eLIFE
elifesciences.org

Figures and figure supplements

The inherent mutational tolerance and antigenic evolvability of influenza hemagglutinin

Bargavi Thyagarajan, Jesse D Bloom

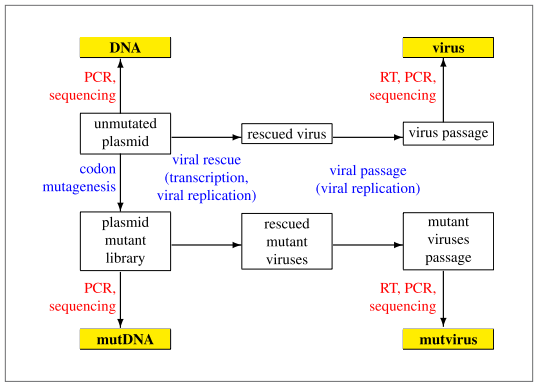

Figure 1. Schematic of the deep mutational scanning experiment. The Illumina deep-sequencing samples are shown in yellow boxes (**DNA**, **mutDNA**, **virus**, **mutvirus**). Experimental steps and associated sources of mutations are shown in blue text, while sources of error during Illumina sample preparation and sequencing are shown in red text. This entire process was performed in biological triplicate.
DOI: 10.7554/eLife.03300.003

