An experimentally determined evolutionary model dramatically improves phylogenetic fit

Jesse D. Bloom[1*]

[1]Division of Basic Sciences and Computational Biology Program,
Fred Hutchinson Cancer Research Center, Seattle, WA 98109, USA

*To whom correspondence should be addressed; E-mail: jbloom@fhcrc.org.

All modern approaches to molecular phylogenetics require a quantitative model for how genes evolve. Unfortunately, existing evolutionary models do not realistically represent the site-heterogeneous selection that governs actual sequence change. Attempts to remedy this problem have involved augmenting these models with a burgeoning number of free parameters. Here I demonstrate an alternative: experimental determination of a parameter-free evolutionary model via mutagenesis, functional selection, and deep sequencing. Using this strategy, I create an evolutionary model for influenza nucleoprotein that describes the gene phylogeny far better than existing models with dozens or even hundreds of free parameters. High-throughput experimental strategies such as the one employed here provide fundamentally new information that has the potential to transform the sensitivity of phylogenetic analyses.

The phylogenetic analysis of gene sequences is one of the most widely used computational procedures in biology. All modern phylogenetic algorithms require an evolutionary model that specifies the rate at which each site substitutes from one identity to another. These evolutionary models can be used to calculate the statistical likelihood of the sequences given a particular phylogenetic tree (*1*). Phylogenetic relationships are typically inferred by finding the tree that maximizes this likelihood (*2*) or by combining the likelihoods with a prior to compute posterior probabilities (*3*).

Actual sequence evolution is governed by the rates at which mutations arise and the subsequent selection that acts upon them (*4, 5*). Unfortunately, neither of these aspects of the evolutionary process are traditionally known *a priori*. The standard approach in molecular phylogenetics is therefore to assume that sites evolve independently and identically, and then construct an evolutionary model that contains free parameters designed to represent features of mutation and selection (*6, 7, 8, 9*). This approach suffers from two major problems. First, although adding parameters enhances a model's fit to data, the parameter values must be estimated from the same sequences that are being analyzed phylogenetically – and so complex models can overfit the data (*10*). Second, even complex models do not contain enough parameters to realistically represent selection, which is highly idiosyncratic to specific sites within a protein. Attempts to predict selection from protein structure have had limited success (*11, 12*), probably because even sophisticated computations cannot consistently predict the impact of mutations (*13*). Approximating selection via additional site-specific parameters can improve phylogenetic fit (*14, 15, 16*), but further exacerbates the proliferation of parameters that already plagues simpler models.

Here I demonstrate a radically different approach for constructing evolutionary models: experimental measurement. This approach requires quantifying both the rates of mutations and the probabilities that mutations fix after they occur. Using nucleoprotein (NP) from influenza A virus as a test case, I show that it is now possible to experimentally quantify these factors with sufficient accuracy to create a parameter-free evolutionary model that fits NP phylogenies far

better than existing parameter-rich models.

Mutation rates were experimentally quantified under the assumption that they are uniform across sites. To separate mutation from selection, I utilized influenza viruses that package GFP in the PB1 segment (*17*). Because the GFP is not under functional selection, substitutions in this gene accumulate at the mutation rate. I passaged 24 replicate populations of GFP viruses by limiting dilution in tissue culture, at each passage serially diluting the virus to the lowest concentration capable of infecting target cells (*18*). Because each limiting dilution bottlenecks the population to one or a few infectious virions, mutations fix rapidly. After 25 rounds of passage, the GFP gene was Sanger sequenced for each replicate to identify 24 substitutions (table S1, table S2), for an overall rate of 5.6×10^{-5} mutations per nucleotide per tissue-culture generation. The rates for different types of mutations are in Table 1, and possess expected features such as an elevation of transitions over transversions.

Quantification of selection is more challenging. Even under the assumption that selection is independent among sites, it is necessary to examine all $\approx 10^4$ amino-acid mutations to NP. This can now be done using "deep mutational scanning," a strategy of mutagenesis, selection, and deep-sequencing developed by Fowler and coworkers (*19*) that has been applied to several genes (*19, 20, 21, 22, 23*). Because many amino-acid mutations are not accessible by single-nucleotide changes, I used a PCR-based strategy to construct codon-mutant libraries that contained multi-nucleotide (i.e. GGC $\rightarrow$ ACT) as well as single-nucleotide (i.e. GGC $\rightarrow$ AGC) mutations. I constructed two mutant libraries of NP from the human H3N2 strain A/Aichi/2/1968 (table S3), and two from a variant of this NP with a single amino-acid substitution (N334H) that enhances protein stability (*24, 25*). Each library contained $> 10^6$ unique plasmid clones, with the number of codon mutations per clone following a Poisson distribution with a mean of 2.7 (fig. S1). Most of the $\approx 10^4$ unique amino-acid mutations therefore occur in multiple clones, both individually and in combination with other mutations.

To assess effects of the mutations on viral replication, the plasmid mutant libraries were used to create pools of mutant influenza viruses by reverse genetics (*26*). The viruses were passaged

twice in tissue culture at low multiplicity of infection to enforce a linkage between genotype and phenotype. The NP gene was reverse-transcribed and PCR-amplified from viral RNA after each passage, and similar PCR amplicons were generated from the plasmid mutant libraries and controls designed to quantify errors associated with sequencing, reverse transcription, and viral passage (Fig. 1A). This process was performed independently for all four mutant libraries to complete *replicate A*. The entire viral rescue, passaging, and sequencing process was then repeated for all four libraries in *replicate B*.

Mutation frequencies were quantified by Illumina sequencing, using overlapping paired-end reads to reduce errors (fig. S2). Each sample produced $\approx 10^7$ paired reads that could be aligned to NP, providing an average of $\approx 5 \times 10^5$ calls per codon (fig. S3). Sequencing of unmutated NP plasmid revealed a low rate of errors, which were almost exclusively single-nucleotide changes (Fig. 1B, fig. S4A). The plasmid mutant libraries contained a high frequency of single and multi-nucleotide codon mutations. Mutation frequencies for RNA or viruses produced from unmutated NP plasmid were only slightly above the sequencing error rate, indicating that reverse-transcription and viral replication introduced few mutations relative to the targeted mutagenesis in the plasmid libraries. Mutation frequencies were reduced in the mutant viruses relative to the plasmids used to create these viruses, particularly for nonsynonymous and stop-codon mutations – consistent with selection purging deleterious mutations.

Nearly all multi-nucleotide codon mutations were found numerous times in the plasmid mutant libraries (Fig. 1C, fig. S4B). More than half of all and $> 80\%$ of synonymous multi-nucleotide codon mutations were found multiple times in the mutant viruses. Because even synonymous mutations are sometimes deleterious to influenza (*27*) and so will be purged by selection, this latter number provides a lower bound for the completeness with which codon mutations were sampled by the mutant viruses created by reverse genetics. Because most amino acids are encoded by multiple codons, the fraction of amino-acid mutations sampled is higher.

Qualitatively, it is obvious that changes in mutation frequencies between the plasmid mutant libraries and the resulting mutant viruses reflect selection. But it is less obvious how to quanti-

tatively analyze this information. Selection acts on the full genomes of all viruses in the population. In contrast, the experiments only measure site-independent mutation frequencies averaged over the population. Assume that each site r has an inherent preference $\pi_{r,a}$ for amino-acid a, with $\sum_a \pi_{r,a} = 1$. The motivation for envisioning site-heterogenous but site-independent amino-acid preferences comes from experiments suggesting that the dominant constraint on mutations that fix during NP evolution relates to protein stability (*24*), and that mutational effects on stability tend to be conserved in a site-independent manner (*25*). The expected frequency $f_{r,x}$ of mutant codon x at site r in the mutant viruses is related to the preference $\pi_{r,\mathcal{A}(x)}$ for its encoded amino-acid $\mathcal{A}(x)$ by $f_{r,x} = \epsilon_{r,x} + \rho_{r,x} + \frac{\mu_{r,x} \times \pi_{r,\mathcal{A}(x)}}{\sum_y \mu_{r,y} \times \pi_{r,\mathcal{A}(y)}}$ where $\mu_{r,x}$ is the frequency that site r is mutagenized to codon x in the plasmid mutant library, $\epsilon_{r,x}$ is the frequency the site is erroneously identified as x during sequencing, $\rho_{r,x}$ is the frequency the site is mutated to x during reverse transcription, y is summed over all codons, and the probability that a site experiences multiple mutations or errors in the same clone is taken to be negligibly small. The observed codon counts are multinomially distributed around these expected frequencies, so by placing a symmetric Dirichlet-distribution prior over $\pi_{r,a}$ and jointly estimating the error and mutation rates from the appropriate samples in Fig. 1A, it is possible to infer the posterior mean for all preferences by MCMC (supporting online text).

The amino-acid preferences inferred from the first and second viral passages within each replicate are extremely similar (Fig. 2A,B), indicating that most selection occurs during initial viral creation and passage, and that technical variation (preparation of samples, stochasticity in sequencing, etc) has little impact. The preferences inferred from the two experimental replicates are substantially but less perfectly correlated (Fig. 2C) – probably because the mutant viruses created by reverse genetics independently in each replicate are different incomplete samples of the $> 10^6$ clones in the plasmid mutant libraries. Nonetheless, the substantial correlation between replicates shows that the sampling is sufficient to clearly reveal inherent preferences. Final preferences were inferred by combining the data from the first passage of both replicates (Fig. 2D, file S1). These preferences are consistent with existing knowledge about NP function

and stability. For example, at the conserved residues in NP's RNA binding interface (*28*), the amino acids found in natural sequences tend to be the ones with the highest preferences (table S4). Similarly, for mutations that have been experimentally characterized as having large effects on NP protein stability (*24, 25*), the stabilizing amino acid has the higher preference (table S5).

The preferences can be combined with the measured mutation rates in a parameter-free site-specific but site-independent evolutionary model. Let the rate of substitution $P_{r,xy} = Q_{xy} \times F_{r,xy}$ from codon x to y at site r be the product of the rate Q_{xy} of this mutation with the site-specific probability $F_{r,xy}$ that the mutation fixes once it occurs. The mutation rate Q_{xy} is given by Table 1 if x and y differ by a single-nucleotide mutation, and is zero otherwise. Here I define the probability of fixation in terms of the preferences as

$$F_{r,xy} = \begin{cases} 1 & \text{if } \mathcal{A}(x) = \mathcal{A}(y) \text{ or } \pi_{r,\mathcal{A}(y)} \geq \pi_{r,\mathcal{A}(x)} \\ \frac{\pi_{r,\mathcal{A}(y)}}{\pi_{r,\mathcal{A}(x)}} & \text{otherwise.} \end{cases}$$

This definition envisions mutations to higher preference amino acids always being tolerated, but those to lower preference amino acids only being tolerated in a fraction of genetic backgrounds (for justification and an alternative definition based on selection coefficients (*5*), see supplementary online text).

The $P_{r,xy}$ values define an experimentally determined evolutionary model with a stationary state giving the expected amino-acid frequencies at equilibrium. These evolutionary equilibrium frequencies (fig. S5, file S2) are different than the amino-acid preferences, since they also depend on the mutation rates and genetic code. For example, if arginine and lysine have equal preferences at a site, arginine will be more evolutionarily abundant since it has more codons.

The experimentally determined evolutionary model can be used to compute phylogenetic likelihoods, thereby enabling its comparison to existing models. I used *codonPhyML* (*29*) to infer maximum-likelihood trees (Fig. 3, fig. S6) for NP sequences from human influenza using the Goldman-Yang (*GY94*) (*6*) and the Kosiol *et al.* (*KOSI07*) (*7*) codon substitution models. In the simplest form, *GY94* contains 11 free parameters (9 equilibrium frequencies plus transition-transversion and synonymous-nonsynonymous ratios), while *KOSI07* contains 62 parameters

(60 frequencies plus transition-transversion and synonymous-nonsynonymous ratios). More complex variants add parameters allowing variation in substitution rate (*8*) or synonymous-nonsynonymous ratio among sites or lineages (*9*). For all these models, *HYPHY* (*30*) was used to calculate the likelihood after optimizing branch lengths for the tree topologies inferred by *codonPhyML*.

Comparison of these likelihoods strikingly validates the superiority of the experimentally determined model (Table 2, table S6, table S7). Adding free parameters generally improves a model's fit to data, and this is true within *GY94* and *KOSI07*. But the parameter-free experimentally determined evolutionary model describes the sequence phylogeny with a likelihood far greater than even the most highly parameterized *GY94* and *KOSI07* variants. Comparison using Aikake information content (AIC) to penalize parameters (*10*) even more emphatically highlights the superiority of the experimentally determined model. There is a clear correlation between the quality and volume of experimental data and the phylogenetic fit: models from individual experimental replicates give lower likelihoods than both replicates combined, and the technically superior *replicate A* (Fig. 1C versus fig. S4) gives a better likelihood than *replicate B*. The superiority of the experimentally determined model is due to measurement of site-specific amino-acid preferences: randomizing preferences among sites gives likelihoods far worse than any other model. A model that interprets the preferences in terms of selection coefficients (*5*) also outperforms *GY94* and *KOSI07*, but by a smaller margin (supplementary online text, table S6, table S7).

These results establish the superiority of the experimentally determined model for NP – but what is the generality of the approach? At first blush, it might seem that the experimental data acquired here is unlikely to ever become available for most situations of interest. However, it is worth remembering that the very gene sequences that are the subjects of molecular phylogenetics were once rare pieces of data – now such sequences are so abundant that they easily overwhelm modern computers. The experimental ease of the deep mutational scanning approach used here is on a comparable trajectory: similar approaches have already been applied

to several proteins (*19, 20, 21, 22, 23*), and there continue to be rapid improvements in techniques for mutagenesis (*31, 32*) and sequencing (*33, 34, 35*). It is therefore especially encouraging that the phylogenetic fit of the NP evolutionary model increases with the quality and volume of experimental data from which it is derived. The increasing availability of such data has the potential to transform phylogenetic analyses by greatly increasing the accuracy of evolutionary models, while at the same time replacing a plethora of free parameters with experimentally measured quantities that can be given clear biological and evolutionary interpretations.

References

1. J. Felsenstein, *Systematic Zoology* **22**, 240 (1973).

2. J. Felsenstein, *J. Mol. Evol.* **17**, 368 (1981).

3. J. P. Huelsenbeck, B. Larget, R. E. Miller, F. Ronquist, *Systematic Biology* **51**, 673 (2002).

4. J. L. Thorne, S. C. Choi, J. Yu, P. G. Higgs, H. Kishino, *Mol. Biol. Evol.* **24**, 1667 (2007).

5. A. L. Halpern, W. J. Bruno, *Mol. Biol. Evol.* **15**, 910 (1998).

6. N. Goldman, Z. Yang, *Mol. Biol. Evol.* **11**, 725 (1994).

7. C. Kosiol, I. Holmes, N. Goldman, *Mol. Biol. Evol.* **24**, 1464 (2007).

8. Z. Yang, *J. Mol. Evol.* **39**, 306 (1994).

9. Z. Yang, R. Nielsen, N. Goldman, A.-M. K. Pedersen, *Genetics* **155**, 431 (2000).

10. D. Posada, T. R. Buckley, *Systematic Biology* **53**, 793 (2004).

11. N. Rodrigue, C. L. Kleinman, H. Philippe, N. Lartillot, *Mol. Biol. Evol.* **26**, 1663 (2009).

12. C. L. Kleinman, N. Rodrigue, N. Lartillot, H. Philippe, *Mol. Biol. Evol.* **27**, 1546 (2010).

13. V. Potapov, M. Cohen, G. Schreiber, *Prot. Eng. Des. Sel.* **22**, 553 (2009).

14. N. Lartillot, H. Philippe, *Mol. Biol. Evol.* **21**, 1095 (2004).

15. S. Q. Le, N. Lartillot, O. Gascuel, *Phil. Trans. R. Soc. B* **363**, 3965 (2008).

16. C.-H. Wu, M. A. Suchard, A. J. Drummond, *Mol. Biol. Evol.* **30**, 669 (2013).

17. J. D. Bloom, L. I. Gong, D. Baltimore, *Science* **328**, 1272 (2010).

18. Materials and methods are available as supplementary material on *Science* online.

19. D. M. Fowler, *et al.*, *Nat. Methods* **7**, 741 (2010).

20. D. Melamed, D. L. Young, C. E. Gamble, C. R. Miller, S. Fields, *RNA* **19**, 1537 (2013).

21. M. W. Traxlmayr, *et al.*, *J. Mol. Biol.* (2012).

22. L. M. Starita, *et al.*, *Proc. Natl. Acad. Sci. USA* **110**, E1263 (2013).

23. B. P. Roscoe, K. M. Thayer, K. B. Zeldovich, D. Fushman, D. N. Bolon, *J. Mol. Biol.* (2013).

24. L. I. Gong, M. A. Suchard, J. D. Bloom, *eLife* **2**, e00631 (2013).

25. O. Ashenberg, L. I. Gong, J. D. Bloom, *Proc. Natl. Acad. Sci. USA* **110**, 21071 (2013).

26. E. Hoffmann, G. Neumann, Y. Kawaoka, G. Hobom, R. G. Webster, *Proc. Natl. Acad. Sci. USA* **97**, 6108 (2000).

27. G. A. Marsh, R. Rabadán, A. J. Levine, P. Palese, *J. Virology* **82**, 2295 (2008).

28. Q. Ye, R. M. Krug, Y. J. Tao, *Nature* **444**, 1078 (2006).

29. M. Gil, M. S. Zanetti, S. Zoller, M. Anisimova, *Mol. Biol. Evol.* **30**, 1270 (2013).

30. S. L. Pond, S. D. Frost, S. V. Muse, *Bioinformatics* **21**, 676 (2005).

31. E. Firnberg, M. Ostermeier, *PLoS One* **7**, e52031 (2012).

32. P. C. Jain, R. Varadarajan, *Analytical Biochemistry* (2013).

33. J. B. Hiatt, R. P. Patwardhan, E. H. Turner, C. Lee, J. Shendure, *Nat. Methods* **7**, 119 (2010).

34. M. W. Schmitt, *et al.*, *Proc. Natl. Acad. Sci. USA* **109**, 14508 (2012).

35. D. I. Lou, *et al.*, *Proc. Natl. Acad. Sci. USA* **110**, 19872 (2013).

Acknowledgments

Thanks to D. Fowler, J. Kitzman, A. Adey, O. Ashenberg, and T. Bedford for helpful discussions. This work was supported by grant R01GM102198 from the NIGMS of the National Institutes of Health. Sequencing data are archived in the SRA at accession SRP036064. Links to computer code are provided in the Supporting Online Material.

