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

Nicholas C. Wu^{1,2,7}, Arthur P. Young^{1,7}, Laith Q. Al-Mawsawi¹, C. Anders Olson¹, Jun Feng¹, Hangfei Qi¹, Shu-Hwa Chen³, I-Hsuan Lu³, Chung-Yen Lin³, Robert G. Chin⁴, Harding H. Luan¹, Nguyen Nguyen¹, Stanley F. Nelson^{2,4}, Xinmin Li⁵, Ting-Ting Wu¹, Ren Sun^{1,2,6,*}

¹ Department of Molecular and Medical Pharmacology, David Geffen School of Medicine, University of California, Los Angeles, CA 90095, USA

² Molecular Biology Institute, University of California, Los Angeles, CA 90095, USA

³ Institute of Information Science, Academia Sinica, Taipei, Taiwan

⁴ Department of Human Genetics, David Geffen School of Medicine, University of California, Los Angeles, CA 90095, USA

⁵ Department of Pathology and Laboratory Medicine, David Geffen School of Medicine, University of California, Los Angeles, CA 90095, USA

⁶ AIDS Institute, University of California, Los Angeles, CA 90095, USA

⁷ These authors contributed equally to this work

* E-mail: RSun@mednet.ucla.edu

17 Abstract

18 A novel genome-wide genetics platform is presented in this study, which permits functional
 19 interrogation of all point mutations across a viral genome in parallel. Here we generated the first
 20 fitness profile of individual point mutations across the influenza virus genome. Critical residues
 21 on the viral genome were systematically identified, which provided a collection of subdomain
 22 data informative for structure-function studies and for effective rational drug and vaccine design.
 23 Our data was consistent with known, well-characterized structural features. In addition, we
 24 have achieved a validation rate of 68% for severely attenuated mutations and 94% for neutral
 25 mutations. The approach described in this study is applicable to other viral or microbial genomes
 26 where a means of genetic manipulation is available.

27 Introduction

28 The influenza virus causes several hundred thousand deaths every year, and this number can
 29 reach millions in pandemic years. The huge socioeconomic associated with influenza highlights
 30 the importance of understanding of virus-host interactions [1,2]. The rapidly evolving nature of
 31 influenza challenges the development of anti-influenza drugs and vaccine [3–7]. Consequently,
 32 it is important to develop drugs or vaccines that target indispensable regions on the influenza
 33 virus to maximize the genetic barrier for the emergence of resistance or escape mutations. Nev-
 34 ertheless, genetic research on the influenza virus has largely relied on naturally variants and
 35 individual mutants created in the laboratory. A substantial part of the genome remains unchar-
 36 acterized.

37

38 Traditional genetics studies the relationship of a single genotype-phenotype at a time, and has
 39 been extensively to study panels of influenza mutations. However, the low throughput of tradi-
 40 tional genetics limited the number of mutations being examined. In contrast, high-throughput
 41 genetics interrogates the phenotypic outcomes of multiple mutants in parallel. Genome-wide
 42 insertional mutagenesis is a common high-throughput genetics approach. It has been employed
 43 in the influenza virus to systematically identify regions that are tolerate to mutations [8]. How-
 44 ever, the resolution of insertion-based approach is limited at the protein subdomain level. This
 45 resolution is insufficient to identify residues critical for replication. As a result, there is a demand
 46 for a high-throughput genetics platform at a single-residue resolution.

47

48 Recently, we have developed a high-throughput genetic platform which allowed us to profile
 49 the fitness effect of individual point mutations across the influenza A virus hemagglutinin (HA)
 50 segment [9]. The principle of the high-throughput genetic platform is to utilize a large mutant
 51 library and deep sequencing. Here, we extended this approach to quantify the fitness effects of
 52 each point mutation in 96% of the influenza A virus genome. This technique will enable system-
 53 atic identification of indispensable regions for drug or vaccine targets. More importantly, it can

be applied to any specified growth conditions for any virus that can be genetically manipulated.