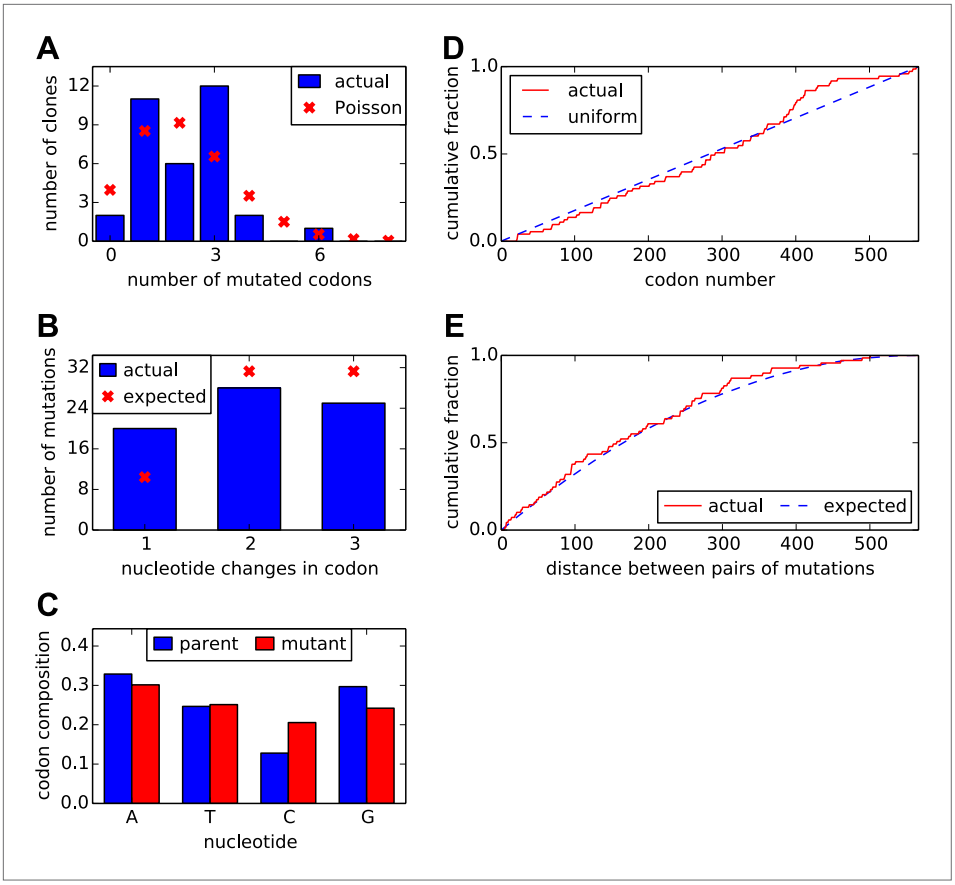

Figure 2. Properties of the HA codon-mutant library as assessed by Sanger sequencing of 34 individual clones drawn roughly evenly from the three experimental replicates. (**A**) There are an average of 2.1 codon mutations per clone, with the number per clone following a roughly Poisson distribution. (**B**) The codon mutations involve a mix of one-, two-, and three-nucleotide mutations. (**C**) The nucleotide composition of the mutant codons is roughly uniform. (**D**) The mutations are distributed uniformly along HA's primary sequence. (**E**) There is no tendency for mutations to cluster in primary sequence. Shown is distribution of observed pairwise distances between mutations in multiply mutated clones vs the expected distribution when the mutations are placed independently in the clones. All plots show results only for substitution mutations; insertion/deletion mutations are not shown. However, only two insertion/deletion mutations (0.06 per clone) were identified. The data and computer code used to generate this figure are at https://github.com/jbloom/SangerMutantLibraryAnalysis/tree/v0.2.

DOI: 10.7554/eLife.03300.004

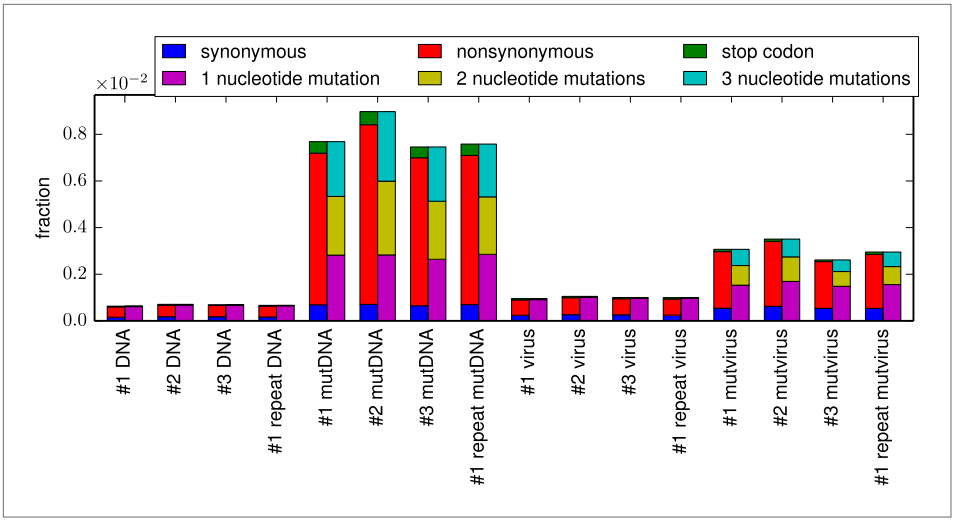

Figure 3. The per-codon frequencies of mutations in the samples. The samples are named as in ***Figure 1***, with the experimental replicate indicated with the numeric label. The **DNA** samples have a low frequency of mutations, and these mutations are composed almost entirely of single-nucleotide codon changes—these samples quantify the baseline error rate from PCR and deep sequencing. The mutation frequency is only slightly elevated in **virus** samples, indicating that viral replication and reverse transcription introduce only a small number of additional mutations. The **mutDNA** samples have a high frequency of single- and multi-nucleotide codon mutations, as expected from the codon mutagenesis procedure. The **mutvirus** samples have a lower mutation frequency, with most of the reduction due to fewer stop-codon and nonsynonymous mutations—consistent with purifying selection purging deleterious mutations. The data and code used to create this plot is available via http://jbloom.github.io/mapmuts/example_WSN_HA_2014Analysis.html; this plot is the file *parsesummary_codon_types_and_nmuts.pdf* described therein. The sequencing accuracy was increased by using overlapping paired-end reads as illustrated in ***Figure 3—figure supplement 1***. The overall number of overlapping paired-end reads for each sample is shown in ***Figure 3—figure supplement 2***. A representative plot of the read depth across the primary sequence is shown in ***Figure 3—figure supplement 3***.