mutation type	rate
A $\rightarrow$ G, T $\rightarrow$ C (transition)	2.4×10^{-5}
G $\rightarrow$ A, C $\rightarrow$ T (transition)	2.3×10^{-5}
A $\rightarrow$ C, T $\rightarrow$ G (transversion)	9.0×10^{-6}
C $\rightarrow$ A, G $\rightarrow$ T (transversion)	9.4×10^{-6}
A $\rightarrow$ T, T $\rightarrow$ A (transversion)	3.0×10^{-6}
G $\rightarrow$ C, C $\rightarrow$ G (transversion)	1.9×10^{-6}

Table 1: Influenza mutation rates. Numbers represent the probability a site that has the parent identity will mutate to the specified nucleotide in a single tissue-culture generation, and are calculated from table S1 and table S2 after adding one pseudocount to each mutation type. Mutations are in pairs because an A$\rightarrow$ G can derive from this mutation on the sequenced strand or a T $\rightarrow$ C on the complementary strand.

model	log likelihood	parameters (optimized + empirical)	AIC
experimental, combined replicates	-12338.9	0 (0 + 0)	24677.8
experimental, replicate A	-12372.8	0 (0 + 0)	24745.7
experimental, replicate B	-12392.0	0 (0 + 0)	24783.9
GY94, branch-specific ω, multiple rates	-12769.1	556 (547 + 9)	26650.1
GY94, multiple ω, multiple rates	-12853.9	13 (4 + 9)	25733.8
KOSI07, branch-specific ω, multiple rates	-12891.7	607 (547 + 60)	26997.3
GY94, multiple ω, one rate	-12935.9	12 (3 + 9)	25895.8
KOSI07, multiple ω, multiple rates	-12999.9	64 (4 + 60)	26127.7
GY94, one ω, multiple rates	-13069.8	12 (3 + 9)	26163.5
KOSI07, multiple ω, one rate	-13154.8	63 (3 + 60)	26435.5
KOSI07, one ω, multiple rates	-13191.5	63 (3 + 60)	26508.9
GY94, one ω, one rate	-13205.0	11 (2 + 9)	26431.9
KOSI07, one ω, one rate	-13404.0	62 (2 + 60)	26932.0
randomized experimental, combined replicates	-14209.4	0 (0 + 0)	28418.8
randomized experimental, replicate A	-14243.7	0 (0 + 0)	28487.4
randomized experimental, replicate B	-14259.1	0 (0 + 0)	28518.2

Table 2: Likelihoods computed using various evolutionary models after optimizing the branch lengths for the tree in Fig. 3. Experimentally determined models vastly outperform *GY94* or *KOSI07* despite having no free parameters. Randomizing the experimentally determined amino-acid preferences among sites makes the models far worse. All variants of *GY94* and *KOSI07* contain empirical equilibrium frequencies plus a transition-transversion ratio and synonymous-nonsynonymous ratio (ω) optimized by likelihood. Some variants allow multiple (four discrete gamma categories) rates or ω values, or a different ω for each branch.

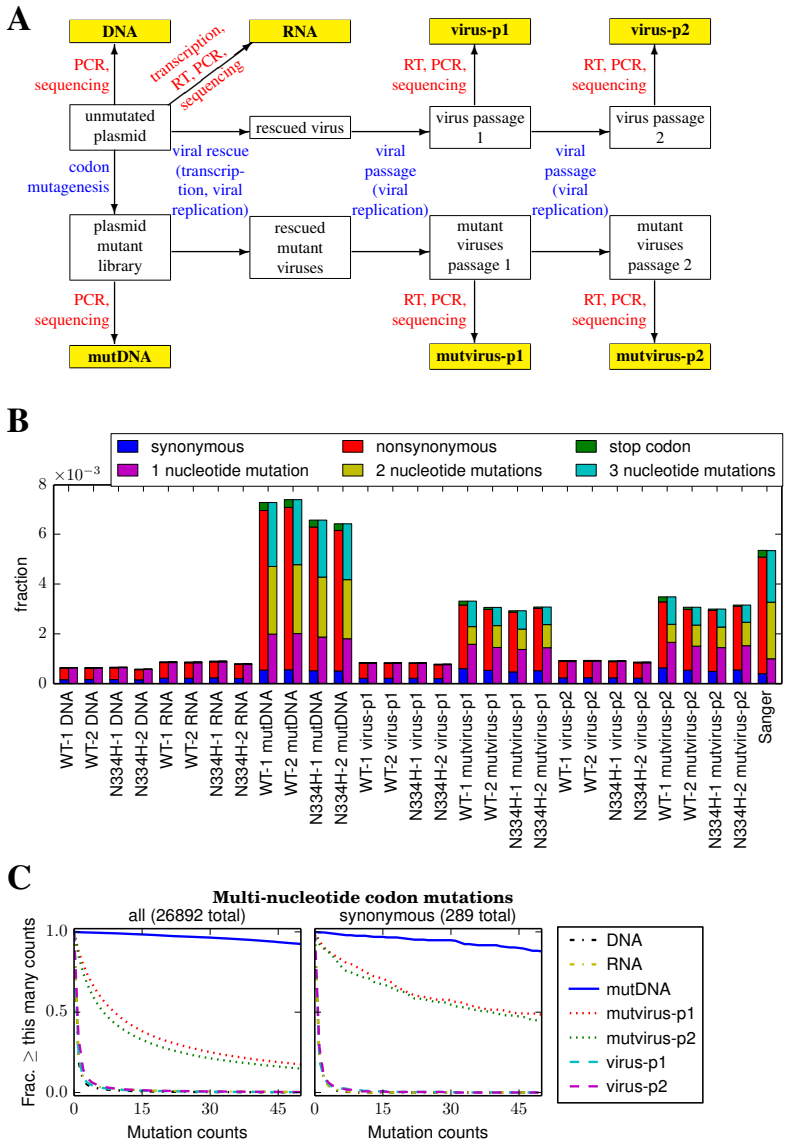


Fig. 1: Measurement of mutational effects. **(A)** Experimental design, with sequenced samples in yellow. Blue indicates sources of mutation and selection; red indicates sources of errors. **(B)** Per-codon mutation frequencies for the four libraries in *replicate A*. Mutations due to sequencing, reverse-transcription, and viral replication are rare, and are mostly single nucleotide. Mutations are reduced in *mutvirus* relative to *mutDNA* due to selection against stop-codon and nonsynonymous mutations. At right are frequencies in the pooled plasmid mutant libraries estimated from Sanger sequencing (fig. S1), which is not affected by Illumina sequencing errors. **(C)** Number of times multi-nucleotide codon mutations (which are rarely generated by errors) are identified in the combined libraries of *replicate A*. Nearly all mutations are sampled in *mutDNA*. Most synonymous mutations are sampled in *mutvirus* – the frequency of all mutations is lower, probably due to selection. Comparable data for *replicate B* are in fig. S4.

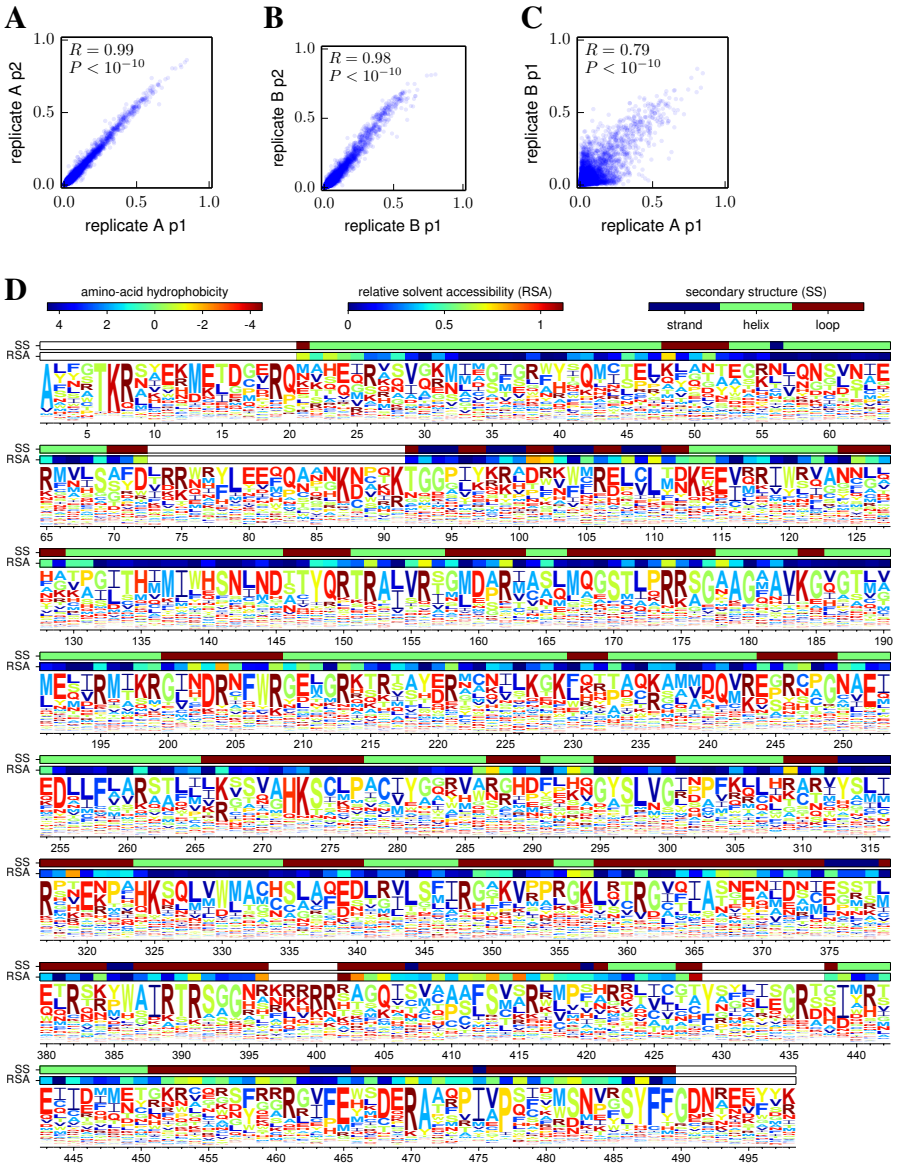


Fig. 2: Amino-acid preferences. **(A), (B)** Preferences inferred from passage 1 and 2 are similar within each replicate, indicating that most selection occurs during initial viral creation and passage, and that technical variation is small. **(C)** Preferences from the two independent replicates are also correlated, but less perfectly. The increased variation is presumably due to stochasticity during the independent viral creation from plasmids for each replicate. **(D)** Preferences for all sites in NP (the N-terminal Met was not mutagenized) inferred from passage 1 of the combined replicates. Letters have heights proportional to the preference for that amino acid, and are colored by hydrophobicity. Relative solvent accessibility and secondary structure are overlaid for residues in crystal structure. Correlation plots show Pearson's R and P-value.

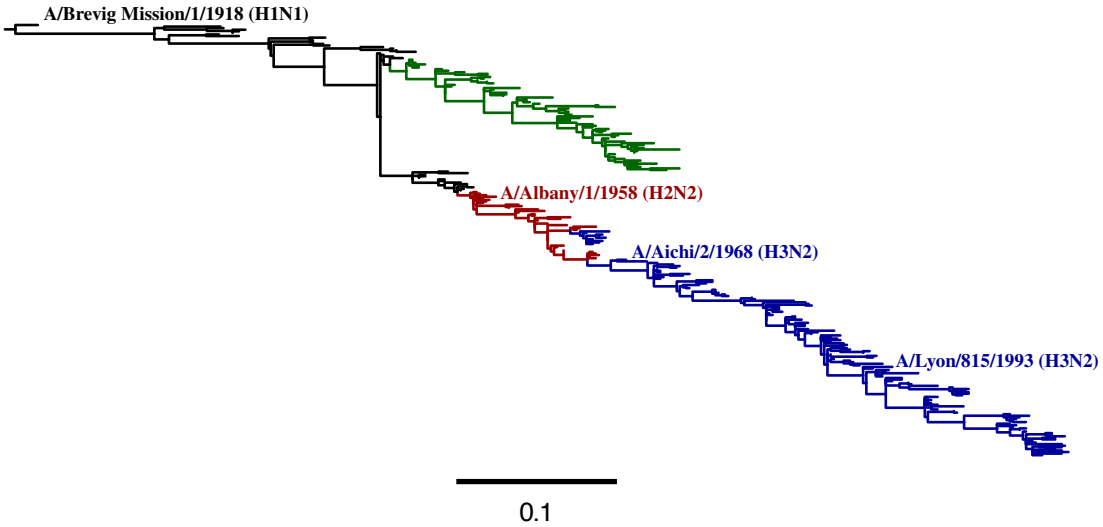


Fig. 3: Phylogenetic tree of NPs from human influenza descended from a close relative of the 1918 virus. Maximum-likelihood tree constructed using *GY94* with up to three sequences per year from each subtype. Black - H1N1 from 1918 lineage; green - seasonal H1N1; red - H2N2; blue - H3N2.

Supplementary materials

Supplementary text

Availability of data and computer code . . . 18
Experimental measurement of mutation rates . . . 19
Construction of NP codon-mutant libraries . . . 21
Viral growth and passage . . . 24
Sample preparation and Illumina sequencing . . . 26
Read alignment and quantification of mutation frequencies . . . 29
Inference of the amino-acid preferences . . . 30
The experimentally determined evolutionary model . . . 35
Phylogenetic analyses . . . 38
References for supplementary material . . . 40

Supplementary tables

table S1 Mutations from limiting-dilution passages . . . 43
table S2 Mutation counts from limiting dilution passages . . . 44
table S3 Sequence of A/Aichi/2/1968 (H3N2) NP . . . 45
table S4 Amino-acid preferences for residues in NP RNA-binding groove . . . 46
table S5 Amino-acid preferences for mutations with known stability effects . . . 47
table S6 Likelihoods for tree estimated with *GY94* . . . 48
table S7 Likelihoods for tree estimated with *KOSI07* . . . 49

Supplementary figures

fig. S1 Sanger sequencing of clones from plasmid mutant libraries . . . 50
fig. S2 Overlapping paired-end Illumina sequencing . . . 51
fig. S3 Illumina deep-sequencing read depth . . . 52
fig. S4 Mutational scanning for *replicate B* . . . 53
fig. S5 Evolutionary equilibrium amino-acid frequencies . . . 54
fig. S6 Tree inferred using *KOSI07* model . . . 55

Supplementary files

file S1 Inferred amino-acid preferences . . . 56
file S2 Evolutionary equilibrium amino-acid frequencies . . . 57

Availability of data and computer code

- The sequencing data described in this manuscript are available at the SRA under accession SRP036064 at `http://www.ncbi.nlm.nih.gov/sra/?term=SRP036064`.
- A detailed description of the computational process used to analyze the sequencing data and infer the amino-acid preferences is at `http://jbloom.github.io/mapmuts/example_2013Analysis_Influenza_NP_Aichi68.html`. The source code itself is at `https://github.com/jbloom/mapmuts`. NOTE: the link to the description is accessible now; access to the source code itself will be made publicly available upon publication of the paper. If you are a reviewer and would like to see the source code, please contact the editor to relay the request for access to the author.
- A detailed description of the computational process used for the phylogenetic analyses is available at `http://jbloom.github.io/phyloExpCM/example_2013Analysis_Influenza_NP_Human_1918_Descended.html`. The source code itself is at `https://github.com/jbloom/phyloExpCM`. NOTE: the link to the description is accessible now; access to the source code itself will be made publicly available upon publication of the paper. If you are a reviewer and would like to see the source code, please contact the editor to relay the request for access to the author.

Experimental measurement of mutation rates

The mutational rates of the influenza polymerase were measured by repeated limiting-dilution passage of replicate viral populations. A challenge in quantifying mutational rates is that the accumulation of substitutions is due to a combination of the inherent mutational process and the selection that acts on mutations after they occur. To avoid the confounding effects of selection, I utilized a previously described (*17*) system for packaging GFP in the PB1 gene. The GFP gene is not under functional selection as it is not needed for viral growth.

I initially generated GFP-carrying viruses with all genes derived from A/WSN/1933 (H1N1) by reverse genetics as described previously (*17*). These viruses were repeatedly passaged at limiting dilution in MDCK-SIAT1-CMV-PB1 cells (*17*) using what will be referred to as *low serum media* (Opti-MEM I with 0.5% heat-inactivated fetal bovine serum, 0.3% BSA, 100 U/ml penicillin, 100 μg/ml streptomycin, and 100 μg/ml calcium chloride) – a moderate serum concentration was retained and no trypsin was added because viruses with the WSN HA and NA are trypsin independent (*36*). These passages were performed for 27 replicate populations. For each passage, a 100 μl volume containing the equivalent of 2 μl of virus collection was added to the first row of a 96-well plate. The virus was then serially diluted 1:5 down the plate such that at the conclusion of the dilutions, each well contained 80 μl of virus dilution, and the final row contained the equivalent of $(0.8 \times 2\,\mu\text{l})\,/5^7 \approx 2 \times 10^{-5}$ μl of the original virus collection. MDCK-SIAT1-CMV-PB1 cells were then added to each well in a 50 μl volume containing 2.5×10^3 cells. The plates were grown for approximately 80 hours, and wells were examined for cytopathic effect indicative of viral growth. Most commonly the last well with clear cytopathic effect was in row E or F (corresponding to an infection volume of between $(0.8 \times 2\,\mu\text{l})\,/5^4 \approx 3 \times 10^{-3}$ and $(0.8 \times 2\,\mu\text{l})\,/5^5 \approx 5 \times 10^{-4}$ μl, although this varied somewhat among the replicates and passages. The last well with cytopathic effect was collected and used as the parent population for the next round of limiting-dilution passage.

