Title
High-Throughput Genetics in Virus Research: Application and Insight

Author
Wu, Nicholas

Publication Date
2015

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA
Los Angeles

High-Throughput Genetics in Virus Research: Application and Insight

A dissertation submitted in partial satisfaction of the
requirements for the degree Doctor of Philosophy
in Molecular Biology

by

Nicholas Ching Hai Wu

2015

ABSTRACT OF THE DISSERTATION

High-Throughput Genetics in Virus Research:
Application and Insight

by

Nicholas Ching Hai Wu

Doctor of Philosophy in Molecular Biology
University of California, Los Angeles, 2015
Professor Ren Sun, Chair

Traditional genetics, which includes forward and reverse genetics, has been employed extensively to study influenza virus. Although traditional genetics is powerful, it has a limited throughput which only focuses on the linkage of one mutation with one phenotype at a time. In my thesis research, a high-throughput genetics platform is being developed to examine the phenotypic outcomes of all point mutations in a viral gene or genome in parallel. The underlying concept is to randomly mutagenize every nucleotide of an entire genome, monitor enrichment or diminishment of all point mutations under specified growth conditions, and perform massive deep-sequencing to determine which mutations contribute to negative, neutral, or positive outcomes under the given conditions. Using this high-throughput genetics platform, the fitness effects of individual point mutations was profiled across influenza A virus hemagglutinin gene. This technique was further applied to identify novel functional residues and interferon-sensitive mutation. The high-throughput genetics platform can potentially be adapted to study any microbes that can be genetically manipulated. My thesis also describes a novel experimental approach, tag linkage sequencing, to monitor viral quasis-

pecies. Tag linkage sequencing utilizes a molecular tag to identify short sequencing reads that are from the same original DNA template. This allows the reconstruction of individual viral genomes within a viral quasispecies from deep sequencing data. This approach was employed to investigate the genetic content of a clinical sample from a patient infected with human immunodeficiency virus (HIV).

The dissertation of Nicholas Ching Hai Wu is approved.

Eleazar Eskin

Matteo Pellegrini

James O. Lloyd-Smith

Ren Sun, Committee Chair

University of California, Los Angeles

2015

This dissertation is dedicated to my wife, Janice Han Yuen Lee, for her love, support and encouragement.

TABLE OF CONTENTS

ACKNOWLEDGEMENTS

First and foremost, I would like to thank God for all the blessings, opportunities, and people that I have met in my life. I would also like to express my sincere gratitude to my dearest parents for their unconditional love and support throughout the years, and allowing me to realize my own potential. I would like to thank my PhD advisor and mentor, Dr. Ren Sun, for providing me with an excellent atmosphere for doing research, and the freedom he has given me to pursue scientific questions of my own interest. I would not have been able to complete my dissertation without his critical guidance. His motivation, enthusiasm, and immense knowledge have truly inspired me. I am also very grateful to my committee members, Dr. Eleazar Eskin, Dr. Matteo Pellegrini, Dr. David Eisenberg, and Dr. James Lloyd-Smith. Thank you to Dr. Eleazar Eskin for providing me the opportunity to collaborate with him. I truly enjoyed this experience and have learned a lot about algorithm development. Thank you to Dr. Matteo Pellegrini for helpful advice on transcriptome analysis. Thank you to Dr. David Eisenberg for generously granting me the opportunity to rotate with him in my first year and let me utilize his computational resouce for research. Thank you to Dr. James Lloyd-Smith for all the useful comments on my research and helping me to develop my background in evolutionary biology.

I would like to thank all the past and present Sun lab members who have created such a dynamic and joyful research environment. Thank you to Dr. Arthur Young, Dr. Anders Olson, Dr. Laith Al-Mawsawi, Dr. Jun Feng, Dr. Hangfei Qi, Dr. Jiaying Feng, Dr. Roland Remenyi, Dr. Danyang Gong, Dr. Shuai Le, Dr. Xiaojuan Zheng, Yong Hoon Kim, Yushen Du, Harding Luan, Nguyen Nguyen, Kevin Tran, and Tian-Hao Zhang for all the technical and intellectual supports. A special thank you to Dr. Arthur Young for teaching me all the experimental techniques and providing important intellectual input into my research. I would also like to thank all my collaborators. Thank you to Dr. Martha Lewis, Dr. Otto Yang, and Dr. Justin De La Cruz for providing HIV clinical samples for my viral quasispecies project. Thank you to Dr. Alexander Zelikovsky, Dr. Serghei Mangul, Dr. Nicholas Mancuso, and Alex Artyomenko for the wonderful collaboration on algorithm development. Thank you to Dr. Claude Loverdo and Dr. Ruain Ke for the discussion and insight on viral evolution. Thank you to Dr. Stanley Nelson, Traci Toy, Dr. Xinmin Li, Jamie Zhou, and

Dr. Janice Yoshizawa for their help on performing Illumina sequencing. Thank you to Sugandha Dandekar and Hemani Wijersuriya for their help on performing 454 sequencing. Thank you Dr. Lin Jiang for his help and advice on protein modeling.

Lastly, I would also like to thank my family and friends for supporting and encouraging me throughout my life. Without them, I may never have gotten to where I am today.

Chapter 2 is adopted from the following paper: Nicholas C. Wu*, Arthur P. Young*, Laith Q. Al-Mawsawi, C. Anders Olson, Jun Feng, Hangfei Qi, Shu-Hwa Chen, I-Hsuan Lu, Chung-Yen Lin, Harding H. Luan, Nguyen Nguyen, Stanley F. Nelson, Xinmin Li, Ting-Ting Wu, and Ren Sun. (2014) High-throughput profiling of influenza A virus hemagglutinin gene at single-nucleotide resolution. Scientific Reports 4: 4942. *equal contributors

Chapter 3 is adopted from the following manuscript: Nicholas C. Wu, C. Anders Olson, Yushen Du, Shuai Le, Kevin Tran, Danyang Gong, Laith Q. Al-Mawsawi, Hangfei Qi, Ting-Ting Wu, and Ren Sun. Functional constraint profiling of PA influenza virus polymerase subunit at single-amino acid resolution.

Chapter 4 is adopted from the following paper: Nicholas C. Wu, Arthur P. Young, Laith Q. Al-Mawsawi, C. Anders Olson, Jun Feng, Hangfei Qi, Harding H. Luan, Xinmin Li, Ting-Ting Wu, and Ren Sun. (2014) High-throughput identification of amino acid residues essential for anti-interferon function in influenza A virus NS segment. Journal of Virology 88(17): 10157-10164.

Chapter 5 is adopted from the following paper: Nicholas C. Wu, Justin De La Cruz, Laith Q. Al-Mawsawi, C. Anders Olson, Hangfei Qi, Harding H. Luan, Nguyen Nguyen, Yushen Du, Shuai Le, Ting-Ting Wu, Xinmin Li, Martha J. Lewis, Otto O. Yang, and Ren Sun. (2014) HIV-1 quasispecies delineation by tag linkage deep sequencing. PLoS One 9(5): e97505.

VITA

2010 B.S., Chemistry - Biochemistry Specialization
University of Virginia
Charlottesville, Virginia

2013 Advancement to Candidacy for Ph.D.
University of California, Los Angeles
Los Angeles, California

PUBLICATIONS

1. Laith Q. Al-Mawsawi, **Nicholas C. Wu**, C. Anders Olson, Vivian Cai Shi, Hangfei Qi, Xiaojuan Zheng, Ting-Ting Wu, and Ren Sun. (2014) High-throughput profiling of point mutations across the HIV-1 genome. Retrovirology 11(1): 124.

2. C. Anders Olson, **Nicholas C. Wu**, and Ren Sun. (2014) A comprehensive biophysical description of pairwise epistasis throughout an entire protein domain. Current Biology 24(22): 2643-2651. (Featured Article)

3. Roland Remenyi*, Hangfei Qi*, Sheng-Yao Su, Zugen Chen, **Nicholas C. Wu**, Vaithilingaraja Arumugaswami, Shawna Truong, Virginia Chu, Tamar Stokelman, Hung-Hao Lo, C. Anders Olson, Ting-Ting Wu, Shu-hwa Chen, Chungyen Lin, and Ren Sun. (2014) A comprehensive functional map of the hepatitis C virus genome provides a resource for probing viral proteins. mBio 5(5): e01469-14. *equal contributors

4. Danyang Gong, **Nicholas C. Wu**, Yafang Xie, Jun Feng, Leming Tong, Kevin F. Brulois, Harding Luan, Yushen Du, Jae U. Jung, Cun-Yu Wang, Mo Kwan Kang, No-Hee Park, Ren Sun, and Ting-Ting Wu. (2014) Kaposi's sarcoma-associated herpesvirus ORF18 and ORF30 are essential for late gene expression during lytic replication. Journal of Virology 88(19): 11369-11382.

5. **Nicholas C. Wu**, Arthur P. Young, Laith Q. Al-Mawsawi, C. Anders Olson, Jun Feng, Hangfei Qi, Harding H. Luan, Xinmin Li, Ting-Ting Wu, and Ren Sun. (2014) High-throughput identification of amino acid residues essential for anti-interferon function in influenza A virus NS segment. Journal of Virology 88(17): 10157-10164.

6. Serghei Mangul*, **Nicholas C. Wu***, Nicholas Mancuso, Alex Zelikovsky, Ren Sun, and Eleazar Eskin. (2014) Accurate viral population assembly from ultra-deep sequencing data. Bioinformatics 30 (12): i329-i337. *equal contributors

7. Laith Q. Al-Mawsawi*, **Nicholas C. Wu***, Justin De La Cruz, Vivian Cai Shi, Ting-Ting Wu, Martha J. Lewis, Otto O. Yang, and Ren Sun. (2014) HIV-1 gag genetic variation in a single acutely infected participant defined by high-resolution deep sequencing. AIDS Research and Human Retroviruses 30(8): 806-811. *equal contributors

8. **Nicholas C. Wu**, Justin De La Cruz, Laith Q. Al-Mawsawi, C. Anders Olson, Hangfei Qi, Harding H. Luan, Nguyen Nguyen, Yushen Du, Shuai Le, Ting-Ting Wu, Xinmin Li, Martha J. Lewis, Otto O. Yang, and Ren Sun. (2014) HIV-1 quasispecies delineation by tag linkage deep sequencing. PLoS One 9(5): e97505.

9. **Nicholas C. Wu***, Arthur P. Young*, Laith Q. Al-Mawsawi, C. Anders Olson, Jun Feng, Hangfei Qi, Shu-Hwa Chen, I-Hsuan Lu, Chung-Yen Lin, Harding H. Luan, Nguyen Nguyen, Stanley F. Nelson, Xinmin Li, Ting-Ting Wu, and Ren Sun. (2014) High-throughput profiling of influenza A virus hemagglutinin gene at single-nucleotide resolution. Scientific Reports 4: 4942. *equal contributors

10. Jun Feng, Paul D De Jesus, Victoria Su, Stephanie Han, Danyang Gong, **Nicholas C. Wu**, Yuan Tian, Xudong Li, Ting-Ting Wu, Sumit K Chanda, and Ren Sun. (2014) RIOK3 is an adaptor protein required for IRF3-mediated type I interferon production. Journal of Virology 88(14): 7987-7997. (Articles of Significant Interest)

11. Shuai Le, Xinyue Yao, Shuguang Lu, Yinling Tan, Xiancai Rao, Ming Li, Xiaolin Jin, Jing Wang, Yan Zhao, **Nicholas C. Wu**, Renate Lux, Xuesong He, Wenyuan Shi, and Fuquan Hu. (2014) Chromosomal DNA deletion confers phage resistance to Pseudomonas aeruginosa. Scientific Reports 4: 4738.

12. Hangfei Qi, C. Anders Olson, **Nicholas C. Wu**, Ruian Ke, Claude Loverdo, Roland Remenyi, Virginia Chu, Shawna Truong, Zugen Chen, Yushen Du, Sheng-Yao Su, Laith Q. Al-Mawsawi, Ting-Ting Wu, Shu-Hua Chen, Chungyen Lin, Weidong Zhong, James O. Lloyd-Smith, and Ren Sun. (2014) A quantitative high-resolution genetic profile rapidly identifies sequence determinants of viral fitness and drug sensitivity. PLOS Pathogens 10(4): e1004064.

13. Ruian Ke, Claude Loverdo, Hangfei Qi, C. Anders Olson, **Nicholas C. Wu**, Ren Sun, and James O. Lloyd-Smith. (2014) Modelling clinical data shows active tissue concentration of daclatasvir is 10-fold lower than its plasma concentration. Journal of Antimicrobial Chemotherapy 69(3): 724-727.

14. **Nicholas C. Wu***, Arthur P. Young*, Sugandha Dandekar, Hemani Wijersuriya, Laith Q. Al-Mawsawi, Ting-Ting Wu, and Ren Sun. (2013) Systematic identification of H274Y compensatory mutations in influenza A Virus neuraminidase by high-throughput screening. Journal of Virology 87(2): 1193-1199. *equal contributors

15. **Nicholas Wu**, Xiaozeng Yang, and Lei Li. (2011) Identification of feed forward loops composed of microRNAs and transcription factors in Arabidopsis. Journal of Biochemistry and Molecular Biology in the Post-Genomic Era 1(1): 79-85.

16. Hang He, Huiyong Zhang, Xiangfeng Wang, **Nicholas Wu**, Xiaozeng Yang, Runsheng Chen, Yi Li, Xing Wang Deng, and Lei Li. (2010) Development of a versatile, target-oriented tiling microarray assay for measuring allele-specific gene expression. Genomics 96(5): 308-315.

CHAPTER 1

INTRODUCTION TO HIGH-THROUGHPUT GENETICS

1.1 BACKGROUND

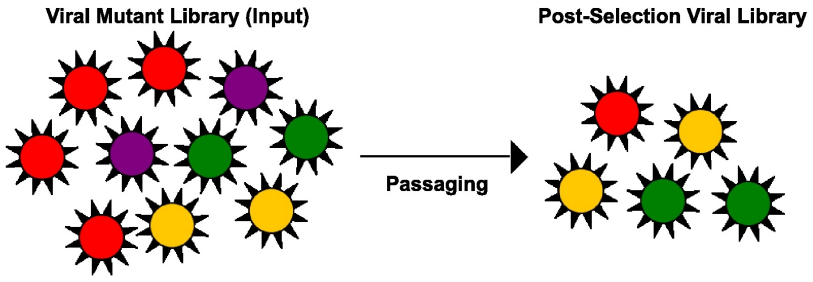
Viral infection is one of the largest global socio-economic burdens. Most, if not all, people experience illness caused by viral infections at least once in their life. Viral infection may result in hospitalization or even death depending on the virus strains and species. Several detrimental and most concerned viruses include human immunodeficiency virus (HIV), hepatitis C virus (HCV), hepatitis B virus (HBV) and influenza virus. Together, they caused two to three million deaths annually from either acute infection or secondary complications [1–4]. The mortality rate can be a magnitude higher during influenza pandemic years [5]. In addition, there are 35 million people chronically infected by HIV [1], 150 million by HCV [2], and more than 240 million by HBV [3]. The continued development of anti-viral drug and vaccine becomes utmost important to ameliorate the global cost brought by viral infection.

Despite the current availability of multiple anti-viral drugs and vaccines, resistant and escape mutants remain a challenge. Mutation rate ranges from 10^{-8} to 10^{-6} mutations per nucleotide per cell infection for DNA viruses and from 10^{-6} to 10^{-4} mutations per nucleotide per cell infection for RNA viruses [6]. This high mutation rate and genome flexibility enable rapidly development of drug resistance and vaccine escape in natural circulating virus strains [7–13]. This highlights the necessity of advancing current antiviral strategy. An ideal antiviral drug or vaccine should target a viral region that has a high fitness cost upon mutation to maximize the genetic barrier for the emergence of resistant or escape mutants. Various virus sequence databases are established to document genetic information of different clinical isolates around the world [14–22]. They provide a basis for identifying conserved regions across the viral genome, which are potential target regions for effective antiviral agents [25–30]. The caveat is that sequence conservation is not equivalent to sequence essentialness. New viral variants emerge constantly, implying that a portion of residues that are conserved currently are mutable. In addition, several mutational analyses on influenza virus suggested that conserved residues are not essential for viral replication in cell culture [5, 6, 28]. Therefore, experimental interrogation is often required to validate the essentialness of individual conserved mutants and to examine their biological roles.

Traditional genetics, which focuses on a single genotype-phenotype relationship at a time, has been extensively applied to experimentally characterize viral mutants of interest. It provides the foundation in understanding the virus life cycle and virulence factors. However, this process has a low throughput and thus, restricts the number of mutants being tested. With the advert of sequencing technology, the concept of high-throughput genetics has started to emerge (reviewed in [31]). High-throughput genetics aims to study the phenotypic outcome of multiple mutants simultaneously. A general strategy is to employ a mutant library consisting of a large number of mutants and to monitor the enrichment or diminishment, hence the phenotypic outcome, of individual mutants (Fig. 1-1). This strategy has been adapted to many different biological systems (reviewed in [40]), and also becomes more popular in virus research.

1.2 INSERTIONAL HIGH-THROUGHPUT GENETICS

Prior to the existence of high-throughput genetic approach at single-residue resolution, transposon insertional mutagenesis is one of the standard techniques for high-throughput genetic screening. It utilizes the random insertion property of transposon to construct a mutant library. Different transposon systems have been employed in high-throughput genetics, such as Ty1 [33], Mu [34], MoMLV [35]. Following transposon insertion, a digestion and religation steps are usually performed to minimize the length of insertion and the disruption of the genome. For example, by coupling MuA transposase with NotI restriction sites only gives a final insertion of 15 nucleotides [36–38]. To describe the genetic content of an insertion mutant library, it is necessary to quantify individual insertion mutants present in the library. Genetic footprinting is a classic approach for examining an insertion mutant library. The underlying concept of genetic footprinting is to use a PCR-based approach along with a radioactive primer complement to the insert. The insertion site for individual mutants can be proximate by the PCR product size. The genetic content of the mutant library is therefore determined by analyzing the PCR product using electrophoresis gel and quantified through the relative abundance of individual bands [36–38]. This approach was further advanced by a fluorescence-based capillary electrophoresis profiling [22], which improved the accuracy in insertion site calling. The development of next generation sequencing (NGS) has sophisticated the monitoring of an insertion mutant library both qualitatively



Genotype	WT	Mutant A	Mutant B	Mutant C
Input Frequency	20%	40%	20%	20%
Post-Selection Frequency	40%	20%	0%	40%
Enrichment Ratio	2	0.5	0	2
Relative Fitness	1	0.25	0	1

Figure 1-1. Overview of fitness profiling using high-throughput genetics. The viral mutant library is selected by passaging. The selection pressure is replication fitness. Each circle represents an individual viral particle. Different colors represent different genotypes (WT or mutants). The fitness of each mutant inferred from the profiling is described in the lower panel. Input frequency represents the frequency of the mutant in the pre-selected mutant library. Post-selection frequency represents the frequency of the mutant in the post-selected mutant library. Enrichment ratio is calculated by the ratio of post-selection frequency to pre-selection frequency. Relative fitness for each mutant is calculated by the normalizing the enrichment ratio to that of the WT.

and quantitatively [23, 40]. The conceptual detail of integrating NGS with insertion mutant library is well-described by van Opijnen et al. and will not further discussed here [42].

Insertional-based high-through genetics has been applied to study organisms with a small genome such as yeast and bacteria with at a single gene-resolution level [43, 44]. It was also adapted to identify essential genes in murine gamma-herpesvirus 68 (MHV-68, a double-stranded DNA virus). A higher genetic resolution, at a protein subdomain level, can be achieved in RNA virus, which has a more compact genome. Insertional-based high-throughput genetics has aided the identification of flexible protein regions in the murine norovirus [45], venezuelan equine encephalitis virus [24], influenza virus [23], and HCV [22, 40]. However, the maximum resolution is limited at a protein subdomain level due to the significant structural disruption of the amino acid insertion. Consequently, the resolution of insertional-based approach is not sufficient for identification of critical residues.

1.3 HIGH-THROUGHPUT GENETICS AT SINGLE-RESIDUE RESOLUTION

With the advert of the sequencing throughput in NGS, it becomes possible to perform high-throughput genetics at single-residue resolution. Mutant library construction for high-throughput genetic study of virus is mainly based on two different strategies, namely error-prone PCR [43, 46, 48, 50], and oligo-based approach [50–52]. Nonetheless, alternative approaches such as “doped” oligonucleotide [54] and chemical-induced mutagenesis [51] have also been described for single-residue mutant library construction and potentially can be adapted for high-throughput genetic study of virus.

Regardless of the experimental strategies for mutant library construction, the occurrence frequency of each mutation is often lower than the sequencing error rate in NGS. Therefore, one of the major challenges of high-throughput genetics is to distinguish true point mutations from sequencing errors in NGS. Several strategies of sequencing library preparation have been developed to provide a solution. The most straight-forward approach is to take advantage of the paired-end property of NGS read [54]. This approach requires a short insert length of the paired-

end read such that the forward read overlaps with the reverse read . Consequently, it increases the confidence of base calling by sequencing the same nucleotide twice. A more sensitive approach is to add a molecular tag to individual nucleotide templates for sequencing error correction [21, 54, 57, 59, 60]. The underlying conceptual basis is to assign a unique tag to individual nucleotide templates. Individual tagged templates would be sequenced multiple times. True mutations would exist in most, if not all, reads sharing the same tag, while sequencing error would only exist in one or two reads sharing the same tag. Recently, an approach termed "circle sequencing" is also described for sequencing error correction [49], which takes advantage of the increased read lengths in Illumina sequencing technology. CirSeq involves a circularization step such that each sequencing read is a tandem repeat of a nucleotide template. It enables a given nucleotide template to be sequenced multiple times within a single read to distinguish sequencing errors from true mutations. All of the above approaches have been employed for high-throughput genetic study of virus [43, 46, 50, 50–52, 62]. The increase of sensitivity in detecting rare mutations increases the confidence in identifying lethal mutation (mutation that disappear after selection).

High-throughput genetics has potential applications in addressing several important biomedical issues. As high-throughput genetics is achieving single-amino acid resolution, it enables identification of essential surfaces on available protein structures for rational drug design to increase the fitness cost of potential escape mutations. In addition, peptide stretches intolerable to mutations provide potential epitopes for vaccine development. In addition, high-throughput genetics has many other applications which would potentially transform several subfields in virus research. In my thesis, the development and application of high-throughput genetics at single-residue resolution are described.

1.4 BIBLIOGRAPHY

1. World Health Organization (2014) HIV/AIDS Fact sheet N360. URL http://www.who.int/mediacentre/factsheets/fs360/en/.

2. World Health Organization (2014) Hepatitis C Fact sheet N164. URL http://www.who.int/mediacentre/factsheets/fs164/en/index.html.

3. World Health Organization (2014) Hepatitis B Fact sheet N204. URL http://www.who.int/mediacentre/factsheets/fs204/en/.

4. World Health Organization (2014) Influenza (Seasonal) Fact sheet N211. URL http://www.who.int/mediacentre/factsheets/fs211/en/.

5. Taubenberger JK, Morens DM (2006) 1918 influenza: the mother of all pandemics. Emerg Infect Dis 12: 15–22.

6. Sanjun R, Nebot MR, Chirico N, Mansky LM, Belshaw R (2010) Viral mutation rates. J Virol 84: 9733–9748.

7. Hurt AC, Holien JK, Parker MW, Barr IG (2009) Oseltamivir resistance and the h274y neuraminidase mutation in seasonal, pandemic and highly pathogenic influenza viruses. Drugs 69: 2523–2531.

8. Suzuki H, Saito R, Masuda H, Oshitani H, Sato M, et al. (2003) Emergence of amantadine-resistant influenza a viruses: epidemiological study. J Infect Chemother 9: 195–200.

9. Clavel F, Hance AJ (2004) Hiv drug resistance. N Engl J Med 350: 1023–1035.

10. Rong L, Dahari H, Ribeiro RM, Perelson AS (2010) Rapid emergence of protease inhibitor resistance in hepatitis c virus. Sci Transl Med 2: 30ra32.

11. Lok ASF, Zoulim F, Locarnini S, Mangia A, Niro G, et al. (2002) Monitoring drug resistance in chronic hepatitis b virus (hbv)-infected patients during lamivudine therapy: evaluation of performance of inno-lipa hbv dr assay. J Clin Microbiol 40: 3729–3734.

12. Lacombe K, Boyd A, Lavocat F, Pichoud C, Gozlan J, et al. (2013) High incidence of treatment-induced and vaccine-escape hepatitis b virus mutants among human immunodeficiency virus/hepatitis b-infected patients. Hepatology 58: 912–922.

13. Zharikova D, Mozdzanowska K, Feng J, Zhang M, Gerhard W (2005) Influenza type a virus escape mutants emerge in vivo in the presence of antibodies to the ectodomain of matrix protein 2. J Virol 79: 6644–6654.

14. Bao Y, Bolotov P, Dernovoy D, Kiryutin B, Zaslavsky L, et al. (2008) The influenza virus resource at the national center for biotechnology information. J Virol 82: 596–601.

15. Shafer RW, Jung DR, Betts BJ (2000) Human immunodeficiency virus type 1 reverse transcriptase and protease mutation search engine for queries. Nat Med 6: 1290–1292.

16. Kuiken C, Korber B, Shafer RW (2003) Hiv sequence databases. AIDS Rev 5: 52–61.

17. Kuiken C, Yusim K, Boykin L, Richardson R (2005) The los alamos hepatitis c sequence database. Bioinformatics 21: 379–384.

18. Combet C, Penin F, Geourjon C, Delage G (2004) Hcvdb: hepatitis c virus sequences database. Appl Bioinformatics 3: 237–240.

19. Hayer J, Jadeau F, Delage G, Kay A, Zoulim F, et al. (2013) Hbvdb: a knowledge database for hepatitis b virus. Nucleic Acids Res 41: D566–D570.

20. Gnaneshan S, Ijaz S, Moran J, Ramsay M, Green J (2007) Hepseq: International public health repository for hepatitis b. Nucleic Acids Res 35: D367–D370.

21. Shin-I T, Tanaka Y, Tateno Y, Mizokami M (2008) Development and public release of a comprehensive hepatitis virus database. Hepatol Res 38: 234–243.

22. Panjaworayan N, Roessner SK, Firth AE, Brown CM (2007) Hbvregdb: annotation, comparison, detection and visualization of regulatory elements in hepatitis b virus sequences. Virol J 4: 136.

23. Ekiert DC, Bhabha G, Elsliger MA, Friesen RHE, Jongeneelen M, et al. (2009) Antibody recognition of a highly conserved influenza virus epitope. Science 324: 246–251.

24. Throsby M, van den Brink E, Jongeneelen M, Poon LLM, Alard P, et al. (2008) Heterosubtypic neutralizing monoclonal antibodies cross-protective against h5n1 and h1n1 recovered from human igm+ memory b cells. PLoS One 3: e3942.

25. Rolland M, Nickle DC, Mullins JI (2007) Hiv-1 group m conserved elements vaccine. PLoS Pathog 3: e157.

26. Santos AFA, Lengruber RB, Soares EA, Jere A, Sprinz E, et al. (2008) Conservation patterns of hiv-1 rt connection and rnase h domains: identification of new mutations in nrti-treated patients. PLoS One 3: e1781.

27. Depla E, der Aa AV, Livingston BD, Crimi C, Allosery K, et al. (2008) Rational design of a multiepitope vaccine encoding t-lymphocyte epitopes for treatment of chronic hepatitis b virus infections. J Virol 82: 435–450.

28. Keck ZY, Olson O, Gal-Tanamy M, Xia J, Patel AH, et al. (2008) A point mutation leading to hepatitis c virus escape from neutralization by a monoclonal antibody to a conserved conformational epitope. J Virol 82: 6067–6072.

29. Chu C, Fan S, Li C, Macken C, Kim JH, et al. (2012) Functional analysis of conserved motifs in influenza virus pb1 protein. PLoS One 7: e36113.

30. Stewart SM, Pekosz A (2011) Mutations in the membrane-proximal region of the influenza a virus m2 protein cytoplasmic tail have modest effects on virus replication. J Virol 85: 12179–12187.

31. Araya CL, Fowler DM (2011) Deep mutational scanning: assessing protein function on a massive scale. Trends Biotechnol 29: 435–442.

32. Fowler DM, Fields S (2014) Deep mutational scanning: a new style of protein science. Nat Methods 11: 801–807.

33. Smith V, Botstein D, Brown PO (1995) Genetic footprinting: a genomic strategy for determining a gene's function given its sequence. Proc Natl Acad Sci U S A 92: 6479–6483.

34. Haapa S, Taira S, Heikkinen E, Savilahti H (1999) An efficient and accurate integration of mini-mu transposons in vitro: a general methodology for functional genetic analysis and molecular biology applications. Nucleic Acids Res 27: 2777–2784.

35. Singh IR, Crowley RA, Brown PO (1997) High-resolution functional mapping of a cloned gene by genetic footprinting. Proc Natl Acad Sci U S A 94: 1304–1309.

36. Laurent LC, Olsen MN, Crowley RA, Savilahti H, Brown PO (2000) Functional characterization of the human immunodeficiency virus type 1 genome by genetic footprinting. J Virol 74: 2760–2769.

37. Hallet B, Sherratt DJ, Hayes F (1997) Pentapeptide scanning mutagenesis: random insertion of a variable five amino acid cassette in a target protein. Nucleic Acids Res 25: 1866–1867.

38. Kekarainen T, Savilahti H, Valkonen JPT (2002) Functional genomics on potato virus a: virus genome-wide map of sites essential for virus propagation. Genome Res 12: 584–594.

39. Arumugaswami V, Remenyi R, Kanagavel V, Sue EY, Ho TN, et al. (2008) High-resolution functional profiling of hepatitis c virus genome. PLoS Pathog 4: e1000182.

40. Remenyi R, Qi H, Su SY, Chen Z, Wu NC, et al. (2014) A comprehensive functional map of the hepatitis c virus genome provides a resource for probing viral proteins. MBio 5: e01469–e01414.

41. Heaton NS, Sachs D, Chen CJ, Hai R, Palese P (2013) Genome-wide mutagenesis of influenza virus reveals unique plasticity of the hemagglutinin and ns1 proteins. Proc Natl Acad Sci U S A 110: 20248–20253.

42. van Opijnen T, Camilli A (2013) Transposon insertion sequencing: a new tool for systems-level analysis of microorganisms. Nat Rev Microbiol 11: 435–442.

43. Baba T, Ara T, Hasegawa M, Takai Y, Okumura Y, et al. (2006) Construction of escherichia coli k-12 in-frame, single-gene knockout mutants: the keio collection. Mol Syst Biol 2: 2006.0008.

44. Ross-Macdonald P, Coelho PS, Roemer T, Agarwal S, Kumar A, et al. (1999) Large-scale analysis of the yeast genome by transposon tagging and gene disruption. Nature 402: 413–418.

45. Thorne L, Bailey D, Goodfellow I (2012) High-resolution functional profiling of the norovirus genome. J Virol 86: 11441–11456.

46. Beitzel BF, Bakken RR, Smith JM, Schmaljohn CS (2010) High-resolution functional mapping of the venezuelan equine encephalitis virus genome by insertional mutagenesis and massively parallel sequencing. PLoS Pathog 6: e1001146.

47. Al-Mawsawi LQ, Wu NC, Olson C, Shi V, Qi H, et al. (2014) High-throughput profiling of point mutations across the hiv-1 genome. Retrovirology 11: 124.

48. Wu NC, Young AP, Dandekar S, Wijersuriya H, Al-Mawsawi LQ, et al. (2013) Systematic identification of h274y compensatory mutations in influenza a virus neuraminidase by high-throughput screening. J Virol 87: 1193–1199.

49. Wu NC, Young AP, Al-Mawsawi LQ, Olson CA, Feng J, et al. (2014) High-throughput profiling of influenza a virus hemagglutinin gene at single-nucleotide resolution. Sci Rep 4: 4942.

50. Wu NC, Young AP, Al-Mawsawi LQ, Olson CA, Feng J, et al. (2014) High-throughput identification of loss-of-function mutations for anti-interferon activity in the influenza a virus ns segment. J Virol 88: 10157–10164.

51. Bloom JD (2014) An experimentally determined evolutionary model dramatically improves phylogenetic fit. Mol Biol Evol 31: 1956–1978.

52. Thyagarajan B, Bloom JD (2014) The inherent mutational tolerance and antigenic evolvability of influenza hemagglutinin. Elife 3.

53. Qi H, Olson CA, Wu NC, Ke R, Loverdo C, et al. (2014) A quantitative high-resolution genetic profile rapidly identifies sequence determinants of hepatitis c viral fitness and drug sensitivity. PLoS Pathog 10: e1004064.

54. Fowler DM, Araya CL, Fleishman SJ, Kellogg EH, Stephany JJ, et al. (2010) High-resolution mapping of protein sequence-function relationships. Nat Methods 7: 741–746.

55. Robins WP, Faruque SM, Mekalanos JJ (2013) Coupling mutagenesis and parallel deep sequencing to probe essential residues in a genome or gene. Proc Natl Acad Sci U S A 110: E848–E857.

56. Kinde I, Wu J, Papadopoulos N, Kinzler KW, Vogelstein B (2011) Detection and quantification of rare mutations with massively parallel sequencing. Proc Natl Acad Sci U S A 108: 9530–9535.

57. Schmitt MW, Kennedy SR, Salk JJ, Fox EJ, Hiatt JB, et al. (2012) Detection of ultra-rare mutations by next-generation sequencing. Proc Natl Acad Sci U S A 109: 14508–14513.

58. Jabara CB, Jones CD, Roach J, Anderson JA, Swanstrom R (2011) Accurate sampling and deep sequencing of the hiv-1 protease gene using a primer id. Proc Natl Acad Sci U S A 108: 20166–20171.

59. Flaherty P, Natsoulis G, Muralidharan O, Winters M, Buenrostro J, et al. (2012) Ultrasensitive detection of rare mutations using next-generation targeted resequencing. Nucleic Acids Res 40: e2.

60. Gout JF, Thomas WK, Smith Z, Okamoto K, Lynch M (2013) Large-scale detection of in vivo transcription errors. Proc Natl Acad Sci U S A 110: 18584–18589.

61. Lou DI, Hussmann JA, McBee RM, Acevedo A, Andino R, et al. (2013) High-throughput dna sequencing errors are reduced by orders of magnitude using circle sequencing. Proc Natl Acad Sci U S A 110: 19872–19877.

62. Acevedo A, Brodsky L, Andino R (2013) Mutational and fitness landscapes of an rna virus revealed through population sequencing. Nature .