DOI: 10.7554/eLife.03300.005

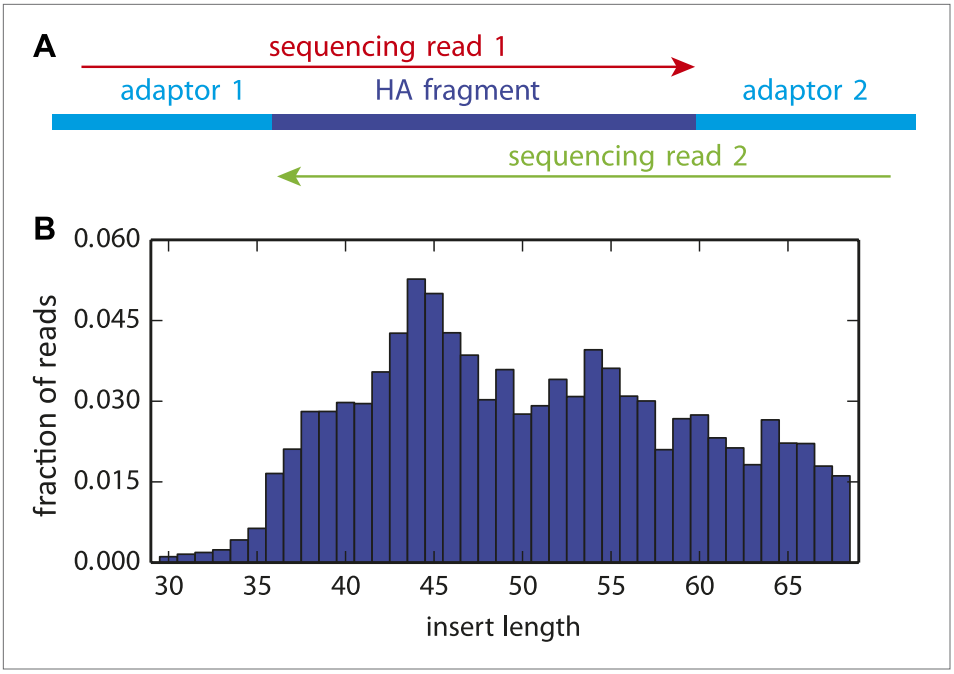

Figure 3—figure supplement 1. The overlapping paired-end Illumina sequencing strategy.
DOI: 10.7554/eLife.03300.006

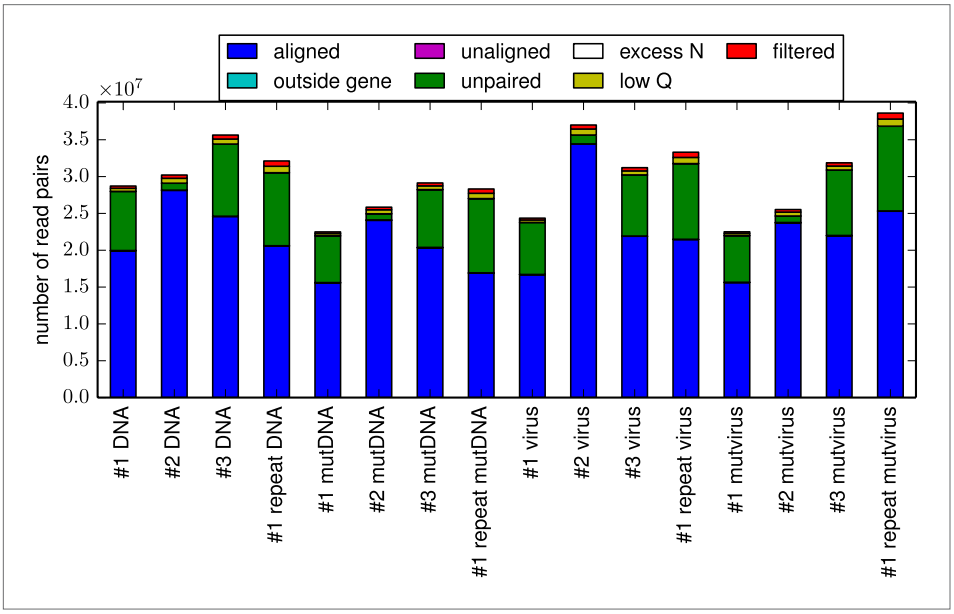

Figure 3—figure supplement 2. The total number of reads for each sample.
DOI: 10.7554/eLife.03300.007