The goal of the limiting-dilution passages was to repeatedly bottleneck the population so mutations would rapidly go to fixation. After 25 limiting-dilution passages, 10 of the 27 viral populations no longer caused any visible GFP expression in the cells in which they caused cytopathic effect, indicating fixation of a mutation that ablated GFP fluorescence. The 17 remaining populations all caused fluorescence in infected cells, although in some cases the intensity was visibly reduced – these populations therefore must have retained at least a partially functional GFP. Total RNA was purified from each viral population, the PB1 segment was reverse-transcribed using the primers `CATGATCGTCTCGTATTAGTAGAAACAAGGCATTTTTT CATGAAGGACAAGC` and `CATGATCGTCTCAGGGAGCGAAAGCAGGCAAACCATTTGATTGG`, and the reverse-transcribed cDNA was amplified by conventional PCR using the same primers. For 22 of the 27 replicate viral populations, this process amplified an insert with the length expected for the full GFP-carrying PB1 segment. For 2 of the replicates, this amplified inserts between 0.4 and 0.5 kb shorter than the expected length, suggesting an internal deletion in part of the segment. For 3 replicates, this failed to amplify any insert, suggesting total loss of the

GFP-carrying PB1 segment, a very large internal deletion, or rearrangement that rendered the reverse-transcription primers ineffective. For the 24 replicates from which an insert could be amplified, the GFP coding region of the PCR product was Sanger sequenced to determine the consensus sequence. The results are in table S1.

Overall, the 24 replicate viral populations accumulated 24 nucleotide substitutions in the 720-nucleotide GFP over the 25 passages (table S2). This corresponds to a mutation rate of 5.6×10^{-5} mutations per nucleotide per tissue-culture generation – very close to the estimate of 7.6×10^{-5} mutations per nucleotide per tissue-culture generation reported by Parvin, Palese, and coworkers (*37*). There is a strong bias for transition mutations over transversions.

The data in table S2 can be used to calculate the rate of mutation from nucleotide m to nucleotide n, given that the parent nucleotide is already m. Let m_c to denote the complement of nucleotide m (so for example A_c is T). An observed change of m to n on the template strand can arise from either a mutation of m to n on this strand or a mutation of m_c to n_c on the complementary strand during viral replication – that is, A $\rightarrow$ G is indistinguishable from T $\rightarrow$ C, since a change of A $\rightarrow$ G on the template strand could also be induced by a T $\rightarrow$ C change on the complementary strand. Therefore, it is only possible to estimate the combined rate $R_{m\rightarrow n} + R_{m_c\rightarrow n_c}$. Assuming that the same mutational processes operate on the template and complementary strand, then the rate at which these two changes will appear on the template strand are equal, so $R_{m\rightarrow n} = R_{m_c\rightarrow n_c}$.

To estimate $R_{m\rightarrow n}$ from table S2, it is necessary to normalize by the nucleotide composition of the GFP gene. The numbers of each of the four types of nucleotides in this gene are: $N_A = 175$, $N_T = 103$, $N_C = 241$, and $N_G = 201$. Given that the counts in table S2 come after 25 passages of 24 replicates:

$$R_{m\rightarrow n} = R_{m_c\rightarrow n_c} = \frac{1}{24 \times 25 \times 2} \times \frac{N_{m\rightarrow n} + N_{m_c\rightarrow n_c} + \mathcal{C}}{N_m + N_{m_c}} \tag{1}$$

where $\mathcal{C}$ is a pseudocount that is added to the observed counts to avoid estimating rates of zero and $N_{m\rightarrow n}$ is the number of observed mutations from m to n in table S2. The values of $R_{m\rightarrow n}$ estimated from Equation 1 give the probability that a given nucleotide that is already m will mutate to n in a single tissue-culture generation. Using a value of $\mathcal{C} = 1$ (a pseudocount of one) gives the rates shown in Table 1. Note that the measured mutation rates are very close to being reversible, since $R_{m\rightarrow n} + R_{m_c\rightarrow n_c} \approx R_{n\rightarrow m} + R_{n_c\rightarrow m_c}$ for all nucleotides. For the evolutionary models described below, I enforce reversibility by defining the rates to be the average of the two reversible mutations: $R_{m\rightarrow n} + R_{m_c\rightarrow n_c} = R_{n\rightarrow m} + R_{n_c\rightarrow m_c} = \left(R_{m\rightarrow n} + R_{m_c\rightarrow n_c} + R_{n\rightarrow m} + R_{n_c\rightarrow m_c}\right)/2$.

Construction of NP codon-mutant libraries

The goal was to construct a mutant library with an average of two to three random codon mutations per gene. Most techniques for creating mutant libraries of full-length genes, such as error-prone PCR (*38*) and chemical mutagenesis (*39*), introduce mutations at the nucleotide level, meaning that codon substitutions involving multiple nucleotide changes occur at a negligible rate. There are established methods for mutating specific sites at the codon level with synthetic oligonucleotides that contain triplets of randomized NNN nucleotides (*40*). However, these oligo-based codon mutagenesis techniques are typically only suitable for mutating a small number of sites. Recently, several groups have developed strategies for introducing codon mutations along the lengths of entire genes (*31, 32, J. Kitzman and J. Shendure - personal communication*). Most of these strategies are designed to create exactly one codon mutation per gene. For the purpose of my experiments, it was actually desirable to introduce a distribution of around one to four codon mutations per gene to examine the average effects of mutations in a variety of closely related genetic backgrounds. Therefore, I devised a codon-mutagenesis protocol specifically for this purpose.

This technique involved iterative rounds of low-cycle PCR with pools of mutagenic synthetic oligonucleotides that each contain a randomized NNN triplet at a specific codon site. Two replicate libraries each of the wildtype and N334H variants of the Aichi/1968 NP were prepared in full biological duplicate, beginning each with independent preps of the plasmid template. To avoid cross-contamination, all purification steps used an independent gel for each sample, with the relevant equipment thoroughly washed to remove residual DNA.

First, for each codon except for that encoding the initiating methionine in the 498-residue NP gene (table S3), I designed an oligonucleotide that contained a randomized NNN nucleotide triplet preceded by the 16 nucleotides upstream of that codon in the NP gene and followed by the 16 nucleotides downstream of that codon in the NP gene. I ordered these oligonucleotides in 96-well plate format from Integrated DNA Technologies, and combined them in equimolar quantities to create the *forward-mutagenesis* primer pool. I also designed and ordered the reverse complement of each of these oligonucleotides, and combined them in equimolar quantities to create the *reverse-mutagenesis* pool. The primers for the N334H variants differed only for those that overlapped the N334H codon. I also designed end primers that annealed to the termini of the NP sequence and contained sites appropriate for BsmBI cloning into the influenza reverse-genetics plasmid pHW2000 (*26*). These primers are *5'-BsmBI-Aichi68-NP* (catgatcgtctcagggagcaaaagcagggtagataatcactcacag) and *3'-BsmBI-Aichi68-NP* (catgatcgtctcgtattagtagaaacaagggtatttttcttta).

I set up PCR reactions that contained 1 μl of 10 ng/μl template pHWAichi68-NP plasmid (*24*), 25 μl of 2X KOD Hot Start Master Mix (product number 71842, EMD Millipore), 1.5 μl each of 10 μM solutions of the end primers *5'-BsmBI-Aichi68-NP* and *3'-BsmBI-Aichi68-NP*, and 21 μl of water. I used the following PCR program (referred to as the *amplicon PCR program* in the remainder of this paper):

1. 95 °C for 2 minutes
2. 95°C for 20 seconds
3. 70°C for 1 second
4. 50°C for 30 seconds cooling to 50°C at 0.5°C per second.
5. 70°C for 40 seconds
6. Repeat steps 2 through 5 for 24 additional cycles
7. Hold 4°C

The PCR products were purified over agarose gels using ZymoClean columns (product number D4002, Zymo Research) and used as templates for the initial codon mutagenesis fragment PCR.

Two fragment PCR reactions were run for each template. The forward-fragment reactions contained 15 μl of 2X KOD Hot Start Master Mix, 2 μl of the *forward mutagenesis* primer pool at a total oligonucleotide concentration of 4.5 μM, 2 μl of 4.5 μM *3'-BsmBI-Aichi68-NP*, 4 μl of 3 ng/μl of the aforementioned gel-purified linear PCR product template, and 7 μl of water. The reverse-fragment reactions were identical except that the *reverse mutagenesis* pool was substituted for the *forward mutagenesis* pool, and that *5'-BsmBI-Aichi68-NP* was substituted for *3'-BsmBI-Aichi68-NP*. The PCR program for these fragment reactions was identical to the *amplicon PCR program* except that it utilized a total of 7 rather than 25 thermal cycles.

The products from the fragment PCR reactions were diluted 1:4 in water. These dilutions were then used for the joining PCR reactions, which contained 15 μl of 2X KOD Hot Start Master Mix, 4 μl of the 1:4 dilution of the forward-fragment reaction, 4 μl of the 1:4 dilution of the reverse-fragment reaction, 2 μl of 4.5 μM *5'-BsmBI-Aichi68-NP*, 2 μl of 4.5 μM *3'-BsmBI-Aichi68-NP*, and 3 μl of water. The PCR program for these joining reactions was identical to the *amplicon PCR program* except that it utilized a total of 20 rather than 25 thermal cycles. The products from these joining PCRs were purified over agarose gels.

The purified products of the first joining PCR reactions were used as templates for a second round of fragment reactions followed by joining PCRs. These second-round products were used as templates for a third round. The third-round products were purified over agarose gels, digested with BsmBI (product number R0580L, New England Biolabs), and ligated into a dephosphorylated (Antarctic Phosphatase, product number M0289L, New England Biolabs) BsmBI digest of pHW2000 (*26*) using T4 DNA ligase. The ligations were purified using ZymoClean columns, electroporated into ElectroMAX DH10B T1 phage-resistant competent cells (product number 12033-015, Invitrogen) and plated on LB plates supplemented with 100 μg/ml of ampicillin. These transformations yielded between 400,000 and 800,000 unique transformants per plate, as judged by plating a 1:4,000 dilution of the transformations on a second set of plates. Transformation of a parallel no-insert control ligation yielded approximately 50-fold fewer colonies, indicating that self ligation of pHW2000 only accounts for a small fraction of the transformants. For each library, I performed three transformations, grew the plates overnight, and then scraped the colonies into liquid LB supplemented with ampicillin and mini-prepped several hours later to yield the plasmid mutant libraries. These libraries each contained in excess

of 10^6 unique transformants, most of which will be unique codon mutants of the NP gene.

I sequenced the NP gene for 30 individual clones drawn from the four mutant libraries. As shown in fig. S1, the number of mutations per clone was approximately Poisson distributed and the mutations occurred uniformly along the primary sequence. If all codon mutations are made with equal probability, 9/63 of the mutations should be single-nucleotide changes, 27/63 should be two-nucleotide changes, and 27/63 should be three-nucleotide changes. This is approximately what was observed in the Sanger-sequenced clones. The nucleotide composition of the mutated codons was roughly uniform, and there was no tendency for clustering of multiple mutations in primary sequence. The full results of the Sanger sequencing and analysis are `https://github.com/jbloom/SangerMutantLibraryAnalysis/tree/v0.1`. The results of this Sanger sequencing are compatible with the mutation frequencies obtained from deep sequencing the **mutDNA** samples after subtracting off the sequencing error rate estimated from the **DNA** samples (Fig. 1B), especially considering that the statistics from the Sanger sequencing are subject to sampling error due to the limited number of clones analyzed.

Viral growth and passage

As described in the main text, two independent replicates of viral growth and passage were performed (*replicate A* and *replicate B*). The experimental procedures were mostly identical between replicates, but there were a few small differences. In the actual experimental chronology, *replicate B* was performed first, and the modifications in *replicate A* were designed to improve the sampling of the mutant plasmids by the created mutant viruses. These modifications may be the reason why *replicate A* slightly outperforms *replicate B* by two objective measures: the viruses more completely sample the codon mutations (Fig. 1 versus fig. S4), and the evolutionary model derived solely from *replicate A* gives a higher likelihood than the evolutionary model derived solely from *replicate B* (Table 2, table S6, table S7).

For *replicate B*, I used reverse genetics to rescue viruses carrying the Aichi/1968 NP or one of its derivatives, PB2 and PA from the A/Nanchang/933/1995 (H3N2), a PB1 gene segment encoding GFP, and HA / NA / M / NS from A/WSN/1933 (H1N1) strain. With the exception of the variants of NP used, these viruses are identical to those described in (*24*), and were rescued by reverse genetics in 293T-CMV-Nan95-PB1 and MDCK-SIAT1-CMV-Nan95-PB1 cells as described in that reference. The previous section describes four NP codon-mutant libraries, two of the wildtype Aichi/1968 gene (WT-1 and WT-2) and two of the N334H variant (N334H-1 and N334H-2). I grew mutant viruses from all four mutant libraries, and four paired unmutated viruses from independent preps of the parent plasmids. A major goal was to maintain diversity during viral creation by reverse genetics – the experiment would obviously be undermined if most of the rescued viruses derived from a small number of transfected plasmids. I therefore performed the reverse genetics in 15 cm tissue-culture dishes to maximize the number of transfected cells. Specifically, 15 cm dishes were seeded with 10^7 293T-CMV-Nan95-PB1 cells in D10 media (DMEM with 10% heat-inactivated fetal bovine serum, 2 mM L-glutamine, 100 U/ml penicillin, and 100 μg/ml streptomycin). At 20 hours post-seeding, the dishes were transfected with 2.8 μg of each of the eight reverse-genetics plasmids. At 20 hours post-transfection, about 20% of the cells expressed GFP (indicating transcription by the viral polymerase of the GFP encoded by pHH-PB1flank-eGFP), suggesting that many unique cells were transfected. At 20 hours post-transfection, the media was changed to the *low serum media* described above. At 78 hours post-transfection, the viral supernatants were collected, clarified by centrifugation at 2000xg for 5 minutes, and stored at $4^\circ\mathrm{C}$. The viruses were titered by flow cytometry as described previously (*24*) – titers were between 1×10^3 and 4×10^3 infectious particles per μl. A control lacking the NP gene yielded no infectious virus as expected.

The virus was then passaged in MDCK-SIAT1-CMV-Nan95-PB1 cells. These cells were seeded into 15 cm dishes, and when they had reached a density of 10^7 per plate they were infected with 10^6 infectious particles (MOI of 0.1) of the transfectant viruses in *low serum media*. After 18 hours, 30-50% of the cells were green as judged by microscopy, indicating viral spread. At 40 hours post-transfection, 100% of the cells were green and many showed clear signs of cytopathic effect. At this time the viral supernatants were again collected, clarified,

and stored at $4°\mathrm{C}$. NP cDNA isolated from these viruses was the source the deep-sequencing samples **virus-p1** and **mutvirus-p1** in Fig. 1A. The virus was then passaged a second time exactly as before (again using an MOI of 0.1). NP cDNA from these twice-passaged viruses constituted the source for the samples **virus-p2** and **mutvirus-p2** in Fig. 1A.

For *replicate A*, all viruses (both the four mutant viruses and the paired unmutated controls) were re-grown independently from the same plasmids preps used for *replicate B*. The experimental process was identical to that used for *replicate B* except for the following:

- Standard influenza viruses (rather than the GFP-carrying variants) were used, so plasmid pHW-Nan95-PB1 (*24*) was substituted for pHH-PB1flank-eGFP during reverse genetics, and 293T and MDCK-SIAT1 cells were substituted for the PB1-expressing variants.

- Rather than creating the viruses by transfecting a single 15-cm dish, each of the eight virus samples was created by transfecting two 12-well dishes, with the dishes seeded at 3×10^5 293T and 5×10^4 MDCK-SIAT1 cells prior to transfection. The passaging was then done in four 10 cm dishes for each sample, with the dishes seeded at 4×10^6 MDCK-SIAT1 cells 12-14 hours prior to infection. The passaging was still done at an MOI of 0.1. These modifications were designed to increase diversity in the viral population.

- The viruses were titered by TCID50 rather than flow cytometry.

Sample preparation and Illumina sequencing

For each sample, a PCR amplicon was created to serve as the template for Illumina sequencing. The steps used to generate the PCR amplicon for each of the seven sample types (Fig. 1A) are listed below. Once the PCR template was generated, for all samples the PCR amplicon was created using the *amplicon PCR program* described above in 50 μl reactions consisting of 25 μl of 2X KOD Hot Start Master Mix, 1.5 μl each of 10 μM of *5'-BsmBI-Aichi68-NP* and *3'-BsmBI-Aichi68-NP*, the indicated template, and ultrapure water. A small amount of each PCR reaction was run on an analytical agarose gel to confirm the desired band. The remainder was then run on its own agarose gel without any ladder (to avoid contamination) after carefully cleaning the gel rig and all related equipment. The amplicons were excised from the gels, purified over ZymoClean columns, and analyzed using a NanoDrop to ensure that the absorbance at 260 nm was at least 1.8 times that at 230 nm and 280 nm. The templates were as follows:

- **DNA**: The templates for these amplicons were 10 ng of the unmutated independent plasmid preps used to create the codon mutant libraries.