Results

Quantification of the fitness effect of individual point mutation

Our high-throughput genetics platform aims to randomly mutagenize each nucleotide of the genome, monitor the changes in occurrence frequency for individual point mutations under specified growth conditions using massive deep-sequencing [9]. The changes in occurrence frequency of each point mutation (such as diminishment or enrichment) allow us to quantify the mutational fitness outcomes under the given growth conditions. The mutant libraries were created by error-prone PCR on the eight-plasmid reverse genetics system influenza A/WSN/1933 (H1N1) [10] (see materials and methods). Subsequently, eight viral mutant libraries were generated by transfection, each with one of the eight segments mutagenized. All viral mutant libraries were passaged for two 24-hour rounds in A549 cells (human lung epithelial carcinoma cells). The plasmid library and the passaged viral library were each sequenced by Illumina HiSeq 2000. Here, a relative fitness index (RF index) is used to estimate the mutational fitness effect. The RF index is calculated as:

$$\text{RF index} = \frac{\text{occurrence frequency in passaged library}}{\text{occurrence frequency in plasmid library}}$$

The occurrence frequency of individual mutations was expected to be lower than the sequencing error rate ($\sim 0.1\%$ - 1%) in next generation sequencing (NGS). Therefore, we utilized a two-step PCR approach for sequencing library preparation to distinguish true mutations from sequencing errors. In the first PCR, a unique tag was assigned to individual molecules. The second PCR generated multiple identical copies for individual tagged molecules. The input copy number for the second PCR was well-controlled such that individual tagged molecules would be sequenced ~ 10 times. True mutations would exist in all sequencing reads sharing the same tag, whereas

79 sequencing errors would not. Individual molecules, each carrying a unique tag, have an average
80 copy number of ~ 10 in the sequencing data, which validated the sequencing library preparation
81 design (Fig. S1).

82 **Point mutation fitness profiling of influenza A virus genome**

83 The RF indices for individual point mutations were profiled across 96% of nucleotide positions
84 in the influenza A/WSN/1933 virus genome (Fig. 1). The remaining 4% of nucleotide were from
85 the termini of each gene segment due to PCR amplification difficulty. As expected, a positive
86 correlation exists between RF index and the degree of amino acid conservation of missense
87 mutations (Fig. S2). In addition, the fitness data for well-characterized mutants were consistent
88 with their phenotypes reported in the literature. Examples include a critical salt bridge for
89 viral replication on nucleoprotein (NP) [11] (Fig. S3A), replication enhancement mutation on
90 polymerase subunit (PB2) [12] (Fig. S3B), attenuation of oseltamivir resistance mutation on
91 neuraminidase (NA) [13] (Fig. S3C), low fitness cost of amantadine/rimantadine resistance
92 mutations on ion channel (M2) [5,14,15] (Fig. S3D), and the basic stretch on matrix protein (M1)
93 required for assembly [16] (Fig. S4). Furthermore, comparison between our fitness data with
94 the polymerase activity on 19 PB1 mutants previously reported showed an 80% correlation [17].
95 Mutants that displayed a severely attenuated (RF index < 0.05) or neutral (RF index > 0.4)
96 phenotype were randomly selected across the genome, individually constructed and tested. The
97 replication phenotype of each single mutant validated the profiling data with a confirmation rate
98 of 68% for severely attenuated mutations and 94% for neutral mutations (Fig. 2). These data
99 taken together provides validity to our fitness profiling data set.

100 **Structural analysis and identification of indispensable protein surface**

101 Our high-throughput profiling technique provides a basis to identify essential protein surfaces
102 for drug targeting and indispensable regions for vaccine epitopes. We have performed a struc-
103 tural analysis on NA, a major influenza vaccine antigen. Here we identified a cluster of essential

residues at the tetramer formation interface, suggesting that it bears functional importance and can possibly be a drug targeting site. In contrast, such a large cluster of essential residues could not be found in any other part of the NA surface. The lack of essential residues on the NA surface explain the functional basis of antigenic drift.