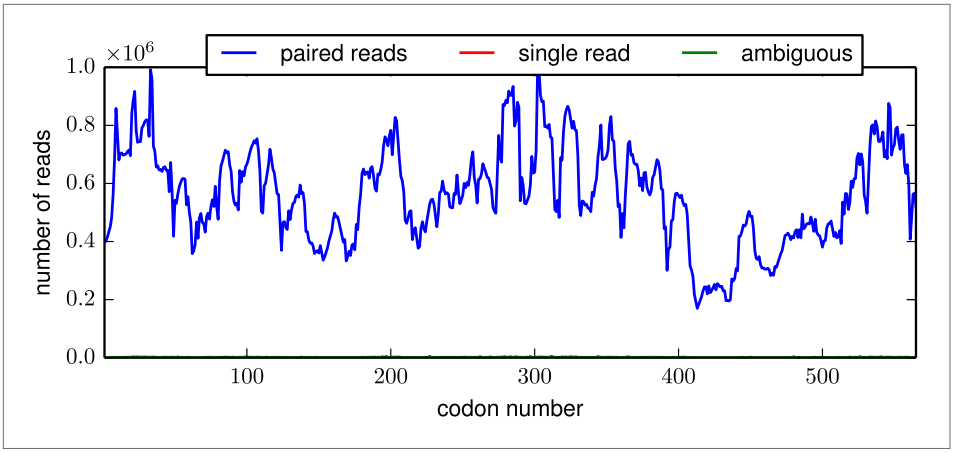

Figure 3—figure supplement 3. The per-codon read depth as a function of primary sequence.
DOI: 10.7554/eLife.03300.008

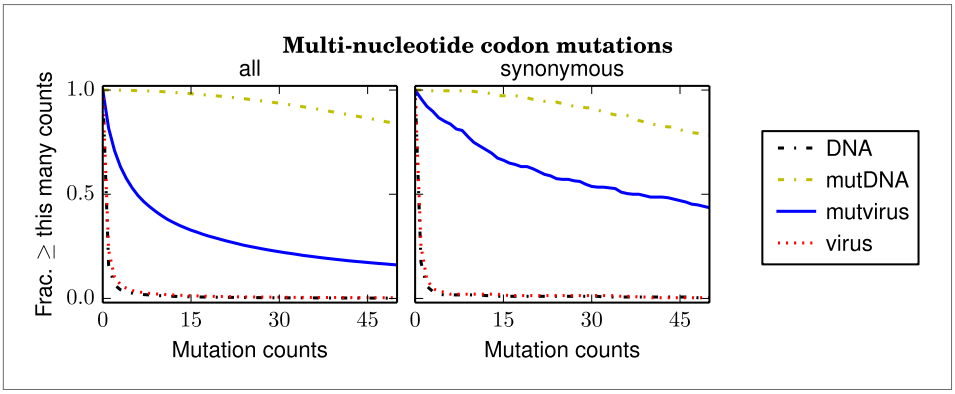

Figure 4. The number of times that each possible multi-nucleotide codon mutation was observed in each sample after combining the data for the three biological replicates. Nearly all mutations were observed many times in the **mutDNA** samples, indicating that the codon mutagenesis was comprehensive. Only about half of the mutations were observed at least five times in the **mutvirus** samples, indicating either a bottleneck during virus generation or purifying selection against many of the mutations. If the analysis is restricted to synonymous multi-nucleotide codon mutations, then about 85% of mutations are observed at least five times in the **mutvirus** samples. Since synonymous mutations are less likely to be eliminated by purifying selection, this latter number provides a lower bound on the fraction of codon mutations that were sampled by the mutant viruses. The redundancy of the genetic code means that the fraction of amino-acid mutations sampled is higher. The data and code used to create this figure are available via http://jbloom.github.io/mapmuts/example_WSN_HA_2014Analysis.html; this plot is the file *countparsedmuts_multi-nt-codonmutcounts.pdf* described therein. Similar plots for the individual replicates are shown in ***Figure 4—figure supplement 1***.
DOI: 10.7554/eLife.03300.009

