Institute, B. MsigDB. *http://software.broadinstitute.org/gsea/msigdb/genesets.jsp?collection=H*

50. Maassab, H. F. & Bryant, M. L. The development of live attenuated cold-adapted influenza virus vaccine for humans. *Rev. Med. Virol.* **9,** 237–44 (1999).

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. *J. Virol.* **78,** 995–998 (2004).

52. Sasaki, S. *et al.* Distinct cross-reactive B-cell responses to live attenuated and inactivated influenza vaccines. *J. Infect. Dis.* **210,** 865–874 (2014).

53. Rockman, S. *et al.* Neuraminidase-inhibiting antibody is a correlate of cross-protection against lethal H5N1 influenza virus in ferrets immunized with seasonal influenza vaccine. *J Virol* **87,** 3053–3061 (2013).

54. Carragher, D. M., Kaminski, D. A., Moquin, A., Hartson, L. & Randall, T. D. A novel role for non-neutralizing antibodies against nucleoprotein in facilitating resistance to influenza virus. *J. Immunol.* **181,** 4168–4176 (2008).

55. Doud, M. B., Hensley, S. E. & Bloom, J. D. Complete mapping of viral escape from neutralizing antibodies. *PLoS Pathog.* **13,** 1–20 (2016).

56. Yager, E. J. *et al.* Age-associated decline in T cell repertoire diversity leads to holes in the repertoire and impaired immunity to influenza virus. *J. Exp. Med.* **205,** 711–723 (2008).

57. Mueller, S. *et al.* Live attenuated influenza virus vaccines by computer-aided rational design. *Nat. Biotechnol.* **28,** 723–6 (2010).

58. Si, L. *et al.* Generation of influenza A viruses as live but replication-incompetent virus vaccines. *Science (80-.).* **354,** 1–5 (2016).

59. Wu, C.-Y. *et al.* Influenza A surface glycosylation and vaccine design. *Proc. Natl. Acad. Sci.* **114,** 201617174 (2016).

60. Wang, L. *et al.* Generation of a Live Attenuated Influenza Vaccine that Elicits Broad Protection in Mice and Ferrets. *Cell Host Microbe* **21,** 334–343 (2017).

61. Steel, J. *et al.* Live attenuated influenza viruses containing NS1 truncations as vaccine candidates against H5N1 highly pathogenic avian influenza. *J. Virol.* **83,** 1742–53 (2009).

62. Russell, S. J. RNA viruses as virotherapy agents. *Cancer Gene Ther.* **9,** 961–966 (2002).

63. Bergmann, M. *et al.* A Genetically Engineered Influenza A Virus with ras-Dependent. *Animals* 8188–8193 (2001).

64. Melamed, D., Young, D. L., Gamble, C. E., Miller, C. R. & Fields, S. Deep mutational scanning of an RRM domain of the Saccharomyces cerevisiae poly(A)-binding protein. *RNA* **19,** 1537–51 (2013).

65. Fowler, D. M. & Fields, S. Deep mutational scanning: a new style of protein science. *Nat. Methods* **11,** 801–7 (2014).

66. Melnikov, A., Rogov, P., Wang, L., Gnirke, A. & Mikkelsen, T. S. Comprehensive mutational scanning of a kinase in vivo reveals substrate-dependent fitness landscapes. *Nucleic Acids Res.* **42,** 1–8 (2014).

67. Sun, S. *et al.* An extended set of yeast-based functional assays accurately identifies human disease mutations. *Genome Res.* **26,** 670–680 (2016).

68. Chan, W., Zhou, H., Kemble, G. & Jin, H. The cold adapted and temperature sensitive influenza A/Ann Arbor/6/60 virus, the master donor virus for live attenuated influenza vaccines, has multiple defects in replication at the restrictive temperature. *Virology* **380,** 304–11 (2008).

69. Jin, H. *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* **306,** 18–24 (2003).

70. York, A. G. *et al.* Limiting Cholesterol Biosynthetic Flux Spontaneously Engages Type i IFN Signaling. *Cell* **163,** 1716–1729 (2015).

71. Travanty, E. *et al.* Differential Susceptibilities of Human Lung Primary Cells to H1N1 Influenza Viruses. *J. Virol.* **89,** 11935–11944 (2015).

72. Reich, S. *et al.* Structural insight into cap-snatching and RNA synthesis by influenza polymerase. *Nature* **516,** 361–366 (2014).

73. Pflug, A., Guilligay, D., Reich, S. & Cusack, S. Structure of influenza A polymerase bound to the viral RNA promoter. *Nature* **516,** 355–60 (2014).