CHAPTER 2

HIGH-THROUGHPUT PROFILING OF INFLUENZA A VIRUS HEMAGGLUTININ GENE AT SINGLE-NUCLEOTIDE RESOLUTION

2.1 ABSTRACT

Genetic research on influenza virus biology has been informed in large part by nucleotide variants present in seasonal or pandemic samples, or individual mutants generated in the laboratory, leaving a substantial part of the genome uncharacterized. Here, we have developed a single-nucleotide resolution genetic approach to interrogate the fitness effect of point mutations in 98% of the amino acid positions in the influenza A virus hemagglutinin (HA) gene. Our HA fitness map provides a reference to identify indispensable regions to aid in drug and vaccine design as targeting these regions will increase the genetic barrier for the emergence of escape mutations. This study offers a new platform for studying genome dynamics, structure-function relationships, virus-host interactions, and can further rational drug and vaccine design. Our approach can also be applied to any virus that can be genetically manipulated.

2.2 INTRODUCTION

The broad field of systems biology was significantly advanced in the past decade due to many technological improvements, such as the invention of DNA microarray, next generation sequencing, mass-spectrometry and other applications permitting high-throughput screenings [1,2]. These technical advancements have enabled large scale studies including interactomics, proteomics, transcriptomics, genomics, epigenomics, and metagenomics, which have revolutionized biomedical research [3–8]. A multitude of structure-function information is embedded in these studies that is valuable for rational drug and vaccine design. In addition, the continued development of *in silico* approaches to protein structural modeling, prediction, and design further complements the impact of high-throughput biological data [9, 11, 12, 32].

High-throughput tools have also influenced the advancement of genetic approaches. Traditional genetic methods focus on a single genotype-phenotype relationship at a time, and has been extensively employed to analyze individual mutations. In contrast, high-throughput genetic methods examine the phenotypic outcomes of multiple mutations simultaneously. Genome-wide insertional mutagenesis is a common high-throughput genetic approach. It has been employed to characterize bacterial genomes at a single-gene resolution level [13, 14]. A higher resolution has been achieved in two medically important RNA viruses, HCV and influenza [22, 23]. However, the maximum resolution of the insertional mutagenic approach is limited to a protein subdomain level and thus is insufficient to identify critical amino acid residues. Therefore, there is a demand for a high-throughput genetic platform at a single-residue resolution.

In this study, we developed a single-nucleotide resolution genetic approach using a large mutant library and a sensitive deep sequencing technique to annotate the influenza A virus hemagglutinin (HA) gene, which carries critical roles in receptor binding, viral entry, host shifts, and immune escape mechanisms. Here, we probe for fitness effects of individual substitutions in 98% of all amino acid positions across HA. Our results provide a comprehensive structure-function description of HA and offer a reference to identify potential vaccine epitope. More importantly, the high-throughput profiling platform established in this study can be applied to any genetically

manipulable viral gene or genome to probe mutational fitness effects under any specified growth condition.

2.3 RESULTS

High-throughput genetic approach at single-nucleotide resolution

The conceptual basis of our high-throughput genetic platform is to randomly mutagenize each position of the genome, monitor the enrichment or diminishment of each point mutation under a specified growth condition, and perform massive deep-sequencing to determine which mutations are associated with negative, neutral, or positive fitness outcomes under the given growth condition. The mutant library was created on influenza A/WSN/1933 (H1N1) hemagglutinin (HA) gene by performing error-prone PCR on the eight-plasmid reverse genetics system [27] (see materials and methods). Subsequently, the viral mutant library was generated by transfection and passaged for two 24-hour replication selection rounds in A549 cells (human lung epithelial carcinoma cells) (Fig. 2-1A). The plasmid library and the passaged viral library were each sequenced by Illumina HiSeq 2000. Individual mutants would experience an identical selection pressure with other mutants in the pool during the course of transfection and infection. Therefore, comparing the genetic compositions of the plasmid library and the passaged viral library reflects the variation in replication rates for each mutation. Here, we use a relative fitness index (RF index) as a proxy for the fitness effect of individual mutations. The RF index is calculated as:

RF index = (occurrence frequency in passaged library)/(occurrence frequency in plasmid library)

The occurrence frequency of individual mutations was largely expected to be lower than the sequencing error rate of 0.1% in the Illumina next generation sequencing (NGS). Therefore, we utilized a two-step PCR approach for library preparation to distinguish true mutations from sequencing errors (Fig. 2-1B). In the first PCR, the HA gene was divided into 12 amplicons for amplification with a unique tag assigned to individual molecules. In the second PCR, multiple identical copies for individual tagged molecules were generated. The input copy number for the second PCR was well-controlled such that after a sub-saturation PCR, individual tagged molecules would be sequenced $\sim$10 times. True mutations would exist in most, if not all, sequencing reads sharing the same tag, whereas sequencing errors would not. This error-correction approach is based on a valid assumption that occurrence of sequencing error is independent of the identity of

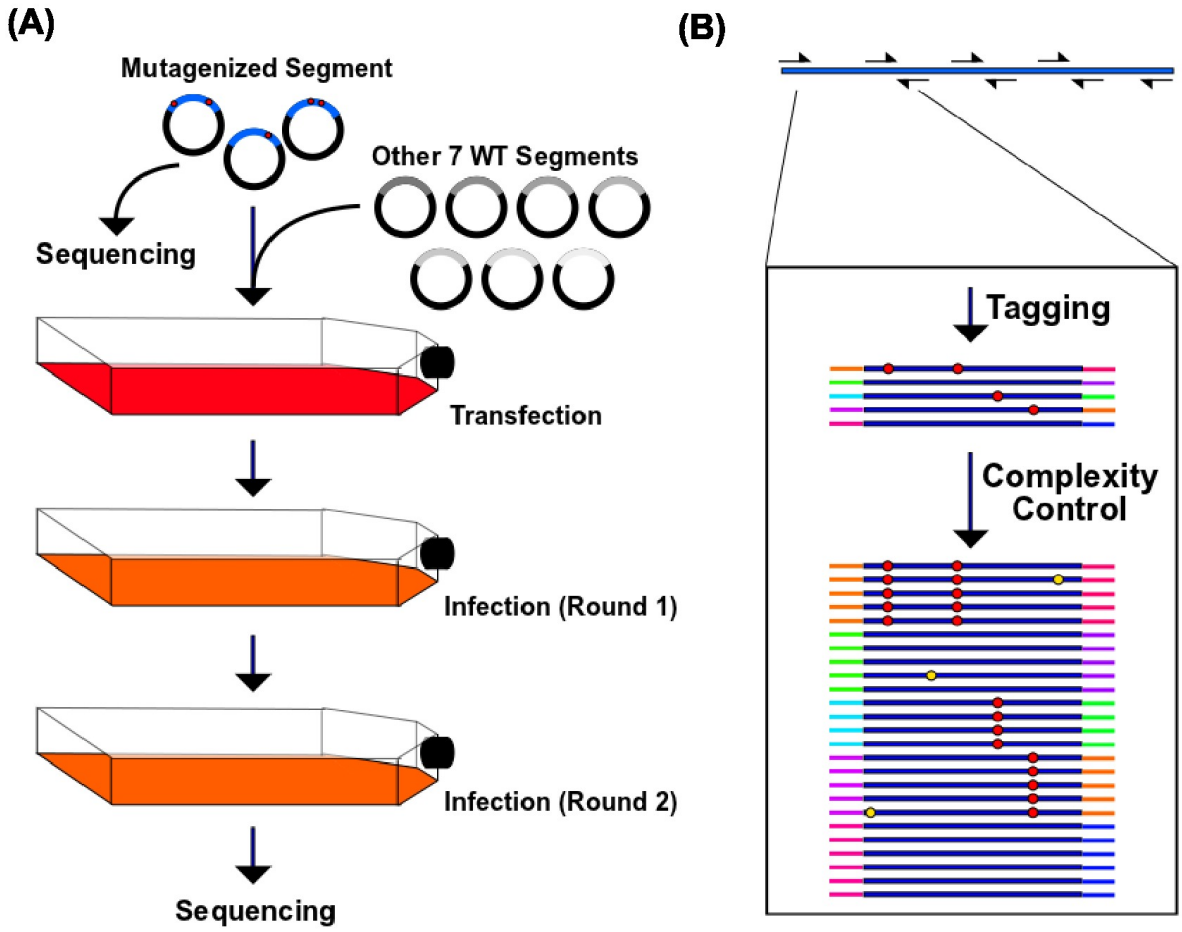


Figure 2-1. Mutant library passaging and sequencing library preparation. (A) The HA segment was randomized by error-prone PCR. The randomized segment with the remaining seven wild type segments were transfected into C227 cells to generate the viral mutant library. Two rounds of 24-hour infections were performed using A549 cells with an MOI of 0.05. Both the plasmid library and the passaged viral library were subjected to sequencing using the Illumina HiSeq 2000 machine. (B) The HA gene was divided into 12 amplicons for the first PCR. Unique tags were assigned to both ends of the individual molecules during the amplification process. The second PCR generated identical copies of individual molecules linked with unique tags. Red circles represent true mutations; Yellow circles represent sequencing errors.

the nucleotide tag [21]. Therefore, sequencing errors could be distinguished from true mutations. Individual molecules, each carrying a unique tag, have an average copy number of $\sim$10 (median = 10) in the sequencing data, which verified the sequencing library preparation design.

Point mutation fitness profiling of hemagglutinin

The RF indices of individual point mutations were profiled across 98% of amino acid positions of HA in biological duplicate (Spearman correlation = 0.78) (Fig. 2-2A). The remaining 2% of amino acid positions not observed were from the termini of HA, where the first and last amplicon primers are located. Silent mutations and nonsense mutations provided an internal control to access the data quality. In principle, silent mutations, which alter the nucleotide sequence but not the amino acid sequence, rarely impose a fitness cost. On the other hand, nonsense mutations, which result in a truncated protein product, are lethal to the virus. Indeed, our data is consistent with this notion. Silent mutations have a significantly higher RF index than nonsense mutations ($P < 2e^{-16}$, two-tailed Student's t-test) (Fig. 2-2B). In addition, the RF index distributions of silent mutations and nonsense mutations are well separated, which validated the reliability of our approach. However, several silent mutations with a low RF index were observed, which may be indicative of their roles in codon usage, RNA structure, and other functions beyond protein-coding.

Furthermore, the fitness data is consistent with the reported phenotypes of mutants that have been previously characterized in the literature. Examples include a temperature sensitive substitution (Y174H) [20], a host switching substitution (D238G) [21], two thermodynamic stabilizing substitutions (D111E and Q299R) [22], and four HA cleavage site substitutions (Y342H, Y342C, Y342N and Y342F) [23]. Y174H, D238G, Y342H, Y342C, and Y342N, which are expected to be deleterious under our experimental condition, have a relatively low RF index (ranging from 0.04 to 0.23). On the other hand, D111E, Q299R, and Y342F, which are expected to be neutral under our experimental condition, have a relatively high RF index (ranging from 0.37 to 1.03). These comparisons show the consistency between our dataset and the experimental results reported in the literature.

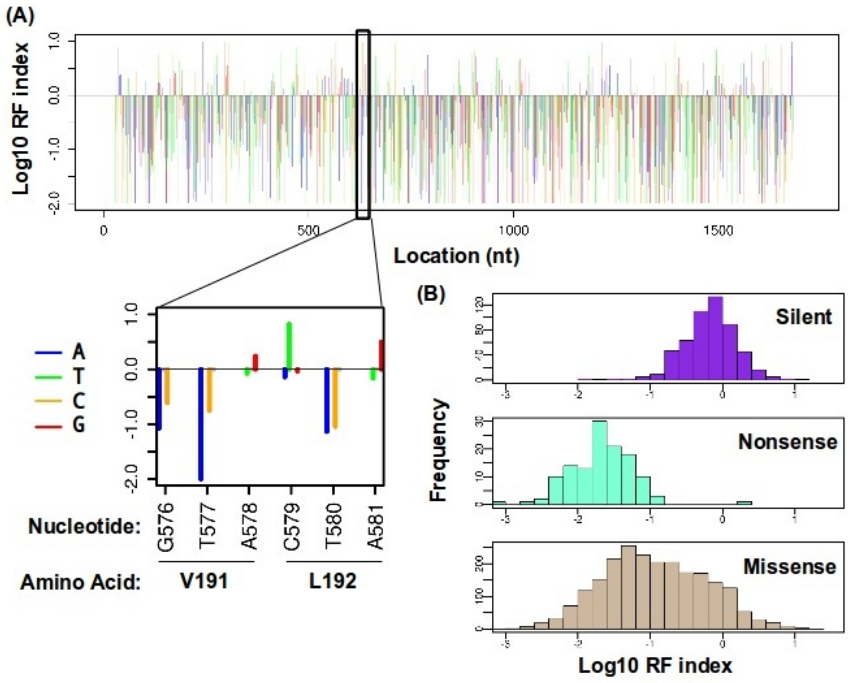


Figure 2-2. Single-nucleotide resolution fitness profiling. (A) The RF index for individual point mutations across the HA gene was computed. Log_{10} of the RF index is plotted on the y-axis. Each nucleotide position is represented by four consecutive lines for the RF indices that correspond to mutating to A (blue), T (green), C (orange), or G (red). The Log_{10} RF index of wild type (WT) nucleotides is set as zero. Only point mutations with a coverage of $\geq$ 30 tag-conflated reads in the plasmid library are shown. Otherwise, point mutations are plotted as a gray circle on the zero baseline. A short region is shown as an inset to demonstrate the resolution of our dataset. (B) The distributions of the $\log_{10}$ RF indices for silent substitutions, nonsense substitutions and missense substitutions are displayed as histograms. Mutations located at the 5' terminal 200 bp and 3' terminal 200 bp regions are not included in this analysis to avoid confounding by the vRNA packaging signal [19].

Independent experimental validation also confirmed our dataset. Six randomly selected point mutations were individually reconstructed and analyzed. RF indices of each mutation have a positive correlation with the $TCID_{50}$ value measured from a rescue experiment (Fig. 2-3A-B). Overall, these analyses verified the reliability of the fitness profiling data and demonstrated our platform to be comprehensive and at high resolution.

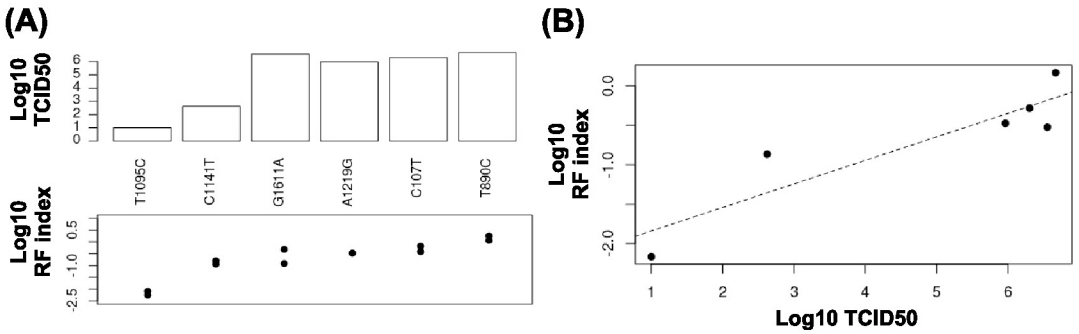


Figure 2-3. Experimental validation.(A) The top panel displays the $\log_{10}$ $TCID_{50}$ value of mutant virus rescued from transfection. The bottom panel represents their $\log_{10}$ RF indices from the biological duplicate. (B) A Pearson correlation of 0.9 is obtained between $\log_{10}$ $TCID_{50}$ from transfection (x-axis) and $\log_{10}$ RF index (y-axis)

Structural analysis of hemagglutinin

Our platform has a high sensitivity for monitoring negative selection in addition to positive selection and therefore enables the identification of deleterious mutations that disappear throughout viral passaging. The availability of the influenza HA crystal structure allowed us to further extrapolate structural insights from our dataset. A weak, yet significant spearman correlation of 0.30 was observed between the RF index and the relative solvent accessible surface area (SASA) of HA ($P < 2e^{-16}$). This indicates that surface residues are more tolerant to substitutions than core residues, which is consistent with observations in cellular proteins [24, 25]. We also analyzed the fitness effects of mutations in different types of structural elements, namely α-helices (mean $\log_{10}$ RF index = -1.19), β-strands (mean $\log_{10}$ RF index = -0.97), turns (mean $\log_{10}$ RF index = -0.98) and coils (mean $\log_{10}$ RF index = -1.01). Interestingly, mutations in α-helices are more deleterious than mutations in β-strands ($P = 1e^{-4}$), turns ($P = 1e^{-3}$) and coils ($P = 2e^{-3}$). In contrast, the

fitness effects of mutations in β-strands, turns and coils are not significantly different from each other (P $>$ 0.4). This result implies that most functional elements in HA are contained within α-helices.

We further investigated each α-helix by computing their individual mean $\log_{10}$ RF index (Fig. 2-4A). As expected from the SASA analysis, the α-helices located at the core of HA_1 are the least tolerant to mutations (red and pink, mean $\log_{10}$ RF index = -1.52 and -1.40 respectively). The other α-helix in HA_1 is also relatively intolerable to mutations (orange, mean $\log_{10}$ RF index = -1.11), which is consistent with its role in receptor binding for viral entry. [26]. In HA_2, the two α-helices located at the stem-loop region are relatively intolerable to mutations (green and cyan, mean $\log_{10}$ RF index = -1.11 and -1.22 respectively), which can be attributed to their functional role in membrane fusion during viral entry [27]. In fact, all of the mean $\log_{10}$ RF indices reported above are lower than that of the entire HA (mean $\log_{10}$ RF index = -1.04). Together, these findings demonstrated that α-helices in HA are important for different functional mechanisms.

Interestingly, the non-structural loop region (blue) that interspaces the aforementioned helices (green and cyan) is more tolerant to mutations compared to its neighboring α-helices (mean $\log_{10}$ RF index = -0.76) (Fig. 2-4A). This region undergoes a transition from a non-structural loop to an α-helix during membrane fusion. Nonetheless, the relatively high RF index in this region suggests that the structural requirement for this transition is not stringent. This is further evidenced by a proline substitution analysis (Fig. 2-4B). Among all 20 standard amino acids, proline has the poorest α-helix formation propensity as its presence would result in a break or a kink of an α-helix [28]. Therefore, it is expected that proline substitutions in an α-helix would carry a low RF index (deleterious). Indeed, all proline substitutions in the HA α-helices have a $\log_{10}$ RF index $<$ -1. In contrast, two out of three proline substitutions in the non-structural loop have a $\log_{10}$ RF index $>$ -1 (-0.81 and -0.19 respectively). This result suggests that the formation of a continuous α-helix in this region is not a strict requirement during membrane fusion.

We also performed an in depth analysis on the α-helix that is important for homotrimer formation (colored in cyan in Fig. 2-4A). Helix wheel projection showed that high hydrophobicity was

critical at heptad position d (Fig. 2-4C). We further investigated the RF index of those amino acid substitutions at heptad position d (Fig. 2-4D). Silent mutation at G430 had the lowest RF index (0.24) among all silent mutations at this heptad position. This RF index was employed as a reference to identify substitutions that has a relatively neutral fitness effect. Only three out of 27 amino acid substitutions at this heptad position has an RF index $\geq$0.24, namely Y437F (RF index = 0.35), V465I (RF index = 0.40) and V465A (RF index = 0.30). These three substitutions are conserved in volume and hydrophobicity, which suggests that residues at heptad position d has a stringent structural constraint in side chain conformation and hydrophobicity for homotrimer formation.

Identification of essential regions

Our profiling also provides information to identify possible essential protein surfaces and indispensable regions useful for vaccine epitopes. The RF indices of the most destructive substitutions in our dataset can be projected on the HA structure to identify putative functional regions that cannot tolerate certain amino acid substitutions (Fig. 2-5A-B). Whereas the RF indices of the least destructive substitutions for HA is projected on the HA structure to identify essential regions that are intolerable to any substitution (Fig. 2-5C). As expected, the trimer formation surface (Fig. 2-5A) and the stem domain (Fig. 2-5B-C), which is the major functional component of the membrane fusion machinery in HA, show as essential regions in our profiling data. In addition, our dataset identified the cross-subtype conserved influenza HA stalk region as an indispensable region (Fig. 2-5C-D), which is at the binding site of the proposed influenza universal antibody, CR6261 [29,30]. Although several missense substitutions in the binding site are allowed, they are conservative substitutions (N389D and T392S) unlikely to disrupt antibody recognition (Fig. 2-5C-D). It confirms the promising aspect of the proposed universal antibody [30]. In addition, the main antigenic sites on the globular head of HA were largely tolerable to substitutions (Fig. 2-5C). This observation suggests a functional basis for the tendency of this domain to rapidly undergo genetic drift, which adversely affects both natural and vaccine-induced immunity [31]. Overall, our work details the genetic cost for individual point mutations across HA – the primary target of anti-influenza neutralizing antibodies [29–33]. This dataset therefore provides a valuable reference for rational vaccine design.

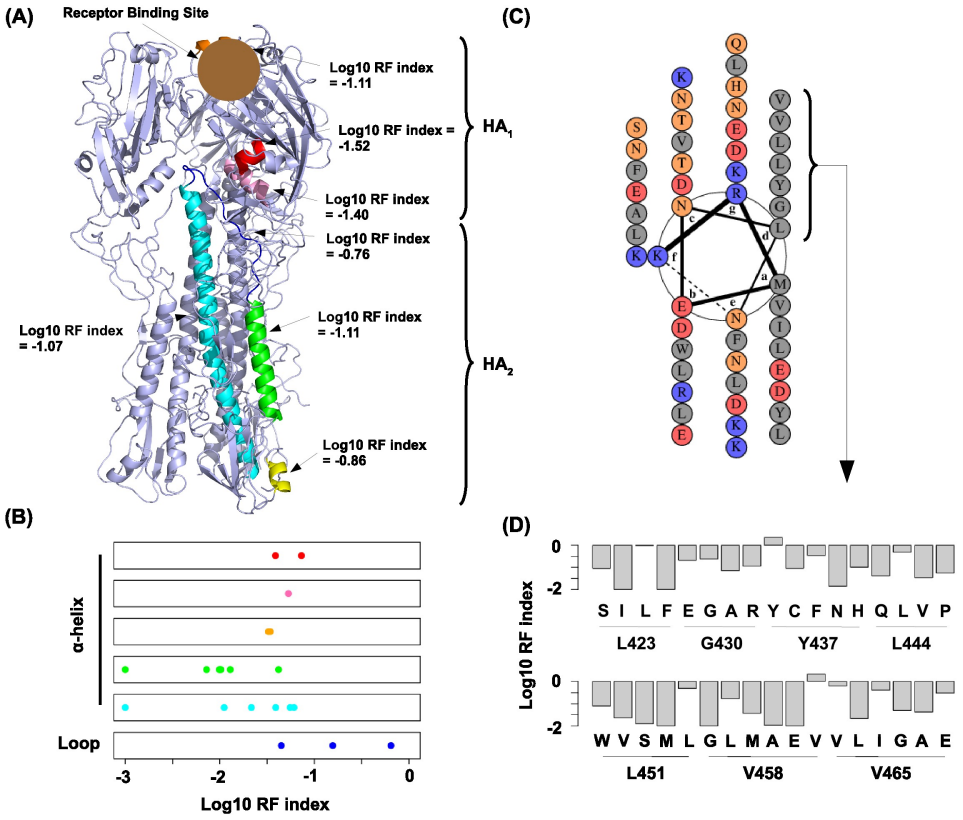


Figure 2-4. Structural analysis on hemagglutinin. (A) All α-helices (orange, red, pink, cyan, green, yellow) and a non-structural loop (blue) in HA are highlighted. Mean $\log_{10}$ RF indices for individual highlighted structural elements are shown. (B) The $\log_{10}$ RF indices for all observed X $\rightarrow$ P mutations (where X can be any amino acids but P) in individual highlighted structural elements are plotted as stripcharts. The colors of the stripcharts match the highlight colors of the corresponding structural elements in panel A. The bottom stripchart represents the non-structural loop that undergoes α-helix formation during membrane fusion. (C) Helical wheel was constructed by DrawCoil 1.0 (http://www.grigoryanlab.org/drawcoil/). Amino acid property of each residue is color coded. Polar: orange; Hydrophobic: grey; Positively charged: red; Negatively charged: blue. (D) The bar chart represents the RF indices of all profiled amino acid substitutions at heptad position d. RF indices of silent mutations are also included for comparison.

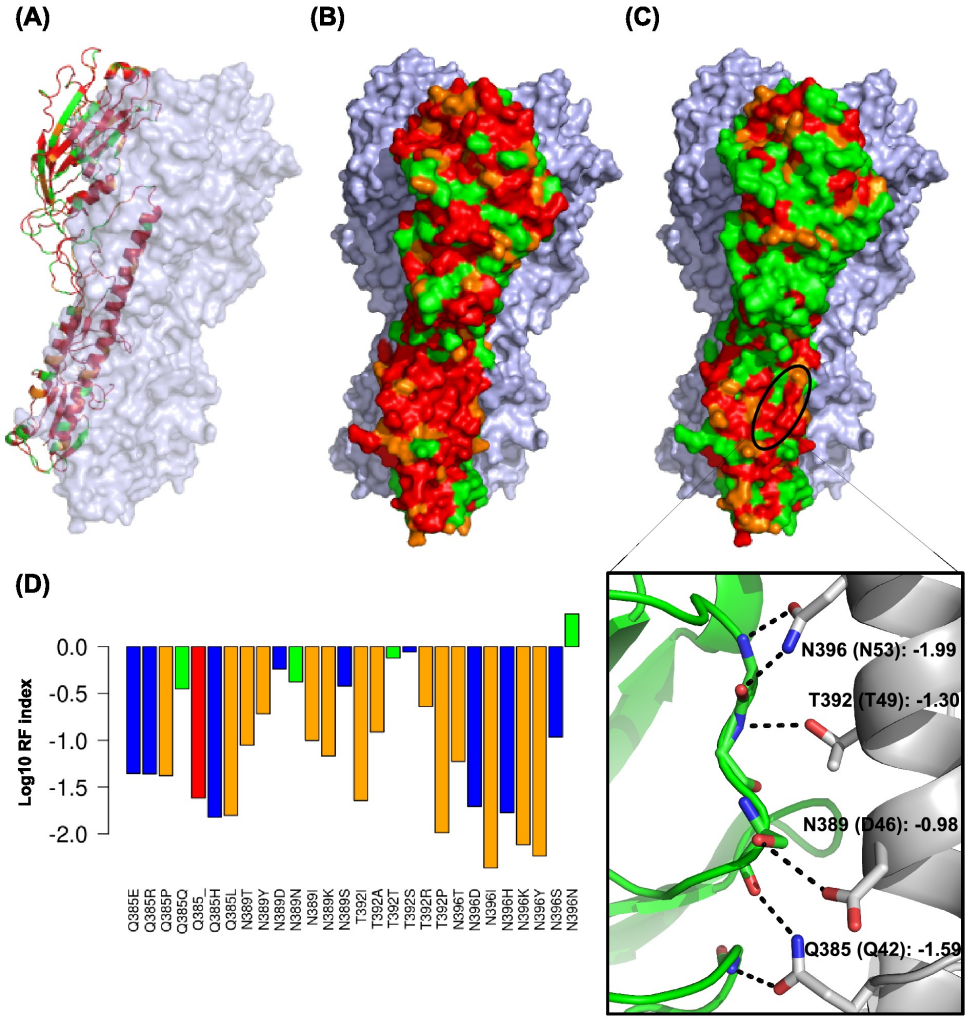


Figure 2-5. Essential regions on hemagglutinin. The RF indices of (A-B) the most destructive missense substitutions or (C) the least destructive missense substitutions in the profiling data for individual amino acids are projected on the HA protein structure to identify essential regions intolerable to mutations. The inset represents the side chain interaction between HA (grey) and the proposed influenza universal antibody CR6261 (green) (PDB: 3GBN) [29]. Parentheses represent the residue naming according to HA_2 [29]. The mean $\log_{10}$ RF indices of nonconservative mutations for each residue are shown. All hydrogen bonds (black dotted lines) are displayed as described [29]. (A-C) Red: RF index < 0.05; Orange: RF index < 0.1; Green: other. The structure is based on PDB: 1RUZ [34]. (D) The RF indices for missense mutations within the universal antibody recognition sites are shown. Red: nonsense substitution; orange: nonconservative substitution; blue: conservative substitution; green: silent mutation.

2.4 DISCUSSION

Traditionally, critical residues on a viral genome are discovered by testing individual mutants and requires multiple assays to dissect the associated biological functions. The low throughput nature of this process limits the number of mutants tested. In this study, we have developed a comprehensive strategy using the influenza A virus as a model system to profile the fitness effects of individual point mutations and to identify essential residues throughout the HA gene in a high-throughput manner.

Recently, two studies that describe the development of a deep sequencing-based high-throughput genetic platform at single-nucleotide resolution have been reported in the literature [51,52]. Robins et al. probed for essential residues in T7 bacteriophage and T7-like virus JSF7 of *Vibrio cholerae* using mutant libraries constructed by chemical-induced transition of a GC base pair to an AT base pair [51]. Acevedo et al., on the other hand, interrogated the fitness effects of individual point mutations that naturally emerged in an evolving poliovirus population which has a high mutation rate, rather than employing any engineering strategy of introducing mutations [52]. In this study, we have developed a novel strategy which utilizes a saturated point mutation library together with a sensitive sequencing approach. When compared to the two aforementioned approaches, our method is more comprehensive and unbiased due to the mutant library construction strategy, which is independent of spontaneous mutations. This application can be extended to other influenza genes and to other genetically manipulable viruses under any applied selection condition at a single-nucleotide resolution level.

Identification of residues essential for viral replication is often inferred by sequence conservation. Observed sequence conservation derives from the viral sequences that initiated the endemic, and is influenced by the host genetic background and the specific immune responses associated with the host. Conservation is not equivalent to essentialness for viral replication in cells. Mutational analysis of conserved amino acid residues on influenza A virus has revealed that a significant fraction of conserved residues are dispensable in viral replication [5,6,38]. In addition, new mutations emerge every flu season, implying that a certain portion of residues that are conserved currently

are still capable to mutate in the natural environment and provide a fitness advantage under future unforeseen selection pressures. This also suggests that a conserved amino acid may not necessarily be essential to viral replication. Additionally, analyses of conserved sequences provide information on viral genetic elements that survived in the selected human population in recent history, but does not provide much information on viral genetic elements that were unable to survive the selection process, nor about which host factor was responsible for exerting the selection. Our approach provides a complementary, yet more direct approach to identify amino acid residues that are critical for viral replication in a defined cellular environment. Nonetheless, to be more comprehensive, similar studies should be performed with strains across subtypes and include different selection conditions.

In summary, the platform described here enabled the simultaneous functional profiling of point mutations across the entire influenza HA at single-nucleotide resolution to determine their roles in viral replication. Our platform provides an efficient tool to address several important biomedical questions. The fitness profiling data allows the study of structure-function relationships at single-amino acid resolution. It enables the search for essential protein surfaces on available structures and thus offers a reference for drug design approaches that aim to increase the genetic barrier for the emergence of escape mutations [40–42]. Essential peptide stretches could also provide potential targets for drug and vaccine development [43]. Our genetic platform can be applied to study viral genome dynamics and identify critical residues for virus-host interactions in a specific cellular responses (such as apoptosis, autophagy, inflammasome induction, ER stress, etc.) and immune responses (such as NK cells, T cells, antibodies, macrophages, cytokines, etc.) [44, 45]. The current development of a live attenuated influenza vaccine has been based on the modification of NS1 to increase interferon sensitivity [37]. However, this study provides a platform to explore alternative strategies. Comparing the *in vitro* fitness profile with an *in vivo* profile could also permit the identification of mutants that replicate efficiently *in vitro* but not *in vivo*. The resultant information when coupled with known mutants that are sensitive to a specified immune response could help achieve a higher titer during vaccine production, but exhibit an attenuated phenotype after injection into the human body where an intact immune system is present. Most importantly, our platform is applicable to other viral or microbial genomes where genetic manipulation is available in the

laboratory. The sensitivity of our platform will increase as NGS technology improves. With the continued development of NGS technology, we foresee that our platform will be further advanced and can be applied at a much lower cost.

2.5 MATERIALS AND METHODS

Viral mutant library and point mutations

The plasmid mutant library was created by performing error-prone PCR on the HA segment of the eight-plasmid reverse genetics system of influenza A/WSN/1933 (H1N1) [27]. We PCR-amplified the HA gene insert with error-prone polymerase Mutazyme II (Stratagene, La Jolla, CA). The mutation rate of the error-prone PCR was optimized by adjusting the input template amount to avoid the accumulation of deleterious mutations. The restriction enzyme site BsmBI was present in the PCR primers, and used to clone into a BsmBI-digested parental vector pHW2000. Ligations were carried out with high concentration T4 ligase (Life Technologies, Carlsbad, CA). Transformations were carried out with electrocompetent MegaX DH10B T1R cells (Life Technologies), and $>$ 200,000 colonies were scraped and directly processed for plasmid DNA purification (Qiagen Sciences, Germantown, MD). As extensive trans-complementation was expected during the transfection step, $>$ 35 million cells were used for transfection to average out any bias or artifact generated from possible trans-complementation. Point mutants for the validation experiment were constructed using the QuikChange XL Mutagenesis kit (Stratagene) according to the manufacturer's instructions.

Transfections, infections, and titering

C227 cells, a dominant negative IRF-3 stably expressing cell line derived from human embryonic kidney (293T) cells, were transfected with Lipofectamine 2000 (Life Technologies) using the HA mutant library plasmid plus 7 other wildtype plasmids. Supernatant was replaced with fresh cell growth medium at 24 hrs and 48 hrs post-transfection. At 72 hrs post-transfection, supernatant containing infectious virus was harvested, filtered through a 0.45 um MCE filter, and stored at -80 degree Celsius. The $TCID_{50}$ was measured on A549 cells (human lung carcinoma cells).

Virus from the C227 transfection was used to infect A549 cells at an MOI of 0.05. Infected cells were washed three times with PBS followed by the addition of fresh cell growth medium at 2 hrs post-infection. Virus was harvested at 24 hrs post-infection. For the mutant library profiling, HA mutant library was passaged for two 24-hour rounds in A549 cells. Our pilot experiments as well as our previous study revealed that two rounds of passaging were suffcient for profiling [48]. The

biological duplicate was performed by an independenly transfected viral library, followed by two rounds of passaging as described above.

Sequencing library preparation

Viral RNA was extracted from the passaged viral mutant library using QIAamp Viral RNA Mini Kit (Qiagen Sciences) and was reverse transcribed to cDNA using Superscript III reverse transcriptase (Life Technologies). DNA from the plasmid library or cDNA from the passaged viral mutant library were amplified with both forward and reverse primers each flanked with a 6 "N" tag and the Illumina flow cell adapter region. Flanking region for 5' primer: 5'-CTA CAC GAC GCT CTT CCG ATC TNN NNN N-3', Flanking region for 3' primer: 5'-TGC TGA ACC GCT CTT CCG ATC TNN NNN N-3'. Following PCR, 12 amplicon products were pooled together. 1.5 million copies of the pooled product were used as the input for the second PCR, which was equivalent to 10 paired-end reads per molecule if 15 million paired-end reads were sequenced. 5'-AAT GAT ACG GCG ACC ACC GAG ATC TA CAC TCT TTC CCT ACA CGA CGC TCT TCC G-3' and 5'-CAA GCA GAA GAC GGC ATA CGA GAT CGG TCT CGG CAT TCC TGC TGA ACC GCT CTT CCG-3' were used as the primers for the second PCR. Products of the second PCR were submitted for next generation sequencing. The error-correction technique described in this study shared the same philosophy as described for detecting rare mutations in human cells [21]. However, this study included the fine restraint of limiting the input tagged template copy number and PCR efficiency during the second step PCR to accurately control the distribution of cluster size in the sequencing output to a median of 10. Raw sequencing data have been submitted to the NIH Short Read Archive under accession number: PRJNA243038.