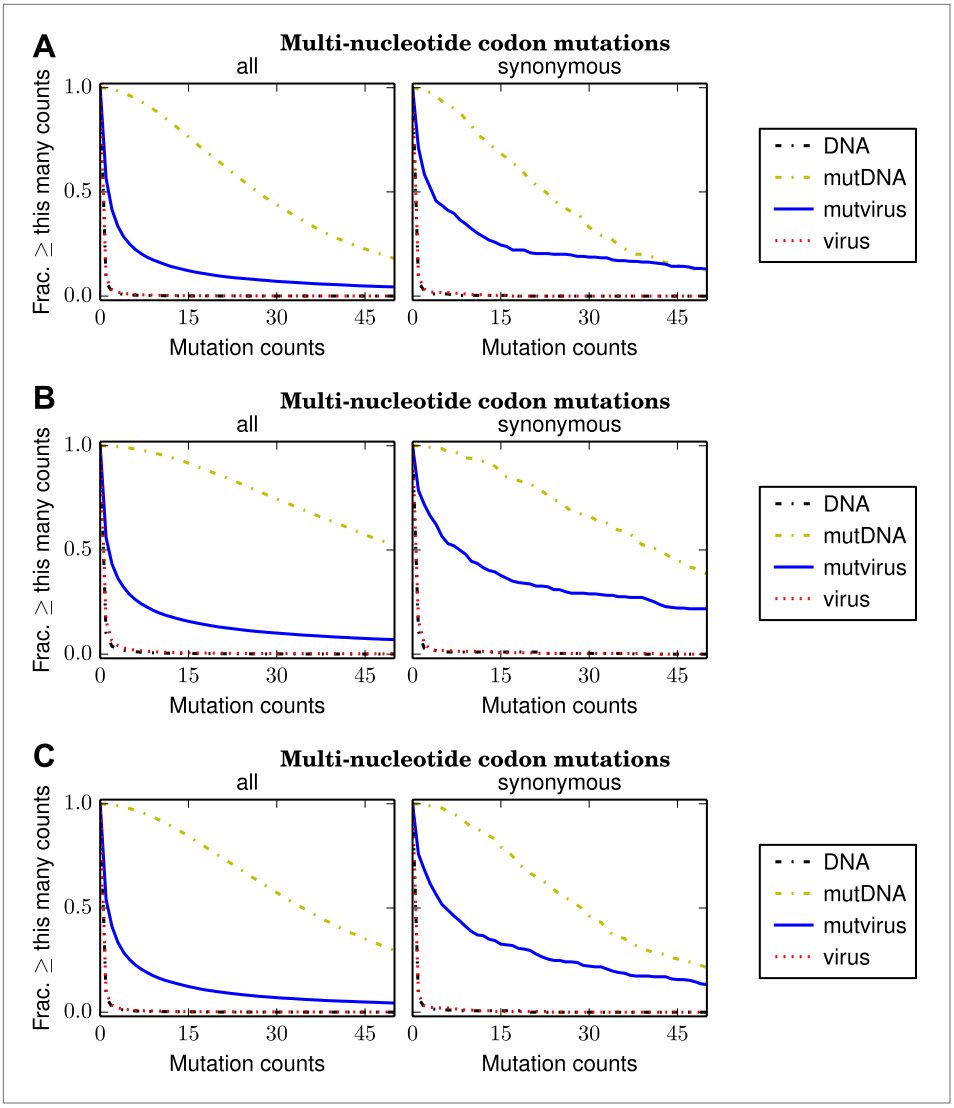

Figure 4—figure supplement 1. Plots like those in ***Figure 4*** for the individual biological replicates. (**A**) replicate 1, (**B**) replicate 2, and (**C**) replicate 3. These plots are the files *replicate_1/countparsedmuts_multi-nt-codonmutcounts.pdf*, *replicate_2/countparsedmuts_multi-nt-codonmutcounts.pdf*, and *replicate_3/countparsedmuts_multi-nt-codonmutcounts.pdf* described at http://jbloom.github.io/mapmuts/example_WSN_HA_2014Analysis.html.