74. Hengrung, N. *et al.* Crystal structure of the RNA-dependent RNA polymerase from influenza C virus. *Nature* **527,** 114–117 (2015).

75. Gamblin, S. *et al.* The structure and receptor binding properties of the 1918 influenza hemagglutinin. *Science (80-.).* **303,** 1838–42 (2004).

76. Tao, Y., Farsetta, D. L., Nibert, M. L. & Harrison, S. C. RNA synthesis in a cage--structural studies of reovirus polymerase lambda3. *Cell* **111,** 733–745 (2002).

77. Arzt, S. *et al.* Combined results from solution studies on intact influenza virus M1 protein and from a new crystal form of its N-terminal domain show that M1 is an elongated monomer. *Virology* **279,** 439–446 (2001).

78. Lutz, A., Dyall, J., Olivo, P. D. & Pekosz, A. Virus-inducible reporter genes as a tool for detecting and quantifying influenza A virus replication. *J. Virol. Methods* **126,** 13–20 (2005).

79. Gong, D. *et al.* Kaposi's sarcoma-associated herpesvirus ORF18 and ORF30 are essential for late gene expression during lytic replication. *J. Virol.* **88,** 11369–11382 (2014).

80. Trapnell, C., Pachter, L. & Salzberg, S. L. TopHat: Discovering splice junctions with RNA-Seq. *Bioinformatics* **25,** 1105–1111 (2009).

81. Kim, D. *et al.* TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biol.* **14,** R36 (2013).

82. Robinson, M. D., McCarthy, D. J. & Smyth, G. K. edgeR: A Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26,** 139–140 (2009).

83. Tripathi, S. *et al.* Meta- and Orthogonal Integration of Influenza 'oMICs' Data Defines a Role for UBR4 in Virus Budding. *Cell Host Microbe* **18,** 723–735 (2015).

CHAPTER 7

CONCLUSIONS AND PERSPECTIVES

The studies in multiple biological systems in my thesis have provide evidence to support the conclusion that the quantitative high-throughput genomic platform is a powerful system with broad applications. It enabled us to investigate fundamental evolution problems, identify functional residues and new functions of target protein, and facilitate drug development. With the development of other technologies, including DNA syhthesis, sequencing and nanotechnologies, we foresee the further improvement of this system. Based on my own understanding, I want to list the following three trends of utilizing the system as the prespective remarks of my thesis.

Functional profiling of cellular genes

Although the system is developed through manipulating microbe genes, including yeast, bacteria and virus, it is currently quickly applied to mammalian cells. The information provided by this system enables us to understand the key residues of a specific cellular protein and further understand their functions. Moreover, it might help us to pin-point disease related mutations and guide the design of possible inhibitors. Just to provide a concrete example: we can study the proteins related to cell proliferation using cell growth rate as a read out or selection condition. Applying saturated mutagenesis, we can learn which mutation is related to abnormal cellular growth rate and can further use flow cytometry to differentiate cells in different phases. We can also investigate the role of mutant proteins on cancer metastasis through transwell migration assay in vitro or using mouse xenograft models in vivo. The structures of the target protein or homologous proteins can be linked to a genetic profile and further facilitate the understanding of sequence-function relationship at the single amino acid resolution.

Investigation of genetic interactions and epistasis

In the previous chapters, all the results are based on profiling of mutations of single nucleotide or single amino acid changes. The scale of library complexity quickly goes up with combinations of mutations or genes. However, interactions and epistasis among amino acids and genes are critical issues that need further investigation to understand the functions of an orgamism or evelotion of orgamisms. Often times, the combinational impact of two (or multiple) mutations/gene functions/environmental conditions is non-linear, which leads to synergistic or antagonistic effects. For example, some low fitness drug resistant mutations can be rescued by complementary secondary mutations, or the defect of one gene function can be buffered by upregulation of other pathways. Efforts have started to investigate this issue, but due to the vast amount of sequence space, most of the studies still focused on double mutations or a handful of multiple mutations. Even with limited information, researchers were still able to extrapolate interesting information and build interaction networks. With the rapid increasing of sequencing power and the maturation of DNA synthesis technology, we will be able to look at the genetic interactions with a deeper and broader view. Co-evolution experiments can also be set up with the saturated mutagenesis on two moving targets (ligand and receptor, protein interaction complex, inhibitor and target protein, et al). This will be a growing research area.

Combination with single cell sequencing and drop-let sequencing

The rapid improvement of single cell and drop-let sequencing technologies will be powerful to be combined with the genetic profile system. The combination will provide detailed information of single-cell behavior upon mutations, and facilitate the investigation of mutational effect on diverse cellular background. The utilization of nanodroplets will enable us to scale up our system several logs to accelerate biological research to another level.