We have also performed a structural analysis using the PA subunit of the influenza virus RNA polymerase as an example to search for indispensable regions to aid in rational drug design. Increasing evidence suggests PA is a valuable target for drug development due to its polyfunctionality [18–20]. Our fitness data provided an informative reference for rational drug design. It captured several critical interactions between PA and PB1, such as the hydrogen bond between PA E617 and PB1 K11 (Fig. 3A), and the hydrophobic interaction between PA and PB1 via the volume-filling residues L666 and F710 (Fig. 3B). It has also revealed a cluster of essential residues on the PA surface consisting of eight amino acids (Fig. 3C), including K539 and K574, which were previously shown to be part of a lead compound binding pocket [19]. This patch of amino acids may be involved in an essential protein-protein interaction for viral replication. Similar analyses using our dataset have been applied to PA endonuclease domain and the M2 ion channel, which are plausible targets in drug development (Fig. S5-6). By projecting the fitness profiling data on three dimensional protein structures, it enables the identification of novel putative essential structural motifs that are surface exposed but not necessarily sequential in the primary sequence. This type of analysis reveals biological targets useful for rational drug and vaccine design. We propose that future antiviral drug design can incorporate the technique described in this study with *in silico* drug screening to increase the efficiency of therapeutic identification.

Discussion

Sequence conservation was often taken as the sole parameter for identifying residues essential for viral replication, although conservation is not equivalent to essentialness for viral replication.

130 It has been suggested that a significant fraction of conserved residues that are conserved in the
131 influenza A virus are dispensable in viral replication [17,21,22]. In addition, new mutations were
132 observed in every flu season, implying that residues that are naturally conserved currently may
133 still be able to mutate under future unforeseen selection pressures. Therefore, a high-throughput
134 fitness profiling complements the shortcoming in the sequence conservation analysis and allows
135 identification of amino acid residues that are critical for viral replication in a defined cellular
136 environment.

137

138 Here we provided a proof-of-concept study to profile the entire influenza A virus genome at
139 single-nucleotide resolution. The fitness effects of individual point mutations were interrogated
140 in a high-throughput manner by coupling a large mutant library with NGS. However, the quan-
141 tifiability of our platform can be further improved as sequencing technology advance. Similar
142 experiments should be performed with strains across subtypes to identify mutations that display
143 a genetic background-dependent fitness effect. These results would provide valuable information
144 to dissect the evolutionary process of the influenza A virus. In addition, this platform can be
145 applied to study the virus-host interaction under different cellular responses (such as apopto-
146 sis, autophagy, inflammasome induction, ER stress, etc.) and immune responses (such as NK
147 cells, T cells, antibodies, macrophages, cytokines, etc.) that influence the viral replication in
148 nature [23,24]. Such results will significantly improve our understanding of the biological role
149 of each residue on the genome of the influenza A virus. They will also help improve the design
150 of a live attenuated influenza vaccine by minimizing the virulence. More importantly, it can
151 potentially be adapted to other virus and microbes that can be genetically manipulated in the
152 laboratory.

153 Acknowledgments

154 We would like to thank J. Zhou, J. Yoshizawa, T. Toy and Z. Chen for performing the high-
155 throughput sequencing experiment, K. Squire for support on data analysis, Y. Liu for advice on

the PA structural analysis, Y. Liang and J. Bloom for valuable discussions.

Materials and Methods

Viral mutant library and point mutations

The plasmid mutant libraries were created by performing error-prone PCR on the eight-plasmid reverse genetics system of influenza A/WSN/1933 (H1N1) [10]. We PCR-amplified the flu insert with error-prone polymerase Mutazyme II (Stratagene, La Jolla, CA). 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 sites BsmBI and/or BsaI were added to the PCR primers, and used to clone into a BsmBI-digested parental vector pHW2000. Ligations were carried out with high concentration T4 ligase (Invitrogen, Grand Island, NY). Transformations were carried out with electrocompetent MegaX DH10B T1R cells (Invitrogen), and > 100,000 colonies for each segment library 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 (Invitrogen) using 7 wildtype plasmids plus 1 mutant (library) plasmid. 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°C. The TCID50 was measured on A549 cells (human lung carcinoma cells).

180

181 Virus from C227 transfection was used to infect A549 at an MOI of 0.05. Infected cells were
182 washed three times with PBS followed by the addition of fresh cell growth medium at 2 hrs
183 post-infection. Virus was harvested at 24 hrs post-infection. For the mutant library profiling,
184 all viral mutant libraries were passaged for two 24-hour rounds in A549 cells. Our pilot exper-
185 iments as well as our previous study revealed that two rounds of passaging were sufficient for
186 profiling [25].