DOI: 10.7554/eLife.03300.010

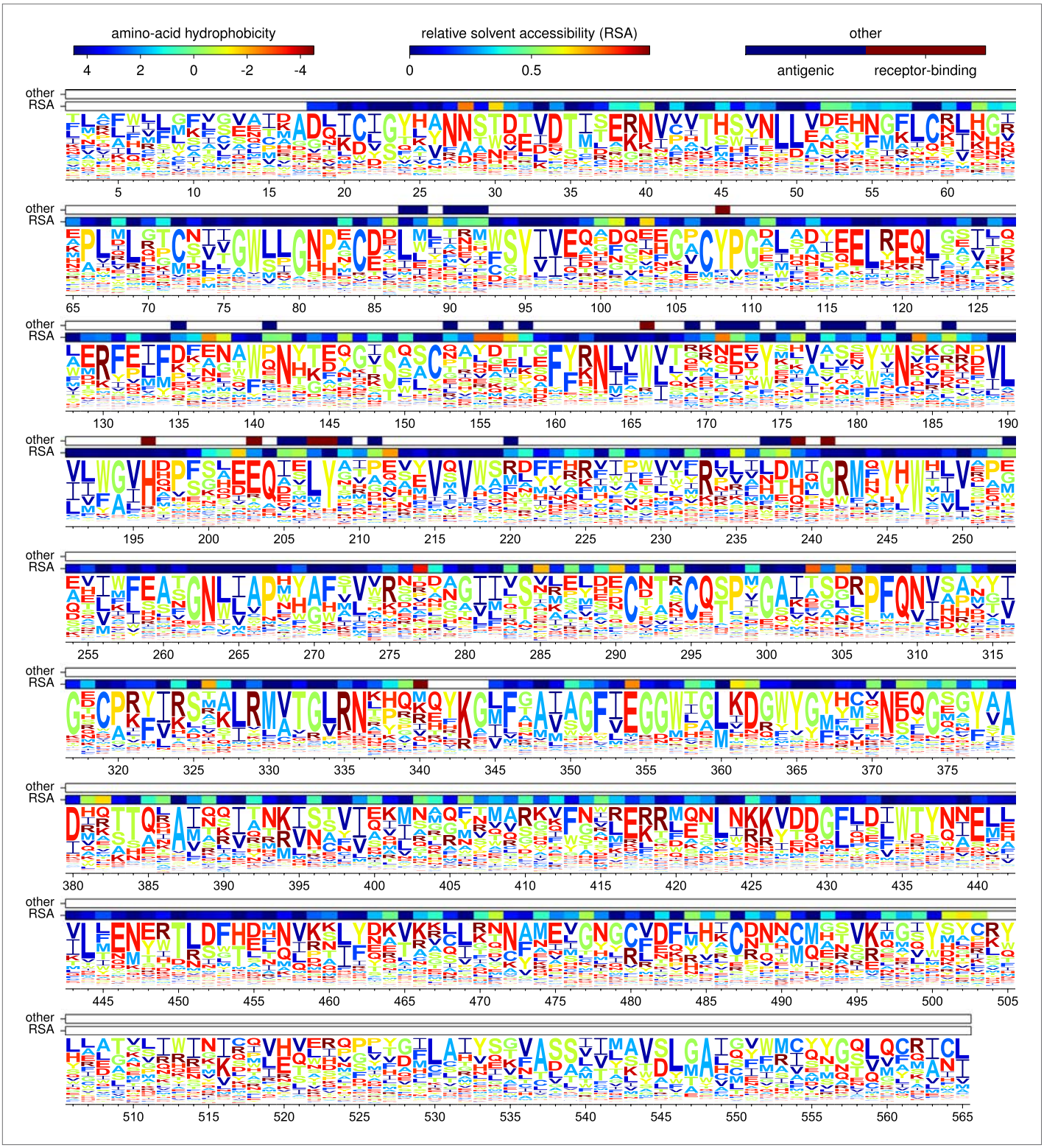

Figure 5. The amino-acid preferences inferred using the combined data from the three biological replicates. The letters have heights proportional to the preference for that amino acid, and are colored by hydrophobicity. The first overlay bar shows the relative solvent accessibility (RSA) for residues in the HA crystal structure. The second overlay bar indicates Caton et al. antigenic sites or conserved receptor-binding residues. The sequence is numbered sequentially beginning with 1 at the N-terminal methionine—however, this first methionine is not shown as it was not mutagenized. ***Figure 5—figure supplement 1*** shows the same data with H3 numbering of the sequence. The data and code used to create this figure are available via http://jbloom.github.io/mapmuts/example_WSN_HA_2014Analysis.html; this plot is the file *sequentialnumbering_site_preferences_logoplot.pdf* described therein.