- **mutDNA**: The templates for these amplicons were 10 ng of the plasmid mutant libraries.

- **RNA**: This amplicon quantifies the net error rate of transcription and reverse transcription. Because the viral RNA is initially transcribed from the reverse-genetics plasmids by RNA polymerase I but the bidirectional reverse-genetics plasmids direct transcription of RNA by both RNA polymerases I and II (*26*), the RNA templates for these amplicons were transcribed from plasmids derived from pHH21 (*41*), which only directs transcription by RNA polymerase I. The unmutated WT and N334H NP genes were cloned into this plasmid to create pHH-Aichi68-NP and pHH-Aichi68-NP-N334H. Independent preparations of these plasmids were transfected into 293T cells, transfecting 2 μg of plasmid into 5×10^5 cells in 6-well dishes. After 32 hours, total RNA was isolated using Qiagen RNeasy columns and treated with the Ambion TURBO DNA-free kit (Applied Biosystems AM1907) to remove residual plasmid DNA. This RNA was used as a template for reverse transcription with AccuScript (Agilent 200820) using the primers *5'-BsmBI-Aichi68-NP* and *3'-BsmBI-Aichi68-NP*. The resulting cDNA was quantified by qPCR specific for NP (see below), which showed high levels of NP cDNA in the reverse-transcription reactions but undetectable levels in control reactions lacking the reverse transcriptase, indicating that residual plasmid DNA had been successfully removed. A volume of cDNA that contained at least 2×10^6 NP cDNA molecules (as quantified by qPCR) was used as template for the amplicon PCR reaction. Control PCR reactions using equivalent volumes of template from the no reverse-transcriptase control reactions yielded no product.

- **virus-p1**: This amplicon was derived from virus created from the unmutated plasmid and collected at the end of the first passage. Clarified virus supernatant was ultracentrifuged

at 64,000 x g for 1.5 hours at 4 °C, and the supernatant was decanted. Total RNA was then isolated from the viral pellet using a Qiagen RNeasy kit. This RNA was used as a template for reverse transcription with AccuScript using the primers *5'-BsmBI-Aichi68-NP* and *3'-BsmBI-Aichi68-NP*. The resulting cDNA was quantified by qPCR, which showed high levels of NP cDNA in the reverse-transcription reactions but undetectable levels in control reactions lacking the reverse transcriptase. A volume of cDNA that contained at least 10^7 NP cDNA molecules (as quantified by qPCR) was used as template for the amplicon PCR reaction. Control PCR reactions using equivalent volumes of template from the no reverse-transcriptase control reactions yielded no product.

- **virus-p2, mutvirus-p1, mutvirus-p2**: These amplicons were created as for the **virus-p1** amplicons, but used the appropriate virus as the initial template as outlined in Fig. 1A.

An important note: it was found that the use of relatively new RNeasy kits with β-mercaptoethanol (a reducing agent) freshly added per the manufacturer's instructions was necessary to avoid what appeared to be oxidative damage to purified RNA.

The overall experiment only makes sense if the sequenced NP genes derive from a large diversity of initial template molecules. Therefore, qPCR was used to quantify the molecules produced by reverse transcription to ensure that a sufficiently large number were used as PCR templates to create the amplicons. The qPCR primers were *5'-Aichi68-NP-for* (`gcaacagctggtctgactcaca`) and *3'-Aichi68-NP-rev* (`tccatgccggtgcgaacaag`). The qPCR reactions were performed using the SYBR Green PCR Master Mix (Applied Biosystems 4309155) following the manufacturer's instructions. Linear NP PCR-ed from the pHWAichi68-NP plasmid was used as a quantification standard – the use of a linear standard is important, since amplification efficiencies differ for linear and circular templates (*42*). The standard curves were linear with respect to the amount of NP standard over the range from 10^2 to 10^9 NP molecules. These standard curves were used to determine the absolute number of NP cDNA molecules after reverse transcription. Note that the use of only 25 thermal cycles in the *amplicon PCR program* provides a second check that there are a substantial number of template molecules, as this moderate number of thermal cycles will not lead to sufficient product if there are only a few template molecules.

In order to allow the Illumina sequencing inserts to be read in both directions by paired-end 50 nt reads (fig. S2), it was necessary to us an Illumina library-prep protocol that created NP inserts that were roughly 50 nt in length. This was done via a modification of the Illumina Nextera protocol. First, concentrations of the PCR amplicons were determined using PicoGreen (Invitrogen P7859). These amplicons were used as input to the Illumina Nextera DNA Sample Preparation kit (Illumina FC-121-1031). The manufacturer's protocol for the tagmentation step was modified to use 5-fold less input DNA (10 ng rather than 50 ng) and two-fold more tagmentation enzyme (10 μl rather than 5 μl), and the incubation at 55 °C was doubled from 5 minutes to 10 minutes. Samples were barcoded using the Nextera Index Kit for 96-indices (Illumina FC-121-1012). For index 1, the barcoding was: **DNA** with N701, **RNA** with N702, **mutDNA** with N703, **virus-p1** with N704, **mutvirus-p1** with N705, **virus-p2** with N706, and **mutvirus-p2**

with N707. After completion of the Nextera PCR, the samples were subjected to a ZymoClean purification rather than the bead cleanup step specified in the Nextera protocol. The size distribution of these purified PCR products was analyzed using an Agilent 200 TapeStation Instrument. If the NP sequencing insert is exactly 50 nt in size, then the product of the Nextera PCR should be 186 nt in length after accounting for the addition of the Nextera adaptors. The actual size distribution was peaked close to this value. The ZymoClean-purified PCR products were quantified using PicoGreen and combined in equal amounts into pools: a WT-1 pool of the seven samples for that library, a WT-2 pool of the seven samples for that library, etc. These pools were subjected to further size selection by running them on a 4% agarose gel versus a custom ladder containing 171 and 196 nt bands created by PCR from a GFP template using the forward primer `gcacggggccgtcgccg` and the reverse primers `tggggcacaagctggagtacaac` (for the 171 nt band) and `gacttcaaggaggacggcaacatcc` (for the 196 nt band). The gel slice for the sample pools corresponding to sizes between 171 and 196 nt was excised, and purified using a ZymoClean column. A separate clean gel was run for each pool to avoid cross-contamination.

Library QC and cluster optimization were performed using Agilent Technologies qPCR NGS Library Quantification Kit (Agilent Technologies, Santa Clara, CA, USA). Libraries were introduced onto the flow cell using an Illumina cBot (Illumina, Inc., San Diego, CA, USA) and a TruSeq Rapid Duo cBot Sample Loading Kit. Cluster generation and deep sequencing was performed on an Illumina HiSeq 2500 using an Illumina TruSeq Rapid PE Cluster Kit and TruSeq Rapid SBS Kit. A paired-end, 50 nt read-length (PE50) sequencing strategy was performed in rapid run mode. Image analysis and base calling were performed using Illumina's Real Time Analysis v1.17.20.0 software, followed by demultiplexing of indexed reads and generation of FASTQ files, using Illumina's CASAVA v1.8.2 software (`http://www.illumina.com/software.ilmn`). These FASTQ files were uploaded to the SRA under accession SRP036064 (see `http://www.ncbi.nlm.nih.gov/sra/?term=SRP036064`).

Read alignment and quantification of mutation frequencies

A custom Python software package was created to quantify the frequencies of mutations from the Illumina sequencing. A detailed description of the algorithm is at `http://jbloom.github.io/mapmuts/example_2013Analysis_Influenza_NP_Aichi68.html`. The source code is at `https://github.com/jbloom/mapmuts`. Briefly:

1. Reads were discarded if either read in a pair failed the Illumina chastity filter, had a mean Q-score less than 25, or had more than two ambiguous (`N`) nucleotides.

2. The remaining paired reads were aligned to each other, and retained only if they shared at least 30 nt of overlap, disagreed at no more than one site, and matched the expected terminal Illumina adaptors with no more than one mismatch.

3. The overlap of the paired reads was aligned to NP, disallowing alignments with gaps or more than six nucleotide mismatches. A small fraction of alignments corresponded exclusively to the noncoding termini of the viral RNA; the rest contained portions of the NP coding sequence.

4. For every paired read that aligned with NP, the codon identity was called if both reads concurred for all three nucleotides in the codon. If the reads disagreed or contained an ambiguity in that codon, the identity was not called.

A summary of alignment statistics and read depth is in fig. S3; The mutation frequencies are summarized in Fig. 1B and fig. S4. Full data can be accessed via `http://jbloom.github.io/mapmuts/example_2013Analysis_Influenza_NP_Aichi68.html`. The apparent mutation frequency in the **DNA** samples quantifies the combination of the PCR and sequencing errors that affects all samples. The apparent mutation frequencies are not much higher in the **RNA**, **virus-p1**, and **virus-p2** samples – indicating that reverse transcription and viral replication introduce only a low frequency of additional errors. The sum of all of these errors is much less than the targeted mutagenesis rate in the **mutDNA** samples, meaning that selection can be clearly distinguished.

The completeness with which mutations were sampled is most easily discerned by examining the number of times that multi-nucleotide codon mutations are identified (this is more straightforward than looking at single-nucleotide codon mutations, since the latter are confounded by sequencing errors). Fig. 1C and fig. S4 show that nearly all mutations were present in the **mutDNA** samples. When examining the **mutvirus** samples, the analysis is confounded by the fact that mutations could be absent because they were never packaged into viruses, or because they were purged by selection. It is therefore informative to look at synonymous mutations which are usually – but not always (*27*) – relatively neutral. This examination shows that most synonymous mutations are sampled repeatedly by the mutant viruses, and that *replicate A* seems to be superior to *replicate B* in this respect – perhaps because of the experimental modifications designed to improve the viral diversity during reverse genetics for *replicate A*.

Inference of the amino-acid preferences

The approach described here is based on the assumption that there is an inherent preference for each amino acid at each site in the protein. This assumption is clearly not completely accurate, as the effect of a mutation at one site can be influenced by the identities of other sites. However, experimental work with NP (*24*) and other proteins (*43, 44, 45, 46*) suggests that at an evolutionary level, sites interact mostly through generic effects on stability and folding. Furthermore, the effects of mutations on stability and folding tend to be conserved during evolution (*25, 43*). So one justification for assuming site-specific but site-independent preferences is that selection on a mutation is mostly determined by whether the protein can tolerate its effect on stability or folding, so stabilizing amino acids will be tolerated in most genetic backgrounds while destabilizing amino acids will only be tolerated in some backgrounds, as has been described experimentally (*24*) and theoretically (*47*). A more pragmatic justification is that the work here builds off this assumption to create evolutionary models that are much better than existing alternatives.

Assume that the preferences are entirely at the amino-acid level and are indifferent to the specific codon (the study of preferences for synonymous codons is an interesting area for future work). Denote the preference of site r for amino-acid a as $\pi_{r,a}$, where

$$\sum_a \pi_{r,a} = 1. \tag{2}$$

Define $\pi_{r,a}/\pi_{r,a'}$ as the expected ratio of amino-acid a to a' after viral growth if both are initially introduced into the mutant library at equal frequency. Mutations that enhance viral growth will have larger values of $\pi_{r,a}$, while mutations that hamper growth will have lower values of $\pi_{r,a}$. However, $\pi_{r,a}/\pi_{r,a'}$ cannot be simply interpreted as the fitness effect of mutating site r from a to a': because most clones have multiple mutations, this ratio summarizes the effect of a mutation in a variety of related genetic backgrounds. A mutation can therefore have a ratio greater than one due to its inherent effect on viral growth or its effect on the tolerance for other mutations (or both). This analysis does not separate these factors, but experimental work (*24*) has shown that it is fairly common for one mutation to NP to alter the tolerance to a subsequent one.

The most naive approach is to set $\pi_{r,a}$ proportional to the frequency of amino-acid a in **mutvirus-p1** divided by its frequency in **mutDNA**, and then apply the normalization in Equation 2. However, such an approach is problematic for several reasons. First, it fails to account for errors (PCR, reverse-transcription) that inflate the observed frequencies of some mutations. Second, estimating ratios by dividing finite counts is notoriously statistically biased (*48, 49*). For example, in the limiting case where a mutation is counted once in **mutvirus-p1** and not at all in **mutDNA**, the ratio is infinity – yet in practice such low counts give little confidence that enough variants have been assayed to estimate the true effect of the mutation.

To circumvent these problems, I used an approach that explicitly accounts for the sampling statistics. The approach begins with prior estimates that the $\pi_{r,a}$ values are all equal, and that the error and mutation rates for each site are equal to the library averages. Multinomial likelihood

functions give the probability of observing a set of counts given the $\pi_{r,a}$ values and the various error and mutation rates. The posterior mean of the $\pi_{r,a}$ values is estimated by MCMC.

Use the counts in **DNA** to quantify errors due to PCR and sequencing. Use the counts in **RNA** to quantify errors due to reverse transcription. Assume that transcription of the viral genes from the reverse-genetics plasmids and subsequent replication of these genes by the influenza polymerase introduces a negligible number of new mutations. The second of these assumptions is supported by the fact that the mutation frequency in **virus-p1** is close to that in **RNA** (Fig. 1B and fig. S4A). The first of these assumptions is supported by the fact that stop codons are no more frequent in **RNA** than in **virus-p1** (Fig. 1B and fig. S4A) – deleterious stop codons arising during transcription will be purged during viral growth, while those arising from reverse-transcription and sequencing errors will not.

At each site r, there are n_{codon} codons, indexed by $i = 1, 2, \ldots n_{\text{codon}}$. Let $\text{wt}\,(r)$ denote the wildtype codon at site r. Let N_r^{DNA} be the total number of sequencing reads at site r in **DNA**, and let $n_{r,i}^{\text{DNA}}$ be the number of these reads that report codon i at site r, so that $\sum_i n_{r,i}^{\text{DNA}} = N_r^{\text{DNA}}$. Similarly, let N_r^{mutDNA}, N_r^{RNA}, and N_r^{mutvirus} be the total number of reads at site r and let $n_{r,i}^{\text{mutDNA}}$, $n_{r,i}^{\text{RNA}}$, and $n_{r,i}^{\text{mutvirus}}$ be the total number of these reads that report codon i at site r in **mutDNA**, **RNA**, and **mutvirus-p1**, respectively.

First consider the rate at which site r is erroneously read as some incorrect identity due to PCR or sequencing errors. Such errors are the only source of non-wildtype reads in the sequencing of **DNA**. For all $i \neq \text{wt}\,(r)$, define $\epsilon_{r,i}$ as the rate at which site r is erroneously read as codon i in **DNA**. Define $\epsilon_{r,\text{wt}(r)} = 1 - \sum_{i \neq \text{wt}(r)} \epsilon_{r,i}$ to be the rate at which site r is correctly read as its wildtype identity of $\text{wt}\,(r)$ in **DNA**. Then $\epsilon_{r,i} = \mathbb{E}\left[n_{r,i}^{\text{DNA}}/N_r^{\text{DNA}}\right]$ where $\mathbb{E}$ denotes the expectation value. Define $\overrightarrow{\epsilon_r} = (\epsilon_{r,1}, \ldots, \epsilon_{r,n_{\text{codon}}})$ and $\overrightarrow{n_r^{\text{DNA}}} = \left(n_{r,1}^{\text{DNA}}, \ldots, n_{r,n_{\text{codon}}}^{\text{DNA}}\right)$ as vectors of the $\epsilon_{r,i}$ and $n_{r,i}^{\text{DNA}}$ values, so the likelihood of observing $\overrightarrow{n_r^{\text{DNA}}}$ given $\overrightarrow{\epsilon_r}$ and N_r^{DNA} is

$$\Pr\left(\overrightarrow{n_r^{\text{DNA}}} \mid N_r^{\text{DNA}}, \overrightarrow{\epsilon_r}\right) = \text{Mult}\left(\overrightarrow{n_r^{\text{DNA}}}; N_r^{\text{DNA}}, \overrightarrow{\epsilon_r}\right) \tag{3}$$

where Mult denotes the multinomial distribution.

Next consider the rate at which site r is erroneously copied during reverse transcription. These reverse-transcription errors combine with the PCR / sequencing errors defined by $\overrightarrow{\epsilon_r}$ to create non-wildtype reads in **RNA**. For all $i \neq \text{wt}\,(r)$, define $\rho_{r,i}$ as the rate at which site r is miscopied to i during reverse transcription. Define $\rho_{r,\text{wt}(r)} = 1 - \sum_{i \neq \text{wt}(r)} \rho_{r,i}$ as the rate at which site r is correctly reverse transcribed. Ignore as negligibly rare the possibility that a site is subject to both a reverse-transcription and sequencing / PCR error within the same clone (a reasonable assumption as both $\epsilon_{r,i}$ and $\rho_{r,i}$ are very small for $i \neq \text{wt}\,(r)$). Then $\epsilon_{r,i} + \rho_{r,i} - \delta_{i,\text{wt}(r)} = \mathbb{E}\left[n_{r,i}^{\text{RNA}}/N_r^{\text{RNA}}\right]$ where $\delta_{i,\text{wt}(r)}$ is the Kronecker delta (equal to one if $i = \text{wt}\,(r)$ and

zero otherwise). The likelihood of observing $\overrightarrow{n_r^{\mathrm{RNA}}}$ given $\overrightarrow{\rho_r}$, $\overrightarrow{\epsilon_r}$, and N_r^{RNA} is

$$\Pr\left(\overrightarrow{n_r^{\mathrm{RNA}}} \mid N_r^{\mathrm{RNA}}, \overrightarrow{\rho_r}, \overrightarrow{\epsilon_r}\right) = \mathrm{Mult}\left(\overrightarrow{n_r^{\mathrm{RNA}}}; N_r^{\mathrm{RNA}}, \overrightarrow{\epsilon_r} + \overrightarrow{\rho_r} - \overrightarrow{\delta_r}\right). \tag{4}$$

where $\overrightarrow{\delta_r} = \left(\delta_{1,\mathrm{wt}(r)}, \ldots, \delta n_{\mathrm{codon},\mathrm{wt}(r)}\right)$ is a vector that is all zeros except for the element $\mathrm{wt}\,(r)$.