187 Sequencing library preparation

188 DNA from the plasmid library or cDNA from the passaged viral mutant library were amplified
189 with both forward and reverse primers each flanked with a 6 “N” tag and the flow cell adapter re-
190 gion. Flanking region for 5’ primer: 5’-CTACACGACGCTCTTCCGATCTNNNNNN-3’, Flank-
191 ing region for 3’ primer: 5’-TGCTGAACCGCTCTTCCGATCTNNNNNN-3’. Following PCR,
192 93 amplicon products were pooled together. 15 million copies of the pooled product were used
193 as the input for the second PCR, which was equivalent to 10 paired-end reads per molecule if
194 150 million paired-end reads (approximately one lane on an Illumina HiSeq 2000 machine) were
195 sequenced. 5’-AATGATACGGCGACCAACGAGATCTACACTC
196 TTTCCCTACACGACGCTCTTCCG-3’ and 5’-CAAGCAGAAGACGGCATACGAGATCGGTCTCGG
197 CATTCTGCTGAACCGCTCTTCCG-3’ were used as the primers for the second PCR. Prod-
198 ucts from the second PCR were submitted for NGS. The error-correction technique described
199 in this study adapted the philosophy described for detecting rare mutations in human cells [26].
200 Raw sequencing data have been submitted to the NIH Short Read Archive under accession
201 number: SRR1042008 (plasmid mutant library) and SRR1042006 (passaged mutant library).

202 Data Analysis

203 Sequencing reads were mapped by BWA with a maximum of six mismatches and no gap [27].
204 Amplicons with the same tag were collected to generate a read cluster. Since each read cluster

205 was originated from the same template, true mutations were called only if the mutations oc-
 206 curred in 90% of the reads within a read cluster. Read clusters with a size below three reads
 207 were filtered out. Read clusters were further conflated into “error-free” reads. Relative fitness
 208 index (RF index) for individual point mutations was computed by:

209

210 $(\text{occurrence frequency in passaged library}) / (\text{occurrence frequency in plasmid library})$

211

212 For all the downstream analysis, only point mutations covered with ≥ 30 tag-conflated reads
 213 (“error-free” reads) in the plasmid library were included. This arbitrary cutoff filtered out
 214 mutants with low statistical confidence.

215 References

- 216 1. de Wit E, Fouchier RAM (2008) Emerging influenza. *J Clin Virol* 41: 1–6.
- 217 2. Taubenberger JK, Morens DM (2010) Influenza: the once and future pandemic. *Public*
- 218 *Health Rep* 125 Suppl 3: 16–26.
- 219 3. Beigel J, Bray M (2008) Current and future antiviral therapy of severe seasonal and avian
- 220 influenza. *Antiviral Res* 78: 91–102.
- 221 4. Couch RB (2008) Seasonal inactivated influenza virus vaccines. *Vaccine* 26 Suppl 4: D5–
- 222 D9.
- 223 5. Hayden FG, Hay AJ (1992) Emergence and transmission of influenza a viruses resistant
- 224 to amantadine and rimantadine. *Curr Top Microbiol Immunol* 176: 119–130.
- 225 6. Pizzorno A, Abed Y, Boivin G (2011) Influenza drug resistance. *Semin Respir Crit Care*
- 226 *Med* 32: 409–422.
- 227 7. Hayden FG, de Jong MD (2011) Emerging influenza antiviral resistance threats. *J Infect*
- 228 *Dis* 203: 6–10.
- 229 8. Heaton NS, Sachs D, Chen CJ, Hai R, Palese P (2013) Genome-wide mutagenesis of
- 230 influenza virus reveals unique plasticity of the hemagglutinin and ns1 proteins. *Proc Natl*
- 231 *Acad Sci U S A* .
- 232 9. Wu NC, Young AP, Al-Mawsawi LQ, Olson CA, Feng J, et al. (2014) High-throughput
- 233 profiling of influenza a virus hemagglutinin gene at single-nucleotide resolution. *Sci Rep*
- 234 4: 4942.
- 235 10. Neumann G, Watanabe T, Ito H, Watanabe S, Goto H, et al. (1999) Generation of in-
- 236 fluenza a viruses entirely from cloned cdnas. *Proc Natl Acad Sci U S A* 96: 9345–9350.