Data analysis

Sequencing reads were mapped by BWA with a maximum of six mismatches and no gap [55]. Amplicons with the same tag were collected to generate a read cluster. Since each read cluster was originated from the same template, true mutations were called only if the mutations occurred in 90% of the reads within a read cluster. We acknowledged that this error-correction approach would only correct errors that occured during the deep sequencing process but not those that were

introduced during the reverse transcription process. Read clusters with a size below three reads were filtered out. Read clusters were further conflated into "error-free" reads. Average coverages in terms of "error-free" reads were 177028 per nucleotide in the plasmid mutant library, 112355 per nucleotide in replicate 1 of passaged viral mutant library, and 161773 per nucleotide in replicate 2 of passaged viral mutant library. Relative fitness index (RF index) for individual point mutations was computed by:

(occurrence frequency in passaged library)/(occurrence frequency in plasmid library)

For all the downstream analysis, only point mutations covered with $\geq$ 30 tag-conflated reads ("error-free" reads) in the plasmid library were included. This arbitrary cutoff filtered out mutants with low statistical confidence, which is $\sim$16% of all possible point mutations. In addition, all C $\rightarrow$ A and G $\rightarrow$ T mutations are not included in the reported dataset due to an observed DNA oxidative damage during library preparation [49].

Structural analysis

The solvent accessible surface area (SASA) for individual residues was computed from PyMOL using the default "get_area" function. SASA obtained from the folded structure was then normalized with the SASA calculated from an unfolded structure to obtain the relative SASA. Secondary structure assignment was performed by STRIDE [50]. The structural analysis was based on PDB: 1RUZ [34]. A two-tailed Student's t-test was employed to compare the $\log_{10}$ RF indices in different types of structural elements. Only missense mutations are included in the analysis unless otherwise stated.

2.6 BIBLIOGRAPHY

1. Mardis ER (2008) Next-generation dna sequencing methods. Annu Rev Genomics Hum Genet 9: 387–402.

2. Schena M, Shalon D, Davis RW, Brown PO (1995) Quantitative monitoring of gene expression patterns with a complementary dna microarray. Science 270: 467–470.

3. Barski A, Cuddapah S, Cui K, Roh TY, Schones DE, et al. (2007) High-resolution profiling of histone methylations in the human genome. Cell 129: 823–837.

4. Chen K, Pachter L (2005) Bioinformatics for whole-genome shotgun sequencing of microbial communities. PLoS Comput Biol 1: 106–112.

5. Mavromatis K, Land ML, Brettin TS, Quest DJ, Copeland A, et al. (2012) The fast changing landscape of sequencing technologies and their impact on microbial genome assemblies and annotation. PLoS One 7: e48837.

6. Wang Z, Gerstein M, Snyder M (2009) Rna-seq: a revolutionary tool for transcriptomics. Nat Rev Genet 10: 57–63.

7. Hann MM, Oprea TI (2004) Pursuing the leadlikeness concept in pharmaceutical research. Curr Opin Chem Biol 8: 255–263.

8. Sanchez C, Lachaize C, Janody F, Bellon B, Rder L, et al. (1999) Grasping at molecular interactions and genetic networks in drosophila melanogaster using flynets, an internet database. Nucleic Acids Res 27: 89–94.

9. Brooks BR, Brooks CL, Mackerell AD, Nilsson L, Petrella RJ, et al. (2009) Charmm: the biomolecular simulation program. J Comput Chem 30: 1545–1614.

10. Kellogg EH, Leaver-Fay A, Baker D (2011) Role of conformational sampling in computing mutation-induced changes in protein structure and stability. Proteins 79: 830–838.

11. Koga N, Tatsumi-Koga R, Liu G, Xiao R, Acton TB, et al. (2012) Principles for designing ideal protein structures. Nature 491: 222–227.

12. Whitehead TA, Chevalier A, Song Y, Dreyfus C, Fleishman SJ, et al. (2012) Optimization of affinity, specificity and function of designed influenza inhibitors using deep sequencing. Nat Biotechnol 30: 543–548.

13. Christen B, Abeliuk E, Collier JM, Kalogeraki VS, Passarelli B, et al. (2011) The essential genome of a bacterium. Mol Syst Biol 7: 528.

14. van Opijnen T, Camilli A (2010) Genome-wide fitness and genetic interactions determined by tn-seq, a high-throughput massively parallel sequencing method for microorganisms. Curr Protoc Microbiol Chapter 1: Unit1E.3.

15. Arumugaswami V, Remenyi R, Kanagavel V, Sue EY, Ho TN, et al. (2008) High-resolution functional profiling of hepatitis c virus genome. PLoS Pathog 4: e1000182.

16. Heaton NS, Sachs D, Chen CJ, Hai R, Palese P (2013) Genome-wide mutagenesis of influenza virus reveals unique plasticity of the hemagglutinin and ns1 proteins. Proc Natl Acad Sci U S A 110: 20248–20253.

17. Neumann G, Watanabe T, Ito H, Watanabe S, Goto H, et al. (1999) Generation of influenza a viruses entirely from cloned cdnas. Proc Natl Acad Sci U S A 96: 9345–9350.

18. Kinde I, Wu J, Papadopoulos N, Kinzler KW, Vogelstein B (2011) Detection and quantification of rare mutations with massively parallel sequencing. Proc Natl Acad Sci U S A 108: 9530–9535.

19. Marsh GA, Hatami R, Palese P (2007) Specific residues of the influenza a virus hemagglutinin viral rna are important for efficient packaging into budding virions. J Virol 81: 9727–9736.

20. Nakajima S, Brown DJ, Ueda M, Nakajima K, Sugiura A, et al. (1986) Identification of the defects in the hemagglutinin gene of two temperature-sensitive mutants of a/wsn/33 influenza virus. Virology 154: 279–285.

21. Leung HSY, Li OTW, Chan RWY, Chan MCW, Nicholls JM, et al. (2012) Entry of influenza a virus with a 2,6-linked sialic acid binding preference requires host fibronectin. J Virol 86: 10704–10713.

22. Bloom JD, Glassman MJ (2009) Inferring stabilizing mutations from protein phylogenies: application to influenza hemagglutinin. PLoS Comput Biol 5: e1000349.

23. Sun X, Tse LV, Ferguson AD, Whittaker GR (2010) Modifications to the hemagglutinin cleavage site control the virulence of a neurotropic h1n1 influenza virus. J Virol 84: 8683–8690.

24. Tokuriki N, Stricher F, Schymkowitz J, Serrano L, Tawfik DS (2007) The stability effects of protein mutations appear to be universally distributed. J Mol Biol 369: 1318–1332.

25. Guo HH, Choe J, Loeb LA (2004) Protein tolerance to random amino acid change. Proc Natl Acad Sci U S A 101: 9205–9210.

26. White CL, Janakiraman MN, Laver WG, Philippon C, Vasella A, et al. (1995) A sialic acid-derived phosphonate analog inhibits different strains of influenza virus neuraminidase with different efficiencies. J Mol Biol 245: 623–634.

27. Bullough PA, Hughson FM, Skehel JJ, Wiley DC (1994) Structure of influenza haemagglutinin at the ph of membrane fusion. Nature 371: 37–43.

28. Pace CN, Scholtz JM (1998) A helix propensity scale based on experimental studies of peptides and proteins. Biophys J 75: 422–427.

29. Ekiert DC, Bhabha G, Elsliger MA, Friesen RHE, Jongeneelen M, et al. (2009) Antibody recognition of a highly conserved influenza virus epitope. Science 324: 246–251.

30. Throsby M, van den Brink E, Jongeneelen M, Poon LLM, Alard P, et al. (2008) Heterosubtypic neutralizing monoclonal antibodies cross-protective against h5n1 and h1n1 recovered from human igm+ memory b cells. PLoS One 3: e3942.

31. Chen JR, Ma C, Wong CH (2011) Vaccine design of hemagglutinin glycoprotein against influenza. Trends Biotechnol 29: 426–434.

32. Sui J, Hwang WC, Perez S, Wei G, Aird D, et al. (2009) Structural and functional bases for broad-spectrum neutralization of avian and human influenza a viruses. Nat Struct Mol Biol 16: 265–273.

33. Corti D, Voss J, Gamblin SJ, Codoni G, Macagno A, et al. (2011) A neutralizing antibody selected from plasma cells that binds to group 1 and group 2 influenza a hemagglutinins. Science 333: 850–856.

34. Gamblin SJ, Haire LF, Russell RJ, Stevens DJ, Xiao B, et al. (2004) The structure and receptor binding properties of the 1918 influenza hemagglutinin. Science 303: 1838–1842.

35. Robins WP, Faruque SM, Mekalanos JJ (2013) Coupling mutagenesis and parallel deep sequencing to probe essential residues in a genome or gene. Proc Natl Acad Sci U S A 110: E848–E857.

36. Acevedo A, Brodsky L, Andino R (2014) Mutational and fitness landscapes of an rna virus revealed through population sequencing. Nature 505: 686–690.

37. Chu C, Fan S, Li C, Macken C, Kim JH, et al. (2012) Functional analysis of conserved motifs in influenza virus pb1 protein. PLoS One 7: e36113.

38. Li Z, Watanabe T, Hatta M, Watanabe S, Nanbo A, et al. (2009) Mutational analysis of conserved amino acids in the influenza a virus nucleoprotein. J Virol 83: 4153–4162.

39. Stewart SM, Pekosz A (2011) Mutations in the membrane-proximal region of the influenza a virus m2 protein cytoplasmic tail have modest effects on virus replication. J Virol 85: 12179–12187.

40. Boltz DA, Aldridge JR, Webster RG, Govorkova EA (2010) Drugs in development for influenza. Drugs 70: 1349–1362.

41. Memoli MJ, Morens DM, Taubenberger JK (2008) Pandemic and seasonal influenza: therapeutic challenges. Drug Discov Today 13: 590–595.

42. Pinto LH, Lamb RA (2007) Controlling influenza virus replication by inhibiting its proton channel. Mol Biosyst 3: 18–23.

43. Tan PT, Khan AM, August JT (2011) Highly conserved influenza a sequences as t cell epitopes-based vaccine targets to address the viral variability. Hum Vaccin 7: 402–409.

44. Ehrhardt C, Seyer R, Hrincius ER, Eierhoff T, Wolff T, et al. (2010) Interplay between influenza a virus and the innate immune signaling. Microbes Infect 12: 81–87.

45. Rossman JS, Lamb RA (2009) Autophagy, apoptosis, and the influenza virus m2 protein. Cell Host Microbe 6: 299–300.

46. Richt JA, Garca-Sastre A (2009) Attenuated influenza virus vaccines with modified ns1 proteins. Curr Top Microbiol Immunol 333: 177–195.

47. Wu NC, Young AP, Dandekar S, Wijersuriya H, Al-Mawsawi LQ, et al. (2013) Systematic identification of h274y compensatory mutations in influenza a virus neuraminidase by high-throughput screening. J Virol 87: 1193–1199.

48. Li H, Durbin R (2009) Fast and accurate short read alignment with burrows-wheeler transform. Bioinformatics 25: 1754–1760.

49. Lou DI, Hussmann JA, McBee RM, Acevedo A, Andino R, et al. (2013) High-throughput dna sequencing errors are reduced by orders of magnitude using circle sequencing. Proc Natl Acad Sci U S A 110: 19872–19877.

50. Heinig M, Frishman D (2004) Stride: a web server for secondary structure assignment from known atomic coordinates of proteins. Nucleic Acids Res 32: W500–W502.

CHAPTER 3

FUNCTIONAL CONSTRAINT PROFILING OF A VIRAL PROTEIN REVEALS DISCONCORDANCE OF EVOLUTIONARY CONSERVATION AND FUNCTIONALITY

3.1 ABSTRACT

Viruses often encode proteins with multiple functions due to their compact genomes. Existing approaches to identify functional residues largely rely on sequence conservation analysis. Inferring functional residues from sequence conservation can produce false positives, in which the conserved residues are functionally silent, or false negatives, where functional residues are not identified since they are species-specific and therefore non-conserved. Furthermore, the tedious process of constructing and analyzing individual mutations limits the number of residues that can be examined in a single study. Here, we developed a systematic approach to identify the functional residues of a viral protein by coupling experimental fitness profiling with protein stability prediction using the PA influenza virus polymerase subunit as the target protein. We identified a significant number of functional residues that were influenza type-specific and were evolutionarily non-conserved among different influenza subtypes. Our results indicate that type-specific functional residues are prevalent and would not otherwise be identified by sequence conservation analysis alone. More importantly, this technique can be adapted to any viral (and potentially non-viral) protein where structural information is available.

3.2 INTRODUCTION

To comprehensively describe the functional roles of a given protein, which are often diverse for many viral proteins and include catalytic activity, intermolecular interactions, and/or cofactor binding, it is necessary to identify the individual functional residues that carryout the biochemical mechanism. Sequence conservation analysis is a common strategy to search for functional residues and is facilitated by the availability of public protein sequence databases [1–3]. The underlying logic is composed of two parts. First, functional residues are essential. Second, essential residues are conserved. However, the reverse may not hold true – conserved residues are not necessary essential. With the extensively studied influenza A virus, several groups have experimentally demonstrated that conserved residues need not be essential for viral replication [5,6,55]. In addition, a residue shown to be essential for viral replication can also be the result of stability constraints, where the residue is essential for protein stability and expression levels, rather than due to the functional constraints [7].

Another caveat of sequence conservation analysis is the inefficacy for identifying species-specific functional residues. This issue is often overlooked. During natural evolution, continuous diversification and adaptation leads to the acquisition of new functions. For example, NS1 from influenza B but not influenza A interacts with ISG15 [8]; NS1 from influenza A but not influenza B interacts with CPSF30 [9]. Furthermore, certain phosphorylation sites are not conserved across influenza A and B viruses [10]. In fact, non-conserved functional residues have been demonstrated in various organisms [11–14]. Consequently, when sequence conservation analysis is based on a set of diverse homologs, as is the case when comparing sequence conservation across influenza types A, B, and C, species-specific functional residues appear as non-conserved residues and are classified as non-functional. As a result, development of a sequence conservation-independent approach is needed to provide an unbiased assessment for the functionality of individual residues and to permit a systematic interrogation of the relationship between functionality and evolutionary conservation.

The influenza A virus PA polymerase subunit consists of a $\sim$25 kD N-terminal domain and a $\sim$55

kD C-terminal domain [15, 16]. Structural information for both domains is available [17–20]. PA forms a heterotrimer complex with two other influenza virus proteins, PB1 and PB2. Together, they function as an RNA-dependent RNA polymerase. The three subunits perform distinct functions, which contribute to the replication and transcription of the viral RNA genome. PB1 binds to the viral promoter and is the catalytic subunit for viral RNA synthesis [21]. PB2 is essential for the transcription of viral RNA and can bind to the 5' cap of host pre-mRNAs for "cap-snatching" [22–24]. PA is required for both replication and transcription of the viral RNA and contains an endonuclease catalytic site for cleaving the capped RNA primer [25–28]. It has also been reported that PA may be involved in other viral processes, such as viral assembly [29, 30], and may possess protease activity [31, 32]. Recently, several groups have proposed targeting the influenza PA polymerase subunit for antiviral drug development as it is an essential component for viral replication [33–39].

In this study, we have developed a systematic approach that is independent of any prior knowledge in sequence conservation to identify functional residues at single-amino acid resolution. In this strategy, we coupled a high-throughput fitness profiling platform with an *in silico* mutant stability prediction. We employed the influenza A virus PA polymerase subunit as the target protein, due to the availability of structural information and the extensive information available for natural sequence variants. The fitness effects of amino acid substitutions were profiled across 94% of all PA protein residues using a novel "small library" approach. Computational modeling predicted the stability effect of all individual substitutions, thus uncovering the structural constraints for individual residues. By integrating the fitness and structural information, we identified known functional sites previously documented in the literature and provide additional insight into the structure-function relationship of the influenza PA protein. We further examined the relationship between evolutionary conservation and functional constraints and show that functional residues are not necessarily conserved. This study not only describes a novel functional annotation platform that provides insight into the relationship between functionality and sequence conservation, but also presents valuable information for drug development and future functional studies of the influenza A virus PA protein. More importantly, this approach has the potential to be adapted for any protein where structural information available.

3.3 RESULTS

Design of a high-throughput genetic platform for fitness profiling

High-throughput genetic approaches have been applied to the study of various proteins (reviewed in [40]), which include several from influenza virus and HIV [42–46, 48]. Generally, a mutant library is monitored using deep sequencing, and the relative fitness of each mutation can be inferred by changes in the frequency of mutation occurrence throughout passaging. Mutant library construction represents a key step in these high-throughput genetic approaches. An ideal mutant library should contain only one point mutation per genome, which poses a challenge for high-throughput mutagenic strategies. Existing approaches have used viral genomes that contain multiple mutations within the mutant library. However, the short read length in current deep sequencing technologies disallows the examination of any possible linkage between distantly placed mutations within each genome. Consequently, genetic interactions between mutations may exist during the selection process, but are not accounted for during the fitness calculation for individual point mutations.

To resolve this drawback in existing high-throughput genetic approaches, we have developed a "small library" strategy (Fig. 3-1A). Each mutant library contains a mutated region that can be covered by a single sequencing read. Here, we generated a 240 bp mutated amplicon by error-prone PCR, which is then cloned into a PCR-generated vector using type IIs restriction enzymes (BsaI or BsmBI). The resulting plasmid mutant library was constructed from $\sim$50,000 clones. A total of nine different "small libraries" for influenza A/WSN/33 PA were constructed. Together, these nine "small libraries" covered the entire PA gene. Each viral mutant library was rescued by transfecting the plasmid mutant library with the other seven wild type (WT) plasmids of the influenza A/WSN/33 eight-plasmid reverse genetic system [27], and then passaged for 24 hours in A549 cells.

The plasmid mutant libraries (DNA library), pre-passaged viral mutant libraries (transfection), and post-passaged viral mutant libraries (infection) were subjected to deep sequencing. In this study, we included a technical replicate for sequencing the DNA library, a biological replicate for transfection, and a biological replicate for infection to estimate the reproducibility of individual steps. In

addition, we also sequenced the WT PA plasmid as a control.

The amplicon sequencing library was prepared for the Illumina MiSeq 250 bp paired-end sequencing, using either DNA (DNA library or WT plasmid) or cDNA (transfection or infection) (Fig. 3-1B). For each "small library", the 240 bp mutated region was amplified by a primer pair that contained a BpmI restriction site. A subsequent BpmI digestion excised the primer region from the PCR amplicon. As a result, the entire 240 bp mutated region would be covered by both forward and reverse reads. This enabled sequencing error correction by read-pairing. We obtained a coverage of at least 20,000 (range = 20,128 to 965,488) for each sequencing library.

Point mutation fitness profiling of influenza PA

The design of our high-throughput genetic platform enables us to examine the mutation in individual genomes. On average, 44% (range = 25% to 76%) of viral genomes contain no mutation (i.e. WT), 33% (range = 20% to 36%) of viral genomes contain a single mutation, and 23% (range = 3% to 42%) of viral genomes contain at least two or more mutations. Occurrence frequency for each point mutation was computed from genomes that contained only one mutation. This allowed a precise fitness calculation for individual point mutations without complication by genetic interactions that may exist with additional mutations. Individual point mutations exhibited an occurrence frequency of 0.04% (range = 0% to 0.3%) across all DNA libraries. Whereas the mutation frequency obtained from sequencing the WT plasmid, which served as a control for sequencing error rates, was 0.005% (range = 0% to 0.07%).

Comparison of the relative frequency of individual point mutations between replicates was performed to assess the reproducibility of our "small library" high-throughout genetic platform (see Materials and Methods for the calculation of relative frequency). A Pearson's correlation of 0.95 was obtained for the technical replicate of DNA library, 0.76 for the biological replicate of transfection, and 0.96 for the biological replicate of infection (Fig. 3-2A). The strong correlations between replicates validated the design of our high-throughput genetic platform. The relative fitness index (RF index) was used as a proxy to estimate the fitness effect for each point mutation.

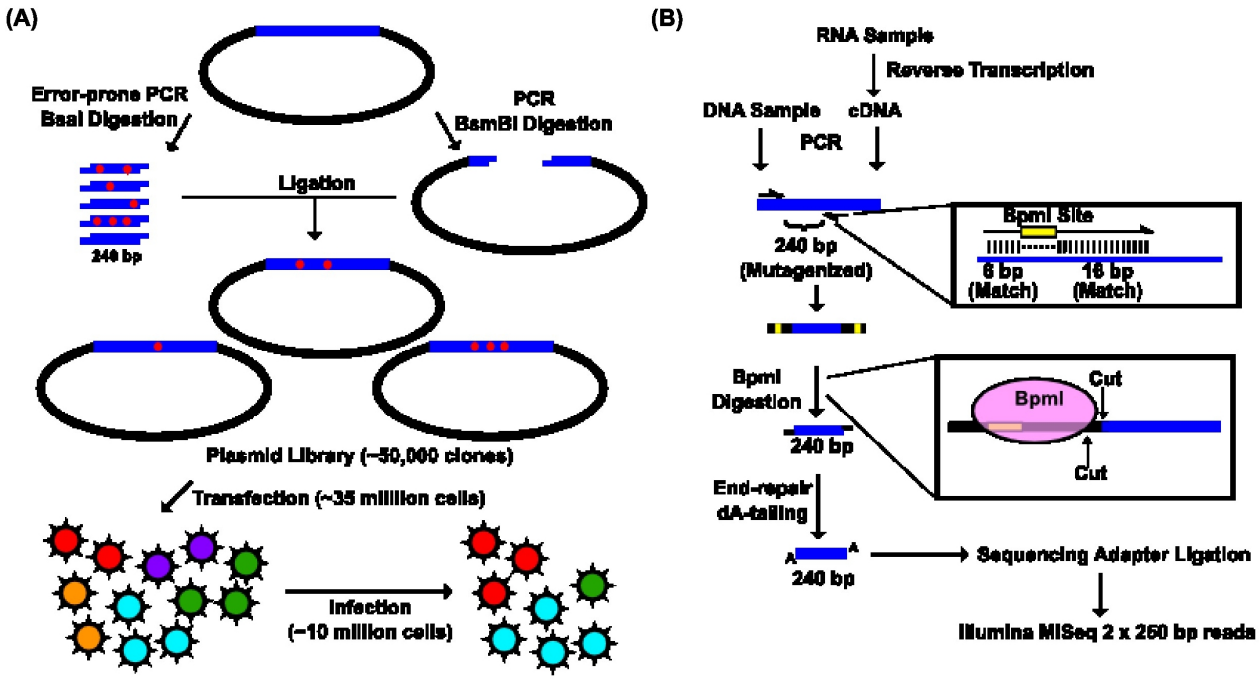


Figure 3-1. Construction of the mutant libraries. (A) A schematic representation of the fitness profiling experiment is shown. A 240 bp insert was generated by error-prone PCR and BsaI digestion. The corresponding vector was generated by high-fidelity PCR and BsmBI digestion. Each of the nine plasmid libraries in this study consist of ~50,000 clones. Each viral mutant library was rescued by transfecting ~35 million 293T cells. Each infection was performed with ~10 million A549 cells. (B) A schematic representation of the sequencing library preparation is shown. DNA plasmid mutant library or viral cDNA was used for PCR so only the 240 bp randomized region was amplified. The amplicon product was then digested with BpmI, end-repaired, dA-tailed, ligated to sequencing adapters, and sequenced using the Illumina MiSeq platform. BpmI digestion removed the primer region in the amplicon PCR, resulting in sequencing reads covering only the barcode for multiplex sequencing and the 240 bp region that was randomized in the mutant library. With this experimental design, the number of mutations carried by individual genomes in the mutant libraries could be precisely determined.

RF index = (Relative frequency$_{infection}$)/(Relative frequency$_{DNA\ library}$)

The RF index of silent mutations (mean = 0.98) was significantly higher than that of nonsense mutations (mean = 0.09) ($P < 2e^{-16}$, two-tailed Student's t-test). Furthermore, the RF index distributions of silent mutations versus nonsense mutations were well-separated (Fig. 3-2B), validating that fitness selection was taking place. The fitness effects of substitutions were profiled across 94% of all amino acid residues in PA. The fitness profiling data is shown in Fig. 3-2C.

Combining high-throughput fitness profiling with mutant protein stability prediction identifies functional sites at single-amino-acid resolution

Next we aimed to identify amino acid residues that were functionally essential, but not structurally important. Essential residues in viral replication can be systematically mapped by high-throughput fitness profiling experiments [43–45, 50–52]. However, fitness profiling only quantifies essentialness, but does not partition the structural versus functional role of individual residues. It has been previously shown that functional residues often carry a suboptimal thermodynamic stability contribution to the proteins in which they reside [53], suggesting the majority of substitutions at functional residues will not affect protein stability. Therefore, functional residues can be identified by substitutions that are deleterious to the virus but not destabilizing to the protein.

Using Rosetta software we predicted the effect of individual substitutions on protein stability; we used the parameters from row 16 of Table I in Kellogg et al., which has been shown to give a correlation of 0.69 with experimental data [32, 54]. In general, substitutions had a lower RF index when the predicted $\Delta\Delta$G increased (Fig. 3-3A). This indicates that viral replication fitness decreases as the PA protein is destabilized. However, we did identify substitutions that had a low RF index, but did not destabilize the protein. We hypothesized that these residues had large functional constraints with little structural effects to the protein upon substitution. To identify the substitutions of interest, a cutoff was set at an RF index < 0.15 (based on the separation point of silent mutations and nonsense mutations) and a predicted $\Delta\Delta G < 0$ (not destabilizing). A total of 32 substitutions

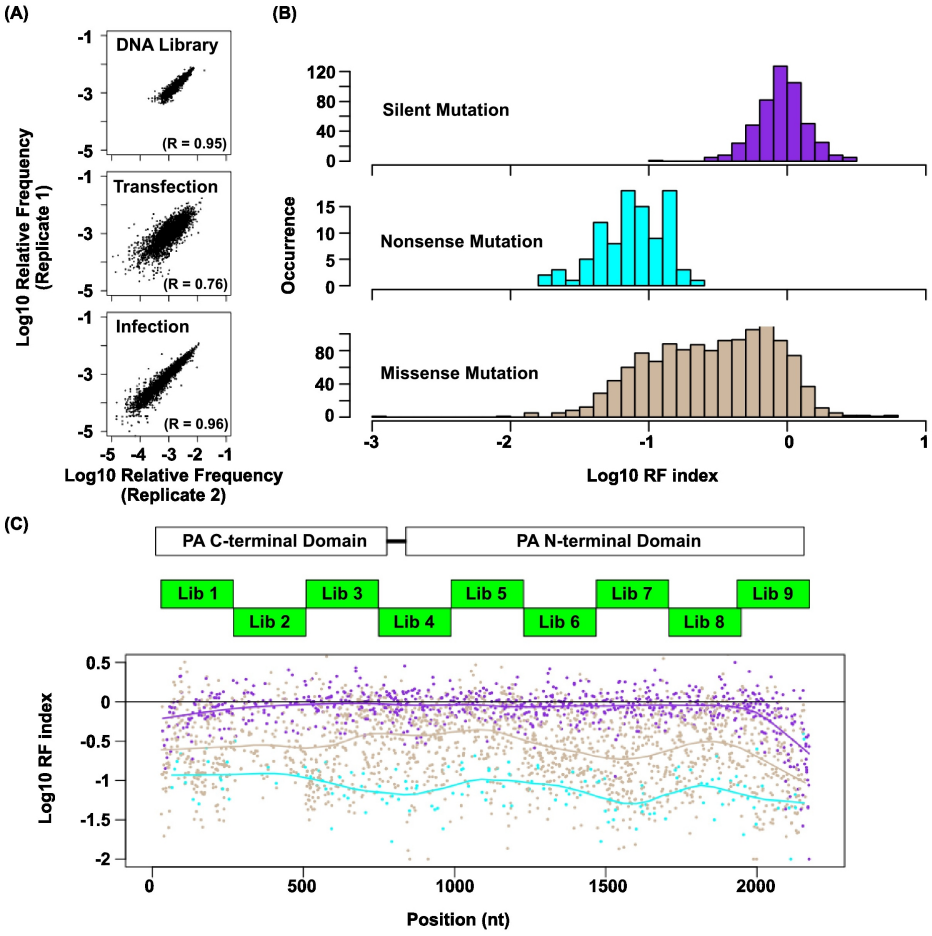


Figure 3-2. Fitness profiling of PA influenza virus polymerase subunit. (A) Correlations of $\log_{10}$ relative frequency of individual point mutations between replicates are shown. Relative frequency$_{mutation\ i}$ = (Occurrence frequency$_{mutation\ i}$)/(Occurrence frequency$_{WT}$) (B) Log_{10} RF indices for silent mutations, nonsense mutations, and missense mutations are shown as histograms. Point mutations located at the 5 terminal 400 bp and 3 terminal 400 bp regions are not included in this analysis to avoid complication by the vRNA packaging signal [48, 49]. (C) The locations of the PA C-terminal domain and the PA N-terminal domain are shown as white boxes. The locations of the mutated regions in each mutant library are shown as green boxes. Log_{10} RF indices for individual point mutations are plotted across the PA gene. Each point mutation is colored coded as in panel B. Purple: silent mutations; Cyan: nonsense mutations; Brown: missense mutations. A smooth curve was fitted by loess and plotted for each point mutation type.

(22 unique residues) in the PA N-terminal domain and 110 substitutions (81 unique residues) in the PA C-terminal domain satisfied this criteria.

A number of functional residues in the PA protein have been characterized in the literature. Out of 32 substitutions of interest in the PA N-terminal domain, eight were at residues that carried known biological functions. This included five substitutions in the endonuclease active site (E80V, E80G, E80K, E119V, K134) [17, 18], and six substitutions in the cRNA promoter binding site (E166D, R170W, R170M, R170K, T173I, T173A) [15, 56]. We also found multiple residues with known biological functions among the 110 substitutions of interest in the C-terminal domain. This included a substitution at a residue required for endonuclease activity (H510R) [25], a substitution at a residue required for small viral RNA (svRNA) binding (R566W) [57], four substitutions at residues required for viral genome replication (E410V, E524V, K539M, K539E) [25], and six substitutions at the PB1-binding site (N412I, N412Y, Q670R, Q670L, F710I, F710Y) [19, 20]. For all residues that carry a deleterious substitution (RF index < 0.15), residues identified as functional residues ($\Delta\Delta G < 0$) had a larger relative SASA (solvent accessible surface area) than amino acid positions that were not (P = $2.6e^{-8}$, two-tailed Student's t-test) (Fig. 3-3B). This indicates that the identified functional residues were mostly surface exposed, as expected if they mediate possible interactions with biomolecules. These results demonstrated the feasibility of combining high-throughput fitness profiling with mutant stability prediction to identify functional sites at single-amino acid resolution.

Identification of residues in PA C-terminal region with functions unrelated to polymerase activity

Because the PA C-terminal region's structure-function relationship remains largely unclear, we aimed to identify functional residues in this region to provide insight into the role of PA during viral replication. Ten substitutions with an RF index < 0.15 and a predicted $\Delta\Delta G < 0$ were individually reconstructed and analyzed. Their spatial locations were distributed throughout the PA C-terminal domain (Fig. 3-4A). The effect of these substitutions on the influenza polymerase activity was tested using an influenza A virus-inducible luciferase reporter assay [58] (Fig. 3-4B). Three substitutions, K281I, K413M, and F681S, completely abolished the influenza polymerase activity. This

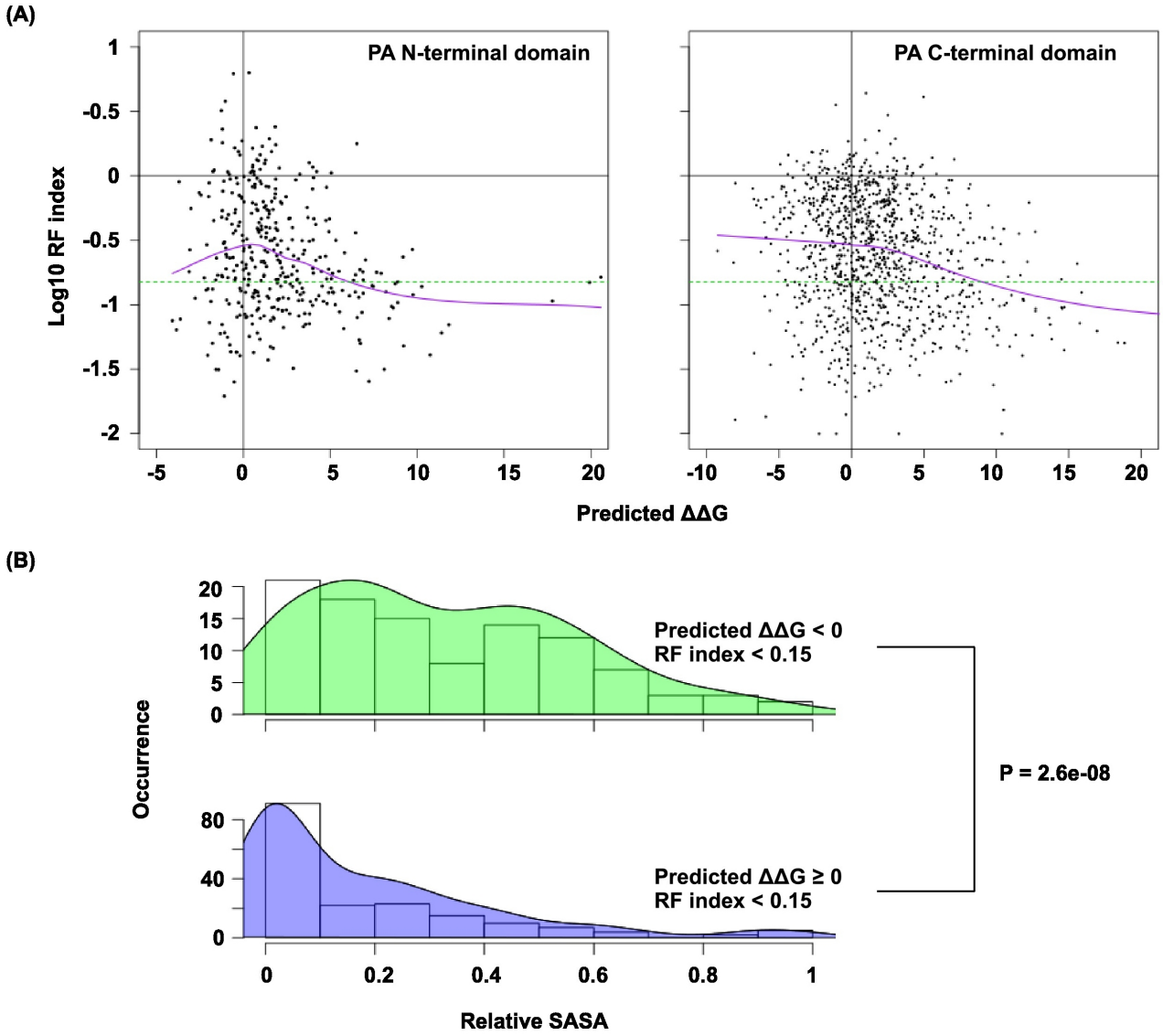


Figure 3-3. Systematic identification of functional residues. (A) Predicted $\Delta\Delta$G for each point mutation is plotted against the $\log_{10}$ RF index. A smooth curve (purple) was fitted by loess and plotted. The horizontal green line represents the RF index cutoff used in this study, RF index = 0.15. (B) The distributions of relative SASA are shown for residues that carried at least one substitutions of interest and for residues that did not carry any substitutions of interest. SASA was computed based on the "get_area" function in PyMOL. SASA obtained from the folded structure was normalized with the SASA calculated from an unfolded Gly-X-Gly tripeptide structure (X = any amino acid) to obtain the relative SASA.

defect is unlikely to be a protein destabilizing effect since all ten mutants analyzed did not alter protein expression levels as compared to WT (Fig. 3-4C).