Next consider the rate at which site r is mutated to some other codon in the plasmid mutant library. These mutations combine with the PCR / sequencing errors defined by $\overrightarrow{\epsilon_r}$ to create non-wildtype reads in **mutDNA**. For all $i \neq \mathrm{wt}\,(r)$, define $\mu_{r,i}$ as the rate at which site r is mutated to codon i in the mutant library. Define $\mu_{r,\mathrm{wt}(r)} = 1 - \sum_{i \neq \mathrm{wt}(r)} \mu_{r,i}$ as the rate at which site r is not mutated. Ignore as negligibly rare the possibility that a site is subject to both a mutation and a sequencing / PCR error within the same clone. Then $\mu_{r,i} + \epsilon_{r,i} - \delta_{i,\mathrm{wt}(r)} = \mathbb{E}\left[n_{r,i}^{\mathrm{mutDNA}} / N_r^{\mathrm{mutDNA}}\right]$. The likelihood of observing $\overrightarrow{n_r^{\mathrm{mutDNA}}}$ given $\overrightarrow{\mu_r}$, $\overrightarrow{\epsilon_r}$, and N_r^{mutDNA} is

$$\Pr\left(\overrightarrow{n_r^{\mathrm{mutDNA}}} \mid N_r^{\mathrm{mutDNA}}, \overrightarrow{\mu_r}, \overrightarrow{\epsilon_r}\right) = \mathrm{Mult}\left(\overrightarrow{n_r^{\mathrm{mutDNA}}}; N_r^{\mathrm{mutDNA}}, \overrightarrow{\mu_r} + \overrightarrow{\epsilon_r} - \overrightarrow{\delta_r}\right). \tag{5}$$

Finally, consider the effect of the preferences of each site r for different amino acids, as denoted by the $\pi_{r,a}$ values. Selection due to these preferences is manifested in **mutvirus**. This selection acts on the mutations in the mutant library ($\mu_{r,i}$), although the actual counts in **mutvirus** are also affected by the sequencing / PCR errors ($\epsilon_{r,i}$) and reverse-transcription errors ($\rho_{r,i}$). Again ignore as negligibly rare the possibility that a site is subject to more than one of these sources of mutation and error within a single clone. Let $\mathcal{A}\,(i)$ denote the amino acid encoded by codon i. Let $\overrightarrow{\pi_r}$ be the vector of $\pi_{r,a}$ values. Define the vector-valued function $\overrightarrow{\mathcal{C}}$ as

$$\overrightarrow{\mathcal{C}}\left(\overrightarrow{\pi_r}\right) = \left(\pi_{r,\mathcal{A}(1)}, \ldots, \pi_{r,\mathcal{A}(n_{\mathrm{codon}})}\right), \tag{6}$$

so that this function returns a n_{codon}-element vector constructed from $\overrightarrow{\pi_r}$. Because the selection in **mutvirus** due to the preferences $\pi_{r,\mathcal{A}(i)}$ occurs after the mutagenesis $\mu_{r,i}$ but before the reverse-transcription errors $\rho_{r,i}$ and the sequencing / PCR errors $\epsilon_{r,i}$, then $\mathbb{E}\left[n_{r,i}^{\mathrm{mutvirus}} / N_r^{\mathrm{mutvirus}}\right] = \epsilon_{r,i} + \rho_{r,i} + \gamma_r \times \pi_{r,\mathcal{A}(i)} \times \mu_{r,i} - 2 \times \delta_{i,\mathrm{wt}(r)}$ where $\gamma_r = \left(\sum_i \pi_{r,\mathcal{A}(i)} \mu_{r,i}\right)^{-1} = \left(\overrightarrow{\mathcal{C}}\left(\overrightarrow{\pi_r}\right) \cdot \overrightarrow{\mu_r}\right)^{-1}$ (where $\cdot$ denotes the dot product) is a normalization factor that accounts for the fact that changes in the frequency of one variant due to selection will influence the observed frequency of other variants. The likelihood of observing $\overrightarrow{n_r^{\mathrm{mutvirus}}}$ given $\overrightarrow{\mu_r}$, $\overrightarrow{\epsilon_r}$, $\overrightarrow{\rho_r}$, $\overrightarrow{\pi_r}$, and N_r^{mutvirus} is therefore

$$\Pr\left(\overrightarrow{n_r^{\mathrm{mutvirus}}} \mid \overrightarrow{\mu_r}, \overrightarrow{\epsilon_r}, \overrightarrow{\rho_r}, \overrightarrow{\pi_r}, N_r^{\mathrm{mutvirus}}\right) = \mathrm{Mult}\left(\overrightarrow{n_r^{\mathrm{mutvirus}}}; N_r^{\mathrm{mutvirus}}, \overrightarrow{\epsilon_r} + \overrightarrow{\rho_r} + \frac{\overrightarrow{\mathcal{C}}\left(\overrightarrow{\pi_r}\right) \circ \overrightarrow{\mu_r}}{\overrightarrow{\mathcal{C}}\left(\overrightarrow{\pi_r}\right) \cdot \overrightarrow{\mu_r}} - 2\overrightarrow{\delta_r}\right). \tag{7}$$

where $\circ$ is the Hademard (entry-wise) product.

Specify priors over $\overrightarrow{\epsilon_r}$, $\overrightarrow{\rho_r}$, $\overrightarrow{\mu_r}$, and $\overrightarrow{\pi_r}$ in the form of Dirichlet distributions (denoted here by Dir). For the priors over the mutation rates $\overrightarrow{\mu_r}$, I choose Dirichlet-distribution parameters such

that the mean of the prior expectation for the mutation rate at each site r and codon i is equal to the average value for all sites, estimated as the frequency in **mutDNA** minus the frequency in **DNA** (Fig. 1B and fig. S4), denoted by $\overline{\mu}$. So the prior is

$$\Pr\left(\overrightarrow{\mu_r}\right) = \mathrm{Dir}\left(\overrightarrow{\mu_r}; n_{\mathrm{codon}} \cdot \sigma_\mu \cdot \overrightarrow{\alpha_{\mu,r}}\right) \tag{8}$$

where $\overrightarrow{\alpha_{\mu,r}}$ is the n_{codon}-element vector with elements $\alpha_{\mu,r,i} = \overline{\mu} + \delta_{i,\mathrm{wt}(r)}\left(1 - n_{\mathrm{codon}}\overline{\mu}\right)$ and σ_μ is the scalar concentration parameter.

For the priors over $\epsilon_{r,i}$ and $\rho_{r,i}$, the Dirichlet-distribution parameters again represent the average value for all sites, but now also depend on the number of nucleotide changes in the codon mutation since sequencing / PCR and reverse-transcription errors are far more likely to lead to single-nucleotide codon changes than multiple-nucleotide codon changes (Fig. 1B and fig. S4). Let $\mathcal{M}\left(\mathrm{wt}\left(r\right), i\right)$ be the number of nucleotide changes in the mutation from codon $\mathrm{wt}\left(r\right)$ to codon i. For example, $\mathcal{M}\left(\mathtt{GCA}, \mathtt{ACA}\right) = 1$ and $\mathcal{M}\left(\mathtt{GCA}, \mathtt{ATA}\right) = 2$. Let $\overline{\epsilon_1}$, $\overline{\epsilon_2}$, and $\overline{\epsilon_3}$ be the average error rates for one-, two-, and three-nucleotide codon mutations, respectively – these are estimated as the frequencies in **DNA**. So the prior is

$$\Pr\left(\overrightarrow{\epsilon_r}\right) = \mathrm{Dir}\left(\overrightarrow{\epsilon_r}; n_{\mathrm{codon}} \cdot \sigma_\epsilon \cdot \overrightarrow{\alpha_{\epsilon,r}}\right) \tag{9}$$

where $\overrightarrow{\alpha_{\epsilon,r}}$ is the n_{codon}-element vector with elements $\alpha_{\epsilon,r,i} = \overline{\epsilon_{\mathcal{M}(\mathrm{wt}(r),i)}}$ where $\overline{\epsilon_0} = 1 - 9 \times \overline{\epsilon_1} - 27 \times \overline{\epsilon_2} - 27 \times \overline{\epsilon_3}$, and where σ_ϵ is the scalar concentration parameter.

Similarly, let $\overline{\rho_1}$, $\overline{\rho_2}$, and $\overline{\rho_3}$ be the average reverse-transcription error rates for one-, two-, and three-nucleotide codon mutations, respectively – these are estimated as the frequencies in **RNA** minus those in **DNA**. So the prior is

$$\Pr\left(\overrightarrow{\rho_r}\right) = \mathrm{Dir}\left(\overrightarrow{\rho_r}; n_{\mathrm{codon}} \cdot \sigma_\rho \cdot \overrightarrow{\alpha_{\rho,r}}\right) \tag{10}$$

where $\overrightarrow{\alpha_{\rho,r}}$ is the n_{codon}-element vector with elements $\alpha_{\rho,r,i} = \overline{\rho_{\mathcal{M}(\mathrm{wt}(r),i)}}$ where $\overline{\rho_0} = 1 - 9 \times \overline{\rho_1} - 27 \times \overline{\rho_2} - 27 \times \overline{\rho_3}$, and where σ_ρ is the scalar concentration parameter.

Specify a symmetric Dirichlet-distribution prior over $\overrightarrow{\pi_r}$ (note that any other prior, such as one that favored wildtype, would implicitly favor certain identities based empirically on the wildtype sequence, and so would not be in the spirit of the parameter-free derivation of the $\pi_{r,a}$ values employed here). Specifically, use a prior of

$$\Pr\left(\overrightarrow{\pi_r}\right) = \mathrm{Dir}\left(\overrightarrow{\pi_r}; \sigma_\pi \cdot \overrightarrow{1}\right) \tag{11}$$

where $\overrightarrow{1}$ is the n_{aa}-element vector that is all ones, and σ_π is the scalar concentration parameter.

It is now possible to writ expressions for the likelihoods and posterior probabilities. Let $\mathcal{N}_r = \left\{\overrightarrow{n_r^{\mathrm{DNA}}}, \overrightarrow{n_r^{\mathrm{mutDNA}}}, \overrightarrow{n_r^{\mathrm{RNA}}}, \overrightarrow{n_r^{\mathrm{mutvirus}}}, N_r^{\mathrm{DNA}}, N_r^{\mathrm{mutDNA}}, N_r^{\mathrm{RNA}}, N_r^{\mathrm{mutvirus}}\right\}$ denote the full set of counts for site r. The likelihood of $\mathcal{N}_r$ given values for the preferences and mutation / error

rates is

$$\begin{aligned}\Pr\left(\mathcal{N}_r \mid \overrightarrow{\pi_r}, \overrightarrow{\epsilon_r}, \overrightarrow{\rho_r}, \overrightarrow{\mu_r}\right) &= \Pr\left(\overrightarrow{n_r^{\mathrm{DNA}}} \mid N_r^{\mathrm{DNA}}, \overrightarrow{\epsilon_r}\right) \times \Pr\left(\overrightarrow{n_r^{\mathrm{RNA}}} \mid N_r^{\mathrm{RNA}}, \overrightarrow{\epsilon_r}, \overrightarrow{\rho_r}\right) \times \\ &\quad \Pr\left(\overrightarrow{n_r^{\mathrm{mutDNA}}} \mid N_r^{\mathrm{mutDNA}}, \overrightarrow{\epsilon_r}, \overrightarrow{\mu_r}\right) \times \\ &\quad \Pr\left(\overrightarrow{n_r^{\mathrm{mutvirus}}} \mid N_r^{\mathrm{mutvirus}}, \overrightarrow{\epsilon_r}, \overrightarrow{\rho_r}, \overrightarrow{\mu_r}, \overrightarrow{\pi_r}\right) \qquad (12)\end{aligned}$$

where the likelihoods that compose Equation 12 are defined by Equations 3, 4, 5, and 7. The posterior probability of a specific value for the preferences and mutation / error rates is

$$\begin{aligned}\Pr\left(\overrightarrow{\pi_r}, \overrightarrow{\epsilon_r}, \overrightarrow{\rho_r}, \overrightarrow{\mu_r} \mid \mathcal{N}_r\right) &= C_r \times \Pr\left(\mathcal{N}_r \mid \overrightarrow{\pi_r}, \overrightarrow{\epsilon_r}, \overrightarrow{\rho_r}, \overrightarrow{\mu_r}\right) \times \qquad (13) \\ &\quad \Pr\left(\overrightarrow{\epsilon_r}\right) \times \Pr\left(\overrightarrow{\rho_r}\right) \times \Pr\left(\overrightarrow{\mu_r}\right) \times \Pr\left(\overrightarrow{\pi_r}\right) \qquad (14)\end{aligned}$$

where C_r is a normalization constant that does not need to be explicitly calculated in the MCMC approach used here. The posterior over the preferences $\overrightarrow{\pi_r}$ can be calculated by integrating over Equation 13 to give

$$\Pr\left(\overrightarrow{\pi_r} \mid \mathcal{N}_r\right) = \int\int\int \Pr\left(\overrightarrow{\pi_r}, \overrightarrow{\epsilon_r}, \overrightarrow{\rho_r}, \overrightarrow{\mu_r} \mid \mathcal{N}_r\right) d\overrightarrow{\epsilon_r} d\overrightarrow{\rho_r} d\overrightarrow{\mu_r}, \qquad (15)$$

where the integration is performed by MCMC. The posterior is summarized by its mean,

$$\langle \overrightarrow{\pi_r} \rangle = \int \overrightarrow{\pi_r} \times \Pr\left(\overrightarrow{\pi_r} \mid \left\{\mathcal{N}_r^k \mid 1 \leq k \leq \mathcal{R}\right\}\right) d\overrightarrow{\pi_r}. \qquad (16)$$

In practice, each replicate consists of four libraries (WT-1, WT-2, N334H-1, and N334H-2) – the posterior mean preferences inferred for each library within a replicate are averaged to give the estimated preferences for that replicate. The preferences within each replicate are highly correlated regardless of whether **mutvirus-p1** or **mutvirus-p2** is used as the **mutvirus** data set (Fig. 2A,B). This correlation between passages is consistent with the interpretation of the preferences as the fraction of genetic backgrounds that tolerate a mutation (if it was a selection coefficient, there should be further enrichment upon further passage). The preferences averaged over both replicates serve as the "best" estimate, and are displayed in Fig. 2D. This figure was created using the WebLogo 3 program (*50, 51*).

Fig. 2D also shows relative solvent accessibility (RSA) and secondary structure for residues present in chain C of NP crystal structure PDB 2IQH (*28*). The total accessible surface area (ASA) and the secondary structure for each residue in this monomer alone was calculated using DSSP (*52, 53*). The RSAs are the total ASA divided by the maximum ASA defined in (*54*). The secondary structure codes returned by DSSP were grouped into three classes: helix (DSSP codes G, H, or I), strand (DSSP codes B or E), and loop (any other DSSP code).

A full description of the algorithm used to infer the preferences along with complete data can be accessed at `http://jbloom.github.io/mapmuts/example_2013Analysis_Influenza_NP_Aichi68.html`.

The experimentally determined evolutionary model

The foregoing sections described measurement of two quantities: the inherent preference $\pi_{r,a}$ for amino-acid a at site r, and the rates $R_{m\rightarrow n}$ of mutations from one nucleotide to another. Assuming that premature stop codons are lethal, evolution can traverse 61 codons at each site. If mutations occur a single nucleotide at a time, the rate of mutation from codon x to codon y is

$$Q_{xy} = \begin{cases} 0 & \text{if } x \text{ and } y \text{ differ by more than one nucleotide} \\ R_{m\rightarrow n} & \text{if } x \text{ differs from } y \text{ by a single nucleotide change } m \text{ to } n. \end{cases} \tag{17}$$

where the rates for a mutation and its reversal are enforced to be equal as described in the section on mutation rates.

Let $F_{r,xy}$ denote the probability that a mutation of site r from x to y goes to fixation, and let $\mathcal{A}(x)$ to denote the amino acid encoded by x. Then the rate of substitution $P_{r,xy}$ from x to y is

$$P_{r,xy} = \begin{cases} Q_{xy} & \text{if } \mathcal{A}(x) = \mathcal{A}(y) \\ Q_{xy} \times F_{r,xy} & \text{if } \mathcal{A}(x) \neq \mathcal{A}(y). \end{cases} \tag{18}$$

Intuitively, it is obvious that the $\pi_{r,a}$ values provide information about the fixation probabilities $F_{r,xy}$. For instance, it seems reasonable to expect that a mutation from x to y at site r will be more likely to fix (relatively larger value of $F_{r,xy}$) if amino-acid $\mathcal{A}(y)$ is preferred to $\mathcal{A}(x)$ at this this site (if $\pi_{r,\mathcal{A}(y)} > \pi_{r,\mathcal{A}(x)}$) and less likely to fix if $\pi_{r,\mathcal{A}(y)} < \pi_{r,\mathcal{A}(x)}$. However, the exact relationship between $\pi_{r,a}$ and $F_{r,xy}$ is unclear. A rigorous derivation would require knowledge of unknown and probably unmeasurable population-genetics parameters for the both the deep-sequencing experiment and the naturally evolving population. So instead, I provide two heuristic relationships between $\pi_{r,a}$ and $F_{r,xy}$. Both relationships satisfy detailed balance (reversibility), such that $\pi_{r,\mathcal{A}(x)} \times F_{r,xy} = \pi_{r,\mathcal{A}(y)} \times F_{r,yx}$, meaning that $F_{r,xy}$ defines a Markov process with $\pi_{r,a}$ as its stationary state when all amino-acid interchanges are equally probable.