DOI: 10.7554/eLife.03300.011

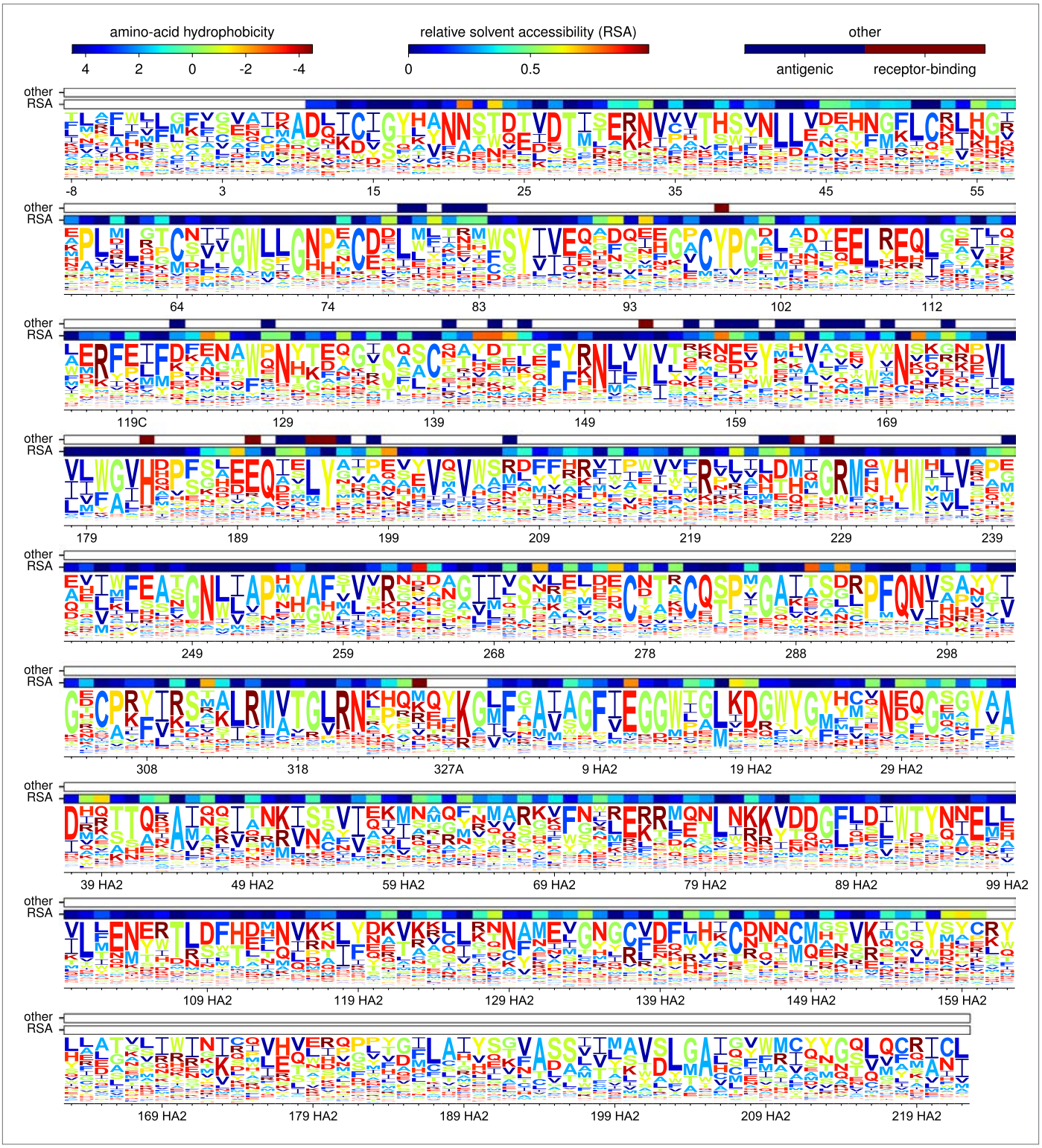

Figure 5—figure supplement 1. A plot matching that shown in ***Figure 5*** except that the HA sequence is numbered using the H3 numbering scheme.