Interestingly, we found six substitutions (D426G, E427V, G429E, E430G, L470H, and R512W) that retained $>$10% of the WT influenza polymerase activity (Fig. 3-4B). A rescue experiment was performed using the influenza A/WSN/33 eight-plasmid reverse genetic system [27]. L470H, G429E, and R512W, which had $\sim$10%-60% of the WT polymerase activity, completely abolished the production of viral particles (Fig. 3-4D). In addition, E430G, which had a WT-level polymerase activity, displayed a four-log drop in virus titer as compared to WT. In contrast, although D426G and E427V displayed a polymerase activity that was only $\sim$10%-20% of WT, each could produce a much higher amount of infectious virus as compared to other substitutions in this set (one-log to two-log higher titers as compared to E430G). Our results suggest that the L470H, G429E, E430G, and R512W substitutions each had a defect that is unrelated to the polymerase activity.

Structural analysis of the single-residue functional profile

When this study was initiated, PA was the only influenza polymerase subunit with structural information available. The structural information for the other two influenza polymerase subunits, PB1 and PB2, were largely unknown. Nonetheless, after the completion of this study, the crystal structure of the complete influenza A virus polymerase complex bound to the viral RNA promoter has been published [59], which provides an independent reference to validate and interpret our data.

Our functional profile identified the PA residues that interact with PB1, and the viral RNA promoter. Moreover, six out of the 10 validated functional residues participate in these interaction interfaces: – D426, E427, and F681 interacted with PB1; L470 interacted with PB2; K281 and R512 interacted with the viral RNA promoter. Our data also identified functional residues that were not involved in polymerase complex formation or RNA binding activity. For example, E430 did not interact with either PB1, PB2, or the viral RNA promoter. This is consistent with our data that E430 is involved in a non-polymerase function. In addition, a putative functional subdomain

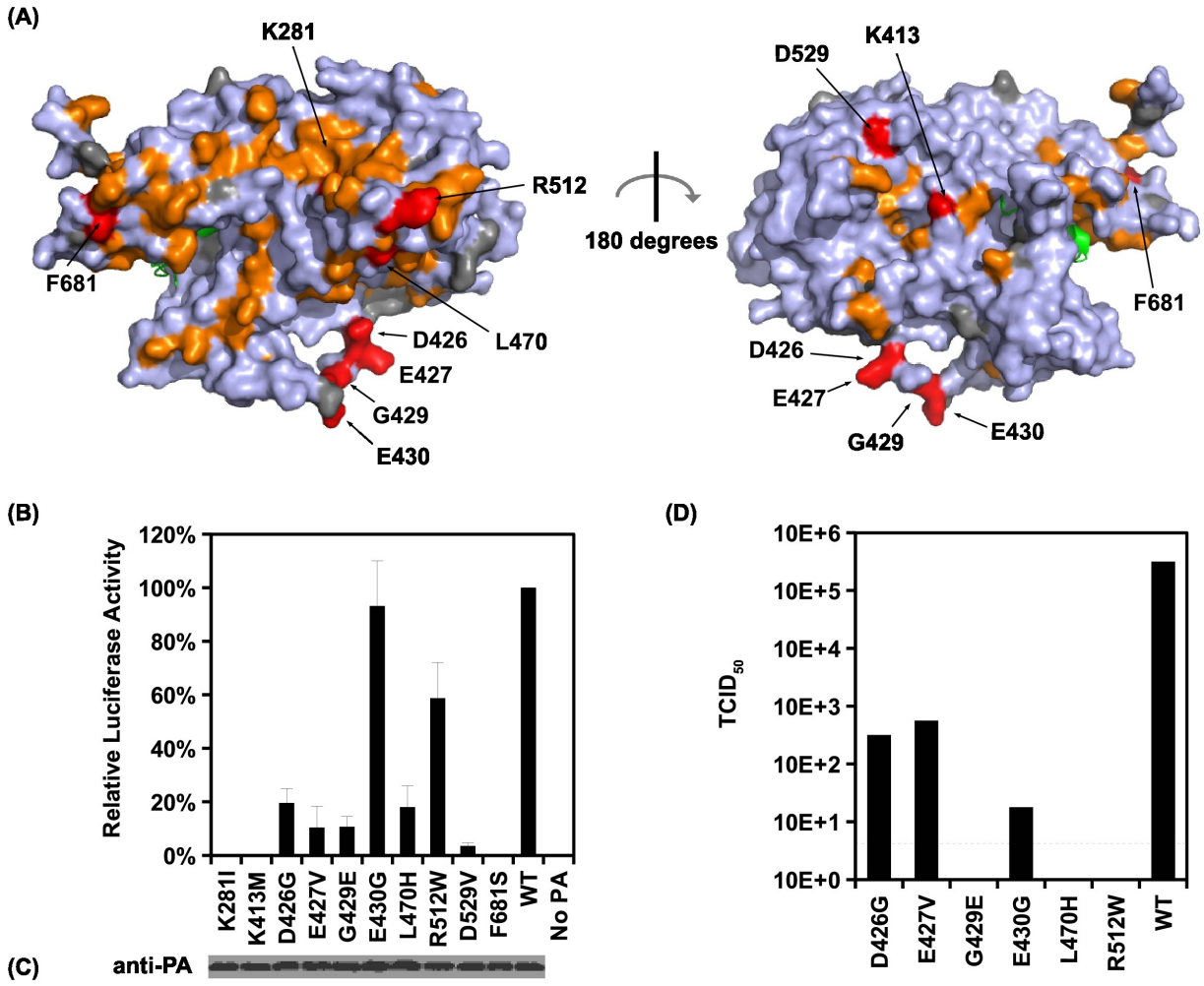


Figure 3-4. Identification of PA residues that carry non-polymerase functions. (A) Locations of substitution with an RF index < 0.15 and a predicted $\Delta\Delta G < 0$ are colored in orange or red respectively. Mutations that were individually reconstructed and analyzed in this study are labeled and colored in red. Residues that were not covered in our profiling data are colored in grey. PB1 is colored in green. PDB ID: 2ZNL [20]. (B) The effects of different PA point mutations on influenza polymerase activity were measured using an influenza A virus-inducible luciferase reporter assay [58]. Error bar represents the standard deviation of three biological replicates. (C) The expression level of each PA mutant was tested by immunoblot analysis. (D) $TCID_{50}$ of the rescued mutant or WT viruses was measured.

independent of the polymerase-interacting surface was identified in our functional profiling data. This putative functional subdomain is composed of a series of charged or polar residues – D286, N412, K413, R454, D529, K559, and K635. Interestingly, this patch of functional residues was adjacent to residue 552, which has been shown to be a host-specific determinant [60]. This indicates a possible biological significance of the putative functional subdomain we identified. Consistently, substitutions at positions D286, N412, K413, R454, D529, and K635 were shown to abolish the polymerase activity in our validation experiment (Fig. 3-4C and 3-5B-C), further confirming the functional importance of this subdomain in viral replication.

Overall, our profiling data is consistent with the polymerase complex-viral RNA promotor complex structural data, which provides an independent validation of our approach.

Relationship between functional constraints, structural constraints, and evolutionary conservation

Phylogenetic analysis indicates that PA displays a high inter-type diversity, while the intra-type diversity is limited. The huge divergence among different types of influenza viruses leads us to hypothesize that a significant number of functional residues are type-specific and are non-conserved across different influenza types. Consequently, we aimed to interrogate the relationship between functional constraints, structural constraints and evolutionary conservation. In this study, sequence conservation for each residue was computed using Shannon's entropy [61]. The higher the entropy, the less conserved a residue is. Here, we divided all profiled residues into three groups: 1) Functional residues, which had at least one substitution that displayed an RF index < 0.15 and a predicted $\Delta\Delta G < 0$. 2) Structural residues, which did not satisfy the condition of functional residues but had at least one substitution that displayed an RF index < 0.15. 3) Neutral residues, which contained all other profiled residues that were neither functional nor structural residues (i.e. all profiled substitutions at a neutral residue displayed an RF index ≥ 0.15).

The entropy calculation was performed for three different influenza type groupings, type A only (Fig. 3-6A), type A and type B (Fig. 3-6B), and for type A, B and C influenza virus (Fig. 3-

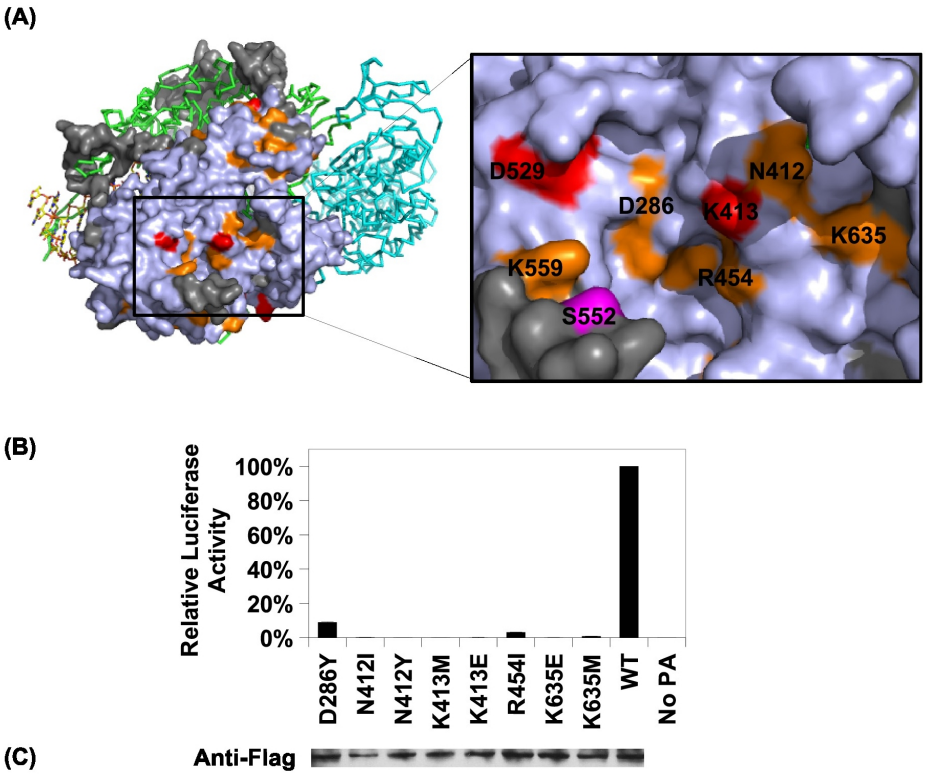


Figure 3-5. Structural analysis of putative functional residues. (A) The location of a putative functional subdomain is shown on the structure of the influenza polymerase heterotrimeric complex (PDB: 4WSB) [59]. For PA, residues were colored as according to the scheme presented in Fig. 3-4. A putative host determinant residue, S552, is colored in magenta. Note, residue 559 carries an arginine [R] instead of a lysine [K] on the PA of A/WSN/33. (B) The effects of different PA point mutations on influenza polymerase activity were measured using an influenza A virus-inducible luciferase reporter assay [58]. Error bar represents the standard deviation of three biological replicates. (C) The expression level of each C-terminal Flag-tagged PA mutant or WT was tested by immunoblot analysis.

6C). In general, functional residues were more conserved than structural residues (P=0.019 to 0.17, Wilcoxon rank-sum test), and structural residues were more conserved than neutral residues (P=7.7e^{-13} to 4.4e^{-5}, Wilcoxon rank-sum test). When the entropy calculation was solely based on type A sequences, 98% of functional residues, 96% of structural residues, and 81% of neutral residues were highly conserved (entropy < 0.1). When the entropy calculation included the sequences from both type A and B influenza virus, 56% of functional residues, 43% of structural residues, and 26% of neutral residues were highly conserved. The fraction of highly conserved residues further decreased when sequences from type C were included $-$ 23% of structural residues, 18% of structural residues, and 9% of neutral residues were highly conserved. These results indicate that a significant number of functional residues are only conserved among type A influenza virus but not across the other types of influenza virus.

We next examined individual residues validated in this study. Among the 13 validated functional residues, six (K281, D286, K413, D426, E430, R454) were only conserved within influenza type A virus but not across type B and C (Fig. 3-6D), and only two (E427, F681) were conserved across different types of influenza viruses. In fact, a similar conservation pattern was observed in the residues of the PB1-binding site. Out of 15 residues that interacted with PB1, six (F411, M595, W619, E623, L639, L640) were conserved only within influenza type A virus, and only one (L666) was conserved across different types of influenza viruses. These results further confirm that functional residues may not necessarily be evolutionarily conserved.

Structural basis of type-specific functional residues

We aimed to further investigate the structural basis of type-specific functional residues. The RNA binding function is required for viral replication and is conserved among type A and B influenza viruses. In the validation above, substituting lysine [K] to isoleucine [I] at residue 281 completely abolished the polymerase activity. This highlights the importance of the hydrogen bond formed between K281 and the RNA phosphate backbone in the influenza A virus (Fig. 3-7A). However, PA K281 is not conserved between type A and B influenza viruses. All influenza B viruses carry an alanine [A] at residue 281, which is unable to form a hydrogen bond with the RNA backbone.

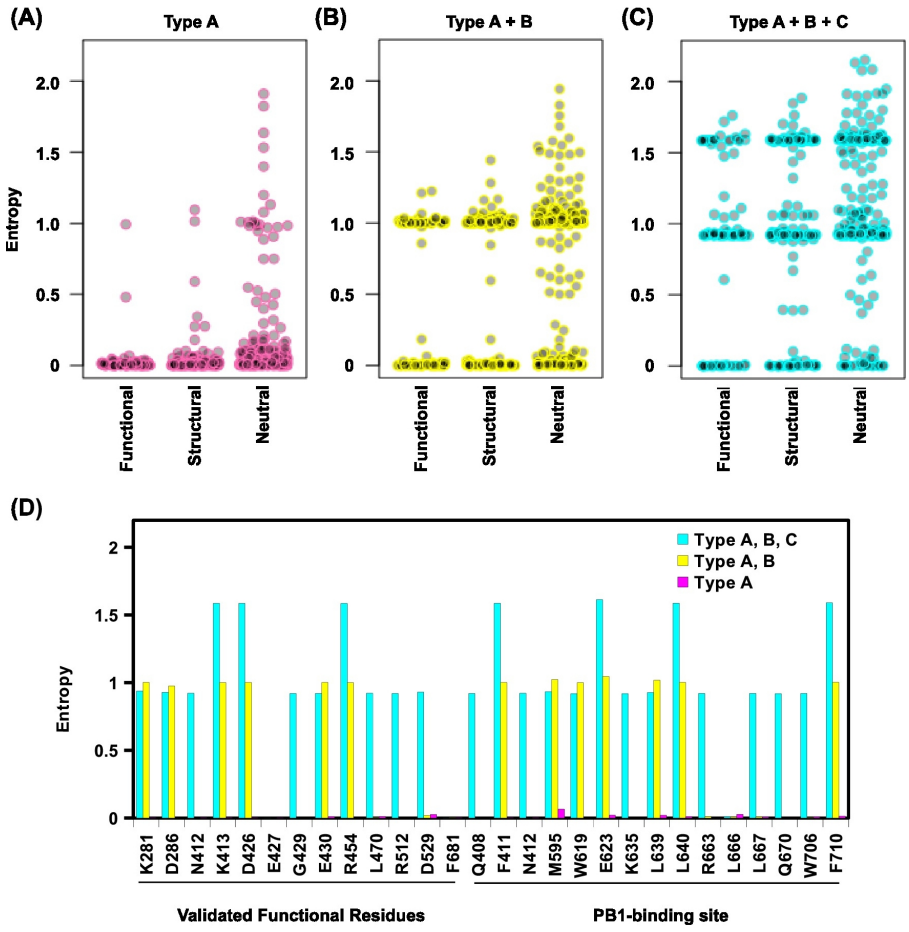


Figure 3-6. Sequence entropy analysis. (A) Distribution of sequence entropy for functional residues. (B) Distribution of sequence entropy for structural residues. (C) Distribution of sequence entropy for neutral residues. (D) Sequence entropy for different residues, including the validated functional residues in this study and residues in the PB1-binding site. Residues in the PB1-binding site are defined as those residues on the PA C-terminal domain that are in contact with PB1 based on PDB: 2ZNL [20].

The critical hydrogen bond mediated by K281 in influenza A virus is replaced by the main chain of G569 in the influenza B virus (Fig. 3-7B). Thus, conserved functions may not necessarily require conserved functional residues.

Together, these analyses show that while certain functional residues were completely conserved among different types of influenza viruses, a significant number of residues that mediate critical viral functions may not be conserved, and suggests that some residues may have acquired functionality in recent evolutionary history.

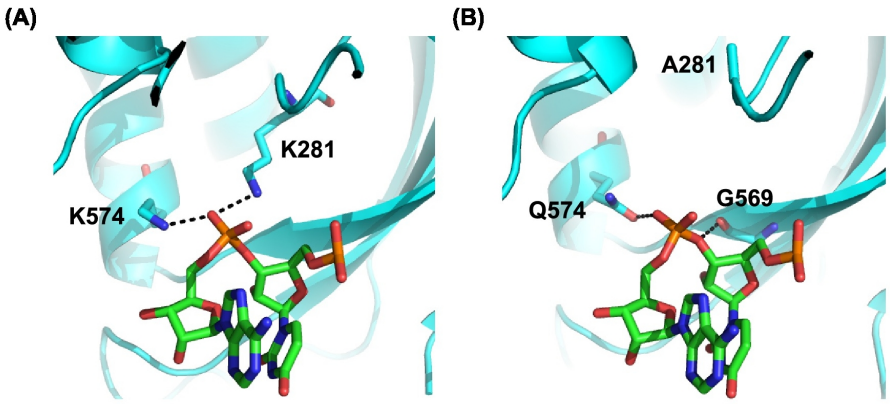


Figure 3-7. Structure-function relationship of residue 281. (A) The interaction of influenza A PA with the RNA phosphate backbone located between base 3 and 4 is shown. RNA is colored in green. PA is colored in cyan. Hydrogen bonds are represented by dotted black lines. Numbering of residue position is based on A/WSN/33. (B) The interaction of influenza B PA with the RNA phosphate backbone located between base 3 and 4 is shown. RNA is colored in green. PA is colored in cyan. Hydrogen bonds are represented by dotted black lines.

3.4 DISCUSSION

Traditionally, sequence conservation is the common approach for identifying functional residues. In this study, we coupled two high-throughput techniques, experimental fitness profiling and *in silico* mutant stability prediction, to systematically identify functional residues in the influenza A virus PA protein. This strategy provided a direct measure of essentialness and enabled the partitioning of functional constraints versus structural constraints at each residue position. This approach is independent of any prior knowledge of sequence conservation. Therefore, it is devoid of the caveats associated with sequence conservation analysis and possesses the power to identify species-specific functional residues. A number of functional residues identified in this study, including eleven that were validated, are not completely conserved across different types of influenza viruses, suggesting that even functional residues may not be conserved. This disparity between conservation and function highlights the power of our approach to identify functional residues that would not be identified by traditional sequence conservation analysis alone.

During natural evolution, continuous accumulation of protein mutations drives speciation and divergence from the common ancestor. The genomic plasticity of an evolving species permits the acquisition of new function through mutations [62]. Evolution of a new function has been demonstrated in bacteriophage λ within an experimental timescale [63], and a long-term evolution experiment on *Escherichia coli* [64]. Therefore, it is not surprising to see species-specific function even in recently separated species. Based on the sequence comparison of hemagglutinin, it was estimated that type A and B influenza virus diverged from type C $\sim$8,000 years ago, whereas type A influenza virus diverged from type B $\sim$4,000 years ago [65]. This length of time is sufficient for the influenza virus to develop a type-specific function as exemplified by type-specific virus-host interactions in NS1 [8, 9]. Furthermore, conservation of protein function does not necessarily support that sequence conservation exists at the primary sequence level, which is evidenced by the differences between the nuclear localization signal of influenza A and B NP proteins [66, 67]. In fact, this study reveals that type-specific functional residues are prevalent in the influenza virus PA protein. These results not only provide insight into how functional residues evolve through species diversification, but also highlights the caveats encountered when identifying functional sites from

conservation-based approaches.

In the past decade, proteins from different medically important viruses, such as influenza, HIV, and HCV, have been crystallized [68–70]. The approach described in this study systematically integrates the available structural information with mutation fitness information to examine the structure-function relationship of a viral protein of interest and to map functional subdomains. Profiling datasets will facilitate functional characterization of the protein of interest, and will promote drug target discovery and rational drug design. The emergence of drug resistant mutations is a major challenge for antiviral drug development. Therefore, it is important to target functional subdomains that are less tolerable to substitution in order to increase the genetic barrier for developing drug resistant mutations. Our profiling technique can help locate such functional subdomains that are suitable for drug development. More importantly, our technique can potentially be adapted to study any protein, provided the relevant structural information is available.

Construction of mutant libraries and individual point mutations

The PA plasmid mutant libraries were created by performing error-prone PCR on the PA segment of the eight-plasmid reverse genetics system of influenza A/WSN/1933 (H1N1) [27]. To generate the mutated insert, We PCR-amplified regions of the PA gene with error-prone polymerase Mutazyme II (Stratagene, La Jolla, CA) using the following primers:

Library 1 insert: 5'-CAG GTC TCA TCA AAA TGG AAG ATT TTG TGC GA-3' and 5'-CAG GTC TCA ATA CTG TTT ATT ACT GTC CAG GC-3'
Library 2 insert: 5'-CAG GTC TCA TCG AGG GAA GAG ATC GCA CAA TA-3' and 5'-CAG GTC TCA CTG GTT TTG ATC CTA GCC CTG CT-3'
Library 3 insert: 5'-CAG GTC TCA CCG ACT ACA CTC TCG ATG AAG AA-3' and 5'-CAG GTC TCA TTT ACT TCT TTG GAC ATT TGA GA-3'
Library 4 insert: 5'-CAG GTC TCA ACG GCT ACA TTG AGG GCA AGC TT-3' and 5'-CAG GTC TCA TAA TTT GGA TTT ATT CCC TTT TC-3'
Library 5 insert: 5'-CAG GTC TCA AAC CCA ATG TTG TTA AAC CAC AC-3' and 5'-CAG GTC TCA GCC TTG TTG AAC TCA TTC TGA AT-3'
Library 6 insert: 5'-CAG GTC TCA AAT TGA GGT CGC TTG CAA GTT GG-3' and 5'-CAG GTC TCA CCC TCC TTA GTT CTA CAC TTG CT-3'
Library 7 insert: 5'-CAG GTC TCA ATT TCC AAT TAA TTC CAA TGA TA-3' and 5'-CAG GTC TCA TTA ATT TTT GAG GTT CCA TTT GT-3'
Library 8 insert: 5'-CAG GTC TCA GGC CTA TGT TCT TGT ATG TGA GG-3' and 5'-CAG GTC TCA TGT GGA GAT GCA TAC AAG CTG TT-3'
Library 9 insert: 5'-CAG GTC TCA GAA GGT CTG CAG AAC TTT ATT GG-3' and 5'-CAG GTC TCA GGA CAG TAT GGA TAG CAA ATA GT-3'

The corresponding vector for each of the nine mutant library was generated by PCR with KOD DNA polymerase (EMD Millipore, Billerica, MA) using the following primers:

Library 1 vector: 5'-CAC GTC TCT TTG AAT CAG TAC CTG CTT TCG CT-3' and 5'-CAC GTC

TCA GTA TTT GCA ACA CTA CAG GGG CT-3'

Library 2 vector: 5'-CAC GTC TCC TCG ATT ATT TCA AAT CTG TGC TT-3' and 5'-CAC GTC TCA CCA GGC TAT TCA CCA TAA GAC AA-3'

Library 3 vector: 5'-CAC GTC TCG TCG GCC TTT GTG GCC ATT TCC TC-3' and 5'-CAC GTC TCG TAA ATG CTA GAA TTG AAC CTT TT-3'

Library 4 vector: 5'-CAC GTC TCG CCG TTC GGT TCG AAT CCA TCC AC-3' and 5'-CAC GTC TCA ATT ATC TTC TGT CAT GGA AGC AA-3'

Library 5 vector: 5'-CAC GTC TCG GGT TCC TTC CAT CCA AAG AAT GT-3' and 5'-CAC GTC TCA AGG CAT GTG AAC TGA CCG ATT CA-3'

Library 6 vector: 5'-CAC GTC TCC AAT TCT GGT TCA TCA CTA TCA TA-3' and 5'-CAC GTC TCG AGG GAA GGC GAA AGA CCA ATT TG-3'

Library 7 vector: 5'-CAC GTC TCG AAA TCA TCC ATT GCT GCA CAG GA-3' and 5'-CAC GTC TCA TTA AAA TGA AAT GGG GGA TGG AA-3'

Library 8 vector: 5'-CAC GTC TCA GGC CTT GAC ACA TGG CCT ATG GC-3' and 5'-CAC GTC TCC CAC AAC TAG AAG GAT TTT CAG CT-3'

Library 9 vector: 5'-CAC GTC TCC CTT CCC AAT GGA ACC TTC CTC CA-3' and 5'-CAC GTC TCT GTC CAA AAA GTA CCT TGT TTC TA-3'

The insert was then digested by BsaI (New England Biolabs, Ipswich, MA), whereas the vector was digested by BsmBI (New England Biolabs). Ligation was performed for each of the nine libraries with T4 DNA ligase (Life Technologies, Carlsbad, CA) using the corresponding insert and vector. Transformations were carried out with electrocompetent MegaX DH10B T1R cells (Life Technologies). For each of the nine mutant libraries, $\sim$50,000 colonies were scraped and directly processed for plasmid DNA purification (Qiagen Sciences, Germantown, MD). Point mutations for the validation experiment were constructed using the QuikChange XL Mutagenesis kit (Stratagene) according to the manufacturer's instructions.

Transfections, infections and titering

$\sim$35 million 293T (human embryonic kidney) cells were used for transfection to rescue each viral mutant library from the plasmid mutant library as described [42, 43, 48]. Transfections were performed using Lipofectamine 2000 (Life Technologies) according to the manufacturer's instructions. Supernatant was replaced with fresh cell growth medium at 24 hours and 48 hours post-transfection. At 72 hours post-transfection, supernatant containing infectious virus was harvested, filtered through a 0.45 um MCE filter, and stored at -80^{o}C. The $TCID_{50}$ was measured on A549 cells (human lung carcinoma cells). To passage each viral mutant library, $\sim$10 million A549 cells were used for infection at an MOI of 0.05. At 2 hours post-infection, infected cells were washed three times with PBS followed by the addition of fresh cell growth medium. Virus was harvested at 24 hrs post-infection.

Sequencing library preparation

Viral RNA was extracted using QIAamp Viral RNA Mini Kit (Qiagen Sciences) and treated with DNaseI (Life Technologies) to digest any residual plasmid DNA from transfection. The DNA-free RNA was then 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 the following primers:

Library 1: 5'-CTG ATT CTG GAG GGA AGA TTT TGT GCG A-3' and 5'-TGC AAA CTG GAG TTA TTA CTG TCC AGG C-3'
Library 2: 5'-AAT AAT CTG GAG AAG AGA TCG CAC AAT A-3' and 5'-ATA GCC CTG GAG TGA TCC TAG CCC TGC T-3'
Library 3: 5'-AAA GGC CTG GAG CAC TCT CGA TGA AGA A-3' and 5'-TAG CAT CTG GAG CTT TGG ACA TTT GAG A-3'
Library 4: 5'-ACC GAA CTG GAG CAT TGA GGG CAA GCT T-3' and 5'-GAA GAT CTG GAG GAT TTA TTC CCT TTT C-3'
Library 5: 5'-GAA GGA CTG GAG TGT TGT TAA ACC ACA C-3' and 5'-CAC ATG CTG GAG TGA ACT CAT TCT GAA T-3'

Library 6: 5'-ACC AGA CTG GAG GTC GCT TGC AAG TTG G-3' and 5'-GCC TTC CTG GAG TAG TTC TAC ACT TGC T-3'
Library 7: 5'-GGA TGA CTG GAG ATT AAT TCC AAT GAT A-3' and 5'-TCA TTT CTG GAG TTG AGG TTC CAT TTG T-3'
Library 8: 5'-GTC AAG CTG GAG GTT CTT GTA TGT GAG G-3' and 5'-CTA GTT CTG GAG ATG CAT ACA AGC TGT T-3'
Library 9: 5'-ATT GGG CTG GAG TGC AGA ACT TTA TTG G-3' and 5'-TTT TTG CTG GAG ATG GAT AGC AAA TAG T-3'

The resulting PCR amplicons were digested with BpmI (New England Biolabs). End repair and 3' dA-tailing were performed by end repair module and dA-tailing module respectively (New England BioLabs). dA-tailed amplicons were ligated to sequencing adapters using T4 DNA ligase (Life Technologies). Adapters were generated by annealing two oligos: 5'-ACA CT CTT TCC CTA CAC GAC GCT CTT CCG ATC TNN NT-3' and 5'-/5Phos/NNN AGA TCG GAA GAG CGG TTC AGC AGG AAT GCC GAG-3'. The location of multiplex ID for distinguishing different samples is underlined. The adapter-ligated products were enriched by a final PCR using primers: 5'-AAT GAT ACG GCG ACC ACC GAG ATC TAC ACT CTT TCC CTA CAC GAC-3' and 5'-CAA GCA GAA GAC GGC ATA CGA GAT CGG TCT CGG CAT TCC TGC TGA ACC-3'. Deep sequencing was performed using two lanes of the Illumina MiSeq with 250 bp paired-end reads. Raw sequencing data have been submitted to the NIH Short Read Archive (SRA) under accession number: BioProject PRJNA254185.

Sequencing data analysis

Sequencing data were de-multiplexed by the three-nucleotide barcode. A paired-end read was filtered and removed if the corresponding forward and reverse reads did not match. Each mutation was called by comparing individual reads to the WT reference sequence. All analysis was performed by custom python scripts, which are available upon request. For the RF index calculation, only mutants that carried a single mutation were considered. RF index for a given mutation was computed as follow:

RF index = (Relative frequency in passaged library)/(Relative frequency in plasmid library), where Relative frequency$_{mutation\ i}$ = (Occurrence frequency$_{mutation\ i}$)/(Occurrence frequency$_{WT}$)

To avoid fitness calculations being obscured by sequencing errors, only the point mutations with an occurrence frequency of $\geq$0.03% in the DNA library were included in the downstream analysis unless otherwise stated.

$\Delta\Delta$G predictions for single amino acid substitutions

PDB: 4M5Q (PA N-terminal endonuclease domain) [34] and PDB: 2ZNL (PA C-terminal domain) [20] were used for $\Delta\Delta$G prediction of single amino-acid substitution. $\Delta\Delta$G prediction was performed by the ddg_monomer application in Rosetta software [54]. Parameters from row 16 of Table I in Kellogg *et al.* were used [32]. Briefly, a "soft-rep" energy function was used for sidechain repacking for all residues, in which the Lennard-Jones repulsive interactions at short atomic separations were damped. After repacking, a restrained quasi-Newton minimization step was performed for both sidechain and backbone using a "hard-rep" energy function, in which the repulsive interactions were not damped. All options followed the high resolution protocol flags of the ddg_monomer application. The $\Delta\Delta$G prediction result along with the RF index for individual substitutions are shown in Supplemental Dataset 1. Minimal, if any, destabilizing effect is expected if predicted $\Delta\Delta$G is < 0.

Luciferase reporter assay for influenza polymerase activity

An influenza A virus-inducible luciferase reporter assay was used to measure the virus polymerase activity [58]. 293T cells seeded on 48-well plates were transfected with 100 ng each of PB2, PB1, PA, NP, 50 ng of vLuciferase reporter plasmid and 5 ng of PGK-renilla-luciferase using Lipofectamine 2000 (Life Technologies) according to the manufacturer's instructions. Luciferase activity measurement was performed at 24 hours post-transfection using Promega Dual-Luciferase Assay Kit according to the manufacturers instructions (Promega, Madison, WI). Relative luciferase activity was calculated by normalizing the firefly-luciferase activities to their internal renilla luciferase

controls.

Protein expression analysis

293T cells seeded on a 12-well plate were transfected with pHW2000-PA plasmid using Lipofectamine 2000 (Life Technologies) according to the manufacturer's instructions. At 24 hours post-transfection, cells were lysed and heated with SDS loading buffer for five minutes. Lysates were loaded onto a 10% polyacrylamide gel and subjected to immunoblot analysis. Rabbit anti-PA antibody (catalog number: GTX125932, GeneTex, Irvine, CA), mouse anti-Flag antibody (Sigma), sheep horseradish peroxidase-conjugated anti-mouse Immunoglobulin G (GE Healthcare, Pasadena, CA), and donkey horseradish peroxidase-conjugated anti-rabbit Immunoglobulin G (GE Healthcare) were used for protein detection.

Real-time reverse-transcription PCR (RT-qPCR)

Viral RNA was extracted using QIAamp Viral RNA Mini Kit (Qiagen Sciences) and treated with DNaseI (Life Technologies) to digest any residual plasmid DNA from transfection. The DNA-free RNA was then reverse transcribed to cDNA using Superscript III reverse transcriptase (Life Technologies). The cDNA was subjected to qPCR analysis. qPCR was performed on a DNA Engine OPTICON 2 system (Bio-Rad, Irvine, CA) using SYBR Green (Life Technologies) with primers: 5'-GAC GAT GCA ACG GCT GGT CTG-3' and 5'-ACC ATT GTT CCA AC TCC TTT-3'.