- 237 11. Shen YF, Chen YH, Chu SY, Lin MI, Hsu HT, et al. (2011) E339...r416 salt bridge of
238 nucleoprotein as a feasible target for influenza virus inhibitors. *Proc Natl Acad Sci U S*
239 *A* 108: 16515–16520.
- 240 12. Bussey KA, Bousse TL, Desmet EA, Kim B, Takimoto T (2010) Pb2 residue 271 plays a
241 key role in enhanced polymerase activity of influenza a viruses in mammalian host cells.
242 *J Virol* 84: 4395–4406.
- 243 13. Ives JAL, Carr JA, Mendel DB, Tai CY, Lambkin R, et al. (2002) The h274y mutation in
244 the influenza a/h1n1 neuraminidase active site following oseltamivir phosphate treatment
245 leave virus severely compromised both in vitro and in vivo. *Antiviral Res* 55: 307–317.
- 246 14. Abed Y, Goyette N, Boivin G (2005) Generation and characterization of recombinant
247 influenza a (h1n1) viruses harboring amantadine resistance mutations. *Antimicrob Agents*
248 *Chemother* 49: 556–559.
- 249 15. Krumbholz A, Schmidtke M, Bergmann S, Motzke S, Bauer K, et al. (2009) High preva-
250 lence of amantadine resistance among circulating european porcine influenza a viruses. *J*
251 *Gen Virol* 90: 900–908.
- 252 16. Das SC, Watanabe S, Hatta M, Noda T, Neumann G, et al. (2012) The highly conserved
253 arginine residues at positions 76 through 78 of influenza a virus matrix protein m1 play
254 an important role in viral replication by affecting the intracellular localization of m1. *J*
255 *Virol* 86: 1522–1530.
- 256 17. Chu C, Fan S, Li C, Macken C, Kim JH, et al. (2012) Functional analysis of conserved
257 motifs in influenza virus pb1 protein. *PLoS One* 7: e36113.
- 258 18. DuBois RM, Slavish PJ, Baughman BM, Yun MK, Bao J, et al. (2012) Structural and bio-
259 chemical basis for development of influenza virus inhibitors targeting the pa endonuclease.
260 *PLoS Pathog* 8: e1002830.

- 261 19. Li L, Chang S, Xiang J, Li Q, Liang H, et al. (2012) Screen anti-influenza lead compounds
262 that target the pa(c) subunit of h5n1 viral rna polymerase. PLoS One 7: e35234.
- 263 20. Nakazawa M, Kadowaki S, Watanabe I, Kadowaki Y, Takei M, et al. (2008) Pa
264 subunit of rna polymerase as a promising target for anti-influenza virus agents. Antiviral
265 Res 78: 194–201.
- 266 21. Li Z, Watanabe T, Hatta M, Watanabe S, Nanbo A, et al. (2009) Mutational analysis of
267 conserved amino acids in the influenza a virus nucleoprotein. J Virol 83: 4153–4162.
- 268 22. Stewart SM, Pekosz A (2011) Mutations in the membrane-proximal region of the influenza
269 a virus m2 protein cytoplasmic tail have modest effects on virus replication. J Virol 85:
270 12179–12187.
- 271 23. Ehrhardt C, Seyer R, Hrincius ER, Eierhoff T, Wolff T, et al. (2010) Interplay between
272 influenza a virus and the innate immune signaling. Microbes Infect 12: 81–87.
- 273 24. Rossman JS, Lamb RA (2009) Autophagy, apoptosis, and the influenza virus m2 protein.
274 Cell Host Microbe 6: 299–300.
- 275 25. Wu NC, Young AP, Dandekar S, Wijersuriya H, Al-Mawsawi LQ, et al. (2013) Systematic
276 identification of h274y compensatory mutations in influenza a virus neuraminidase by
277 high-throughput screening. J Virol 87: 1193–1199.
- 278 26. Kinde I, Wu J, Papadopoulos N, Kinzler KW, Vogelstein B (2011) Detection and quan-
279 tification of rare mutations with massively parallel sequencing. Proc Natl Acad Sci U S
280 A 108: 9530–9535.
- 281 27. Li H, Durbin R (2009) Fast and accurate short read alignment with burrows-wheeler
282 transform. Bioinformatics 25: 1754–1760.