It is helpful to first consider what the $\pi_{r,a}$ values actually represent. Most NP variants in the libraries contain multiple mutations (an average of about three per gene). The $\pi_{r,a}$ values therefore represent the relative preferences for amino acids at a site averaged over the nearby mutational neighborhood of the parent protein – they do *not* simply represent the effect of each mutation in the parent. (Note that experimental work does support the idea of conserved site preferences of this sort (*25*)). Therefore, one interpretation is that $\pi_{r,a}$ is proportional to the fraction of genetic backgrounds in which a mutation is tolerated. In this interpretation, a mutation from x to y is always tolerated if $\pi_{r,\mathcal{A}(y)} > \pi_{r,\mathcal{A}(x)}$, but is only sometimes tolerated if $\pi_{r,\mathcal{A}(x)} < \pi_{r,\mathcal{A}(y)}$. In this scenario, there is strong selection during initial viral growth on whether the mutation is tolerated in the particular genetic background in which it occurs, and there should be little further enrichment or depletion during subsequent viral passages. In contrast, if the $\pi_{r,a}$ values are related to selection coefficients, the enrichment / depletion of beneficial / deleterious mutations should increase upon further viral passage. Fig. 1A,B shows that the

enrichment of mutations after two viral passages is very similar to the enrichment after one viral passage. These data are therefore loosely consistent with interpreting the $\pi_{r,a}$ values as proportional to the fraction of genetic backgrounds in which mutations are tolerated. Note that this interpretation can be related to the the selection-threshold evolutionary dynamics described in (*47*). The equation that describes this scenario is

$$F_{r,xy} = \begin{cases} 1 & \text{if } \mathcal{A}(x) = \mathcal{A}(y) \text{ or } \pi_{r,\mathcal{A}(y)} \geq \pi_{r,\mathcal{A}(x)} \\ \frac{\pi_{r,\mathcal{A}(y)}}{\pi_{r,\mathcal{A}(x)}} & \text{otherwise.} \end{cases} \tag{19}$$

An alternative interpretation is that $F_{r,xy}$ reflects the selection coefficient for the mutation from $\mathcal{A}(x)$ to $\mathcal{A}(y)$. Mathematically formalizing this interpretation requires relating the $\pi_{r,a}$ values to selection coefficients to determine $F_{r,xy}$. One relationship was suggested by Halpern and Bruno (*5*). Specifically, if the $\pi_{r,a}$ values represent the expected amino-acid equilibrium frequencies in a hypothetical evolving population in which all amino-acid interchanges are equally likely, and assuming (probably unrealistically) that this hypothetical population and the actual population in which NP evolves are in the weak-mutation limit and have identical constant effective population sizes, then Halpern and Bruno (*5*) derive

$$F_{r,xy} \propto \begin{cases} 1 & \text{if } \mathcal{A}(x) = \mathcal{A}(y) \text{ or } \pi_{r,\mathcal{A}(x)} = \pi_{r,\mathcal{A}(y)} \\ \frac{\ln\left(\frac{\pi_{r,\mathcal{A}(y)}}{\pi_{r,\mathcal{A}(x)}}\right)}{1-\frac{\pi_{r,\mathcal{A}(x)}}{\pi_{r,\mathcal{A}(y)}}} & \text{otherwise.} \end{cases} \tag{20}$$

Once $F_{r,xy}$ has been defined by either Equation 19 or 20, it is possible to compute the codon substitution probabilities $P_{r,xy}$ using Equation 18. The additional definition of $P_{r,xx}$ as

$$P_{r,xx} = -\sum_{y \neq x} P_{r,xy}, \tag{21}$$

makes it possible to construct a 61×61 matrix

$$\mathbf{P_r} = [P_{r,xy}] \tag{22}$$

that gives the codon substitution rates for site r. If the frequencies of the individual codons at site r at time zero are given by the 61-element vector $\mathbf{p_{r,0}}$ (normalized such that its entries sum to one), then the frequencies at some future time t are given by

$$\mathbf{p_{r,t}} = \mathbf{p_{r,0}} \exp\left(ut\mathbf{P_r}\right) \tag{23}$$

where u is a scaling constant that defines the rate of sequence change for the particular units used for t. For reasonable values of $R_{m\to n}$ and $\pi_{r,a}$, $\mathbf{P_r} + \mathbf{I}$ is an irreducible and acyclic stochastic

matrix (where $\mathbf{I}$ is the identity matrix). The Perron-Frobenius theorems therefore guarantee that for each site r, there is a unique stationary distribution $\mathbf{p_{r,\infty}}$ satisfying the eigenvector equation

$$\mathbf{p_{r,\infty}} = \mathbf{p_{r,\infty}} \left(\mathbf{P_r} + \mathbf{I} \right). \tag{24}$$

These expected evolutionary equilibrium frequencies $\mathbf{p_{r,\infty}}$ are not identical to the preferences $\pi_{r,a}$ because of infuence of the genetic code and the mutation rates. The evolutionary equilibrium frequencies defined by using Equation 19 for $F_{r,xy}$ are plotted in fig. S5.

The $P_{r,xy}$ values define a parameter-free codon substitution model that can be used in phylogenetic algorithms as described in the next section. With the choices for $F_{r,xy}$ and Q_{xy} described above, this substitution model is reversible.

Phylogenetic analyses

A detailed description of the phylogenetic analyses is available at `http://jbloom.github.io/phyloExpCM/example_2013Analysis_Influenza_NP_Human_1918_Descended.html`. The source code is at `https://github.com/jbloom/phyloExpCM`.

A set of NP coding sequences was assembled for human influenza lineages descended from a close relative the 1918 virus (H1N1 from 1918 to 1957, H2N2 from 1957 to 1968, H3N2 from 1968 to 2013, and seasonal H1N1 from 1977 to 2008). All full-length NP sequences from the Influenza Virus Resource (*55*) were downloaded, and up to three unique sequences per year from each of the four lineages described above were retained. These sequences were aligned using EMBOSS needle (*56*). Outlier sequences that correspond to heavily lab-adapted strains, lab recombinants, mis-annotated sequences, or zoonotic transfers (for example, a small number of human H3N2 strains are from zoonotic swine variant H3N2 rather than the main human H3N2 lineage) were removed. This was done by first removing known outliers in the influenza databases (*57*), and then using an analysis with RAxML (*58*) and Path-O-Gen (`http://tree.bio.ed.ac.uk/software/pathogen/`) to remove remaining sequences that were extreme outliers from the molecular clock. The final alignment after removing outliers consisted of 274 unique NP sequences.

Maximum-likelihood phylogenetic trees were constructed using *codonPhyML* (*29*). Two substitution models were used. The first was *GY94* (*6*) using *CF3x4* equilibrium frequencies (*59*), a single transition-transversion ratio optimized by maximum likelihood, and a synonymous-nonsynonymous ratio drawn from four discrete gamma-distributed categories with mean and shape parameter optimized by maximum likelihood (*9*). The second was *KOSI07* (*7*) using the *F* method for the equilibrium frequencies, optimizing the relative transversion-transition ratio by maximum likelihood, and letting the relative synonymous-nonsynonymous ratio again be drawn from four gamma-distributed categories with mean and shape parameter optimized by maximum likelihood. The trees produced by *codonPhyML* are unrooted. These trees were rooted using Path-O-Gen (`http://tree.bio.ed.ac.uk/software/pathogen/`), and visualized with FigTree (`http://tree.bio.ed.ac.uk/software/figtree/`) to create the images in Fig. 3 and fig. S6. The tree topologies are extremely similar for both models.

The evolutionary models were compared by using them to optimize the branch lengths of the fixed tree topologies in Fig. 3 and fig. S6 so as to maximize the likelihood using *HYPHY* (*30*) for sites 2 to 498 (site 1 was not included, since the N-terminal methionine is conserved and was not mutated in the plasmid mutant libraries). The results are shown in Table 2, table S6, and table S7. Regardless of which tree topology was used, the experimentally determined evolutionary models outperformed all variants of *GY94* and *KOSI07*. The experimentally determined evolutionary models performed best when using the preferences determined from the combined data from both replicates and using Equation 19 to compute the fixation probabilities. Using the data from just one replicate also outperforms *GY94* and *KOSI07*, although the likelihoods are slightly worse. In terms of the completeness with which mutations are sampled in the mutant viruses,

replicate A is superior to *replicate B* as discussed above – and the former replicate gives higher likelihoods. If the fixation probabilities are instead determined using the method of Halpern and Bruno (*5*) as in Equation 20, the experimentally determined models still outperform *GY94* and *KOSI07* – but the likelihoods are substantially worse. To check that the experimentally determined models really do utilize the site-specific preferences information, the preferences were randomized among sites and likelihoods were computed. These randomized models perform vastly worse than any of the alternatives.

The variants of *GY94* and *KOSI07* that were used for the *HYPHY* analyses are listed in the tables of likelihoods. The parameters were counted as follows: the simplest variants contained equilibrium frequency parameters that were empirically estimated from the sequences under analysis: there are 9 such parameters for *GY94* using *CF3x4* (*6, 59*) and 60 such parameters for *KOSI07* using *F* (*7*). In addition, all variants of the models contain a transition-transversion ratio optimized by likelihood, and at least a single synonymous-nonsynonymous ratio optimized by likelihood. For more complex variants, one or more of the following is also done:

- The overall substitution rate is drawn from a gamma distribution with four discrete categories to add one shape parameter (*8*).
- The nonsynonymous-synonymous ratio is drawn from a gamma distribution with four discrete categories to add one shape parameter (*9*).
- A different nonsynonymous-synonymous ratio is estimated for each branch (*60*) to add 547 parameters.

References for supplementary material

1. J. Felsenstein, *Systematic Zoology* **22**, 240 (1973).
2. J. Felsenstein, *J. Mol. Evol.* **17**, 368 (1981).
3. J. P. Huelsenbeck, B. Larget, R. E. Miller, F. Ronquist, *Systematic Biology* **51**, 673 (2002).
4. J. L. Thorne, S. C. Choi, J. Yu, P. G. Higgs, H. Kishino, *Mol. Biol. Evol.* **24**, 1667 (2007).
5. A. L. Halpern, W. J. Bruno, *Mol. Biol. Evol.* **15**, 910 (1998).
6. N. Goldman, Z. Yang, *Mol. Biol. Evol.* **11**, 725 (1994).
7. C. Kosiol, I. Holmes, N. Goldman, *Mol. Biol. Evol.* **24**, 1464 (2007).
8. Z. Yang, *J. Mol. Evol.* **39**, 306 (1994).
9. Z. Yang, R. Nielsen, N. Goldman, A.-M. K. Pedersen, *Genetics* **155**, 431 (2000).
10. D. Posada, T. R. Buckley, *Systematic Biology* **53**, 793 (2004).
11. N. Rodrigue, C. L. Kleinman, H. Philippe, N. Lartillot, *Mol. Biol. Evol.* **26**, 1663 (2009).
12. C. L. Kleinman, N. Rodrigue, N. Lartillot, H. Philippe, *Mol. Biol. Evol.* **27**, 1546 (2010).
13. V. Potapov, M. Cohen, G. Schreiber, *Prot. Eng. Des. Sel.* **22**, 553 (2009).
14. N. Lartillot, H. Philippe, *Mol. Biol. Evol.* **21**, 1095 (2004).
15. S. Q. Le, N. Lartillot, O. Gascuel, *Phil. Trans. R. Soc. B* **363**, 3965 (2008).
16. C.-H. Wu, M. A. Suchard, A. J. Drummond, *Mol. Biol. Evol.* **30**, 669 (2013).
17. J. D. Bloom, L. I. Gong, D. Baltimore, *Science* **328**, 1272 (2010).
18. Materials and methods are available as supplementary material on *Science* online.
19. D. M. Fowler, *et al.*, *Nat. Methods* **7**, 741 (2010).
20. D. Melamed, D. L. Young, C. E. Gamble, C. R. Miller, S. Fields, *RNA* **19**, 1537 (2013).
21. M. W. Traxlmayr, *et al.*, *J. Mol. Biol.* (2012).
22. L. M. Starita, *et al.*, *Proc. Natl. Acad. Sci. USA* **110**, E1263 (2013).
23. B. P. Roscoe, K. M. Thayer, K. B. Zeldovich, D. Fushman, D. N. Bolon, *J. Mol. Biol.* (2013).

24. L. I. Gong, M. A. Suchard, J. D. Bloom, *eLife* **2**, e00631 (2013).

25. O. Ashenberg, L. I. Gong, J. D. Bloom, *Proc. Natl. Acad. Sci. USA* **110**, 21071 (2013).

26. E. Hoffmann, G. Neumann, Y. Kawaoka, G. Hobom, R. G. Webster, *Proc. Natl. Acad. Sci. USA* **97**, 6108 (2000).

27. G. A. Marsh, R. Rabadán, A. J. Levine, P. Palese, *J. Virology* **82**, 2295 (2008).

28. Q. Ye, R. M. Krug, Y. J. Tao, *Nature* **444**, 1078 (2006).

29. M. Gil, M. S. Zanetti, S. Zoller, M. Anisimova, *Mol. Biol. Evol.* **30**, 1270 (2013).

30. S. L. Pond, S. D. Frost, S. V. Muse, *Bioinformatics* **21**, 676 (2005).

31. E. Firnberg, M. Ostermeier, *PLoS One* **7**, e52031 (2012).

32. P. C. Jain, R. Varadarajan, *Analytical Biochemistry* (2013).

33. J. B. Hiatt, R. P. Patwardhan, E. H. Turner, C. Lee, J. Shendure, *Nat. Methods* **7**, 119 (2010).

34. M. W. Schmitt, *et al.*, *Proc. Natl. Acad. Sci. USA* **109**, 14508 (2012).

35. D. I. Lou, *et al.*, *Proc. Natl. Acad. Sci. USA* **110**, 19872 (2013).

36. H. Goto, Y. Kawaoka, *Proc. Natl. Acad. Sci. USA* **95**, 10224 (1998).

37. J. Parvin, A. Moscona, W. Pan, J. Leider, P. Palese, *J. Virology* **59**, 377 (1986).

38. P. C. Cirino, K. M. Mayer, D. Umeno, *Directed evolution library creation: methods and protocols* (Humana Press, 2003), chap. Generating mutant libraries using error-prone PCR, pp. 3–9.

39. C. Neylon, *Nucleic Acids Research* **32**, 1448 (2004).

40. R. Georgescu, G. Bandara, L. Sun, *Directed evolution library creation* (Springer, 2003), chap. Saturation mutagenesis, pp. 75–83.

41. G. Neumann, *et al.*, *Proc. Natl. Acad. Sci. USA* **96**, 9345 (1999).

42. Y. Hou, H. Zhang, L. Miranda, S. Lin, *PLoS One* **5**, e9545 (2010).

43. L. Serrano, A. G. Day, A. R. Fersht, *J. Mol. Biol.* **233**, 305 (1993).

44. J. D. Bloom, *et al.*, *Proc. Natl. Acad. Sci. USA* **102**, 606 (2005).

45. S. Bershtein, M. Segal, R. Bekerman, N. Tokuriki, D. S. Tawfik, *Nature* **444**, 929 (2006).

46. J. D. Bloom, S. T. Labthavikul, C. R. Otey, F. H. Arnold, *Proc. Natl. Acad. Sci. USA* **103**, 5869 (2006).

47. J. D. Bloom, A. Raval, C. O. Wilke, *Genetics* **175**, 255 (2007).

48. K. Pearson, *Biometrika* **7**, 531 (1910).

49. R. Ogliore, G. Huss, K. Nagashima, *Nuclear Instruments and Methods in Physics Research Section B: Beam Interactions with Materials and Atoms* **269**, 1910 (2011).