Sequence entropy and phylogenetic analysis

PA protein sequences of type A and B influenza virus and P3 protein sequences of type C influenza virus were retrieved from the Influenza Research Database [71]. A total of 3271 PA protein sequences from type A influenza virus, 562 PA protein sequences from type B influenza virus, and 4 P3 protein sequences from type C influenza virus were obtained using the following parameters: human host, all geographical locations, complete segment only, include pH1N1, remove duplicate sequences.

Due the large number of sequences available, sequence entropy was computed using a boot-

strapping approach. Briefly, 100 sequences were sampled with replacement from each of the indicated types of influenza virus. Multiple sequence alignment was performed on the sampled sequences along with the A/WSN/33 PA sequence using MUSCLE (version 3.8.31) [72]. Shannon's entropy for each residue position was calculated by:

Entropy = $-\sum_{i=1}^{M} P_i \log_2(P_i)$ [61], where P_i is the fraction of residues of amino acid type i, and M is the number of amino acid types (i.e. 20).

This whole procedure (starting from the sampling process) was performed 100 times. For each residue position, the entropy was computed as the average value in this 100-time bootstrap.

3.6 BIBLIOGRAPHY

1. Bairoch A, Apweiler R (1996) The swiss-prot protein sequence data bank and its new supplement trembl. Nucleic Acids Res 24: 21–25.

2. Pruitt KD, Tatusova T, Maglott DR (2005) Ncbi reference sequence (refseq): a curated nonredundant sequence database of genomes, transcripts and proteins. Nucleic Acids Res 33: D501–D504.

3. Kanz C, Aldebert P, Althorpe N, Baker W, Baldwin A, et al. (2005) The embl nucleotide sequence database. Nucleic Acids Res 33: D29–D33.

4. Li Z, Watanabe T, Hatta M, Watanabe S, Nanbo A, et al. (2009) Mutational analysis of conserved amino acids in the influenza a virus nucleoprotein. J Virol 83: 4153–4162.

5. Stewart SM, Pekosz A (2011) Mutations in the membrane-proximal region of the influenza a virus m2 protein cytoplasmic tail have modest effects on virus replication. J Virol 85: 12179–12187.

6. Chu C, Fan S, Li C, Macken C, Kim JH, et al. (2012) Functional analysis of conserved motifs in influenza virus pb1 protein. PLoS One 7: e36113.

7. Worth CL, Gong S, Blundell TL (2009) Structural and functional constraints in the evolution of protein families. Nat Rev Mol Cell Biol 10: 709–720.

8. Yuan W, Krug RM (2001) Influenza b virus ns1 protein inhibits conjugation of the interferon (ifn)-induced ubiquitin-like isg15 protein. EMBO J 20: 362–371.

9. Twu KY, Noah DL, Rao P, Kuo RL, Krug RM (2006) The cpsf30 binding site on the ns1a protein of influenza a virus is a potential antiviral target. J Virol 80: 3957–3965.

10. Hutchinson EC, Denham EM, Thomas B, Trudgian DC, Hester SS, et al. (2012) Mapping the phosphoproteome of influenza a and b viruses by mass spectrometry. PLoS Pathog 8: e1002993.

11. Zhang DW, Cole SP, Deeley RG (2001) Identification of a nonconserved amino acid residue in multidrug resistance protein 1 important for determining substrate specificity: evidence

for functional interaction between transmembrane helices 14 and 17. J Biol Chem 276: 34966–34974.

12. Tungtur S, Meinhardt S, Swint-Kruse L (2010) Comparing the functional roles of nonconserved sequence positions in homologous transcription repressors: implications for sequence/function analyses. J Mol Biol 395: 785–802.

13. Genovese NJ, Broker TR, Chow LT (2011) Nonconserved lysine residues attenuate the biological function of the low-risk human papillomavirus e7 protein. J Virol 85: 5546–5554.

14. Freeley M, Kelleher D, Long A (2011) Regulation of protein kinase c function by phosphorylation on conserved and non-conserved sites. Cell Signal 23: 753–762.

15. Hara K, Schmidt FI, Crow M, Brownlee GG (2006) Amino acid residues in the n-terminal region of the pa subunit of influenza a virus rna polymerase play a critical role in protein stability, endonuclease activity, cap binding, and virion rna promoter binding. J Virol 80: 7789–7798.

16. Guu TSY, Dong L, Wittung-Stafshede P, Tao YJ (2008) Mapping the domain structure of the influenza a virus polymerase acidic protein (pa) and its interaction with the basic protein 1 (pb1) subunit. Virology 379: 135–142.

17. Yuan P, Bartlam M, Lou Z, Chen S, Zhou J, et al. (2009) Crystal structure of an avian influenza polymerase pa(n) reveals an endonuclease active site. Nature 458: 909–913.

18. Dias A, Bouvier D, Crpin T, McCarthy AA, Hart DJ, et al. (2009) The cap-snatching endonuclease of influenza virus polymerase resides in the pa subunit. Nature 458: 914–918.

19. He X, Zhou J, Bartlam M, Zhang R, Ma J, et al. (2008) Crystal structure of the polymerase pa(c)-pb1(n) complex from an avian influenza h5n1 virus. Nature 454: 1123–1126.

20. Obayashi E, Yoshida H, Kawai F, Shibayama N, Kawaguchi A, et al. (2008) The structural basis for an essential subunit interaction in influenza virus rna polymerase. Nature 454: 1127–1131.

21. Biswas SK, Nayak DP (1994) Mutational analysis of the conserved motifs of influenza a virus polymerase basic protein 1. J Virol 68: 1819–1826.

22. Li ML, Rao P, Krug RM (2001) The active sites of the influenza cap-dependent endonuclease are on different polymerase subunits. EMBO J 20: 2078–2086.

23. Fechter P, Mingay L, Sharps J, Chambers A, Fodor E, et al. (2003) Two aromatic residues in the pb2 subunit of influenza a rna polymerase are crucial for cap binding. J Biol Chem 278: 20381–20388.

24. Guilligay D, Tarendeau F, Resa-Infante P, Coloma R, Crepin T, et al. (2008) The structural basis for cap binding by influenza virus polymerase subunit pb2. Nat Struct Mol Biol 15: 500–506.

25. Fodor E, Crow M, Mingay LJ, Deng T, Sharps J, et al. (2002) A single amino acid mutation in the pa subunit of the influenza virus rna polymerase inhibits endonucleolytic cleavage of capped rnas. J Virol 76: 8989–9001.

26. Fodor E, Mingay LJ, Crow M, Deng T, Brownlee GG (2003) A single amino acid mutation in the pa subunit of the influenza virus rna polymerase promotes the generation of defective interfering rnas. J Virol 77: 5017–5020.

27. Huarte M, Falcn A, Nakaya Y, Ortn J, Garca-Sastre A, et al. (2003) Threonine 157 of influenza virus pa polymerase subunit modulates rna replication in infectious viruses. J Virol 77: 6007–6013.

28. Kawaguchi A, Naito T, Nagata K (2005) Involvement of influenza virus pa subunit in assembly of functional rna polymerase complexes. J Virol 79: 732–744.

29. Regan JF, Liang Y, Parslow TG (2006) Defective assembly of influenza a virus due to a mutation in the polymerase subunit pa. J Virol 80: 252–261.

30. Liang Y, Danzy S, Dao LD, Parslow TG, Liang Y (2012) Mutational analyses of the influenza a virus polymerase subunit pa reveal distinct functions related and unrelated to rna polymerase activity. PLoS One 7: e29485.

31. Hara K, Shiota M, Kido H, Ohtsu Y, Kashiwagi T, et al. (2001) Influenza virus rna polymerase pa subunit is a novel serine protease with ser624 at the active site. Genes Cells 6: 87–97.

32. Rodriguez A, Prez-Gonzlez A, Nieto A (2007) Influenza virus infection causes specific degradation of the largest subunit of cellular rna polymerase ii. J Virol 81: 5315–5324.

33. Liu Y, Lou Z, Bartlam M, Rao Z (2009) Structure-function studies of the influenza virus rna polymerase pa subunit. Sci China C Life Sci 52: 450–458.

34. Bauman JD, Patel D, Baker SF, Vijayan RSK, Xiang A, et al. (2013) Crystallographic fragment screening and structure-based optimization yields a new class of influenza endonuclease inhibitors. ACS Chem Biol 8: 2501–2508.

35. Li L, Chang S, Xiang J, Li Q, Liang H, et al. (2012) Screen anti-influenza lead compounds that target the pa(c) subunit of h5n1 viral rna polymerase. PLoS One 7: e35234.

36. Muratore G, Goracci L, Mercorelli B, gnes Foeglein, Digard P, et al. (2012) Small molecule inhibitors of influenza a and b viruses that act by disrupting subunit interactions of the viral polymerase. Proc Natl Acad Sci U S A 109: 6247–6252.

37. Tintori C, Laurenzana I, Fallacara AL, Kessler U, Pilger B, et al. (2014) High-throughput docking for the identification of new influenza a virus polymerase inhibitors targeting the pa-pb1 protein-protein interaction. Bioorg Med Chem Lett 24: 280–282.

38. DuBois RM, Slavish PJ, Baughman BM, Yun MK, Bao J, et al. (2012) Structural and biochemical basis for development of influenza virus inhibitors targeting the pa endonuclease. PLoS Pathog 8: e1002830.

39. Kowalinski E, Zubieta C, Wolkerstorfer A, Szolar OHJ, Ruigrok RWH, et al. (2012) Structural analysis of specific metal chelating inhibitor binding to the endonuclease domain of influenza ph1n1 (2009) polymerase. PLoS Pathog 8: e1002831.

40. Fowler DM, Fields S (2014) Deep mutational scanning: a new style of protein science. Nat Methods 11: 801–807.

41. Wu NC, Young AP, Dandekar S, Wijersuriya H, Al-Mawsawi LQ, et al. (2013) Systematic identification of h274y compensatory mutations in influenza a virus neuraminidase by high-throughput screening. J Virol 87: 1193–1199.

42. Wu NC, Young AP, Al-Mawsawi LQ, Olson CA, Feng J, et al. (2014) High-throughput identification of loss-of-function mutations for anti-interferon activity in the influenza a virus ns segment. J Virol 88: 10157–10164.

43. Wu NC, Young AP, Al-Mawsawi LQ, Olson CA, Feng J, et al. (2014) High-throughput profiling of influenza a virus hemagglutinin gene at single-nucleotide resolution. Sci Rep 4: 4942.

44. Bloom JD (2014) An experimentally determined evolutionary model dramatically improves phylogenetic fit. Mol Biol Evol 31: 1956–1978.

45. Thyagarajan B, Bloom JD (2014) The inherent mutational tolerance and antigenic evolvability of influenza hemagglutinin. Elife 3.

46. Al-Mawsawi LQ, Wu NC, Olson C, Shi V, Qi H, et al. (2014) High-throughput profiling of point mutations across the hiv-1 genome. Retrovirology 11: 124.

47. Neumann G, Watanabe T, Ito H, Watanabe S, Goto H, et al. (1999) Generation of influenza a viruses entirely from cloned cdnas. Proc Natl Acad Sci U S A 96: 9345–9350.

48. Liang Y, Hong Y, Parslow TG (2005) cis-acting packaging signals in the influenza virus pb1, pb2, and pa genomic rna segments. J Virol 79: 10348–10355.

49. Liang Y, Huang T, Ly H, Parslow TG, Liang Y (2008) Mutational analyses of packaging signals in influenza virus pa, pb1, and pb2 genomic rna segments. J Virol 82: 229–236.

50. Qi H, Olson CA, Wu NC, Ke R, Loverdo C, et al. (2014) A quantitative high-resolution genetic profile rapidly identifies sequence determinants of hepatitis c viral fitness and drug sensitivity. PLoS Pathog 10: e1004064.

51. Robins WP, Faruque SM, Mekalanos JJ (2013) Coupling mutagenesis and parallel deep sequencing to probe essential residues in a genome or gene. Proc Natl Acad Sci U S A 110: E848–E857.

52. Acevedo A, Brodsky L, Andino R (2014) Mutational and fitness landscapes of an rna virus revealed through population sequencing. Nature 505: 686–690.

53. Cheng G, Qian B, Samudrala R, Baker D (2005) Improvement in protein functional site prediction by distinguishing structural and functional constraints on protein family evolution using computational design. Nucleic Acids Res 33: 5861–5867.

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

55. Kellogg EH, Leaver-Fay A, Baker D (2011) Role of conformational sampling in computing mutation-induced changes in protein structure and stability. Proteins 79: 830–838.

56. Maier HJ, Kashiwagi T, Hara K, Brownlee GG (2008) Differential role of the influenza a virus polymerase pa subunit for vrna and crna promoter binding. Virology 370: 194–204.

57. Perez JT, Zlatev I, Aggarwal S, Subramanian S, Sachidanandam R, et al. (2012) A small-rna enhancer of viral polymerase activity. J Virol 86: 13475–13485.

58. 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.

59. Pflug A, Guilligay D, Reich S, Cusack S (2014) Structure of influenza a polymerase bound to the viral rna promoter. Nature .

60. Mehle A, Dugan VG, Taubenberger JK, Doudna JA (2012) Reassortment and mutation of the avian influenza virus polymerase pa subunit overcome species barriers. J Virol 86: 1750–1757.

61. Shannon CE (1948) The mathematical theory of communication. The Bell system Technical Journal 27: 379–423.

62. Soskine M, Tawfik DS (2010) Mutational effects and the evolution of new protein functions. Nat Rev Genet 11: 572–582.

63. Meyer JR, Dobias DT, Weitz JS, Barrick JE, Quick RT, et al. (2012) Repeatability and contingency in the evolution of a key innovation in phage lambda. Science 335: 428–432.

64. Blount ZD, Borland CZ, Lenski RE (2008) Historical contingency and the evolution of a key innovation in an experimental population of escherichia coli. Proc Natl Acad Sci U S A 105: 7899–7906.

65. Suzuki Y, Nei M (2002) Origin and evolution of influenza virus hemagglutinin genes. Mol Biol Evol 19: 501–509.

66. Whittaker G, Bui M, Helenius A (1996) Nuclear trafficking of influenza virus ribonuleoproteins in heterokaryons. J Virol 70: 2743–2756.

67. Sherry L, Smith M, Davidson S, Jackson D (2014) The n terminus of the influenza b virus nucleoprotein is essential for virus viability, nuclear localization, and optimal transcription and replication of the viral genome. J Virol 88: 12326–12338.

68. Das K, Aramini JM, Ma LC, Krug RM, Arnold E (2010) Structures of influenza a proteins and insights into antiviral drug targets. Nat Struct Mol Biol 17: 530–538.

69. Engelman A, Cherepanov P (2012) The structural biology of hiv-1: mechanistic and therapeutic insights. Nat Rev Microbiol 10: 279–290.

70. Moradpour D, Penin F (2013) Hepatitis c virus proteins: from structure to function. Curr Top Microbiol Immunol 369: 113–142.

71. Squires RB, Noronha J, Hunt V, Garca-Sastre A, Macken C, et al. (2012) Influenza research database: an integrated bioinformatics resource for influenza research and surveillance. Influenza Other Respir Viruses 6: 404–416.

72. Edgar RC (2004) Muscle: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res 32: 1792–1797.

CHAPTER 4

HIGH-THROUGHPUT IDENTIFICATION OF LOSS-OF-FUNCTION MUTATIONS FOR ANTI-INTERFERON ACTIVITY IN INFLUENZA A VIRUS NS SEGMENT

4.1 ABSTRACT

Viral proteins often display several functions which require multiple assays to dissect their genetic basis. Here, we describe a systematic approach to screen for loss-of-function mutations that confer fitness disadvantage in a specified growth condition. Our methodology is achieved by genetically monitoring a mutant library under two growth conditions, with and without interferon, by deep sequencing. We employed a molecular tagging technique to distinguish true mutations from sequencing error. This approach enabled us to identify mutations that were negatively selected against in addition to those that were positively selected for. Using this technique, we identified loss-of-function mutations on the influenza A virus NS segment that were sensitive to type I interferon in a high-throughput fashion. Mechanistic characterization further showed that a single substitution, D92Y, resulted in the inability of NS to inhibit RIG-I ubiquitination. The approach described in this study can be applied under any specified condition for any virus that can be genetically manipulated.

4.2 INTRODUCTION

Type I interferon (IFN) is a major component in the host immune system against influenza A virus infection [1]. Briefly, upon influenza infection, signal transduction via the mitochondrial antiviral-signaling protein (MAVS) is initiated upon RIG-I ubiquitination [2–4]. MAVS signaling leads to phosphorylation of interferon regulatory factor 3 (IRF3) by TANK-binding kinase 1 (TBK1) or IκB kinase-ϵ (IKKϵ) [5]. Subsequently, it results in the activation of IFN expression [6–8]. IFN acts as both a paracrine and an autocrine signaling molecule. Binding of IFN to type I interferon receptors (IFNARs) activates the classical JAK-STAT pathway [9]. It then induces the expression of hundreds of interferon inducible genes (ISGs), which possess varied antiviral functions [1]. The IFN antiviral signal can be further amplified by a positive-feedback mechanism [10]. Influenza A virus non-structural protein (NS1), one of the two protein products encoded by segment 8 (NS segment), has acquired multiple strategies to counteract the IFN system [11]. It has been reported that NS1 suppresses the IFN system during viral replication in multiple ways, including the inhibition of IRF3 activation [12], the JNK/AP-1 pathway [13], NF-κB signaling [14], PKR [15], and OAS/RNase L [16].

To study virus-host interactions, various high-throughput approaches, such as siRNA screening [17–19], yeast two-hybrid screening [20], and mass spectrometry [21], have been employed to identify relevant host genetic elements. However, there is a lack of a high-throughput platform for the viral counterparts of these interactions. Identification of viral genetic elements involved in virus-host interaction often requires the construction and analysis of individual mutants. This process has a very low-throughput that limits the number of mutants to be tested. Insertional mutagenesis does permit a higher throughput for identifying critical viral genetic elements and has been applied to hepatitis C virus, venezuelan equine encephalitis virus, and influenza A virus [22–24]. Nonetheless, the resolution of this approach is limited to the protein subdomain level and does not allow the identification of critical residues.

Our previous study has demonstrated the feasibility of using a point mutation library to screen for compensatory mutations at a single nucleotide resolution [48]. Compensatory mutations that

conferred a fitness advantage increased in relative occurrence frequency throughout viral passaging. Consequently, these mutations could be easily identified using 454 pyrosequencing as previously described [48]. Although this previous approach can rapidly identify gain-of-function mutations that provide fitness benefits, the method lacks the power to distinguish between neutral and deleterious mutations. The majority of point mutations in a mutant library exhibit an occurrence frequency of $<$0.1%, which is lower than the error rate of next-generation sequencing ($\sim$0.1-1%). True mutations can not be distinguished from sequencing error unless there is positive enrichment. As a result, any loss-of-function mutations that confer a fitness disadvantage cannot be identified. For example, a mutation that has a 10-fold decrease in occurrence frequency during passaging (0.01% $\rightarrow$ 0.001%) would be interpreted as a neutral phenotype ($\sim$0.1% $\rightarrow$ $\sim$0.1%) in the next-generation-sequencing data due to the sequencing error rate. This limitation poses a unique challenge in utilizing next-generation sequencing technology to screen for loss-of-function mutations that carry a fitness defect in large mutant pools. However, loss-of-function mutations often provide valuable information to characterize genetic elements involved in virus-host interaction. It is therefore important to establish a platform to identify loss-of-function mutations that undergo negative selection in a high-throughput manner.

In this study, we describe an approach that incorporates a sensitive deep sequencing technique to systematically identify loss-of-function mutations. It allows the identification of mutations that are negatively selected against in addition to those that are positively selected for. We provide a proof-of-concept of our approach by identifying residues in the influenza virus NS segment that are critical for anti-IFN function. This is achieved by monitoring a point mutation library of the NS segment in two growth conditions $-$ with and without IFN treatment. By utilizing a tag-based sequencing strategy, we were able to distinguish true mutations from sequencing error. The relative interferon-sensitivities for 1021 NS missense mutations were then estimated by comparing the mutant profiles between both conditions. Experimental validation led to the identification of a novel interferon-sensitive substitution, NS1 D92Y. Further characterization suggests that NS1 D92Y has a defect in blocking RIG-I ubiquitination, which is a critical step in the interferon signaling pathway. To our knowledge, this is the first example of systematically identifying loss-of-function mutations involved in a virus-host interaction. This approach can potentially be adapted to any specified

condition for any virus that can be genetically manipulated.

4.3 RESULTS

Experimental design of the high-throughput screening

The underlying rationale of the high-throughput screening for loss-of-function mutations is that substituting a residue phenotypically critical in one particular growth condition would display a fitness deviation compared to the control environment (Fig. 4-1A). We propose that a mutant library can be employed for this purpose to identify such substitutions. The fitness deviation of a particular substitution would be reflected by the difference of its occurrence frequency between two conditions. This establishes the conceptual foundation for the systematic identification of loss-of-function mutations that are selected against in a specified growth condition but not in the control growth condition.

In this study, we aimed to provide a proof-of-concept by identifying interferon-sensitive mutations on the influenza A virus NS segment. We constructed a plasmid mutant library of NS segment of influenza A/WSN/1933 using error-prone PCR. A mutation rate of one to two point mutations was achieved to minimize any genetic interactions that could potentially confound the results. As a result, on average, each point mutation had an occurrence frequency of $\sim$0.04-0.08% in the mutant library. This mutation rate was significantly higher than the natural mutation rate of influenza, which has been shown to be in the order of 10^{-6} mutation per site per infectious cycle [26], but was lower than the sequencing error rate in next-generation sequencing ($\sim$0.1-1%). The plasmid mutant library consisted of a collection of $>$200,000 clones, allowing sufficient coverage of individual point mutations. The viral mutant library was rescued in 293T cells using the eight-plasmid system as described [27], and then passaged in two different growth conditions, with or without IFN. The viral mutant library was passaged for two 24-hour rounds in A549 cells. Consequently, two viral populations (passaged with or without IFN) derived from the same library were deep sequenced to quantify the occurrence frequency of individual point mutations. During each step in both viral rescue and passaging, $>$35 million cells were employed to avoid any bottleneck effect that would randomly drift the occurrence frequency of each point mutation in the mutant pool. In addition, a low MOI (MOI = 0.05) was used to minimize transcomplementation. As a result, the relative fitness of each point mutation would be reflected from its occurrence frequency.

Next-generation sequencing error poses a unique challenge to identify loss-of-function mutations that confer a fitness disadvantage. The majority of point mutations in our mutant library has an occurrence frequency below the Illumina sequencing error rate ($\sim$0.1-1%). During selection, a given point mutation can be enriched, remain unchanged or diminished in occurrence frequency depending on its relative fitness under the specified growth condition. However, only mutations that are significantly enriched would be identified by conventional deep sequencing techniques due to the high sequencing error rate. Deleterious or phenotypically neutral mutations, which have a low occurrence in the population, are not distinguishable from each other. Here, a nucleotide tagging strategy was adapted to distinguish true mutations from sequencing errors [21]. Briefly, a two-step PCR approach was employed in the DNA sequencing library preparation for sequencing error correction. The first PCR assigned a 12 "N" nucleotide tag to individual molecules. Six amplicons, each with $\sim$140 bp were generated in the first PCR to cover the NS segment (from nt 33 to 826). The input amount for the second PCR was well-controlled such that each tagged template would be sequenced $\sim$10 times. The complexity of the tag population ($4^{12} \approx$ 17 million) was $>$100-fold higher than the number of tagged template being sequenced ($\sim$0.16 million per amplicon). Therefore, sequencing reads generated from the same template would share a unique tag. True mutations can be distinguished from sequencing errors by clustering reads that share the same tag (read cluster). True mutations will appear in all reads within a read cluster. In contrast, sequencing errors only appear in a small fraction of the reads within a read cluster. Ultimately, an "error-free" read will be generated from each read cluster. This approach allowed us to identify loss-of-function mutations that decreased in frequency during viral passage without being obscured by sequencing errors.

Interferon-sensitivity profiling result

Each read cluster had an average size of $\sim$10 reads in the deep sequencing data. True mutations were called only if the mutation occurred in $>$90% of the reads within a read cluster. Furthermore, read cluster with $<$3 reads were removed to increase the confidence level in distinguishing true mutations from sequencing errors. An even coverage across the NS segment, ranging from

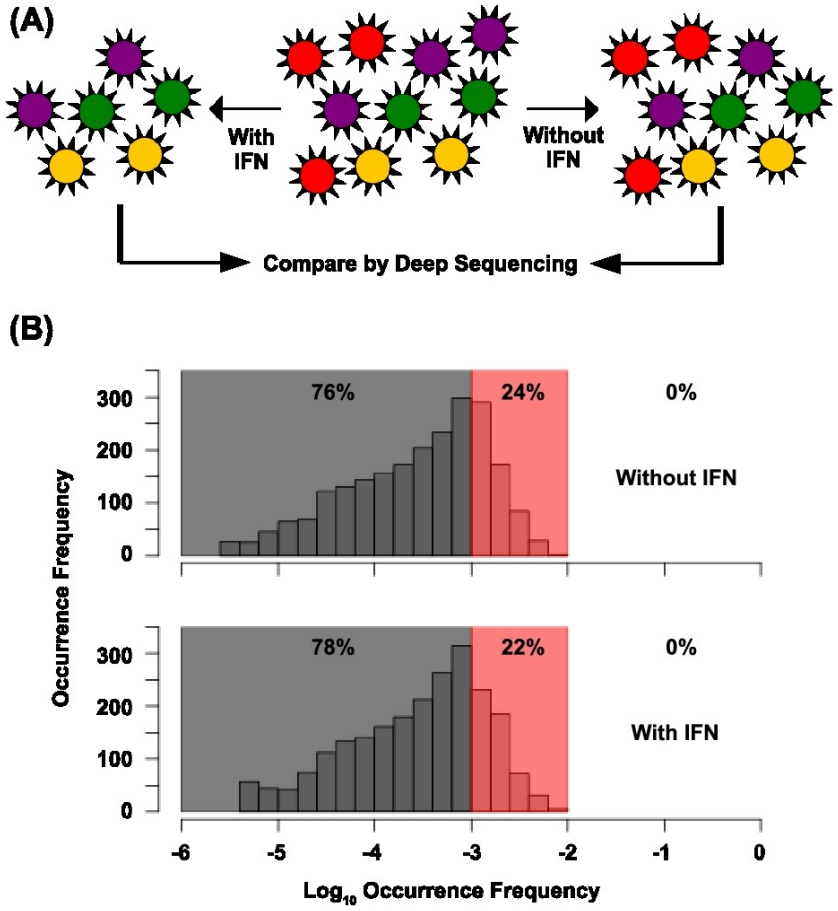


Figure 4-1. Single-nucleotide resolution interferon sensitivity profiling of NS segment. (A) The viral mutant library (middle panel) was passaged in two different growth conditions – with and without IFN. Each circle represents an individual viral particle. Different colors represent mutants with different genotypes. In this depiction, the red virus genotype represents a mutant with a defect in anti-IFN function. Here, red virus has a normal replication phenotype without IFN (right panel), but a deleterious phenotype in the presence of IFN (left panel). (B) Histograms show the distribution of the occurrence frequency of individual point mutations within the mutant library after passage with and without IFN. Sequencing error rate margin (0.1-1%) is shaded in red. Frequency below the sequencing error rate ($<$0.1%) is shaded in grey. Percentages of point mutations within different ranges of occurrence frequency are shown.

$\sim$180000 to $\sim$260000 “error-free” reads, was obtained. This coverage enabled a mutations with an occurrence frequency of the order 10^{-5} to be detected.

As expected, none of the point mutations in the passaged mutant library had an occurrence frequency above the sequencing error margin ($>$1%), whereas 22-24% fell within the sequencing error margin (0.1-1%), and 76-78% were below the sequencing error margin ($<$0.1%) (Fig. 4-1B). This occurrence frequency distribution showed that none of the point mutations conferred significant replication fitness advantage. It also suggested that occurrence frequency of all point mutations within the mutant pool would not be accurately estimated if sequencing error were not corrected. It fully confirmed the necessity to distinguish true mutations from sequencing error using the molecular tag approach. This strategy increased the sensitivity of next-generation sequencing such that the occurrence frequency of individual mutations within the mutant library could be accurately determined.

The occurrence frequency of individual mutations exhibited a Pearson correlation of 0.91 between the two passaging conditions. This indicated that most mutations on the NS segment do not exhibit a relative fitness deviation dependent on IFN. The IFN-sensitivities of individual mutations can be computed by the ratio of their occurrence frequency between the two conditions.

IFN-sensitivity = (occurrence frequency$_{without\ IFN}$)/(occurrence frequency$_{with\ IFN}$)

We aimed to identify loss-of-function mutations that disrupt the anti-IFN function. These loss-of-function mutations would be negatively selected against under IFN treatment but not under the control growth condition. Therefore, they would exhibit a higher relative fitness cost (lower occurrence frequency) in the presence of IFN as compared to the control growth condition. Consequently, loss-of-function mutations would be associated with a high IFN-sensitivity. Our data analysis only included mutations that were sufficiently abundant in the control condition, which allowed IFN-sensitivity to be computed with a higher statistical confidence. A cutoff was set at $>$0.02%, which was reached by $\sim$60% of all possible point mutations. Only mutations that satisfied the confidence cutoff were included in our analysis unless otherwise stated.

Although NS1 is important for counteracting IFN anti-viral effects, it is nonessential for viral replication [29]. As a result, defective NS1 would confer a higher IFN-sensitivity than its functionally intact counterpart. It is expected that mean IFN-sensitivity would be the lowest for silent mutations, followed by missense mutations, and the greatest for nonsense mutations. Indeed, the IFN-sensitivities of nonsense mutations (mean IFN-sensitivity = 1.53) were significantly greater than missense mutations (mean IFN-sensitivity = 1.13) (P = $2e^{-6}$, Wilcoxon rank-sum test), whereas the IFN-sensitivities of missense mutations were significantly greater than silent mutations (mean IFN-sensitivity = 0.97) (P = $1e^{-9}$, Wilcoxon rank-sum test). These results support the validity of the data.

IFN-sensitivity was computed for each of the 1021 missense mutations that satisfied the confidence cutoff (Fig. 4-2). A total of 21 missense mutations displayed an IFN-sensitivity greater than three standard deviations above the mean (IFN-sensitivity $>$ 2.46). Included in this set of IFN-sensitive mutations was an arginine [R] to leucine [L] substitution at NS1 residue 38. R38 has been reported to be absolutely required for NS1 RNA-binding activity [30,31], which facilitates the masking of viral RNA to inhibit IFN activation during viral infection [12–14]. In fact, the mean IFN-sensitivity of all missense mutations at residues critical for RNA-binding activity, P31, D34, R35, R38, K41, G45, R46 and T49, (mean IFN-sensitivity = 1.32) was significantly greater than that of all missense mutations throughout the entire NS segment (P = $5e^{-4}$, Wilcoxon rank-sum test). This result confirms the reliability of our dataset.

Validation of individual interferon-sensitive mutations identified from the screen

Next, we aim to identify novel loss-of-function mutations that disrupt the anti-IFN function of NS1. We randomly selected 9 of the 21 missense mutations in the NS segment that had an IFN-sensitivity greater than three standard deviation above the mean (Fig. 4-3). Individual mutants were constructed and analyzed. The viral copy numbers were compared using qPCR at 24 hours after replication in A549 (MOI of 0.05) with (30 U/ml) and without IFN-α. Six of the nine mutants displayed a $>$2-fold higher IFN-sensitivity than WT in this assay. Three of the six missense mu-

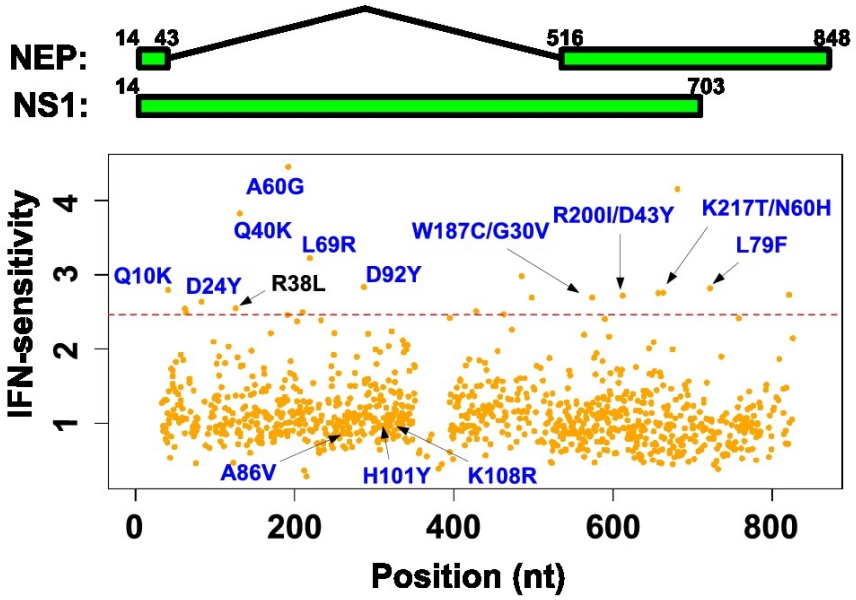


Figure 4-2. IFN-sensitivity profiling. The coding region of NS1 and NEP is depicted as a green rectangle at the top of the figure. The IFN-sensitivity for individual missense mutations is plotted versus nucleotide positions. Only mutations with an occurrence frequency $>$0.02% in the control condition are shown. The red dotted line represents three standard deviations above the mean. Labels indicate the corresponding amino acid substitution. Subsitutions in NEP are underlined. Substitutions in blue label represent mutations that were individually constructed and analyzed in this study.