- 283 28. Russell RJ, Haire LF, Stevens DJ, Collins PJ, Lin YP, et al. (2006) The structure of h5n1
284 avian influenza neuraminidase suggests new opportunities for drug design. *Nature* 443:
285 45–49.
- 286 29. Obayashi E, Yoshida H, Kawai F, Shibayama N, Kawaguchi A, et al. (2008) The structural
287 basis for an essential subunit interaction in influenza virus rna polymerase. *Nature* 454:
288 1127–1131.
- 289 30. Arzt S, Baudin F, Barge A, Timmins P, Burmeister WP, et al. (2001) Combined results
290 from solution studies on intact influenza virus m1 protein and from a new crystal form of
291 its n-terminal domain show that m1 is an elongated monomer. *Virology* 279: 439–446.
- 292 31. Schnell JR, Chou JJ (2008) Structure and mechanism of the m2 proton channel of in-
293 fluenza a virus. *Nature* 451: 591–595.

294 Figure Legends

295 **Figure 1. Single-nucleotide resolution fitness profiling.** The RF index for individual
296 point mutations across the genome was computed. Natural log of RF index, which is the ra-
297 tio of occurrence frequency in the passaged library to the occurrence frequency in the plasmid
298 library, represents the y-axis. Each nucleotide position is represented by four consecutive lines
299 for the RF index that correspond to mutating to A (blue), T (green), C (orange), or G (red).
300 The RF index of WT nucleotides is set as zero. Only point mutations with a coverage of ≥ 30
301 tag-conflated reads in the plasmid library are shown. Point mutations with < 30 tag-conflated
302 reads in the plasmid library is plotted as a gray dot on the zero baseline. The data track for
303 HA is adapted from Wu *et al.* [9].

304

305 **Figure 2. Experimental validation of severely attenuated and neutral mutations.**
306 Based on the data in Fig. 1, mutations that displayed a RF index of < 0.05 were classified as
307 severely attenuated and > 0.4 were classified as neutral. Individual mutants were constructed
308 and compared to the wild type (WT) replication phenotype. Post-transfection titers were plot-
309 ted for lethal and viable mutants. Infection was initiated at an MOI of 0.05. Virus was harvested
310 at 24 hours post infection. For the validated mutations with a RF index < 0.05 , 68% have at
311 least 1 log decrease in titer compared to WT. For the validated mutations with a RF index > 0.4 ,
312 94% have a titer within a 2-fold change as compared to WT. Overall the validation rate is $\sim 80\%$.

313

314 **Figure 3. Structural analysis of the NA homotetramer interface.** The RF index
315 of the least destructive missense mutations for individual amino acids on the NA segment were
316 projected on the protein structure (PDB: 3CL0) to identify for essential regions [28]. The RF
317 index is color coded: RF index < 0.1 , red; $0.1 \leq \text{RF index} < 0.2$, orange; uncovered, grey. Only
318 one monomer of the homotetramer is color coded.

319

320 **Figure 4. Structural analysis of the RNA polymerase PA subunit.** The RF index

of the least destructive missense mutations in the profiling data for individual amino acids on the PA segment are projected on the PA-PB1 complex crystal structure (PDB: 2ZNL) [29]. Most deleterious 10%, red; 10% to 20%, orange; Others, green. Our fitness data is capable to identify several critical interactions and putative functional sites. (D) A hydrogen bond between PA E617 and PB1 K11 is shown. Substitution of PA E617 is deleterious in our fitness data. (E) A hydrophobic interaction is shown between PA L666 and F710 and PB1. Substitution of L666 is deleterious in our fitness data. (F) A cluster of eight essential residues on the surface of PA is shown.

329

Supplemental Figure 1. Distribution of conflated cluster size. Reads from the same amplicon with the same tag was defined as a cluster. The counts (number of reads) for all clusters are displayed as a histogram. Individual molecules, each carrying a unique tag, have an average copy number of ~ 10 in the sequencing data, thus validating the sequencing library preparation design.