50. T. D. Schneider, R. M. Stephens, *Nucleic Acids Res.* **18**, 6097 (1990).

51. G. E. Crooks, G. Hon, J.-M. Chandonia, S. E. Brenner, *Genome research* **14**, 1188 (2004).

52. W. Kabsch, C. Sander, *Biopolymers* **22**, 2577 (1983).

53. R. P. Joosten, *et al.*, *Nucleic Acids Res.* **39**, D411 (2011).

54. M. Tien, A. G. Meyer, S. J. Spielman, C. O. Wilke, *PLoS One* **8**, e80635 (2013).

55. Y. Bao, *et al.*, *J. Virol.* **82**, 596 (2008).

56. P. Rice, I. Longden, A. Bleasby, *Trends in Genetics* **16**, 276 (2000).

57. M. Krasnitz, A. J. Levine, R. Rabadan, *J. Virology* **82**, 8947 (2008).

58. A. Stamatakis, *Bioinformatics* **22**, 2688 (2006).

59. S. K. Pond, W. Delport, S. V. Muse, K. Scheffler, *PLoS One* **5**, e11230 (2010).

60. Z. Yang, R. Nielsen, *J. Mol. Evol.* **46**, 409 (1998).

clone	mutations
Clone 1	G62T (G21V), T693C (synonymous), del153-522 (indel)
Clone 2	None
Clone 3	C29T (T10I)
Clone 4	None
Clone 5	None
Clone 6	G429A (synonymous), C447T (synonymous)
Clone 7	None
Clone 8	None
Clone 9	C646A (R216S)
Clone 10	G471T (K157N), G703A (D235N)
Clone 11	T111C (synonymous), T718G (*240E)
Clone 12	T25C (F9L), T26C (F9S)
Clone 13	C45T (synonymous), C549T (synonymous)
Clone 14	T319C (Y107H), C372T (synonymous), C539T (A180V)
Clone 15	A488C (K163T)
Clone 16	G274T (G92C)
Clone 17	None
Clone 18	None
Clone 19	None
Clone 20	G527A (S176N), A676G (T226A)
Clone 21	G4A (V2I)
Clone 22	T266C (M89T)
Clone 23	None
Clone 24	C30T (synonymous), del45-590 (indel)

Supplementary table S1: Mutations identified by sequencing the 720 nucleotide GFP gene packaged in the PB1 segment after 25 limiting-dilution passages for 24 independent replicates. The numbering is sequential beginning with the first nucleotide of the GFP start codon. For nonsynonymous mutations, the induced amino-acid change is indicated in parentheses.

mutation type	number of occurrences
total substitutions	24
transversions	6
transitions	18
nonsynonymous	15
synonymous	8
stop codons	1
indels	2
T $\rightarrow$ G	1
T $\rightarrow$ C	6
T $\rightarrow$ A	0
G $\rightarrow$ T	3
G $\rightarrow$ C	0
G $\rightarrow$ A	4
C $\rightarrow$ T	7
C $\rightarrow$ G	0
C $\rightarrow$ A	1
A $\rightarrow$ T	0
A $\rightarrow$ G	1
A $\rightarrow$ C	1

Supplementary table S2: Counts for different types of mutations after the 25 limiting-dilution passages. The numbers are calculated from table S1. Given that GFP is 720 nucleotides long, the data suggest a viral mutation rate of 5.6×10^{-5} mutations per nucleotide per tissue-culture generation.

agcaaaagcagggtagataatcactcacagagtgacatcgaaatcATGGCGTCCCAAGGC
ACCAAACGGTCTTATGAACAGATGGAAACTGATGGGGAACGCCAGAATGCAACTGAGATC
AGAGCATCCGTCGGGAAGATGATTGATGGAATTGGACGATTCTACATCCAAATGTGCACT
GAACTTAAACTCAGTGATTATGAGGGGCGACTGATCCAGAACAGCTTAACAATAGAGAGA
ATGGTGCTCTCTGCTTTTGACGAAAGAAGGAATAAATATCTGGAAGAACATCCCAGCGCG
GGGAAGGATCCTAAGAAAACTGGAGGACCCATATACAAGAGAGTAGATAGAAAGTGGATG
AGGGAACTCGTCCTTTATGACAAAGAAGAAATAAGGCGAATCTGGCGCCAAGCCAATAAT
GGTGATGATGCAACAGCTGGTCTGACTCACATGATGATCTGGCATTCCAATTTGAATGAT
ACAACATACCAGAGGACAAGAGCTCTTGTTCGCACCGGCATGGATCCCAGGATGTGCTCT
CTGATGCAGGGTTCGACTCTCCCTAGGAGGTCTGGAGCTGCAGGCGCTGCAGTCAAAGGA
GTTGGGACAATGGTGATGGAGTTGATAAGGATGATCAAACGTGGGATCAATGATCGGAAC
TTCTGGAGAGGTGAAAATGGACGAAAAACAAGGAGTGCTTACGAGAGAATGTGCAACATT
CTCAAAGGAAAATTTCAAACAGCTGCACAAAGGGCAATGATGGATCAAGTGAGAGAAAGT
CGGAACCCAGGAAATGCTGAGATCGAAGATCTCATCTTTCTGGCACGGTCTGCACTCATA
TTGAGAGGGTCAGTTGCTCACAAATCTTGTCTGCCCGCCTGTGTGTATGGACCTGCCGTA
GCCAGTGGCTACGACTTCGAAAAAGAGGGATACTCTTTAGTGGGAATAGACCCTTTCAAA
CTGCTTCAAAACAGCCAAGTATACAGCCTAATCAGACCGAACGAGAATCCAGCACACAAG
AGTCAGCTGGTGTGGATGGCATGCAATTCTGCTGCATTTGAAGATCTAAGAGTATTAAGC
TTCATCAGAGGGACCAAAGTATCCCCAAGGGGGAAACTTTCCACTAGAGGAGTACAAATT
GCTTCAAATGAAAACATGGATGCTATGGAATCAAGTACTCTTGAACTGAGAAGCAGGTAC
TGGGCCATAAGAACCAGAAGTGGAGGAAACACTAATCAACAGAGGGCCTCTGCAGGTCAA
ATCAGTGTGCAACCTGCATTTTCTGTGCAAAGAAACCTCCCATTTGACAAACCAACCATC
ATGGCAGCATTCACTGGGAATACAGAGGGAAGAACATCAGACATGAGGGCAGAAATTATA
AGGATGATGGAAGGTGCAAAACCAGAAGAAATGTCCTTCCAGGGGCGGGGAGTCTTCGAG
CTCTCGGACGAAAGGGCAGCGAACCCGATCGTGCCCTCTTTTGACATGAGTAATGAAGGA
TCTTATTTCTTCGGAGACAATGCAGAGGAGTACGACAATTAAagaaaaataccccttgttt
ctact

Supplementary table S3: The viral RNA sequence for the wildtype A/Aichi/2/1968 (H3N2) nucleoprotein as taken from Bloom lab plasmid pHWAichi68-NP. Shown is the reverse-complement of the full-length negative-sense viral RNA (so that the coding sequence is written in the positive orientation). The coding sequence is in capital letters, while the noncoding termini are in lower case letters. The viral RNA sequence for the N334H A/Aichi2/1968 nucleoprotein (plasmid pHWAichi68-NP-N334H) is identical except that codon 334 is mutated from `AAT` to `CAT`.

residue	frequencies in natural sequences	experimentally measured amino-acid preferences	expected equilibrium evolutionary frequencies from experiments
65	R (0.83), K (0.17)	R (0.40), K (0.10), N (0.06)	R (0.58), S (0.07)
150	R (1.00)	R (0.46), K (0.06), P (0.05), L (0.05)	R (0.63), L (0.07)
152	R (1.00)	R (0.52), K (0.07), Q (0.07)	R (0.71)
156	R (1.00)	R (0.52), Q (0.06)	R (0.69), S (0.06)
174	R (1.00)	R (0.58), N (0.06), T (0.05)	R (0.75)
175	R (1.00)	R (0.46), K (0.16)	R (0.66), K (0.08), S (0.05)
195	R (1.00)	R (0.51)	R (0.69)
199	R (1.00)	R (0.44), M (0.08), Y (0.06), V (0.05)	R (0.64), V (0.05)
213	R (1.00)	R (0.51), N (0.06)	R (0.69)
214	R (0.72), K (0.28)	K (0.24), H (0.09), R (0.09), Q (0.08), M (0.06), N (0.06), A (0.06), I (0.06)	R (0.19), K (0.17), A (0.09), H (0.07), I (0.06), L (0.06), Q (0.06)
221	R (1.00)	R (0.46), E (0.07), K (0.07)	R (0.66), L (0.05)
236	R (0.94), K (0.06)	K (0.32), R (0.30)	R (0.51), K (0.18)
355	R (1.00)	R (0.29), L (0.13), K (0.09)	R (0.43), L (0.19)
357	K (0.56), Q (0.44)	K (0.38), E (0.09), N (0.07), Y (0.05)	K (0.31), R (0.09), E (0.08), N (0.06)
361	R (1.00)	R (0.53), V (0.13)	R (0.68), V (0.11)
391	R (1.00)	R (0.59), K (0.09)	R (0.77)
148	Y (1.00)	Y (0.54), I (0.06)	Y (0.44), I (0.07), T (0.07), P (0.06), S (0.06)

Supplementary table S4: For residues involved in NP's RNA-binding groove, the preferences and expected evolutionary equilibrium frequencies from the experiments correlate well with the amino-acid frequencies in naturally occurring sequences. Shown are the 17 residues in the NP RNA-binding groove in (*28*). The second column gives the frequencies of amino acids in all 21,108 full-length NP sequences from influenza A (excluding bat lineages) in the Influenza Virus Resource as of January-31-2014. The third column gives the experimentally measured amino-acid preferences (Fig. 2D). The fourth column gives the expected evolutionary equilibrium frequency of the amino acids based on the experimentally measured preferences and mutational spectrum (fig. S5). Only residues with frequencies or preferences ≥ 0.05 are listed. In all cases, the most abundant amino acid in the natural sequences has the highest expected evolutionary equilibrium frequency. In 15 of 17 cases, the most abundant amino acid in the natural sequences has the highest experimentally measured preference – in the other two cases, the most abundant amino acid in the natural sequences is among those with the highest preference.

residue	stability measurement	frequencies in natural sequences	experimentally measured amino-acid preferences	expected equilibrium evolutionary frequencies from experiments
259	L259S is destabilizing ($\Delta T_m = -3.9°\mathrm{C}$)	L (0.98), S (0.02)	L (0.23), S (0.04)	L (0.36), S (0.06)
280	V280A is destabilizing ($\Delta T_m = -3.5°\mathrm{C}$)	V (0.89), A (0.10)	V (0.19), A (0.02)	V (0.25), A (0.03)
334	N334H is stabilizing ($\Delta T_m = 4.5°\mathrm{C}$)	H (0.93), N (0.07)	H (0.28), N (0.12)	H (0.23), N (0.10)
384	R384G is destabilizing ($\Delta T_m = -4.8°\mathrm{C}$)	R (0.80), G (0.17)	R (0.22), G (0.04)	R (0.39), G (0.04)

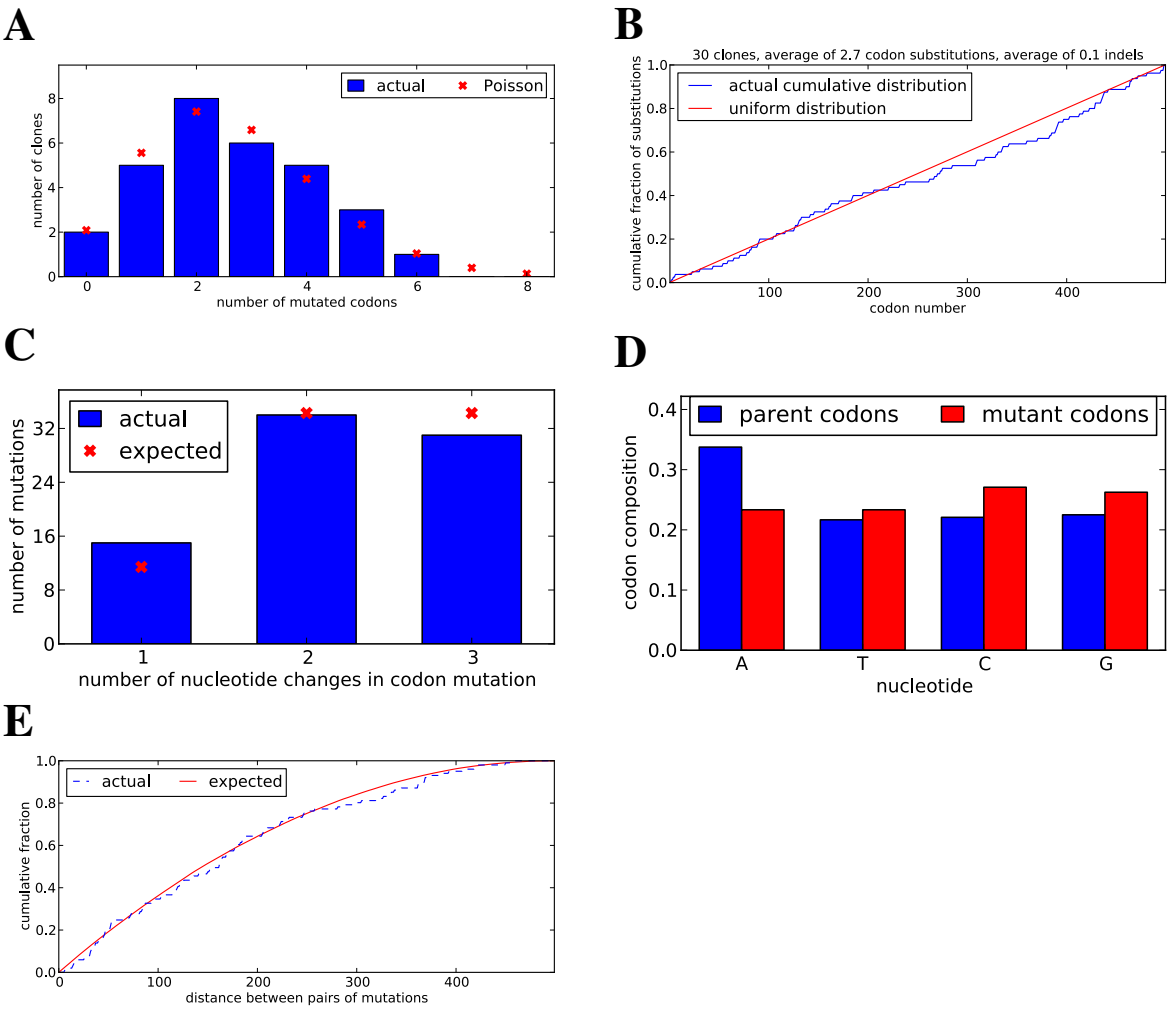
Supplementary table S5: For residues where mutations have previously been characterized as having large effects on the stability of the A/Aichi/2/1968 NP, the more stable amino acid has a higher preference and is also more frequent in actual NP sequences. The second column gives the experimentally measured change in melting temperature (ΔT_m) induced by the mutation to the A/Aichi/2/1968 NP as measured in (*24*) (note that these mutational effects on stability are largely conserved in other NPs (*25*)). The third column gives the frequencies of the amino acids in all 21,108 full-length NP sequences from influenza A (excluding bat lineages) in the Influenza Virus Resource as of January-31-2014. The fourth column gives the experimentally measured amino-acid preferences (Fig. 2D). The fifth column gives the expected evolutionary equilibrium frequency of the amino acids based on the experimentally measured preferences and mutational spectrum (fig. S5).

model	log likelihood	parameters (optimized + empirical)	AIC
experimental, combined replicates	-12338.9	0 (0 + 0)	24677.8
experimental, replicate A	-12372.8	0 (0 + 0)	24745.7
experimental, replicate B	-12392.0	0 (0 + 0)	24783.9
Halpern & Bruno, combined replicates	-12517.9	0 (0 + 0)	25035.7
Halpern & Bruno, replicate A	-12535.4	0 (0 + 0)	25070.8
Halpern & Bruno, replicate B	-12566.7	0 (0 + 0)	25133.3
GY94, branch-specific ω, multiple rates	-12769.1	556 (547 + 9)	26650.1
GY94, multiple ω, multiple rates	-12853.9	13 (4 + 9)	25733.8
KOSI07, branch-specific ω, multiple rates	-12891.7	607 (547 + 60)	26997.3
GY94, multiple ω, one rate	-12935.9	12 (3 + 9)	25895.8
KOSI07, multiple ω, multiple rates	-12999.9	64 (4 + 60)	26127.7
GY94, one ω, multiple rates	-13069.8	12 (3 + 9)	26163.5
KOSI07, multiple ω, one rate	-13154.8	63 (3 + 60)	26435.5
KOSI07, one ω, multiple rates	-13191.5	63 (3 + 60)	26508.9
GY94, one ω, one rate	-13205.0	11 (2 + 9)	26431.9
KOSI07, one ω, one rate	-13404.0	62 (2 + 60)	26932.0
randomized experimental, combined replicates	-14209.4	0 (0 + 0)	28418.8
randomized experimental, replicate A	-14243.7	0 (0 + 0)	28487.4
randomized experimental, replicate B	-14259.1	0 (0 + 0)	28518.2
randomized Halpern & Bruno, combined replicates	-14533.3	0 (0 + 0)	29066.5
randomized Halpern & Bruno, replicate B	-14618.5	0 (0 + 0)	29236.9
randomized Halpern & Bruno, replicate A	-14649.9	0 (0 + 0)	29299.9

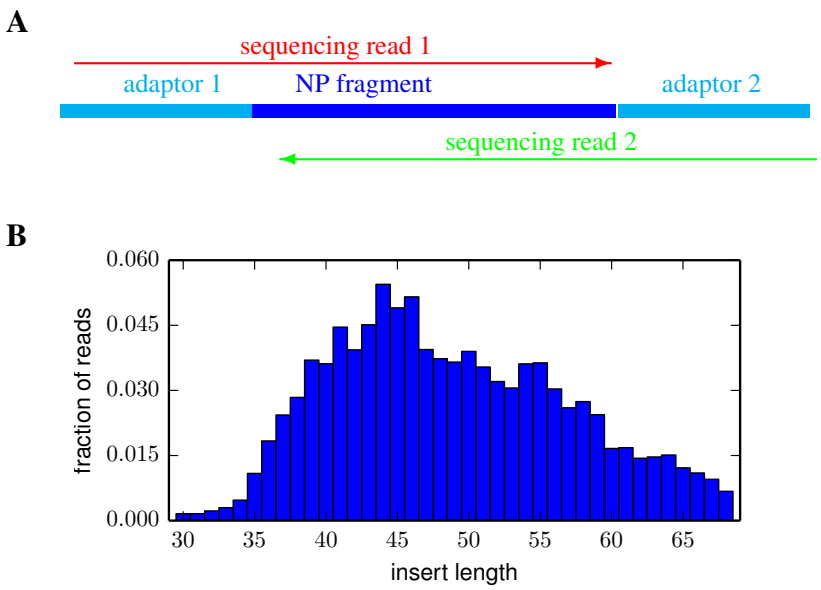
Supplementary table S6: Likelihoods for the various evolutionary models for the tree topology inferred with *codonPhyML* using *GY94* (Fig. 3). This table differs from Table 2 in that it also contains the results from using the experimentally determined evolutionary model defined by the equation of Halpern and Bruno (Equation 20).