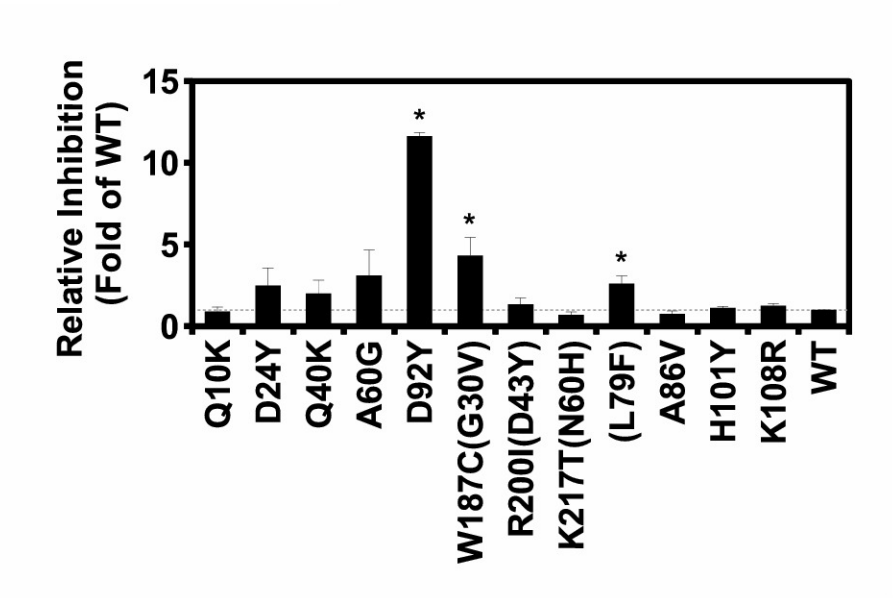


Figure 4-3. Identification of novel IFN-sensitive mutation. The relative viral replication inhibition by IFN-α was measured by RNA copy number using qPCR. Error bars represent the standard error of the mean (SEM) of three independent experiments. The grey dotted line represents the WT level. Substitutions in NEP are showed in parentheses. Mutants with a p-value $<$0.05 are indicated with an asterisk. One-tailed t-test was performed to compute the p-value.

tations resulted in substitutions in the NS1 RNA-binding domain (D24Y, Q40K and A60G), one in each of the NS1 effector domain (D92Y), the NEP (L79F), and the overlapping reading frame of NS1 effect domain (W187C) with NEP (G30V). As a control, three missense mutations on NS1 (A86V, H101Y and K108R) that had an IFN-sensitivity ranging from 0.85 to 0.98 in the profiling data were included in this validation assay. None of them displayed any phenotypic difference from WT in this assay. The validation result verifies the design of our high-throughput screening approach. NS1 D92Y showed the strongest IFN-sensitivity phenotype (12-fold higher than WT, P = 10^{-5}) among the validated mutants and was selected for further characterization.

Mechanistic characterization of D92Y reveals its role in IFN-signaling pathway

Residue 92 on NS1 has been reported as a tumor necrosis factor-α (TNF-α) resistance determinant [33]. Human influenza viruses carry a conserved aspartic acid at this position whereas avian influenza viruses carry a glutamic acid. The D92E NS1 substitution has been shown to contribute to TNF-α resistance while exerting no effect on IFN-sensitivity in human influenza virus [33, 34]. On the other hand, our data suggests that substituting residue 92 to tyrosine significantly increases IFN-sensitivity (Fig. 4-3). We first performed a structural analysis of this mutant by Rosetta software using parameters from row 16 of Table I in Kellogg *et al.* [32]. The energy minimization simulation predicts that D92Y disrupted a hydrophobic pocket on NS1 due to the volume increase from aspartic acid to tyrosine (Fig. 4-4A). We then performed a cycloheximide blocking experiment to examine the protein stability effect of D92Y on NS1. The rates of degradation of D92Y NS1 and WT NS1 did not display a significant difference (Fig. 4-4B), suggesting that D92Y did not affect NS1 protein stability. As a result, residue 92 may be involved in maintaining the conformation of the hydrophobic pocket critical for anti-IFN function. Furthermore, our sequencing data also showed that D92V, D92N, D92G, and D92A had no effect (Fig. 4-4C). Structural modeling demonstrated that these substitutions did not significantly alter the conformation of the hydrophobic pocket (Fig. 4-4A). These results are consistent with the importance of maintaining the hydrophobic pocket for counteracting IFN antiviral effects.

NS1 effector domain has been shown to mediate the inhibition of RIG-I ubiquitination [12], and

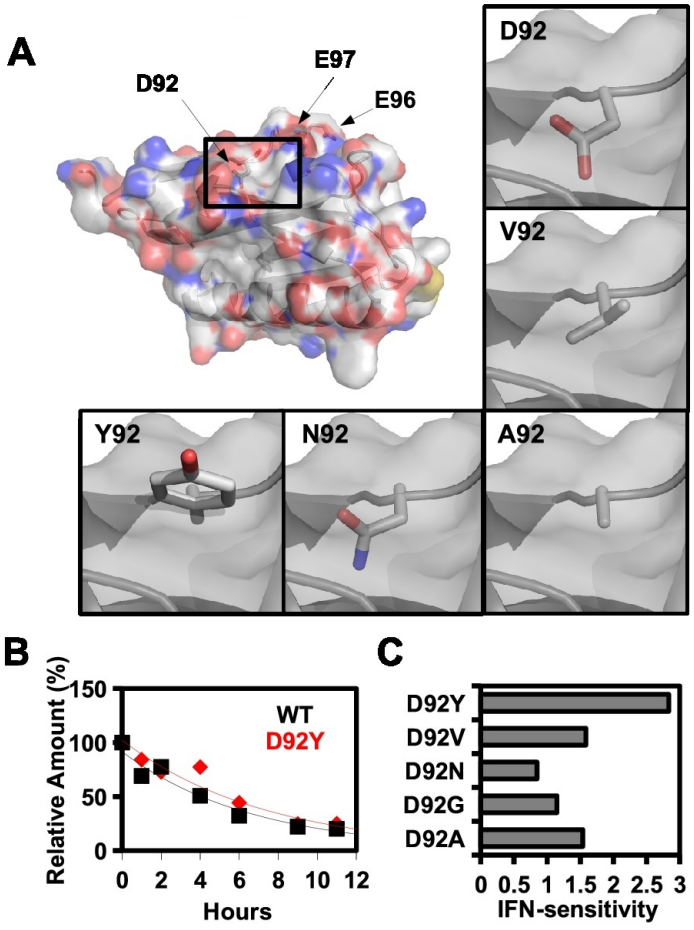


Figure 4-4. Structural and stability effect of substitutions at NS1 residue 92. (A) The structures for different substitutions at residue 92 are displayed. The rotamers for different substitutions were generated by free energy minimization simulation using Rosetta software [32]. Only the D92Y substitution reduced the pocket space. (B) Transiently transfected 293T cells were treated with 40 ug/ml cyclohexamide. The NS1 protein level at different time points were quantified by densitometry to examine the rate of protein degradation. (C) IFN-sensitivities for the profiled missense substitutions at D92 are shown.

hence the inhibition of IFN expression and potentially the positive-feedback mechanism [3, 4, 36]. We hypothesized that D92Y abolished the suppression of RIG-I ubiquitination. We first looked at the downstream effectors of RIG-I ubiquitination, NF-κB and IFN-β [2, 3, 36], using a luciferase reporter assay. Indeed, compared to WT, D92Y lost the ability to suppress NF-κB promoter activation (Fig. 4-5A) and IFN-β promoter activation (Fig. 4-5B) induced by Sendai virus (SeV). These results are consistent with our hypothesis that suppression of RIG-I ubiquitination can be abolished by D92Y. Consequently, we performed a co-immunoprecipitation assay to test the effect of D92Y on the inhibition of SeV-induced RIG-I ubiquitination (Fig. 4-5C). HA-tagged ubiquitin and Flag-tagged RIG-I were cotransfected in the presence of NS1 WT or D92Y. Flag-tagged RIG-I was immmunoprecipitated 18 hours after SeV infection. Ubiquitination of RIG-I was then detected by HA-antibody. Indeed, ubiquitination of RIG-I was inhibited by NS1 WT but not NS1 D92Y. Taken together, our results indicate that residue 92 is critical for the inhibition of RIG-I ubiquitination, an important step for IFN signaling [3], likely by maintaining the conformation of the hydrophobic pocket.

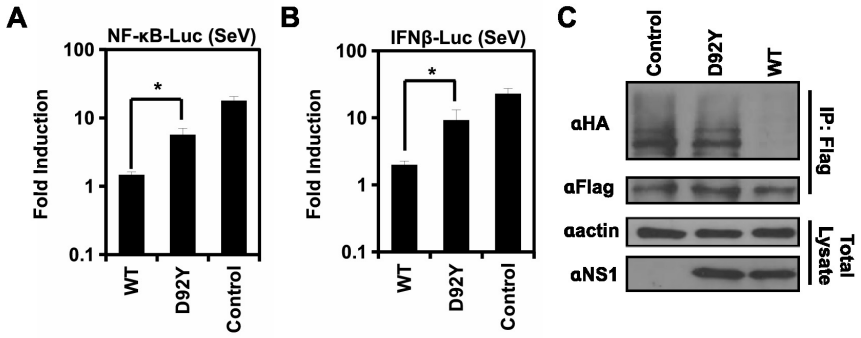


Figure 4-5. Characterization of the NS1 D92Y substitution. (A) SeV mediated NK-κB promoter activities were measured by luciferase assay. (B) SeV mediated IFN-β promoter activities were measured by luciferase assay. (A-B) Mean value is plotted. Error bars represent the standard error of the mean (SEM) of three independent experiments. (C) SeV mediated ubiquitination of RIG-I was measured by immunoprecipitation and western blot. All differences between D92Y and WT were significant (p-value $<$0.05) as indicated with an asterisk. Two-tailed t-test was performed to compute the p-value.

4.4 DISCUSSION

Identification of loss-of-function viral mutation is critical for understanding virus-host interactions. They often represent essential mechanisms for viral replication under a specified growth condition. Mechanistic interrogation of identified mutations offers further structure-function insight. The impact can be translated to vaccine and antiviral drug development. For example, temperature-sensitive mutations and anti-interferon actions provide the foundation for influenza vaccine development [37, 38]. In addition, the cell entry mechanism of influenza virus has been suggested as a potent antiviral target [39–41]. This and many other examples highlight the medical importance of characterization of viral genetic elements. However, there is a lack of high-throughput platform to screen for these functional elements.

In vitro culture system and reverse genetics offer a powerful tool to study virus-host interaction in laboratory settings. They are well-established for several medically important viruses, such as HCV [42–44], HIV [45], and influenza virus [27, 46]. Functional characterization of viral genetic elements often relies on constructing and analyzing individual mutants with multiple assays. The low-throughput of this process limits the number of mutants to be analyzed. In this study, we overcame this challenge by coupling saturation mutagenesis with a sensitive deep sequencing technique. This allowed us to monitor negative selection in addition to positive selection. Functionality of each mutation was inferred by comparing the mutational profile of a point mutation library under different growth conditions. This approach is rapid, unbiased and comprehensive. We provided a proof-of-concept for this differential profiling technique by identifying IFN-sensitive mutations in the influenza A NS segment.

To our knowledge, this is the first example of utilizing a high-throughput genetic platform at single-nucleotide resolution to identify loss-of-function mutations under a specific growth condition. Our previous study described a similar technique, which utilized deep sequencing to monitor a mutant library to identify compensatory mutations [48]. However, the approach described in our previous study was only capable of identifying mutations that were enriched during selection. This is because without significant enrichment during viral passaging, occurrence frequencies of most

point mutations, if not all, would remain below the sequencing error rate. The phenotypic effects of non-enriched point mutations could not be inferred as it was impossible to partition them from sequencing errors. In this study, an error-correction technique was implemented to distinguish true mutations from sequencing errors. This critical step allowed the monitoring of individual point mutations even if their occurrence frequency was below the sequencing error rate. Therefore, we are now capable of identifying loss-of-function mutations that decreased in frequency during the IFN treatment.

We anticipate that the power of the screening technique described in this study will increase as sequencing technology advances. Increasing sequencing depth will minimize the sampling error of individual point mutations, hence increase the accuracy in computing the occurrence frequency of individual point mutations and improve the precision of estimating their fitness deviations between different growth conditions. Additionally, a longer read length will enable the examination of different mutations existing in the same clone, which will allow the study of genetic interactions among point mutations.

A recent study has suggested that besides NS segment, other segments of the influenza A virus also possessed the ability to counteract IFN activity [13]. Although we only applied the screening technique to NS segment, it is worthwhile to extend the analysis to the whole genome or other viral genetic backgrounds in the future. Moreover, this approach is not limited to interferon. It can also be employed to identify temperature sensitive mutations as well as viral genetic elements involved in virus-host interaction under a specific cellular or immune response, such as apoptosis, autophagy, ER stress, NK cells, macrophages, etc. In summary, the genetic approach presented in this study has a wide range of potential applications to identify residues involved in viral-host interactions. More importantly, our methodology can be applied to probe any virus that can be genetically manipulated in the laboratory.

4.5 MATERIALS AND METHODS

Viral mutant library and point mutations

The NS segment point mutation library was constructed as previously described [48] using the eight plasmid reverse genetic system of influenza A/WSN/1933 (H1N1) [27]. Briefly, the NS segment was PCR amplified using the error-prone polymerase Mutazyme II (Stratagene, La Jolla, CA) and cloned into a BsmBI-digested parental vector pHW2000. Ligations were carried out with high concentration T4 ligase (Life Technologies, Carlsbad, CA). Transformations were carried out with electrocompetent MegaX DH10B T1R cells (Life Technologies). Sequencing individual clones showed that the plasmid mutant library had an average of one to two point mutations per clone. The plasmid mutant library was purified from a collection of $>$200,000 clones using the QIAGEN plasmid maxi kit (Qiagen Sciences, Germantown, MD). Point mutations for experimental validation were constructed using the QuikChange XL Mutagenesis kit (Stratagene) according to manufacturer's instructions.

Transfections, infections, and titering

C227 cells, a dominant negative IRF-3 stably expressing cell line derived from human embryonic kidney (293T) cells [48], were transfected with the NS mutant library plasmid (for screening) or NS point mutation plasmid (for validation) plus 7 other wildtype plasmids using Lipofectamine 2000 (Life Technologies). Supernatant was replaced with fresh cell growth medium at 24 hrs and 48 hrs post-transfection. At 72 hrs post-transfection, supernatant containing infectious virus was harvested, filtered through a 0.45 um MCE filter, and stored at -80^{o}C. The $TCID_{50}$ was measured on A549 cells (human lung carcinoma cells).

Virus from C227 cells transfection was used to infect A549 cells at an MOI of 0.05. Infected cells were washed three times with PBS followed by the addition of fresh cell growth medium at 2 hrs post-infection. Virus was harvested at 24 hrs post-infection. For infection under IFN treatment, 30 U/ml of IFN-α (Fitzgerald, Acton, MA) was added 18 hrs pre-infection and the concentration was maintained throughout the course of infection. The viral mutant library was passaged for two rounds in A549 cells before subjected to deep sequencing. Due to the huge complexity of the mu-

tant library, $>$35 million cells were used at both viral rescue and passaging to avoid any bottleneck effect.

Sequencing library preparation

The viral RNA was extracted using QIAmp viral RNA kit (Qiagen Sciences) and reverse-transcribed to cDNA using SuperScript III reverse transcriptase (Life Technologies). A two-step PCR was performed for sequencing library preparation. In the first PCR, the NS segment was divided into six amplicons generated using the primers:

Amplicon 1: 5'-CTA CAC GAC GCT CTT CCG ATC TNN NNN NTA ATG GAT CCA AAC ACT GTG T-3' and 5'-TGC TGA ACC GCT CTT CCG ATC TNN NNN NCC AGC ACG GGT GGC TGT-3'

Amplicon 2: 5'-CTA CAC GAC GCT CTT CCG ATC TNN NNN NTC TTG GTC TGG ACA TCG AA-3' and 5'-TGC TGA ACC GCT CTT CCG ATC TNN NNN NTT TCT GCT TGG GCA TGA GC-3'

Amplicon 3: 5'-CTA CAC GAC GCT CTT CCG ATC TNN NNN NAT GTC AAG GCA CTG GTT CAT-3' and 5'-TGC TGA ACC GCT CTT CCG ATC TNN NNN NCG CCA ACA ATT GTC CCC T-3'

Amplicon 4: 5'-CTA CAC GAC GCT CTT CCG ATC TNN NNN NTA AGG GCC TTC ACC GAA G-3' and 5'-TGC TGA ACC GCT CTT CCG ATC TNN NNN NTC ATT ACT GCT TCT CCA AGC-3'

Amplicon 5: 5'-CTA CAC GAC GCT CTT CCG ATC TNN NNN NAA CAC AGT TCG AGT CTC TGA-3' and 5'-TGC TGA ACC GCT CTT CCG ATC TNN NNN NCT ATT CTC TGT TAT CTT CAG TC-3'

Amplicon 6: 5'-CTA CAC GAC GCT CTT CCG ATC TNN NNN NTG GCG GGA ACA ATT AGG TC-3' and 5'-TGC TGA ACC GCT CTT CCG ATC TNN NNN NAT AAG CTG AAA CGA GAA AGT T-3'

The six amplicon products were then mixed at an equal molar ratio and subjected to the second PCR, which generated multiple copies of each tagged template using the primers: 5'-AAT GAT ACG GCG ACC ACC GAG ATC TAC ACT CTT TCC CTA CAC GAC GCT CTT CCG-3' and 5'-CAA GCA GAA GAC GGC ATA CGA GAT CGG TCT CGG CAT TCC TGC TGA ACC GCT CTT CCG-3'. For individual sample population, $\sim$1 million copies of mixed amplicon products were provided as input for the second PCR. The product from the second PCR was then deep sequenced using Illumina HiSeq 2000.

Data Analysis

Sequencing reads were mapped by BWA with a maximum of six mismatches and no gap [55]. Amplicons with the same tag were grouped into a read cluster and further conflated into an "error-free" read as described in [21]. True mutations were called only if the mutation occurred in $>$90% of the reads within a read cluster. Read clusters with less than three reads were removed. All analyses were performed with custom python scripts that are available upon request. IFN-sensitivity for each point mutation was computed by (occurrence frequency$_{without\ IFN}$)/(occurrence frequency$_{with\ IFN}$).

Real-time reverse-transcription PCR (RT-qPCR)

Supernatant from infected cells was processed with the QIAmp viral RNA kit (Qiagen Sciences). Viral RNA was reverse transcribed with Superscript III (Invitrogen) using random hexamers. qPCR was performed on a DNA Engine OPTICON 2 system (Bio-Rad, Irvine, CA) using SYBR Green (Life Technologies) with primers: 5'-GAC GAT GCA ACG GCT GGT CTG-3' and 5'-ACC ATT GTT CCA AC TCC TTT-3'.

Immunoprecipitation and Immunoblot

293T cells seeded on a 6 well-plate were transfected using Lipofectamine 2000 (Life Technologies). At 24 hours post transfection, cells were infected with Sendai virus (100 U/ml). At 18 hours post infection, cells were lyzed with lysis buffer (50 mM Tris pH 7.5, 150 mM NaCl, 1.0% Nonidet P-40, 10% glycerol, 10 g/ml aprotinin, 10 g/ml pepstatin, 0.5 mM phenylmethylsulfonyl fluoride). Lysates were incubated with EZview red anti-Flag M2 affinity gel (Sigma, St. Louis, MO) for 6 hours. Beads were washed 3 times with lysis buffer. Proteins were eluted with SDS loading buffer (50 mM Tris-HCl pH 6.8, 2% SDS, 10% glycerol, 1% β-mercaptoethanol, 12.5 mM EDTA, 0.02% bromophenol blue) and heated at 90 degree celsius for 5 mins. Mouse anti-Flag antibody (Sigma), mouse anti-β actin antibody (Sigma), mouse anti-HA antibody (Sigma), rabbit anti-NS1 antibody (GeneTex, Irvine, CA), sheep horseradish peroxidase-conjugated anti-mouse Immunoglobulin G (GE Healthcare, Pasadena, CA), and donkey horseradish peroxidase-conjugated anti-rabbit Immunoglobulin G (GE Healthcare) were used for protein detection.

Protein stability

293T cells were transfected with flag-tagged NS1. At 24 hrs post transfection, 40 ug/ml cyclohexamide was added. Transfected cells were harvested at indicated time points. Relative amount of cellular flag-tagged NS1 was quantified by densitometry.

Luciferase Assay

293T cells seeded on 48 well plates were transfected with 50 ng of firefly-luciferase reporter plasmid, 5 ng of PGK_renilla-luciferase, and other indicated expression plasmids using Lipofectamine 2000 (Life Technologies). SeV (100 U/mL) was added at 24 hrs post-transfection and luciferase activity assay was performed at 48 hrs post-transfection using Promega Dual-Luciferase Assay Kit according to manufactuerer's instructions (Promega, Madison, WI). For over-expression mediated promoter activity assay, luciferase activity assay was performed at 24 hrs post-transfection. Adjusted luciferase activity was calculated by normalizing the firefly-luciferase activities to their internal renilla luciferase controls. Fold induction was calculated by dividing the adjusted luciferase activities in the treated samples by that of the untreated samples.

4.6 BIBLIOGRAPHY

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

2. Rehwinkel J, Tan CP, Goubau D, Schulz O, Pichlmair A, et al. (2010) Rig-i detects viral genomic rna during negative-strand rna virus infection. Cell 140: 397–408.

3. Gack MU, Shin YC, Joo CH, Urano T, Liang C, et al. (2007) Trim25 ring-finger e3 ubiquitin ligase is essential for rig-i-mediated antiviral activity. Nature 446: 916–920.

4. Yoneyama M, Fujita T (2009) Rna recognition and signal transduction by rig-i-like receptors. Immunol Rev 227: 54–65.

5. Fitzgerald KA, McWhirter SM, Faia KL, Rowe DC, Latz E, et al. (2003) Ikkepsilon and tbk1 are essential components of the irf3 signaling pathway. Nat Immunol 4: 491–496.

6. Yoneyama M, Suhara W, Fujita T (2002) Control of irf-3 activation by phosphorylation. J Interferon Cytokine Res 22: 73–76.

7. Hou F, Sun L, Zheng H, Skaug B, Jiang QX, et al. (2011) Mavs forms functional prion-like aggregates to activate and propagate antiviral innate immune response. Cell 146: 448–461.

8. Hiscott J, Pitha P, Genin P, Nguyen H, Heylbroeck C, et al. (1999) Triggering the interferon response: the role of irf-3 transcription factor. J Interferon Cytokine Res 19: 1–13.

9. Platanias LC (2005) Mechanisms of type-i- and type-ii-interferon-mediated signalling. Nat Rev Immunol 5: 375–386.

10. Taniguchi T, Takaoka A (2001) A weak signal for strong responses: interferon-alpha/beta revisited. Nat Rev Mol Cell Biol 2: 378–386.

11. Hale BG, Randall RE, Ortn J, Jackson D (2008) The multifunctional ns1 protein of influenza a viruses. J Gen Virol 89: 2359–2376.

12. Talon J, Horvath CM, Polley R, Basler CF, Muster T, et al. (2000) Activation of interferon regulatory factor 3 is inhibited by the influenza a virus ns1 protein. J Virol 74: 7989–7996.

13. Ludwig S, Wang X, Ehrhardt C, Zheng H, Donelan N, et al. (2002) The influenza a virus ns1 protein inhibits activation of jun n-terminal kinase and ap-1 transcription factors. J Virol 76: 11166–11171.

14. Wang X, Li M, Zheng H, Muster T, Palese P, et al. (2000) Influenza a virus ns1 protein prevents activation of nf-kappab and induction of alpha/beta interferon. J Virol 74: 11566–11573.

15. Bergmann M, Garcia-Sastre A, Carnero E, Pehamberger H, Wolff K, et al. (2000) Influenza virus ns1 protein counteracts pkr-mediated inhibition of replication. J Virol 74: 6203–6206.

16. Min JY, Krug RM (2006) The primary function of rna binding by the influenza a virus ns1 protein in infected cells: Inhibiting the 2'-5' oligo (a) synthetase/rnase l pathway. Proc Natl Acad Sci U S A 103: 7100–7105.

17. Panda D, Das A, Dinh PX, Subramaniam S, Nayak D, et al. (2011) Rnai screening reveals requirement for host cell secretory pathway in infection by diverse families of negative-strand rna viruses. Proc Natl Acad Sci U S A 108: 19036–19041.

18. Knig R, Stertz S, Zhou Y, Inoue A, Hoffmann HH, et al. (2010) Human host factors required for influenza virus replication. Nature 463: 813–817.

19. Karlas A, Machuy N, Shin Y, Pleissner KP, Artarini A, et al. (2010) Genome-wide rnai screen identifies human host factors crucial for influenza virus replication. Nature 463: 818–822.

20. Shapira SD, Gat-Viks I, Shum BOV, Dricot A, de Grace MM, et al. (2009) A physical and regulatory map of host-influenza interactions reveals pathways in h1n1 infection. Cell 139: 1255–1267.

21. Shaw ML, Stone KL, Colangelo CM, Gulcicek EE, Palese P (2008) Cellular proteins in influenza virus particles. PLoS Pathog 4: e1000085.

22. Arumugaswami V, Remenyi R, Kanagavel V, Sue EY, Ho TN, et al. (2008) High-resolution functional profiling of hepatitis c virus genome. PLoS Pathog 4: e1000182.

23. Heaton NS, Sachs D, Chen CJ, Hai R, Palese P (2013) Genome-wide mutagenesis of influenza virus reveals unique plasticity of the hemagglutinin and ns1 proteins. Proc Natl Acad Sci U S A 110: 20248–20253.

24. Beitzel BF, Bakken RR, Smith JM, Schmaljohn CS (2010) High-resolution functional mapping of the venezuelan equine encephalitis virus genome by insertional mutagenesis and massively parallel sequencing. PLoS Pathog 6: e1001146.

25. Wu NC, Young AP, Dandekar S, Wijersuriya H, Al-Mawsawi LQ, et al. (2013) Systematic identification of h274y compensatory mutations in influenza a virus neuraminidase by high-throughput screening. J Virol 87: 1193–1199.

26. Nobusawa E, Sato K (2006) Comparison of the mutation rates of human influenza a and b viruses. J Virol 80: 3675–3678.

27. Neumann G, Watanabe T, Ito H, Watanabe S, Goto H, et al. (1999) Generation of influenza a viruses entirely from cloned cdnas. Proc Natl Acad Sci U S A 96: 9345–9350.

28. Kinde I, Wu J, Papadopoulos N, Kinzler KW, Vogelstein B (2011) Detection and quantification of rare mutations with massively parallel sequencing. Proc Natl Acad Sci U S A 108: 9530–9535.

29. Garca-Sastre A, Egorov A, Matassov D, Brandt S, Levy DE, et al. (1998) Influenza a virus lacking the ns1 gene replicates in interferon-deficient systems. Virology 252: 324–330.

30. Wang W, Riedel K, Lynch P, Chien CY, Montelione GT, et al. (1999) Rna binding by the novel helical domain of the influenza virus ns1 protein requires its dimer structure and a small number of specific basic amino acids. RNA 5: 195–205.

31. Donelan NR, Basler CF, Garca-Sastre A (2003) A recombinant influenza a virus expressing an rna-binding-defective ns1 protein induces high levels of beta interferon and is attenuated in mice. J Virol 77: 13257–13266.

32. Kellogg EH, Leaver-Fay A, Baker D (2011) Role of conformational sampling in computing mutation-induced changes in protein structure and stability. Proteins 79: 830–838.

33. Seo SH, Hoffmann E, Webster RG (2002) Lethal h5n1 influenza viruses escape host antiviral cytokine responses. Nat Med 8: 950–954.

34. Hayman A, Comely S, Lackenby A, Hartgroves LCS, Goodbourn S, et al. (2007) Ns1 proteins of avian influenza a viruses can act as antagonists of the human alpha/beta interferon response. J Virol 81: 2318–2327.

35. Gack MU, Albrecht RA, Urano T, Inn KS, Huang IC, et al. (2009) Influenza a virus ns1 targets the ubiquitin ligase trim25 to evade recognition by the host viral rna sensor rig-i. Cell Host Microbe 5: 439–449.

36. Yoneyama M, Kikuchi M, Natsukawa T, Shinobu N, Imaizumi T, et al. (2004) The rna helicase rig-i has an essential function in double-stranded rna-induced innate antiviral responses. Nat Immunol 5: 730–737.

37. Richt JA, Garca-Sastre A (2009) Attenuated influenza virus vaccines with modified ns1 proteins. Curr Top Microbiol Immunol 333: 177–195.

38. Zhou B, Li Y, Speer SD, Subba A, Lin X, et al. (2012) Engineering temperature sensitive live attenuated influenza vaccines from emerging viruses. Vaccine 30: 3691–3702.

39. Whittle JRR, Zhang R, Khurana S, King LR, Manischewitz J, et al. (2011) Broadly neutralizing human antibody that recognizes the receptor-binding pocket of influenza virus hemagglutinin. Proc Natl Acad Sci U S A 108: 14216–14221.

40. Ekiert DC, Wilson IA (2012) Broadly neutralizing antibodies against influenza virus and prospects for universal therapies. Curr Opin Virol 2: 134–141.

41. Edinger TO, Pohl MO, Stertz S (2014) Entry of influenza a virus: host factors and antiviral targets. J Gen Virol 95: 263–277.

42. Blight KJ, McKeating JA, Rice CM (2002) Highly permissive cell lines for subgenomic and genomic hepatitis c virus rna replication. J Virol 76: 13001–13014.

43. Lohmann V, Krner F, Koch J, Herian U, Theilmann L, et al. (1999) Replication of subgenomic hepatitis c virus rnas in a hepatoma cell line. Science 285: 110–113.

44. Lindenbach BD, Evans MJ, Syder AJ, Wlk B, Tellinghuisen TL, et al. (2005) Complete replication of hepatitis c virus in cell culture. Science 309: 623–626.

45. Adachi A, Gendelman HE, Koenig S, Folks T, Willey R, et al. (1986) Production of acquired immunodeficiency syndrome-associated retrovirus in human and nonhuman cells transfected with an infectious molecular clone. J Virol 59: 284–291.

46. Pleschka S, Jaskunas R, Engelhardt OG, Zrcher T, Palese P, et al. (1996) A plasmid-based reverse genetics system for influenza a virus. J Virol 70: 4188–4192.

47. Prez-Cidoncha M, Killip MJ, Oliveros JC, Asensio VJ, Fernndez Y, et al. (2014) An unbiased genetic screen reveals the polygenic nature of the influenza virus anti-interferon response. J Virol 88: 4632–4646.

48. Hwang S, Kim KS, Flano E, Wu TT, Tong LM, et al. (2009) Conserved herpesviral kinase promotes viral persistence by inhibiting the irf-3-mediated type i interferon response. Cell Host Microbe 5: 166–178.

49. Li H, Durbin R (2009) Fast and accurate short read alignment with burrows-wheeler transform. Bioinformatics 25: 1754–1760.

CHAPTER 5

HIV-1 QUASISPECIES DELINEATION BY TAG LINKAGE DEEP SEQUENCING

5.1 ABSTRACT

Trade-offs between throughput, read length, and error rates in high-throughput sequencing limit certain applications such as monitoring viral quasispecies. Here, we describe a molecular-based tag linkage method that allows assemblage of short sequence reads into long DNA fragments. It enables haplotype phasing with high accuracy and sensitivity to interrogate individual viral sequences in a quasispecies. This approach is demonstrated to deduce $\sim$2000 unique 1.3 kb viral sequences from HIV-1 quasispecies *in vivo* and after passaging *ex vivo* with a detection limit of $\sim$0.005$\%$ to $\sim$0.001$\%$. Reproducibility of the method is validated quantitatively and qualitatively by a technical replicate. This approach can improve monitoring of the genetic architecture and evolution dynamics in any quasispecies population.

5.2 INTRODUCTION

Many viruses have such high replication and mutation rates that they exist as a quasispecies *in vivo* [1]. A viral quasispecies population contains a variety of genotypic variants that are related by similar mutations and exist in varying abundance depending on their relative fitness within the host environment. In this report, we refer to viral quasispecies as the whole population of genotypic variants, whereas viral sequence is defined as the individual viral variant within quasispecies population. Viral sequence variation in the quasispecies population can be rapidly generated by point mutation and/or recombination [1, 2]. Mutation rates can be as high as in the order of one per replication cycle, in which the progeny virus is unlikely to be identical to its parental template. This diverse array of viral sequences permits robust adaptation and evolution.

Often, genotypes with a particular set of mutations gain a significant fitness advantage through synergistic phenotypic effect among multiple mutations, which is also known as epistasis. Epistasis has an important role in host adaptation and may drive evolution towards drug resistance and immune evasion [4–7, 14]. In many cases, virus drug resistance requires two or more mutations in concert, especially when multiple drugs are applied simultaneously [7–9]. Therefore, monitoring individual viral haplotypes in the quasispecies populations within patients is important to estimate the risk of viral rebound and further provide customized treatment [10]. Characterizing the population structure of viral quasispecies in the host also helps to understand the evolutionary landscape and *cis*-interactions among genetic elements.

Clonal sequencing has been frequently employed to examine the genetic makeup of individual viruses within a quasispecies population. However, clonal sequencing has a low throughput and a high sequencing cost per nucleotide. It limits the number of viral sequences, hence haplotype variants, being genetically interrogated. On the other hand, next generation sequencing (NGS) technology provides enough throughput and sensitivity to detect very rare viral mutations. Nevertheless, the short read lengths of NGS pose a challenge in reconstruction of individual viral sequences within a viral quasispecies. First of all, it is often difficult to distinguish rare mutations that exist in the quasispecies population with sequencing errors from NGS. Secondly, haplotype

phasing is extremely challenging when mutations are sporadic and are separated by long, highly conserved or even completely identical regions. These technical challenges make it extremely difficult to reconstruct viral quasispecies from NGS data.

Existing methods in reconstructing viral quasispecies from NGS platforms rely heavily on computational tools, including the development of read graph-based or probabilistic-based algorithms that utilize the information from overlapping reads [11–20]. Although they provide an approximation of haplotype information present in a viral quasispecies, the sensitivity and accuracy vary depending on sequencing error rate and quasispecies diversity. As a result, it is critical to develop a viral quasispecies recontruction method with higher sensitivity and accuracy in both mutation calling and haplotype phasing.

In order to genetically define a viral quasispecies population, we developed a novel analytical technique to assemble short Illumina amplicon sequence reads derived from individual viral sequences. In contrast to algorithmic-based methods for quasispecies reconstruction, tag linkage approach is a molecular-based approach. To the best of our knowledge, this is the first experimental approach that specialized in quasispecies reconstruction. The methodology consists of three key steps: 1) Assigning unique tags to individual viral sequences to distinguish each variant within the viral quasispecies, 2) Controlling the complexity of the library during amplification to ensure sufficient coverage for sampled viral sequences, and 3) Using a tag linkage strategy to deduce the full-length templates from non-overlapping amplicons. Here, we provide a proof-of-concept study by utilizing this approach to genetically characterize an HIV-1 quasispecies population under two conditions: an isolated *in vivo* virus population and the virus population derived from the same chronically infected HIV-1 patient passaged *ex vivo* in cell culture. We achieve a detection limit of $\sim$0.005$\%$ to $\sim$0.001$\%$. The reproducibility is validated with a technical replicate. Overall, this approach enables accurate haplotype phasing with very high sensitivity.

5.3 RESULTS

Library preparation for sequencing

The underlying rationale is to assign a unique tag to individual viral sequences within the quasispecies and to distribute the tag to every sequencing read originated from the same viral sequence (Fig. 5-1A). Individual viral sequences within the quasispecies can be assembled by grouping sequencing reads that share the same tag. As a result, the tag linkage approach described in this study permits reconstruction of individual viral sequences from NGS reads despite the lack of overlap.