335

Supplemental Figure 2. Comparison with BLOSUM62-based amino acid conservation. RF index of missense mutations from different segments were extracted and compared to amino acid conservation. The degree of amino acid conservation was quantified by the BLOSUM62 matrix, a substitution matrix based on an implicit model of evolution. The x-axis represents the different cutoffs for BLOSUM62 values. The average RF index value for missense mutations that satisfied the cutoff was plotted against different BLOSUM62 cutoff values. The positive correlation between the RF index and the degree of amino acid conservation of missense mutations indicates that our fitness data shows consistency with the evolutionary trend for missense mutations.

345

Supplemental Figure 3. The RF index of substitutions at different functional sites. (A) E339 and R416 on the NP protein form a salt bridge at the homodimer interface, which is

essential for viral replication [11]. This suggests that it is a feasible drug target. Several small molecules have been identified to target this interface and inhibit viral replication. (B) T271A has been identified as the replication enhancement substitution on PB2. T271A virus showed enhanced growth as compared to the WT strain in mammalian cells *in vitro* [12]. (C) NA 259Y (N1 naming: H274Y), a known oseltamivir drug resistance substitution, was shown to present a strongly attenuated phenotype in WSN [13]. In contrast, H259N (N1 naming: H274N), did not impose a deleterious effect in our fitness profiling data. This substitution is hypothesized to reduce influenza zanamivir sensitivity. Our results suggest further characterization of this substitution is warranted. (D) L26I, L26F, V27A and S31N on M2, the amantadine/rimantadine resistance substitutions [14,15], were shown to impose little effect on viral replication. Our data is consistent with the observation that resistance substitutions emerged rapidly during amantadine/rimantadine drug treatment [5]. Green dotted line represents the average RF index for missense mutation at the indicated segment. Overall, the fitness data was consistent with the phenotypes of functional mutants reported in the literature.

Supplemental Figure 4. Structural analysis of M1. (A) The RF index of the least destructive missense mutations for individual amino acids on the M1 segment were projected on the protein structure (PDB: 1EA3) to identify indispensable regions [30]. The RF index was color coded: RF index < 0.1, red; $0.1 \leq$ RF index < 0.2, orange. (B) The critical residues ${}_{76}\text{RRR}_{78}$ were displayed in stick format as an inset. It has been suggested that this basic amino acid stretch is important for virus assembly and/or budding [16]. Virus substitutions at these positions show an attenuated phenotype. Our data is consistent with the previous observation. The non-structural region at the C-terminal end of ${}_{76}\text{RRR}_{78}$ is also indispensable in our profiling data. This suggests that entire the non-structural region containing the ${}_{76}\text{RRR}_{78}$ basic stretch is functionally important. One possibility for functional importance is that it provides an interface for a protein-protein interaction.

Supplemental Figure 5. Structural analysis of the PA endonuclease domain. The RF index of the least destructive missense mutations in the profiling data for individual amino acids on the PA segment are projected on the PA endonuclease crystal structure (PDB: 4E5G). Most deleterious 10%, red; 10% to 20%, orange; Others, green. A critical helix-helix interface, which consists of T40, V44, M47, I171, R174 and I178, is highlighted. It demonstrates the power of qHRG in identifying residues that are not continuous in the primary sequence.

381

Supplemental Figure 6. Structural analysis of the M2 ion channel. The RF index of the least destructive missense mutations in the profiling data for individual amino acids on the M2 protein are projected on the M2 ion channel crystal structure (PDB: 2RLF) [31]. Most deleterious 10%, red; 10% to 20%, orange; Others, green. An indispensable region on the transmembrane helix is highlighted. Our data captured the essential amino acids W41 and H37, which are critical for M2 ion channel activation [31]. We also identified several adjacent hydrophobic residues, I35, L36, and L38 as critical residues, which can be attributed to their contact with the hydrophobic membrane.

Wu et al. Figure 1

Wu et al. Figure 2

Wu et al. Figure 3

Wu et al. Figure 4

C

Wu et al. Supplemental Figure 1

Wu et al. Supplemental Figure 2

Wu et al. Supplemental Figure 4

Wu et al. Supplemental Figure 5

90 degrees

Wu et al. Supplemental Figure 6