model	log likelihood	parameters (optimized + empirical)	AIC
experimental, combined replicates	-12334.6	0 (0 + 0)	24669.2
experimental, replicate A	-12368.5	0 (0 + 0)	24737.1
experimental, replicate B	-12387.7	0 (0 + 0)	24775.4
Halpern & Bruno, combined replicates	-12513.0	0 (0 + 0)	25026.0
Halpern & Bruno, replicate A	-12530.3	0 (0 + 0)	25060.7
Halpern & Bruno, replicate B	-12562.0	0 (0 + 0)	25124.0
GY94, branch-specific ω, multiple rates	-12769.0	556 (547 + 9)	26650.0
GY94, multiple ω, multiple rates	-12850.8	13 (4 + 9)	25727.5
KOSI07, branch-specific ω, multiple rates	-12889.6	607 (547 + 60)	26993.3
GY94, multiple ω, one rate	-12932.4	12 (3 + 9)	25888.8
KOSI07, multiple ω, multiple rates	-12995.2	64 (4 + 60)	26118.4
GY94, one ω, multiple rates	-13069.1	12 (3 + 9)	26162.3
KOSI07, multiple ω, one rate	-13148.2	63 (3 + 60)	26422.5
KOSI07, one ω, multiple rates	-13188.7	63 (3 + 60)	26503.5
GY94, one ω, one rate	-13204.7	11 (2 + 9)	26431.4
KOSI07, one ω, one rate	-13401.0	62 (2 + 60)	26926.0
randomized experimental, combined replicates	-14205.2	0 (0 + 0)	28410.5
randomized experimental, replicate A	-14239.3	0 (0 + 0)	28478.7
randomized experimental, replicate B	-14255.3	0 (0 + 0)	28510.6
randomized Halpern & Bruno, combined replicates	-14528.4	0 (0 + 0)	29056.8
randomized Halpern & Bruno, replicate B	-14613.6	0 (0 + 0)	29227.1
randomized Halpern & Bruno, replicate A	-14645.0	0 (0 + 0)	29290.0

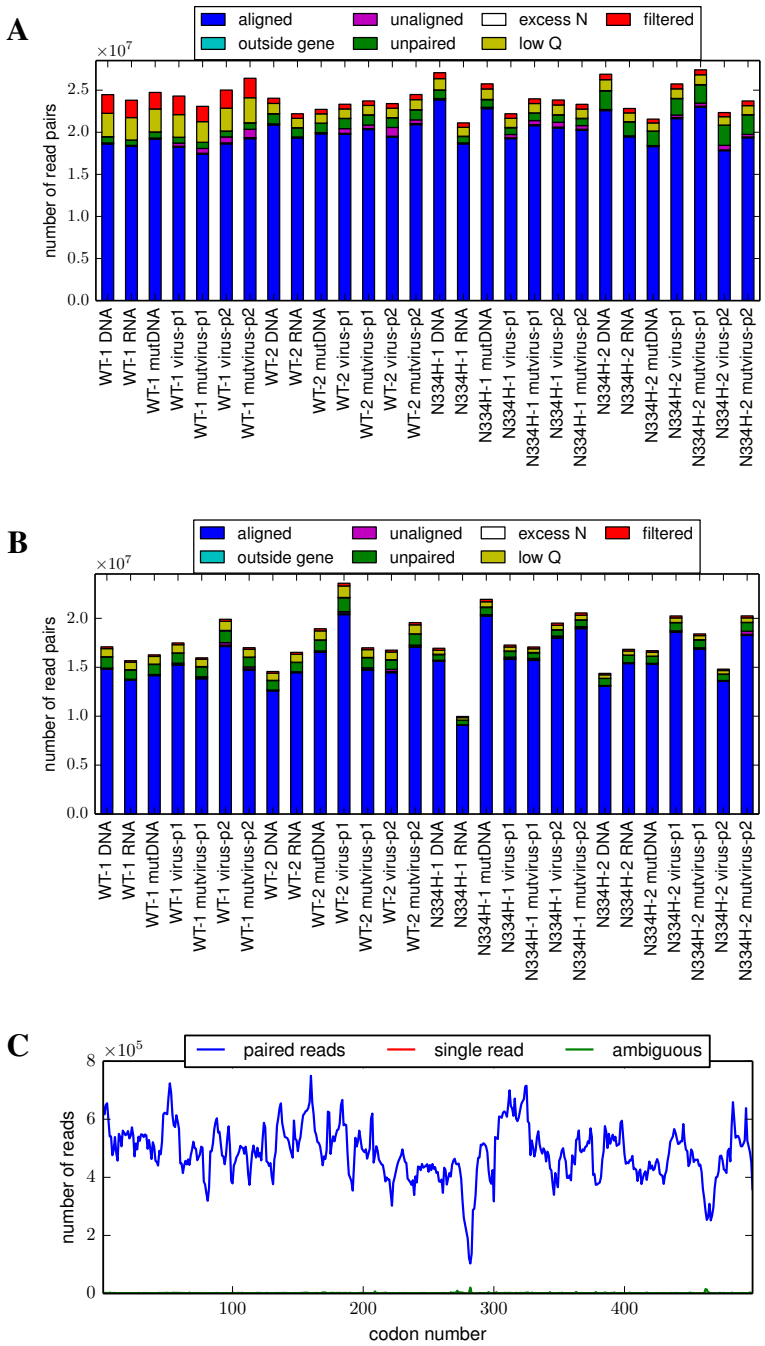
Supplementary table S7: Likelihoods for the various evolutionary models for the tree topology inferred with *codonPhyML* using *KOSI07* (fig. S6). This table differs from table S6 in that it optimizes the likelihoods on the tree topology from fig. S6 rather than the tree topology from Fig. 3.



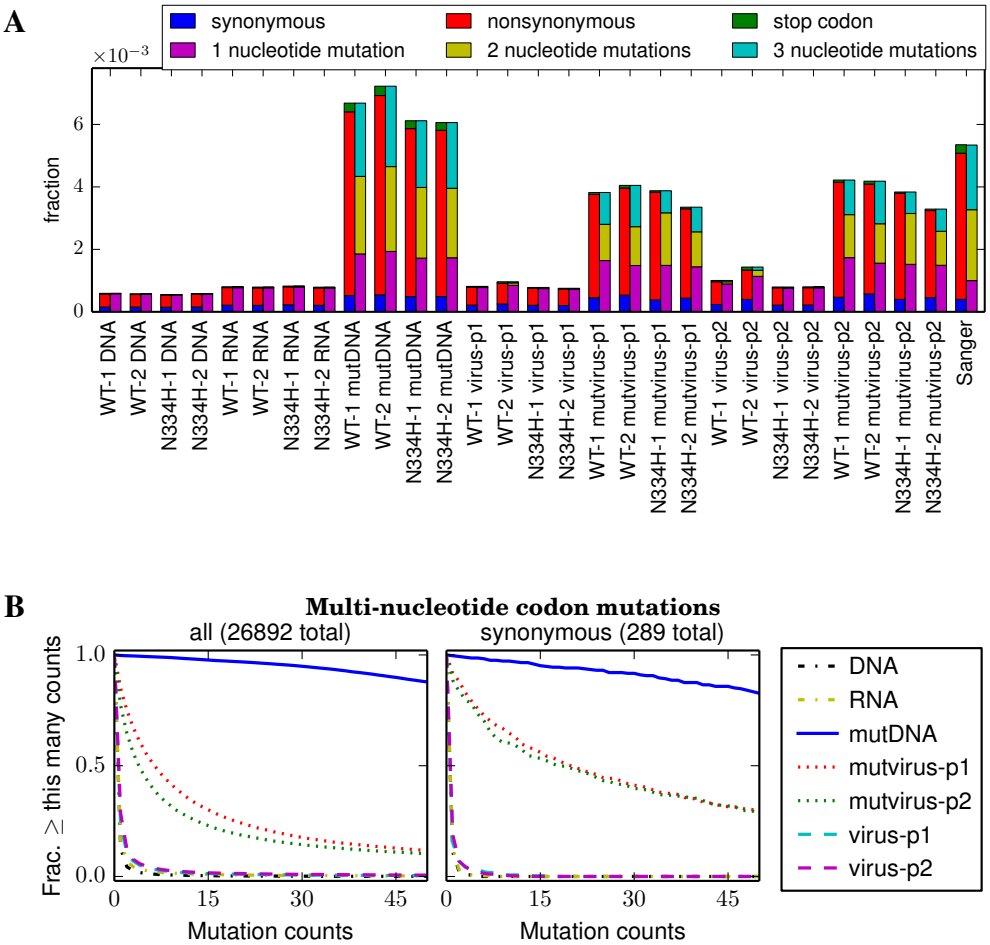
Supplementary fig. S1: Properties of the codon-mutant libraries as assessed by Sanger sequencing of 30 individual clones. The clones have an average of 2.7 codon mutations and 0.1 insertions / deletions per full-length NP coding sequence. **(A)** The number of mutations per gene follows an approximately a Poisson distribution, as expected if mutations occur independently. **(B)** Mutations occur uniformly along the primary sequence of the gene. **(C)** The number of nucleotide changes per codon mutation is roughly as expected if each codon is randomly mutated to any of the other 63 codons, but there is a slight elevation of single-nucleotide mutations. **(D)** The mutant codons have a uniform base composition. **(E)** In genes with multiple mutations, there is no tendency for mutations to cluster in primary sequence. Shown is the actual distribution of pairwise distances between mutations in all multiply mutated clones, as compared to the distribution generated by 1,000 simulations where mutations are placed randomly along the primary sequence of each multiple-mutant clone. The data and code used to generate this figure are available at `https://github.com/jbloom/SangerMutantLibraryAnalysis/tree/v0.1`.



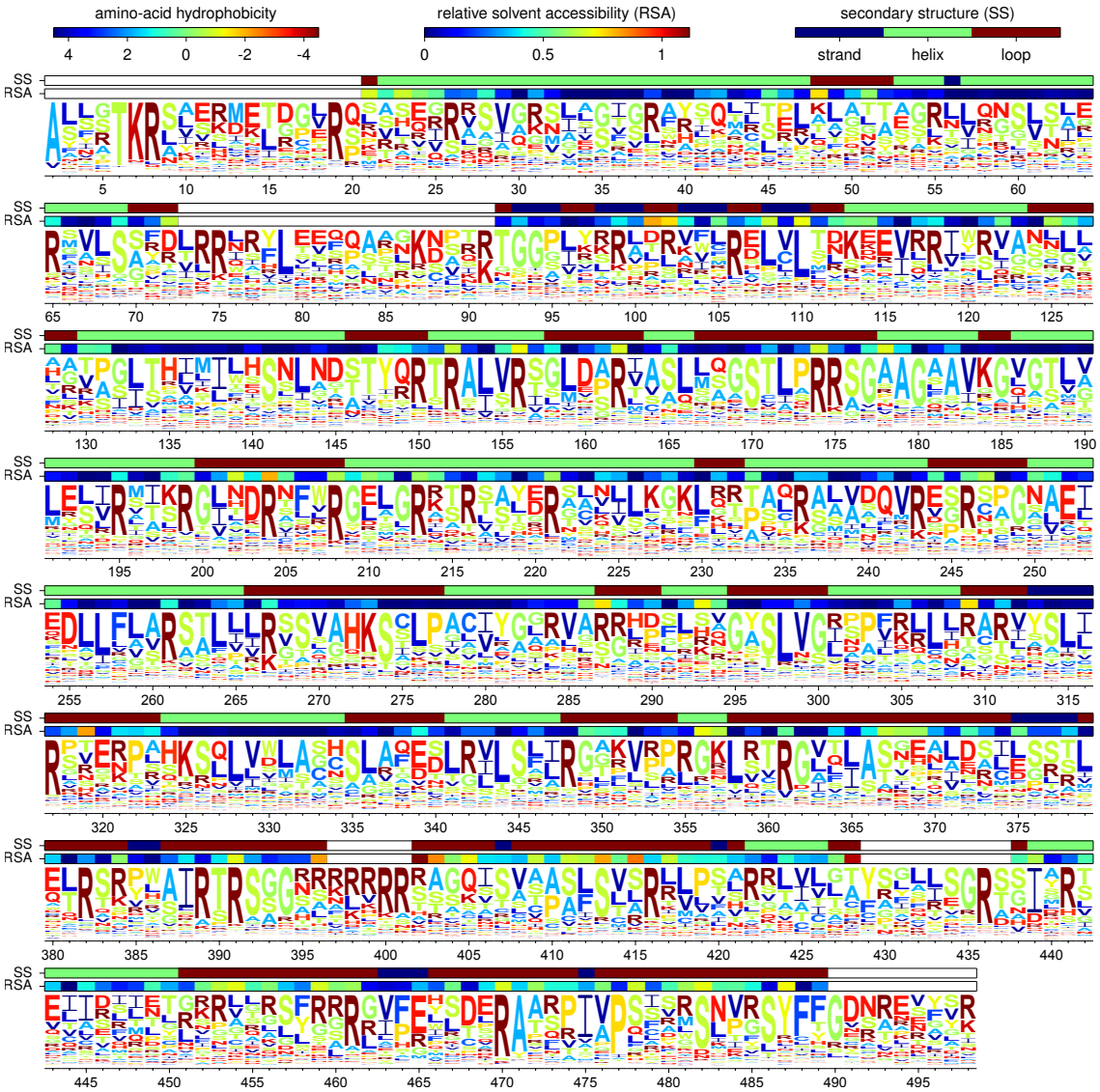
Supplementary fig. S2: Illumina sequencing accuracy was increased by using overlapping paired-end reads. **(A)** The strategy is to shear the NP fragments to about 50 nucleotides in length, and then sequence with overlapping paired-end reads. This provides double coverage, and only codon identities for which both reads agree are called. **(B)** The actual length distribution of alignable paired reads that had at least 30 nucleotides of overlap (lengths between 30 and 70 nucleotides) for a typical sample.



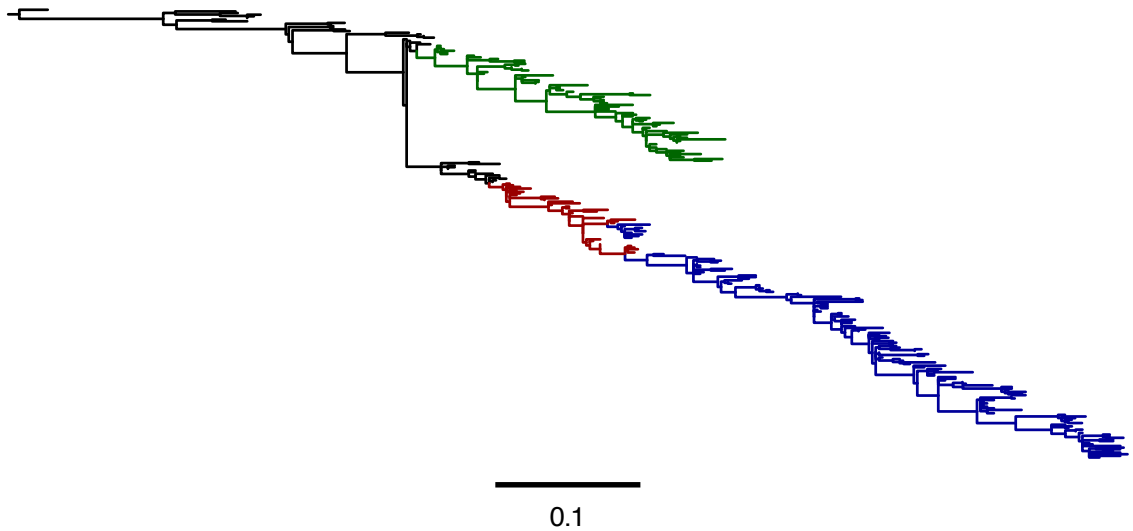
Supplementary fig. S3: The number of alignable paired reads for each sample from **(A)** *replicate A* and **(B)** *replicate B*. Most read pairs passed the various filters, could be overlapped, and could have their overlap aligned to NP. **(C)** The read depth across the primary sequence for a typical sample. The read depth was not entirely uniform probably due to biases in fragmentation locations. A full explanation of the computer code used to generate these figures is available at `http://jbloom.github.io/mapmuts/example_2013Analysis_Influenza_NP_Aichi68.html`.



Supplementary fig. S4: Plots for *replicate B* comparable to those shown for *replicate A* in Fig. 1B,C. Note that comparison of panel **B** here with Fig. 1C indicates that *replicate A* did a better job of sampling mutations in the mutant viruses than did *replicate B*. A full explanation of the computer code used to generate these figures is available at `http://jbloom.github.io/mapmuts/example_2013Analysis_Influenza_NP_Aichi68.html`.



Supplementary fig. S5: The expected frequencies of the amino acids at evolutionary equilibrium as calculated from Equation 24 using the experimentally determined evolutionary model from passage 1 of the combined replicates and Equation 19. Note that these expected frequencies are slightly different than the preferences in Fig. 2D due to the structure of the genetic code and the mutational rates. For instance, when arginine and lysine have equal preferences at a site, arginine will tend to have a higher evolutionary equilibrium frequency because it is encoded by more codons. The code used to generate this plot is available at `http://jbloom.github.io/phyloExpCM/example_2013Analysis_Influenza_NP_Human_1918_Descended.html`.



Supplementary fig. S6: Phylogenetic tree comparable to that in Fig. 3 except constructed with the *KOSI07* model rather than the *GY94* model. Note that the two trees are nearly identical, presumably because the extremely dense sampling of sequences leads to little ambiguity in the tree topology.

Supplementary file S1: The file *Preferences.xlsx* is an Excel table of the inferred amino-acid preferences at each site from the combined passage 1 data for the two replicates. This file contains the numerical data plotted in Fig. 2D. For the computer code and raw data used to generate this file, see `http://jbloom.github.io/mapmuts/example_2013Analysis_Influenza_NP_Aichi68.html`.

Supplementary file S2: The file *EvolutionaryEquilibriumFreqs.xlsx* is an Excel table of the expected equilibrium frequencies of the amino acids during evolution given the mutational spectrum in Table 1, the amino-acid preferences in file S1, and using the interpretation where these preferences are related to the fraction of the time that a given mutation is tolerated. This file contains the numerical data plotted in fig. S5. For the computer code and raw data used to generate this file, see `http://jbloom.github.io/phyloExpCM/example_2013Analysis_Influenza_NP_Human_1918_Descended.html`.