The workflow for sequencing library preparation is summarized in Fig. 5-1B-F. Briefly, individual DNA molecules are assigned a unique tag by PCR (Fig. 5-1B). The tag consists of a 13 "N" sequence that allows distinguishing $4^{13} \approx 70$ million molecules. After tagging individual DNA molecules within the pool, the complexity of the pool is being controlled. Complexity is defined as the number of tagged DNA molecules being processed after the first round of PCR. Thus, the more tagged molecules are being processed, the higher the complexity becomes. If complexity is too high, individual tagged molecules will not be covered repeatedly, leading to a failure in assemble individual DNA molecules. On the other hand, if complexity is too low, sequencing capacity will be wasted due to redundant sequencing coverage of individual tagged DNA molecules being processed. Nonetheless, for quasispecies determination, it is more detrimental if the complexity is too high versus too low because excessive complexity will abolish the sequence assembly process. In general, the relationship between complexity and expected coverage for an individual viral sequence can be calculated with the expected sequencing output:

Coverage = (Sequencing output) / (Complexity x Length of region of interest).

In this formula, sequencing capacity and length of region of interest can be predetermined. Therefore, complexity is estimated solely based on the desired coverage of each tagged DNA molecules. For example, if the region of interest is 1 kb and 1 Gb of sequencing output is expected, then a complexity of 100,000 gives on average 10-fold coverage for individual tagged DNA molecules

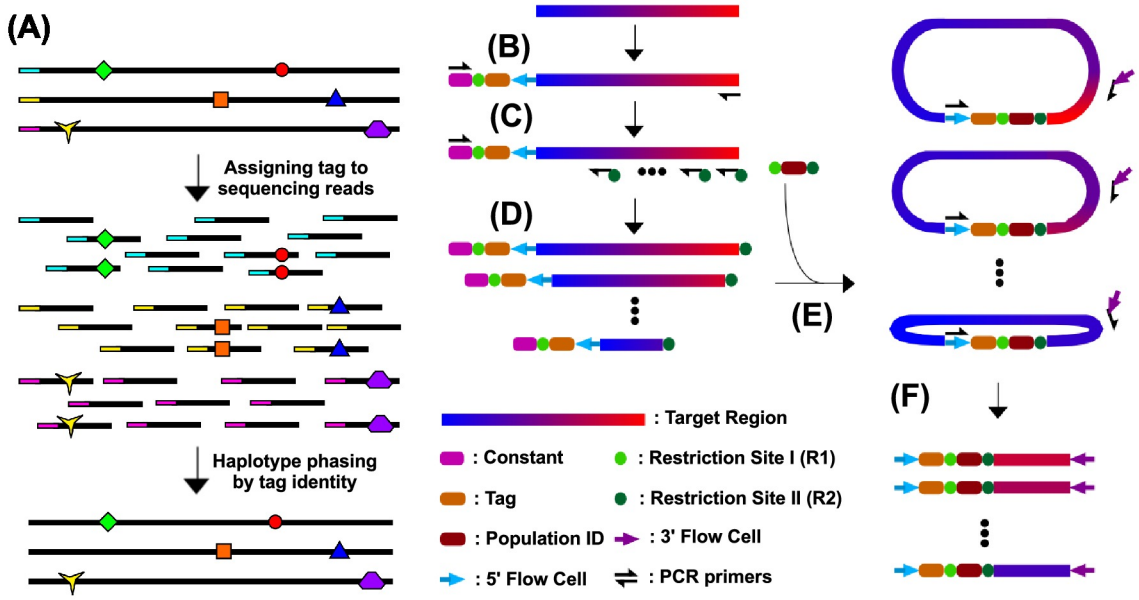


Figure 5-1. Schematic representation of the experimental design. (A) Individual viral sequences within a quasispecies are assigned with a unique tag. In this example, three viral sequences are present in the quasispecies. Horizontal black lines represent individual viral sequences. Red circle represents mutation from the consensus. Colored tag represents unique tag assigned to individual viral sequences. The same tag is distributed to every sequencing reads originated from the same viral sequence. During quasispecies reconstruction, the tag contains the phasing information to reconstruct individual viral sequences. (B) A cassette consisting of a constant region (Constant), a restriction site (R1), a random oligonucleotide (tag) and the forward Illumina adapter (5' FlowCell) is added to the 5' end of the DNA sample. Individual DNA molecules in the resultant pool will each be afforded with a unique tag. (C) The input pool in this PCR step contains a limited number of DNA templates to reduce the complexity of the pool. The resultant PCR amplification generates multiple copies of individually tagged DNA templates. (D) The DNA pool is then divided into a series of PCRs with a second restriction site (R2) added to the 3' end of the reaction product. (E) Ligation with population ID, which is a short specific DNA sequence serving as a barcode for multiplex sequencing, utilizes the two restriction sites, R1 and R2. (F) Amplicons with similar size are generated from different ligated DNA pools. Reverse Illumina adapter (3' FlowCell) is added. Different amplicon pools can then be mixed and subjected to Illumina sequencing.

being processed. With sufficient coverage for an individual viral sequence, we can distinguish sequencing error from true mutation as described previously [21], in addition to haplotype phasing. Therefore, complexity control represents a critical step in our experimental design.

After controlling the complexity, a PCR is performed to generate multiple copies of individually tagged DNA molecules (Fig. 5-1C). The resultant DNA pool is then divided into a series of PCRs to generate products with different lengths (Fig. 5-1D). For every pool, the resultant PCR products contain two different restriction sites on each ends. Next, restriction enzyme digestions generate two sticky ends and remove the constant region for PCR in the earlier step. A self-ligation step follows with the addition of a short insert (Fig. 5-1E). The short insert can serve as a barcode for multiplex sequencing. This ligation step circularizes the DNA, resulting in different sequence regions being proximal to the tag and further allowing linkage formation between any distal region with the tag - another key step in our experimental design. In the final step, a short amplicon ($\sim$200 bp) is recovered for NGS (Fig 5-1F). Each NGS read, from 5' to 3', will cover a tag for short read assembly within a quasispecies sample, a barcode for quasispecies sample identification, and a particular region of interest on the targeted viral sequence. NGS reads sharing the same tag belong to the same DNA molecules. Therefore, haplotypes of individual viral genomes within the quasispecies population can be interrogated.

Assembly of two HIV-1 viral quasispecies

Virus derived from a chronically infected HIV-1 patient was analyzed before (*in vivo*) and after (*ex vivo*) cell culture passaging for 10 weeks. *In vivo* virus sample represented the viral quasispecies within the HIV-1 infected patient. Whereas in *ex vivo* passaging, virus from the same patient was passaged serially in primary $CD4^+$ T lymphocytes from an HIV-1-uninfected donor and reflected the evolution of the viral quasispecies population in the absence of intra-patient selection pressure. We limited the complexity by processing roughly 300,000 viral sequences to ensure sufficient coverage ($\sim$50-fold) in all regions for any given viral sequence (Fig. 5-1B).

50-fold coverage = (18 Gb sequencing output) / (300,000 complexity x 1200 bp region of inter-

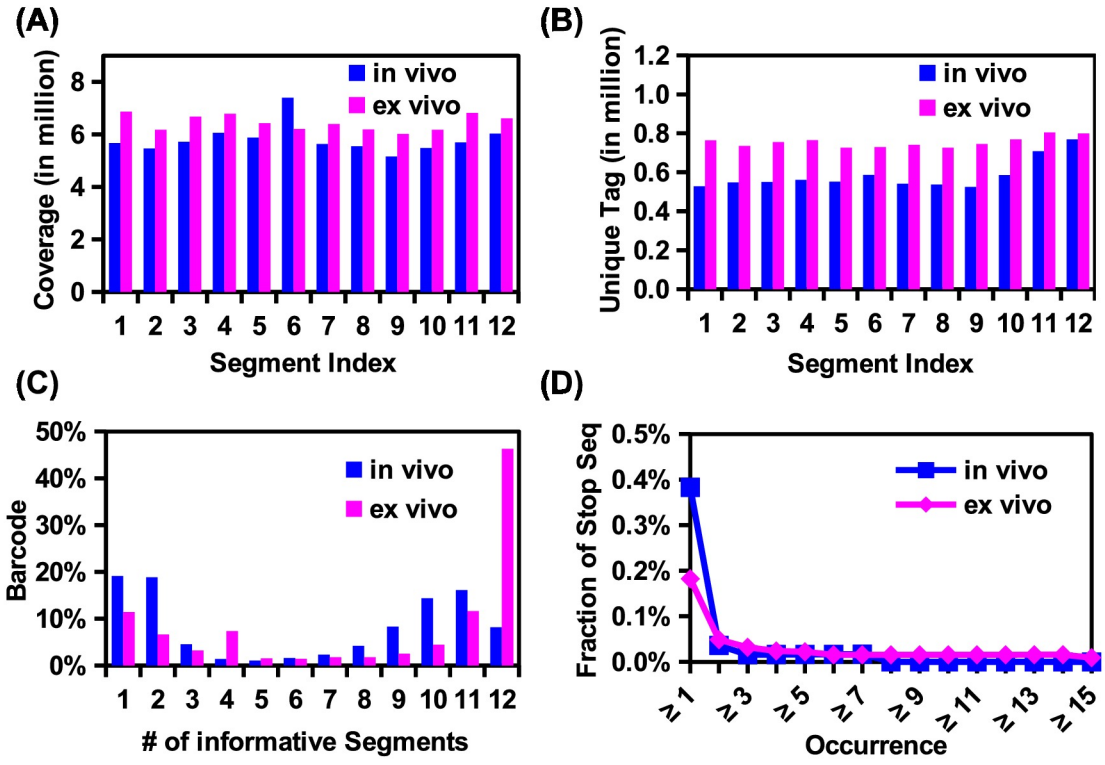


Figure 5-2. Proof-of-concept study using *ex vivo* passaged HIV-1 quasispecies. (A) Sequence coverage for each of the 12 amplicon segments is plotted. (B) Tag coverage for each of the 12 amplicon segments is plotted. Tag coverage is calculated by the number of unique tags present in a given amplicon segment. (C) The assembling successful rate is assessed by a parameter called the 'number of informative segments', which represents the number of amplicon segments a unique tag is present in. For example, in the present study, a tag with 12 informative segments represents a complete assembled contiguous sequence of our HIV-1 DNA target region, while a tag with 11 informative segments indicates that 1 amplicon segment is missing from the assembled contiguous sequence. All tags in the data set are categorized by this parameter. (D) The fraction of sequences containing a stop codon is plotted against different cutoffs for the minimum sequence occurrence.

est length).

Twelve non-overlapping amplicons, which cover a 1,295 nucleotide stretch and encompass most of the *gag* and a portion of the *pol* genes of the HIV-1 genome, were prepared. Sequencing was performed using an Illumina HiSeq 2000 machine. Sequencing coverages in different regions were similar (Fig. 5-2A). The numbers of unique tags in different regions were also comparable (Fig. 5-2B). The absence of apparent coverage bias confirmed the quality of sequencing library preparation. For each region, tags with fewer than three occurrences were filtered and removed to adequately apply the error correction algorithm. This filter eliminated 35-57% of tags depending on region. For a complete viral sequence to be assembled, sequences of all 12 amplicon regions sharing the same tag had to be available. We successfully assembled 54,583 viral sequences in the *in vivo* viral quasispecies and 228,936 viral sequences in the *ex vivo* quasispecies, thus validating the complexity control procedure. However, about $\sim$30-40$\%$ of the tags were present in only one or two regions, which we attributed to PCR or sequencing errors at the tag region (Fig. 5-2C).

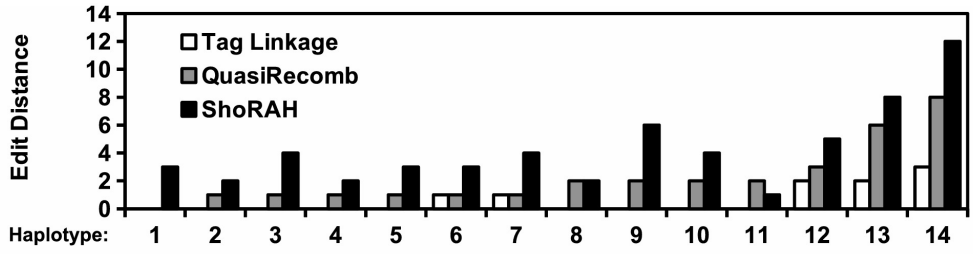
To further evaluate the data quality, the appearance of stop codons in *gag* was examined. Given that viable virus requires translation of a full length Gag polyprotein, stop codons would likely represent PCR errors. While $\sim$0.4$\%$ (*in vivo*) and $\sim$0.2$\%$ (*ex vivo*) of the assembled sequences contained a stop codon, this number dropped dramatically ($<$0.05$\%$) after we filtered-out the sequences with just one occurrence (Fig. 5-2D). Further increasing the cutoff stringency, however, did not significantly suppress the stop codon occurrence frequency. These rare viral sequences were likely to be non-functional virus within the viral quasispecies population generated by hypermutation [22–24]. 47,083 assembled viral sequences from the *in vivo* viral quasispecies and 223,966 assembled viral sequences from the *ex vivo* viral quasispecies passed this quality filter, yielding 2,672 and 1,983 unique viral sequences, respectively. The number of unique viral sequences we successfully assembled represented a $>$ 20 fold increased as compared to that of the previously reported algorithm-based quasispecies assembly method [11–17]. Additionally, the detection limits of rare viral sequences in this study ($\sim$0.005$\%$ and $\sim$0.001$\%$ for the *in vivo* and *ex vivo* viral quasispecies, respectively) also significantly exceeded that reported for the algorithm-

based technique, which was reported to be $\sim$0.1% to $\sim$1% [15–18].

Comparison with algorithmic-based approach

To the best of our knowledge, the existing quasispecies reconstruction approaches are algorithmic-based inference methods. In contrast, tag linkage approach is a molecular-based, direct interrogation method. It is devoid of any inference error that is intrinsic to algorithmic-based approach. Consequently, it enables a much higher accuracy in quasispecies reconstruction than conventional algorithmic-based approach. We compared the performance of our tag linkage method with two algorithmic-based approaches: 1) the state-of-the-art ShoRAH tool [13], and 2) a recently published approach, QuasiRecomb [12], which takes natural recombination event into account. To implement the algorithmic-based approaches, single-read DNA sequencing library of the *in vivo* quasispecies sample was prepared by standard DNA fragmentation. We also employed the tagging strategy here to distinguish true mutations from sequencing error as previously described (see materials and methods) [21]. As a result, quasispecies reconstructions by ShoRAH and QuasiRecomb were minimally confounded by sequencing error. To provide a reference for comparison, we conducted traditional clonal sequencing for the *in vivo* quasispecies population. In this experiment, a 1106 bp region in the *gag* gene was considered. A total of 20 randomly selected clones were sequenced, which represented 14 different haplotypes.

ShoRAH reconstructed 252 viral sequences from the *in vivo* quasispecies sample. However, none of the 14 haplotypes were being reconstructed (Fig. 5-3). For those 14 haplotypes, the respective closest viral sequence deduced by ShoRAH had an edit distance ranging from 1 to 12. QuasiRecomb, on the other hand, reconstructed 1343 viral sequence and was able to identify 1 out of 14 haplotypes from clonal sequencing. This haplotype had an estimated occurrence frequency of 0.8% from QuasiRecomb while it accounted for 7 out of 20 clones in clonal sequencing. It implied that haplotype frequency estimation by QuasiRecomb was inaccurate and that a significant amount of reconstructed haplotype by QuasiRecomb was false positive. QuasiRecomb can also be run in a conservative mode, in which only major haplotypes were reconstructed. Under this running mode, only 6 haplotypes were reconstructed and none of them overlapped with the 14



# of clone	7	1	1	1	1	1	1	1	1	1	1	1	1	1
Tag Linkage	15%	0.02%	0.02%	0.004%	0.002%	4%	0.03%	0.02%	0.009%	0.007%	0.005%	0.002%	0.002%	0.004%
QuasiRecomb	0.8%	0.02%	1%	0.1%	0.02%	0.7%	0.1%	0.01%	0.1%	0.01%	0.02%	0.02%	0.02%	0.2%
ShoRAH	10%	0.6%	7%	10%	0.6%	10%	10%	0.6%	0.2%	0.6%	0.6%	0.6%	7%	7%

Figure 5-3. Comparison of performance in quasispecies reconstruction with ShoRAH and QuasiRecomb. A total of 20 randomly selected clones from the *in vivo* quasispecies population were sequenced using traditional clonal sequencing. They represented 14 different haplotypes. Their edit distances with the corresponding closest reconstructed viral sequences are shown. Edit distance represents the minimum number of substitutions required to change one nucleotide sequence into the other. The estimated fractions of closest reconstructed viral sequences for those 14 haplotypes are displayed in the bottom. The estimated haplotype frequency is displayed in a gray background if there is a complete match (edit distance = 0).

haplotypes being clonal sequenced.

In contrast, 9 out of 14 haplotypes from clonal sequencing were included in the quasispecies reconstructed by our tag linkage approach (Fig. 5-3). The most abundant haplotype from clonal sequencing matched the most abundant reconstructed haplotype from tag linkage approach in this region of interest. The other 8 identified haplotypes were estimated to have an occurrence frequency from 0.002% to 0.02%. It highlighted the sensitivity and accuracy of our tag linkage approach in reconstructing rare haplotypes. The missing five haplotypes were 1-3 edit distances away from their respective closest viral sequence in the quasispecies reconstructed by our tag linkage approach. Overall, tag linkage approach achieved a significant improvement over algorithmic-based approaches in quasispecies reconstruction, both qualitatively and quantitatively.

Diversity comparison between *in vivo* and *ex vivo* HIV-1 quasispecies

We next examined the sequence diversity in both *in vivo* and *ex vivo* quasispecies populations. The most frequent viral sequence represented 8.1% of the *in vivo* viral quasispecies, whereas the most dominant viral sequence represented 32.5% of the *ex vivo* viral quasispecies (Fig. 5-4A). The two most dominant viral sequences in the *ex vivo* sample comprised more than half of the total viral quasispecies while the *in vivo* viral quasispecies was much more diverse. At the amino acid level, 80% of the *in vivo* viral quasispecies were represented by four protein sequences, with a total of 42 unique protein sequences in the population (Fig. 5-4B). In contrast, while only two protein sequences represented 80% of the *ex vivo* viral quasispecies, there were 201 unique protein sequences. A phylogenetic tree analysis demonstrated the effect of differential selection pressures on viral quasispecies evolution from *in vivo* to *ex vivo*, in which two distinct sub-population clusters could be observed (Fig. 5-4C-D).

Recombination pattern of HIV-1 quasispecies

HIV-1, as a diploid retrovirus, is capable of generating recombinant proviral transcript via a template switching event during the reverse transcription step in the viral replication. It facilitates further diversification for adaptation [2]. The depth and comprehensiveness of our data permit an

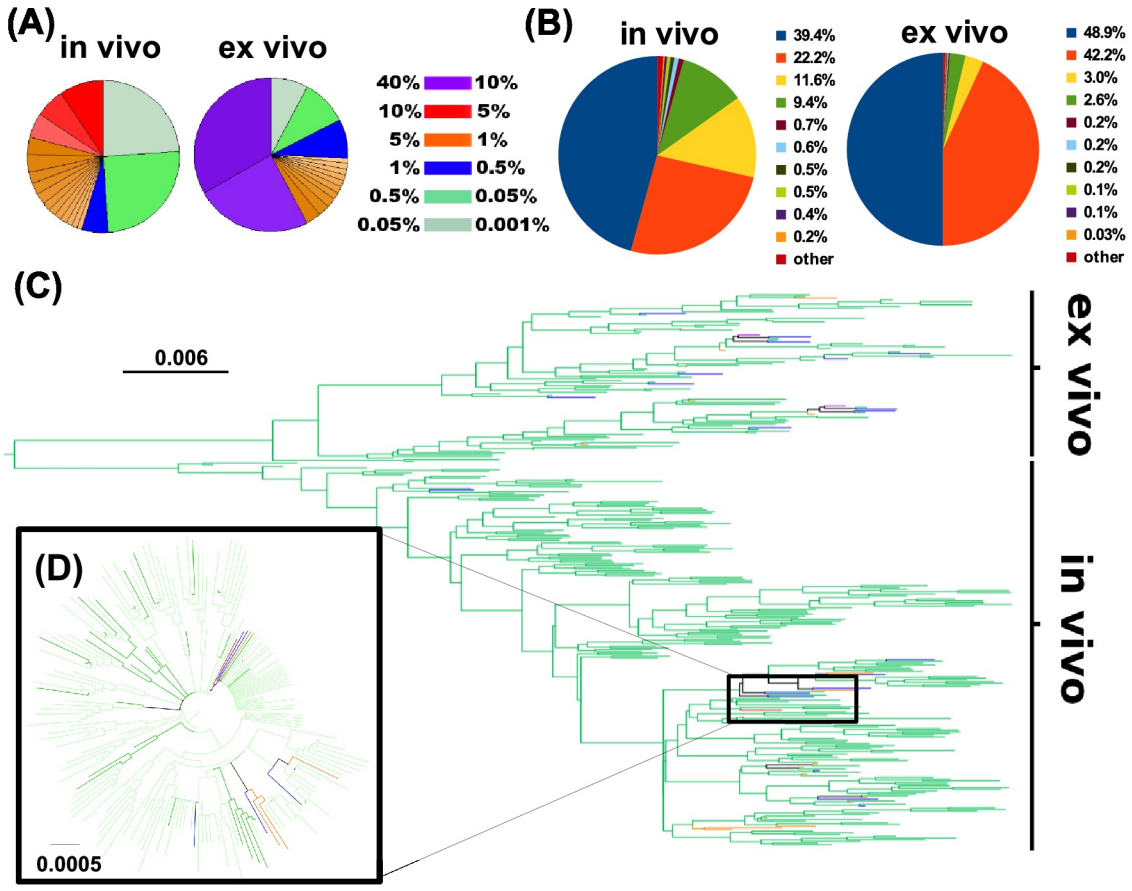


Figure 5-4. Diversity of *in vivo* and *ex vivo* HIV-1 quasispecies. (A) The diversities for both the *in vivo* and *ex vivo* viral quasispecies at the nucleotide level are reflected by the pie chart. The fractions of viral quasispecies for the 10 highest frequency occurring viral sequences are shown. Each color code indicates a range of occurrence frequency as indicated. (B) The diversities for both *in vivo* and *ex vivo* viral quasispecies content on the protein level (aa. 139 to 507 on *gag* protein) are reflected by the pie chart. The fractions of viral quasispecies for the 10 highest frequency occurring viral sequences are shown. (C) A neighbor-joining phylogenetic tree depicting viral nucleotide sequences with occurrence frequencies above 0.05%. The occurrence frequency for each individual node is color coded as described in Fig. 5-4A. (D) A small segment in the phylogenetic tree from Fig. 5-4C is selected. This segment is replotted along with viral sequences with occurrence frequency form 0.001% to 0.05%, which are not included in Fig. 5-4C.

investigation of this viral recombination, as a linkage disequilibrium pattern. Here, we employed the r^2 correlation to measure linkage disequilibrium. r^2 was computed between 38 SNPs that had an occurrence frequency above 0.1% in either the *in vivo* or *ex vivo* viral quasispecies (Fig. 5-5). Several strong correlations ($r^2 > 0.5$) were observed in both the *in vivo* and *ex vivo* viral quasispecies. Nonetheless, the linkage disequilibrium was more pervasive and spanned a larger region in the *in vivo* viral quasispecies than in the *ex vivo* viral quasispecies. From the *in vivo* viral quasispecies, we observed two linkage disequilibrium blocks, a ~200 nucleotide block from position 900 to 1100 and another from nucleotide position 1400 to 1600. The presence of two closely spaced recombination nucleotide blocks suggests that there is a recombination hotspot between position 1100 to 1400, which is located at the p24 region of the gag gene. Another possibility is that certain haplotypes provided a fitness advantage and were positively selected. Further characterization would be needed to dissect the underlying mechanism.

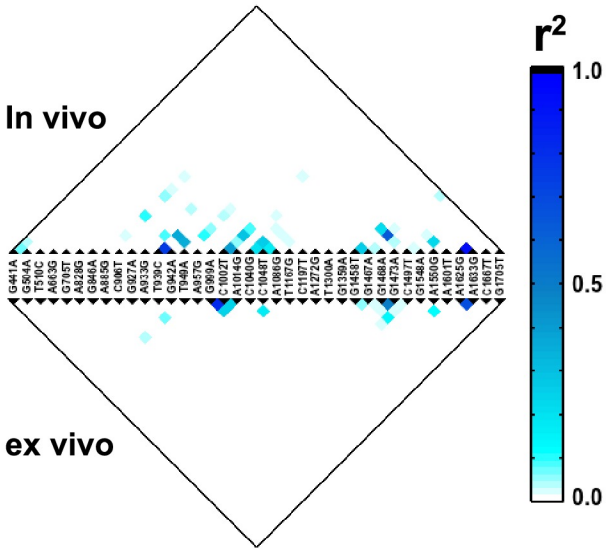


Figure 5-5. Linkage disequilibrium for *in vivo* and *ex vivo* HIV-1 quasispecies. The linkage disequilibrium was measured using r^2 and computed between SNPs that had an occurrence frequency above 0.1% in either the *in vivo* and *ex vivo* viral quasispecies. The stronger the association between two SNPs, the larger value the r^2 is.

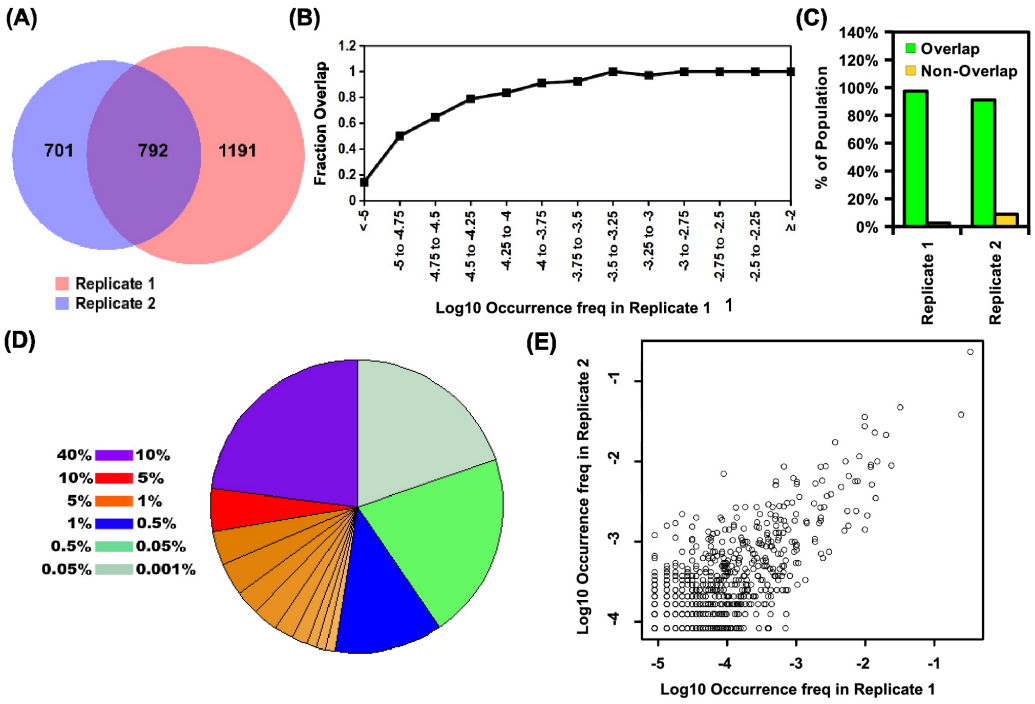


Figure 5-6. Technical Replicate for *ex vivo* viral quasispecies assembly. (A) The reproducibility was assessed by Venn diagrams of the unique viral sequences in both replicates of the *ex vivo* viral quasispecies reassembly. (B) Unique viral sequences were binned into a bin size of 0.25 at log10 scale. The overlapping faction that was covered by replicate 2 were plotted against different bins. (C) The fractions of viral quasispecies population content in replicate 1 that was covered by replicate 2 (Replicate 1) and in replicate 2 that was covered by replicate 1 (Replicate 2) are plotted as a bar chart. (D) The diversity for replicate 2 at the nucleotide level is reflected by the pie chart. The fractions of viral quasispecies for the 10 highest frequency occurring viral sequences are shown. Each color code indicates a range of occurrence frequency as indicated. (E) The occurrence frequency of individual viral sequences was compared between replicate 1 and replicate 2 at $\log_{10}$ scale. There exists a Pearson correlation of 0.87 at normal scale between two replicates.

Reproducibility from a technical replicate

To assess the reproducibility, a technical replicate was performed for the *ex vivo* viral quasispecies population (Fig. 5-1C-E). The technical replicate was repeated for all steps beginning at the stage of generating amplicons of varying length (Fig. 5-1C) - a key step of our approach. Majority of the viral sequences in the replicate (replicate 2) overlapped with the original data set (replicate 1) (Fig. 5-6A). However, a significant fraction of viral sequences was covered by only one of the replicates, but those represented a small fraction, $\sim$3$\%$ to 9$\%$, of the viral quasispecies (Fig. 5-6B-C). Viral sequences that were observed in only one of the two replicates typically had an occurrence frequency $<$ 0.01$\%$ (Fig. 5-6B). It suggests that the difference between replicates was due to sampling limit, where viral sequences with a low occurrence were more likely to be unsampled by one of the replicates. Replicate 2 covered 97$\%$ of the viral quasispecies in the first replicate, whereas replicate 1 covered 91$\%$ of viral quasispecies in the second replicate (Fig. 5-6C). The genetic composition of the viral quasispecies reconstructed from replicate 2 was comparable to that of replicate 1 (Fig. 5-4A and 5-6D). Occurrence frequency for individual viral sequences exhibited a correlation of 0.87 (Pearson correlation at normal scale) between replicates (Fig. 5-6E). These results provided further validation of the tag linkage technique in both qualitative and quantitative manners.

5.4 DISCUSSION

With the advancement of sequencing technology, NGS continues to increase read length and throughput. Nonetheless, the trade-off between read length and throughput still exists [25]. Sequencing platforms with long reads such as Pacific Bio and 454 pyrosequencing have a relatively low throughput. NGS machines with higher throughput such as Illumina and SOLiD do not afford long reads. Despite currently having the highest throughput, the short read length of Illumina creates a challenge in assembling reads into continuous long sequences.

This study describes an amplicon-based tag linkage approach to characterize viral quasispecies population structures and provides a proof-of-concept example showing a very high detection sensitivity. Unlike algorithm-based approaches, the accuracy of our amplicon-based molecular tag approach is independent of viral quasispecies population diversity. In addition, it incorporates an error correction step to identify NGS platform errors, resulting in a dramatic increase in the sensitivity to detect rare haplotypes [21]. Algorithm-based approach for viral quasispecies reconstruction can usually handle 10 to 100 viral sequences at various statistical confidence. In contrast, our tag linkage approach can reconstruct close to 1000 sequences with high confidence as indicated by our replicates. It achieves a significant improvement in accuracy and sensitivity from the algorithm-base approach [11–20].

The major limitation in our approach is the length of deduced sequence, which is restricted by the upper limit of PCR (typically 10 kilobases). Another potential pitfall is PCR recombination. In our protocol, we tried to minimize this artifact by using a high processivity and fidelity DNA polymerase for PCR [26]. In addition, a long PCR extension time was used to ensure extension completion of the amplicon to minimize PCR recombination [27]. Our technical replicate control shows that a majority of the viral quasispecies population content ($>90\%$) are captured in both repetitions, including rare variants, indicating that any artifact by PCR recombination is minimal. Additionally, the high correlation of occurrence frequency for individual viral sequences between each replicate confirms reproducibility. Overall, our control experiments and concurrent analysis validate the amplicon-based tag linkage approach as a highly sensitive methodology for viral qua-

sispecies assembly.

By reconstructing individual sequences within the viral quasispecies, we are able to detect linkage disequilibrium throughout the region of interest. Genome recombination is a frequent process occurring intra-patient for diversification and adaptation [28–32]. Recombinant generation is a non-random process as recombination coldspots and hotspots have been reported in HIV-1 [33–36]. In this study, we observed a more pervasive linkage disequilibrium in the *in vivo* viral quasispecies compared to that of the *ex vivo*, suggesting that there may be genetic interactions within the linkage disequilibrium block that are important for chronic infection. Alternatively, this observation may also be attributed to a higher recombination frequency during *ex vivo* passaging due to an increase in co-infection occurrence. We demonstrate the power of our tag linkage approach in capturing linkage disequilibrium in a viral quasispecies, which can be further utilized to examine genetic interactions and to identify functional residues.

Our technique provides a sensitive and accurate tool to study the evolutionary trajectory of viral quasispecies. It permits the monitoring of a multi-drug resistance (MDR) viral sequence and epistasis within viral quasispecies - an important factor in viral evolution and adaptation [37,38]. Highly active antiretroviral therapy (HAART) therapy is a common treatment to suppress HIV progression by utilizing a drug cocktail designed to target viral proteins at multiple essential stages of the viral life cycle. However, viral rebound can be caused by MDR HIV with extremely low occurrence frequency [9,38–40]. In addition, as most drug resistant mutations compromise viral fitness, drug resistant viruses often carry additional mutations to compensate for this fitness cost [6,10,41–43]. The tag linkage approach provides an important tool to survey the genetic makeup of viral quasispecies and to estimate the risk of viral rebound and virulence by surveillance of pair-wise or even higher-order genetic interactions between mutations.

Although this study is based on HIV quasispecies samples, tag linkage approach is not limited to HIV and can potentially be applied to other viral quasispecies, such as hepatitis B virus (HBV), hepatitis C virus (HCV) and influenza virus. For example, tag linkage approach can be applied to study multi-drug resistance that are also found in naturally occurring HBV as in the case of

HIV [44, 45]. This technique is also suitable for studying *cis*-elements that are prevalent in HCV due to its intrinsic replication property [46]. In addition, tag linkage approach can be utilized to examine permissive and compensatory mutations that are shown to be important in the evolution of influenza virus [47–49].

This technique can also be extended beyond the monitoring of viral quasispecies. One application is to examine the dynamics of $CD4^+$ and $CD8^+$ cells in the immune system during viral infection. They have an active role in virus detection and clearance during both acute and chronic infection. During the establishment of persistent viral infections, the immune system co-evolves with the virus [50]. A complex dynamic occurs between the heterogeneous immune populations and the evolving viral quasispecies. The medical significance of this virus-host dynamic is highlighted by a recent study describing the rise of a broadly neutralizing HIV-1 antibody from co-evolution with acute phase virus [51]. The methodology we describe here offers the research community an approach to understand the dynamic interplay between the host and virus in exquisite detail at the population level.

5.5 MATERIALS AND METHODS

Ethics statement

The study was approved by UCLA IRB. A chronically-infected HIV-1 patient without undergoing antiretroviral therapy was recruited from the Los Angeles area and provided written informed consent.

Subjects and specimen collection

Total peripheral blood mononuclear cells (PBMCs) were isolated from the patient's whole blood sample by standard Ficoll gradient. The plasma viral load at the time of collection was 130,234 viral copies/ml.

Recovery of virus from PBMCs and virus passaging

Ex vivo passaging was conducted as previously described [52]. Briefly, virus was passaged serially in primary $CD4^+$ T lymphocytes from an HIV-1-uninfected donor [53]. After each passage of $\sim$7 days, supernatant virus was collected, titered, and used to infect fresh cells with an MOI of 1.

DNA library preparation for tag linkage assembly

To extract the viral genomic DNA, cell pellets of 200,000 cells were resuspended in PBS and genomic DNA was extracted using the DNeasy Tissue DNA Isolation Kit (Qiagen). DNA was recovered by PCR using the primer set: 5'-GCG GAG GCT AGA AGG AGA GAG ATG G-3' and 5'-CAT CAC CTG CCA TCT GTT TTC CAT A-3'. The forward Illumina sequencing priming site was added to the 5' end of the DNA sample by PCR using the primer set: 5'-AGA TCG GAA GAG CGT CGT GTA GGG GCG GAG GCT AGA AGG AGA GAG ATG-3' and 5'- GTT TAA CTT TTG GGC CAT CCA TTC CTG GC-3'. Then, the constant region, a NotI restriction enzyme site and a 13 nucleotide tag of random 'N' sequence was added to the 5' end of the DNA sample by another PCR using the primer set: 5'-ACA TAG ATA CTA TGC GGC CGC NNN NNN NNN NNN NAG ATC GGA AGA GCG TCG TGT AGG G-3' and 5'- GTT TAA CTT TTG GGC CAT CCA TTC CTG GC-3'. The concentration of the tagged DNA sample was measured using NanoDrop

1000 spectrophotometer (Thermo Fisher Scientific). This concentration was used as a reference to calculate the dilution-fold in the subsequent complexity control step. In the complexity control step, $\sim$300,000 copies of tagged DNA sample were used as the input for PCR using the primer set: 5'-CAC ATA GAT ACT ATG CGG CCG C-3' and 5'-GTT TAA CTT TTG GGC CAT CCA TTC CTG GC-3'. This complexity was calculated based on a $\sim$50-fold coverage for individual viral sequence with 30 Gb expected sequencing output per viral quasispecies sample. This was followed by 12 PCR using the product of the complexity control step as template. Consecutive PCR pools should have a different product size approximately corresponding to the sequencing read length minus 80 bp. From this step forward, the 12 pools were processed independently until sample combination at the high-throughput sequencing step. The products were then subjected to double digestion by NotI and XhoI. NotI and XhoI were chosen because they were not present in the consensus sequence of the target DNA template region. A small insert, which could serve as the population ID, was prepared by annealing 5'-GGC CCG ACG TAA CGA T-3' and 5'-TCG AAT CGT TAC GTC G-3', each with a phosphate group attached at the 5' end. One unit of T4 DNA ligase (Life Technolgies) was used in each ligation reaction. The reaction condition followed manufacturer's instructions. All ligations were performed overnight at $20\,^{\circ}\mathrm{C}$ in 100 uL total reaction volume. The ligated products were used as the templates for PCR to add the 5' flow cell adapters and the reverse read Illumina sequencing priming site. The 3' Illumina flow cell adapters were then added by PCR using the primer set: 5'-AAT GAT ACG GCG ACC ACC G-3' and 5'- CAA GCA GAA GAC GGC ATA CGA GAT CGG TCT CGG CAT TCC TGC TGA ACC GCT CTT CCG-3'. The resultant amplicons from all 12 pools were then mixed. High-throughput sequencing was done by an Illumina HiSeq 2000 machine with an equivalent of 0.75 lane per sample and 2 x 100 bp paired-end reads. All PCRs in this study were performed using KOD DNA polymerase with 1.5 mM $MgSO_4$, 0.2 mM of each dNTP (dATP, dCTP, dGTP, and dTTP) and 0.4 uM of forward and reverse primer. PCR extensions were performed with 50 seconds per kb at $68\,^{\circ}\mathrm{C}$. Annealing temperature for a given PCR was $5\,^{\circ}\mathrm{C}$ below the lowest melting temperature of the pair of primers. All primers in this study were designed to target conserved regions within the quasispecies which were determined by clonal sequencing of the sampled viral sequences. This sequencing library preparation could potentially be adapted to study viral RNA using a reverse transcription primer tag as decribed by Jabara et al [54]. Raw sequencing data have been submitted to the NIH Short

Read Archive under accesion number: SRP032753.

Clonal sequencing

After recovering the DNA by PCR as described above, the amplicon was inserted into target p83-2 plasmid using In-Fusion kit (Clontech). Twenty clones were randomly selected and subjected to capillary sequencing (Laragen).

Data analysis

Sequencing reads were mapped by BWA with 8 mismatches allowed [55]. Pair-end reads containing two or more short inserts (barcodes) were discarded. Error-correction was performed as described previously to distinguish true mutation from sequencing error [21]. The error-correction step grouped all reads sharing the same tag and mapped to the same region into a read cluster that was further conflated into a "error-free" read. As described in Kinde et al. [21], most reads sharing the same tag should share the mutation pattern during mapping. In contrast, a sequencing error would have a low occurrence frequency within a read cluster and could be distinguished from true mutations. Through this process, sequencing error would be corrected to generate an "error-free" read. Read cluster with a size of $<$3 reads were discarded to increase the confidence in generating an "error-free" read. Since intermolecular concatenation at the ligation was observed, a mutation that existed in 45$\%$ of the reads within a conflated read cluster that also shared the same tag was considered as a true mutation. The correlation between technical replicates indicated that intermolecular concatenation did not pose a major barrier in the accuracy of viral quasispecies assembly. Nonetheless, further application should adjust the ligation reaction volume to decrease the intermolecular concatenation during ligation (circularization step). Next, "error-free" reads that shared the same tag were assembled into a contiguous sequence, which represented a single viral sequence. Data processing and analysis were conducted by custom Python scripts. All scripts are available upon request.

Phylogenetic tree construction

ClustalX was used to create the neighbor-joining phylogenetic tree [56]. The phylogenetic tree was mid point-rooted and displayed by FigTree.

Linkage disequilibrium

We used the r^2 correlation to quantify linkage disequilibrium between two SNPs. r^2 was computed as per convention. Briefly, $r^2 = (P_{AB} - P_A \times P_B)^2/(P_A \times P_B \times (1 - P_A) \times (1 - P_B))$, where P_{AB} represented the occurrence frequency of viral sequences that carry both SNP A and SNP B; P_A represented the occurrence frequency of viral sequences that carry SNP A; P_B represented the occurrence frequency of viral sequences that carry SNP B.

DNA library preparation for error-free sequencing

Gag-pol region was PCR amplified using the primer set: 5'- GAC TAG CGG AGG CTA GAA GGA GAG AG-3' and 5'-CAT GTT CTT CTT GGG CCT TAT CTA TTC-3'. The resultant DNA product was sheared to around 200 bp to 600 bp by sonication using the Sonic Dismembrator Model 100 (Fisher Scientific). Dismembrator was set to power level four and samples were pulsed three times for 10 seconds. Samples were kept on ice for 45 seconds in between pulses. End repair and 3' dA-tailing were performed respectively by end repair module and dA-tailing module (New England BioLabs). The DNA product was then ligated to an Y-shape adaptor carrying a nine-nucleotide tag of random 'N' sequence. As a result, each ligated product contained an 18-nucleotide tag, nine from each of the 5' and 3' end. Y-shape adaptor was prepared by annealing two oligonucleotides: 5'-CGC GTA TCC ATG GCA NNN NNN NNN GCC AGA TCG GAA GAG CGG TTC AGC AGG AAT GCC GAG-3' and 5'-ACA CTC TTT CCC TAC ACG ACG CTC TTC CGA TCT GGC-3'. Then, the annealed product was treated with Klenow Fragment (New England BioLabs) and digested with BciVI. An estimated copy of around 10 millions of ligated products were amplified by primer set: 5'-AAT GAT ACG GCG ACC ACC GAG ATC TAC ACT CTT TCC CTA CAC GAC GCT CTT CCG-3' and 5'-CAA GCA GAA GAC GGC ATA CGA GAT CGG TCT CGG CAT TCC TGC TGA ACC GCT CTT CCG-3'. The resultant DNA product was submitted for 2 x 100 bp paired-end sequencing on one lane of Illumina HiSeq 2500 machine.

Quasispecies reconstruction by ShoRAH and QuasiRecomb

"Error-free" reads were generated as described above. Here, a mutation that existed in 95% of the reads within a conflated read cluster that also shared the same tag was considered as a true mutation. Reads were mapped by BWA with 8 mismatches allowed [55]. All reads were treated as single end read. "Error-free" mapped reads were processed by ShoRAH version 0.6 with a window size of 40, a window shift of 1 and default settings for other parameters [13]. Quasispecies reconstruction by QuasiRecomb was performed by default setting [12]. Due to the huge memory requirement of QuasiRecomb, 500,000 mapped reads were randomly sampled and processed. Further increase the number of input reads generated memory error. To limit the false positive rate, a refinement reconstruction was performed using '-refine' option. We employed '-conservative' option for high confidence haplotype reconstruction to identify major haplotypes.

5.6 BIBLIOGRAPHY

1. Wilke CO, Wang JL, Ofria C, Lenski RE, Adami C (2001) Evolution of digital organisms at high mutation rates leads to survival of the flattest. Nature 412: 331–333.

2. Worobey M, Holmes EC (1999) Evolutionary aspects of recombination in rna viruses. J Gen Virol 80 (Pt 10): 2535–2543.

3. Bonhoeffer S, Chappey C, Parkin NT, Whitcomb JM, Petropoulos CJ (2004) Evidence for positive epistasis in hiv-1. Science 306: 1547–1550.

4. da Silva J, Coetzer M, Nedellec R, Pastore C, Mosier DE (2010) Fitness epistasis and constraints on adaptation in a human immunodeficiency virus type 1 protein region. Genetics 185: 293–303.

5. Brockman MA, Schneidewind A, Lahaie M, Schmidt A, Miura T, et al. (2007) Escape and compensation from early hla-b57-mediated cytotoxic t-lymphocyte pressure on human immunodeficiency virus type 1 gag alter capsid interactions with cyclophilin a. J Virol 81: 12608–12618.

6. Dam E, Quercia R, Glass B, Descamps D, Launay O, et al. (2009) Gag mutations strongly contribute to hiv-1 resistance to protease inhibitors in highly drug-experienced patients besides compensating for fitness loss. PLoS Pathog 5: e1000345.

7. Zhang J, Hou T, Wang W, Liu JS (2010) Detecting and understanding combinatorial mutation patterns responsible for hiv drug resistance. Proc Natl Acad Sci U S A 107: 1321–1326.

8. Sanjun R, Moya A, Elena SF (2004) The contribution of epistasis to the architecture of fitness in an rna virus. Proc Natl Acad Sci U S A 101: 15376–15379.

9. Fumero E, Podzamczer D (2003) New patterns of hiv-1 resistance during haart. Clin Microbiol Infect 9: 1077–1084.

10. Verheyen J, Litau E, Sing T, Dumer M, Balduin M, et al. (2006) Compensatory mutations at the hiv cleavage sites p7/p1 and p1/p6-gag in therapy-naive and therapy-experienced patients. Antivir Ther 11: 879–887.

11. Beerenwinkel N, Gnthard HF, Roth V, Metzner KJ (2012) Challenges and opportunities in estimating viral genetic diversity from next-generation sequencing data. Front Microbiol 3: 329.

12. Tpfer A, Zagordi O, Prabhakaran S, Roth V, Halperin E, et al. (2013) Probabilistic inference of viral quasispecies subject to recombination. J Comput Biol 20: 113–123.

13. Zagordi O, Bhattacharya A, Eriksson N, Beerenwinkel N (2011) Shorah: estimating the genetic diversity of a mixed sample from next-generation sequencing data. BMC Bioinformatics 12: 119.

14. Zagordi O, Dumer M, Beisel C, Beerenwinkel N (2012) Read length versus depth of coverage for viral quasispecies reconstruction. PLoS One 7: e47046.

15. Eriksson N, Pachter L, Mitsuya Y, Rhee SY, Wang C, et al. (2008) Viral population estimation using pyrosequencing. PLoS Comput Biol 4: e1000074.

16. Zagordi O, Klein R, Dumer M, Beerenwinkel N (2010) Error correction of next-generation sequencing data and reliable estimation of hiv quasispecies. Nucleic Acids Res 38: 7400–7409.

17. Astrovskaya I, Tork B, Mangul S, Westbrooks K, Mndoiu I, et al. (2011) Inferring viral quasispecies spectra from 454 pyrosequencing reads. BMC Bioinformatics 12 Suppl 6: S1.

18. Henn MR, Boutwell CL, Charlebois P, Lennon NJ, Power KA, et al. (2012) Whole genome deep sequencing of hiv-1 reveals the impact of early minor variants upon immune recognition during acute infection. PLoS Pathog 8: e1002529.

19. Prosperi MCF, Salemi M (2012) Qure: software for viral quasispecies reconstruction from next-generation sequencing data. Bioinformatics 28: 132–133.

20. Skums P, Mancuso N, Artyomenko A, Tork B, Mandoiu I, et al. (2013) Reconstruction of viral population structure from next-generation sequencing data using multicommodity flows. BMC Bioinformatics 14 Suppl 9: S2.

21. Kinde I, Wu J, Papadopoulos N, Kinzler KW, Vogelstein B (2011) Detection and quantification of rare mutations with massively parallel sequencing. Proc Natl Acad Sci U S A 108: 9530–9535.

22. Rose PP, Korber BT (2000) Detecting hypermutations in viral sequences with an emphasis on g –¿ a hypermutation. Bioinformatics 16: 400–401.

23. Janini M, Rogers M, Birx DR, McCutchan FE (2001) Human immunodeficiency virus type 1 dna sequences genetically damaged by hypermutation are often abundant in patient peripheral blood mononuclear cells and may be generated during near-simultaneous infection and activation of cd4(+) t cells. J Virol 75: 7973–7986.

24. Harris RS, Liddament MT (2004) Retroviral restriction by apobec proteins. Nat Rev Immunol 4: 868–877.

25. Loman NJ, Misra RV, Dallman TJ, Constantinidou C, Gharbia SE, et al. (2012) Performance comparison of benchtop high-throughput sequencing platforms. Nat Biotechnol 30: 434–439.

26. Lahr DJG, Katz LA (2009) Reducing the impact of pcr-mediated recombination in molecular evolution and environmental studies using a new-generation high-fidelity dna polymerase. Biotechniques 47: 857–866.

27. Judo MS, Wedel AB, Wilson C (1998) Stimulation and suppression of pcr-mediated recombination. Nucleic Acids Res 26: 1819–1825.

28. Jetzt AE, Yu H, Klarmann GJ, Ron Y, Preston BD, et al. (2000) High rate of recombination throughout the human immunodeficiency virus type 1 genome. J Virol 74: 1234–1240.

29. Froissart R, Roze D, Uzest M, Galibert L, Blanc S, et al. (2005) Recombination every day: abundant recombination in a virus during a single multi-cellular host infection. PLoS Biol 3: e89.

30. Neher RA, Leitner T (2010) Recombination rate and selection strength in hiv intra-patient evolution. PLoS Comput Biol 6: e1000660.

31. Charpentier C, Nora T, Tenaillon O, Clavel F, Hance AJ (2006) Extensive recombination among human immunodeficiency virus type 1 quasispecies makes an important contribution to viral diversity in individual patients. J Virol 80: 2472–2482.

32. Shriner D, Rodrigo AG, Nickle DC, Mullins JI (2004) Pervasive genomic recombination of hiv-1 in vivo. Genetics 167: 1573–1583.

33. Zhuang J, Jetzt AE, Sun G, Yu H, Klarmann G, et al. (2002) Human immunodeficiency virus type 1 recombination: rate, fidelity, and putative hot spots. J Virol 76: 11273–11282.

34. Baird HA, Gao Y, Galetto R, Lalonde M, Anthony RM, et al. (2006) Influence of sequence identity and unique breakpoints on the frequency of intersubtype hiv-1 recombination. Retrovirology 3: 91.

35. Simon-Loriere E, Galetto R, Hamoudi M, Archer J, Lefeuvre P, et al. (2009) Molecular mechanisms of recombination restriction in the envelope gene of the human immunodeficiency virus. PLoS Pathog 5: e1000418.

36. Levy DN, Aldrovandi GM, Kutsch O, Shaw GM (2004) Dynamics of hiv-1 recombination in its natural target cells. Proc Natl Acad Sci U S A 101: 4204–4209.

37. Michalakis Y, Roze D (2004) Evolution. epistasis in rna viruses. Science 306: 1492–1493.

38. Clotet B (2004) Strategies for overcoming resistance in hiv-1 infected patients receiving haart. AIDS Rev 6: 123–130.

39. Palmer S, Boltz V, Maldarelli F, Kearney M, Halvas EK, et al. (2006) Selection and persistence of non-nucleoside reverse transcriptase inhibitor-resistant hiv-1 in patients starting and stopping non-nucleoside therapy. AIDS 20: 701–710.

40. Liu J, Miller MD, Danovich RM, Vandergrift N, Cai F, et al. (2011) Analysis of low-frequency mutations associated with drug resistance to raltegravir before antiretroviral treatment. Antimicrob Agents Chemother 55: 1114–1119.

41. Martinez-Picado J, Martnez MA (2008) Hiv-1 reverse transcriptase inhibitor resistance mutations and fitness: a view from the clinic and ex vivo. Virus Res 134: 104–123.

42. Piana S, Carloni P, Rothlisberger U (2002) Drug resistance in hiv-1 protease: Flexibility-assisted mechanism of compensatory mutations. Protein Sci 11: 2393–2402.

43. Johnson JA, Li JF, Morris L, Martinson N, Gray G, et al. (2005) Emergence of drug-resistant hiv-1 after intrapartum administration of single-dose nevirapine is substantially underestimated. J Infect Dis 192: 16–23.

44. Yim HJ, Hussain M, Liu Y, Wong SN, Fung SK, et al. (2006) Evolution of multi-drug resistant hepatitis b virus during sequential therapy. Hepatology 44: 703–712.

45. Delaney WE, Yang H, Westland CE, Das K, Arnold E, et al. (2003) The hepatitis b virus polymerase mutation rtv173l is selected during lamivudine therapy and enhances viral replication in vitro. J Virol 77: 11833–11841.

46. Moradpour D, Penin F, Rice CM (2007) Replication of hepatitis c virus. Nat Rev Microbiol 5: 453–463.

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

48. Wu NC, Young AP, Dandekar S, Wijersuriya H, Al-Mawsawi LQ, et al. (2013) Systematic identification of h274y compensatory mutations in influenza a virus neuraminidase by high-throughput screening. J Virol 87: 1193–1199.

49. Gong LI, Suchard MA, Bloom JD (2013) Stability-mediated epistasis constrains the evolution of an influenza protein. Elife 2: e00631.

50. Nowak MA, Bangham CR (1996) Population dynamics of immune responses to persistent viruses. Science 272: 74–79.

51. Liao HX, Lynch R, Zhou T, Gao F, Alam SM, et al. (2013) Co-evolution of a broadly neutralizing hiv-1 antibody and founder virus. Nature 496: 469–476.

52. Lewis MJ, Dagarag M, Khan B, Ali A, Yang OO (2012) Partial escape of hiv-1 from cytotoxic t lymphocytes during chronic infection. J Virol 86: 7459–7463.

53. Wong JT, Colvin RB (1987) Bi-specific monoclonal antibodies: selective binding and complement fixation to cells that express two different surface antigens. J Immunol 139: 1369–1374.

54. Jabara CB, Jones CD, Roach J, Anderson JA, Swanstrom R (2011) Accurate sampling and deep sequencing of the hiv-1 protease gene using a primer id. Proc Natl Acad Sci U S A 108: 20166–20171.

55. Li H, Durbin R (2009) Fast and accurate short read alignment with burrows-wheeler transform. Bioinformatics 25: 1754–1760.

56. Larkin MA, Blackshields G, Brown NP, Chenna R, McGettigan PA, et al. (2007) Clustal w and clustal x version 2.0. Bioinformatics 23: 2947–2948.

CHAPTER 6

PERSPECTIVES

6.1 COUPLING HIGH-THROUGHPUT GENETICS WITH OTHER EXPERIMENTAL TOOLS

High-throughput genetics provided an excellent tool to identify genetic determinants responsible for virus-host interaction. While my thesis has developed and demonstrate the usage of high-throughput genetics in virus research, numerous potential applications have not yet been explored. Here the interaction between virus and interferon system will be used as an example to discuss potential strategies. Interferon (IFN) system, a major component of the innate immune system, continues to be a popular field of research. The antagonistic action against IFN seems to be universal across different viruses. It is known that many viruses have developed strategies to counteract the IFN signalling system [1–7]. It is also evidenced that virus can counteract the antiviral action of IFN-stimulated genes (ISG) by physical interaction [8–12].

To dissect the viral genetic determinants that interfere signalling pathway, a possible approach is to combine high-throughput genetics with fluorescent reporters and cell sorting. For example, a green fluorescent protein (GFP) reporter driven by IFN promoter can be used as a readout to screen for mutation that lost the inhibition activity against IFN expression (Fig. 6-1A). Mutation that retains the suppression activity again IFN expression would not turn the infected cells green, whereas those mutation that lost the suppression activity would induce the GFP expression in the infected cells. By applying cell sorting by flow cytometry, these two types of infected cells can be physically separated (Fig. 6-1B). Those mutations that lost the suppression activity can thus be isolated. Subsequently, their genotypes will be revealed by deep sequencing. In fact, the feasibility of combining mutant library, reporter and cell sorting to identify mutation that lost the suppression activity to IFN signalling pathway has been shown [13]. This strategy is generally applicable to study the virus-host interaction in other signalling pathways that are interfered by virus.

To investigate the physical interaction between a viral protein and a host restriction factor (e.g. ISG), a potential strategy is to combine high-throughput genetics with overexpression and knock-out of the host restriction factor (Fig. 6-2). Mutation in the viral genetic determinant that inhibit the host restriction factor would have different fitness effects on the virus depending on the expression level of the host restriction factor. Since the viral determinant is only functionally important for

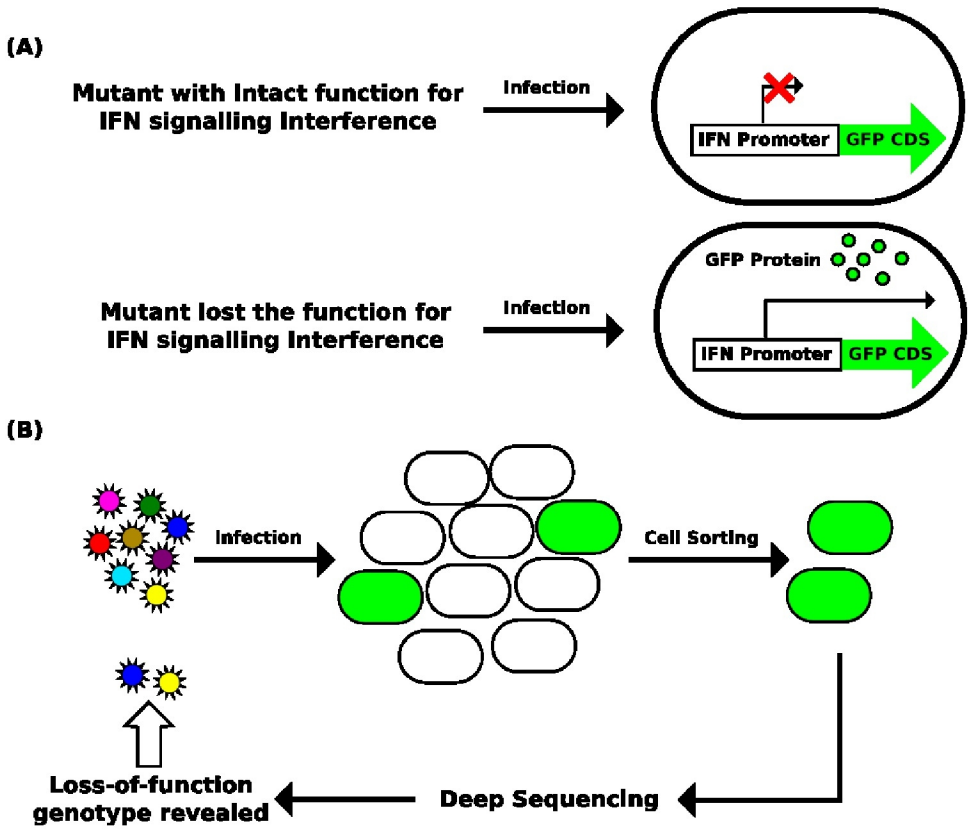


Figure 6-1. Concept of differential profiling. The viral mutant library is passaged under different condition. Control condition means there is no additional selection pressure besides replication capacity. Each circle represents an individual viral particle. Different colors represent different genotypes (WT or mutants). Genotype colored in green represents weak drug resistance mutation, which is moderately enriched after passaging in the presence of drug. Genotype colored in cyan represents strong drug resistance mutation, which is highly enriched after passaging in the presence of drug. Genotype colored in yellow represents interferon sensitive mutation, which disappears from the mutant library after passaging in the presence of interferon.

replication when the host restriction factor is presence, mutation within the viral determinant would impose a high fitness cost when the host restriction factor is expressed at a high level. In contrast, mutation within the viral determinant would may not have any fitness cost when the host restriction factor is absence. As a result, by comparing the fitness profile under different expression level of the host restriction factor, the genetic determinant employed by the virus to suppress the antiviral function of the host restriction factor would be revealed.

In the future, high-throughput genetics can be applied to identify critical viral residues that interplay with other cellular and immune responses, such as apoptosis, autophagy, ER stress, cytokines, to understand how the virus interact with different functional components of the host.

6.2 INVESTIGATING EPISTATIC EFFECT USING HIGH-THROUGHPUT GENETICS

Episatsis, which describes the difference in fitness effect of a mutation under different genetic backgrounds, plays a critical role in viral evolution, such as drug resistance and immune escape [14–18]. The mutational fitness landscapes in different genetic backgrounds can be obtained by applying high-throughput genetics in different viral strains. By comparing the mutational fitness landscapes in different viral strains, mutations that display a genetic-background-dependent fitness can be identified (Fig. 6.3). Such information will also be highly valuable to the modeling of evolutionary landscape and functional sequence space. When a sufficient number of fitness landscapes are obtained, it may even be possible to dissect the genetic relationship among residues to comprehend the functional sequence space and the genotype-phenotype map.

As high-throughput genetic approaches continue to evolve, it might be possible to systematically examine the fitness effect of high-order mutations, hence epistasis. In fact, recently our lab and other groups have used high-throughput genetics to interrogate pairwise epistasis within a domain of a model protein [19–21]. These studies not only quantify the fitness effect of individual single amino acid substitutions, but also that of the double amino acid substitutions. It allows the quantification of epistatic interaction between different mutations. Potentially, those approaches can be adapted to viral system to examine pairwise epistatic landscape within a viral protein domain.

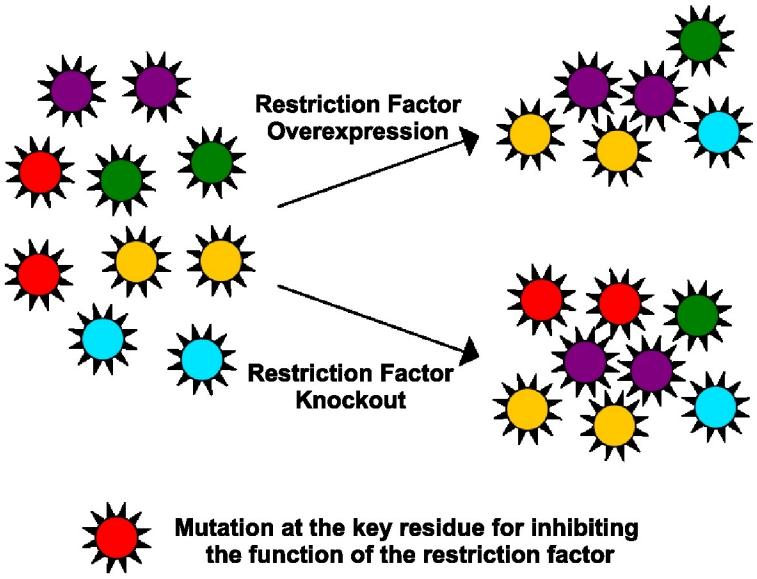


Figure 6-2. Coupling high-throughput genetics with gene knockout and gene overexpression. The genetic interaction between the virus and a restriction host factor of interest is studied. The viral mutant library is passaged under two conditions, namely overexpression of the host restriction factor and knockout of the host restriction factor. Genotype colored in red carries a mutation at the key residue, which is responsible for inhibiting the function of the restriction factor resistance mutation and is the genotype of interest. This genotype of interest will have little, if any, fitness cost when the restriction factor is knocked down in the host cells. However, when the restriction factor is overexpressed in the host cells, this genotype of interest will be disappeared after passaging.

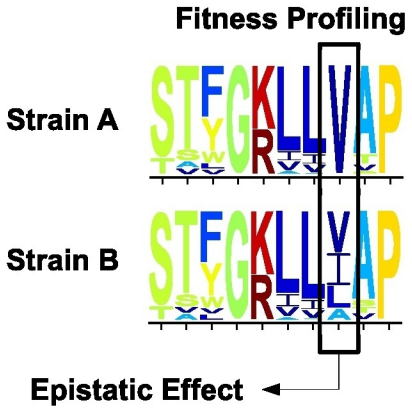


Figure 6-3. Fitness profiling of two viral strains with different genetic backgrounds to identify epistatic effect. Amino acid preference for each residue can be computed from fitness profiling [51]. By comparing fitness profile of different genetic backgrounds, residue with a genotypic-dependent amino acid preference will be identified.

6.3 CONCLUDING REMARKS

High-throughput genetics enables rapid identification of critical genetic elements, such as protein subdomains or nucleotide residues, across the genome under any specified growth conditions. With the continued improvement of deep sequencing technologies and refinement of high-throughput genetics, critical virological questions can be systematically answered by functional genomics on viruses. Previous studies have shown that high-throughput genetics could provide tremendous information that aids the identification of essential genetic elements on the virus genome, and facilitates the understanding of viral evolution and virus-host interaction. High-throughput genetics provide a new strategies in answering biological questions and will significantly accelerate virus research.

6.4 BIBLIOGRAPHY

1. Hale BG, Randall RE, Ortn J, Jackson D (2008) The multifunctional ns1 protein of influenza a viruses. J Gen Virol 89: 2359–2376.

2. Meln K, Fagerlund R, Nyqvist M, Keskinen P, Julkunen I (2004) Expression of hepatitis c virus core protein inhibits interferon-induced nuclear import of stats. J Med Virol 73: 536–547.

3. Lin W, Kim SS, Yeung E, Kamegaya Y, Blackard JT, et al. (2006) Hepatitis c virus core protein blocks interferon signaling by interaction with the stat1 sh2 domain. J Virol 80: 9226–9235.

4. Muoz-Jordan JL, Snchez-Burgos GG, Laurent-Rolle M, Garca-Sastre A (2003) Inhibition of interferon signaling by dengue virus. Proc Natl Acad Sci U S A 100: 14333–14338.

5. Solis M, Nakhaei P, Jalalirad M, Lacoste J, Douville R, et al. (2011) Rig-i-mediated antiviral signaling is inhibited in hiv-1 infection by a protease-mediated sequestration of rig-i. J Virol 85: 1224–1236.

6. Morrison TE, Mauser A, Wong A, Ting JP, Kenney SC (2001) Inhibition of ifn-gamma signaling by an epstein-barr virus immediate-early protein. Immunity 15: 787–799.

7. Vidy A, Chelbi-Alix M, Blondel D (2005) Rabies virus p protein interacts with stat1 and inhibits interferon signal transduction pathways. J Virol 79: 14411–14420.

8. Peters GA, Khoo D, Mohr I, Sen GC (2002) Inhibition of pact-mediated activation of pkr by the herpes simplex virus type 1 us11 protein. J Virol 76: 11054–11064.

9. Mariani R, Chen D, Schrfelbauer B, Navarro F, Knig R, et al. (2003) Species-specific exclusion of apobec3g from hiv-1 virions by vif. Cell 114: 21–31.

10. Taguchi T, Nagano-Fujii M, Akutsu M, Kadoya H, Ohgimoto S, et al. (2004) Hepatitis c virus ns5a protein interacts with 2',5'-oligoadenylate synthetase and inhibits antiviral activity of ifn in an ifn sensitivity-determining region-independent manner. J Gen Virol 85: 959–969.

11. Li S, Min JY, Krug RM, Sen GC (2006) Binding of the influenza a virus ns1 protein to pkr mediates the inhibition of its activation by either pact or double-stranded rna. Virology 349: 13–21.

12. Gack MU, Albrecht RA, Urano T, Inn KS, Huang IC, et al. (2009) Influenza a virus ns1 targets the ubiquitin ligase trim25 to evade recognition by the host viral rna sensor rig-i. Cell Host Microbe 5: 439–449.

13. Prez-Cidoncha M, Killip MJ, Asensio VJ, Fernndez Y, Bengoechea JA, et al. (2014) Generation of replication-proficient influenza virus ns1 point mutants with interferon-hyperinducer phenotype. PLoS One 9: e98668.

14. Bonhoeffer S, Chappey C, Parkin NT, Whitcomb JM, Petropoulos CJ (2004) Evidence for positive epistasis in hiv-1. Science 306: 1547–1550.

15. Campo DS, Dimitrova Z, Mitchell RJ, Lara J, Khudyakov Y (2008) Coordinated evolution of the hepatitis c virus. Proc Natl Acad Sci U S A 105: 9685–9690.

16. Kryazhimskiy S, Dushoff J, Bazykin GA, Plotkin JB (2011) Prevalence of epistasis in the evolution of influenza a surface proteins. PLoS Genet 7: e1001301.

17. Duan S, Govorkova EA, Bahl J, Zaraket H, Baranovich T, et al. (2014) Epistatic interactions between neuraminidase mutations facilitated the emergence of the oseltamivir-resistant h1n1 influenza viruses. Nat Commun 5: 5029.

18. Gong LI, Bloom JD (2014) Epistatically interacting substitutions are enriched during adaptive protein evolution. PLoS Genet 10: e1004328.

19. Araya CL, Fowler DM, Chen W, Muniez I, Kelly JW, et al. (2012) A fundamental protein property, thermodynamic stability, revealed solely from large-scale measurements of protein function. Proc Natl Acad Sci U S A 109: 16858–16863.

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

21. Bank C, Hietpas RT, Jensen JD, Bolon DNA (2015) A systematic survey of an intragenic epistatic landscape. Mol Biol Evol 32: 229–238.