DOI: 10.7554/eLife.03300.012

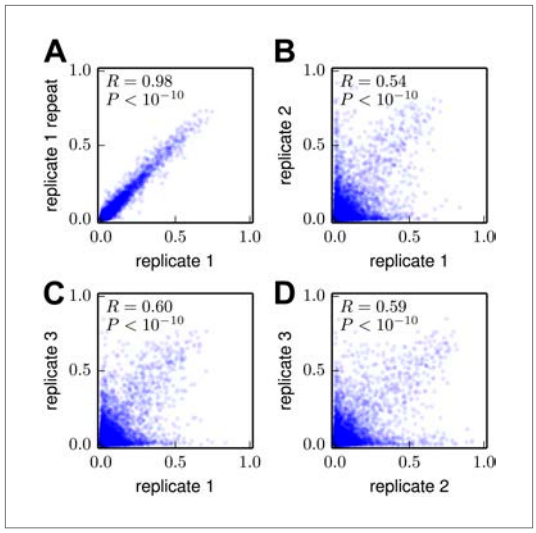

Figure 6. Correlations among the amino-acid preferences inferred using data from the individual biological replicates. (**A**) The preferences from two technical repeats of the sample preparation and deep sequencing of biological replicate #1 are highly correlated. (**B**)–(**D**) The preferences from the three biological replicates are substantially but imperfectly correlated. Overall, these results indicate that technical variation in sample preparation and sequencing is minimal, but that there is substantial variation between biological replicates due to stochastic differences in which mutant viruses predominate during the initial reverse-genetics step. The Pearson correlation coefficient (*R*) and associated p-value are shown in the upper-left corner of each plot. The data and code used to create this figure are available via http://jbloom.github.io/mapmuts/example_WSN_HA_2014Analysis.html; these plots are the files *correlations/replicate_1_vs_replicate_1_repeat.pdf*, *correlations/replicate_1_vs_replicate_2.pdf*, *correlations/replicate_1_vs_replicate_3.pdf*, and *correlations/replicate_2_vs_replicate_3.pdf* described therein.

DOI: 10.7554/eLife.03300.014

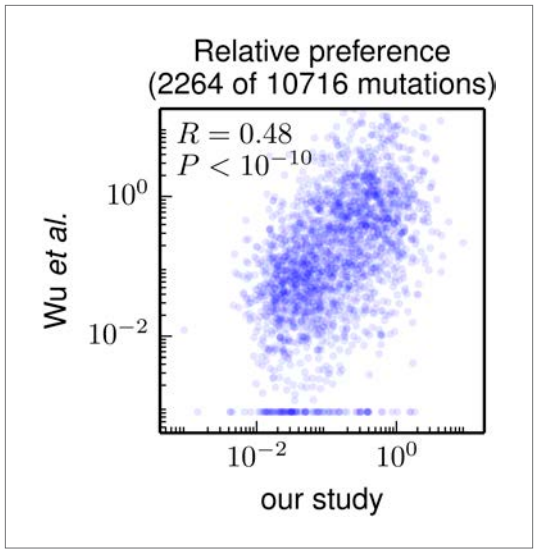

Figure 7. Correlation of the site-specific amino-acid preferences determined in our study with the "relative fitness" (RF) values reported by ***Wu et al. (2014)***.

Wu et al. (2014) report RF values for 2350 of the 564×19 = 10716 possible amino-acid mutations to the WSN HA examined in our study (they only examine single-nucleotide changes and disregard certain types of mutations due to oxidative damage of their DNA). To compare across the data sets, we have normalized their RF values by the RF value for the wildtype amino-acid (which they provide for only 2264 of the 2350 mutations). We then correlate on a logarithmic scale these normalized RF values with the ratio of our measurement of the preference for the mutant amino acid divided by the preference for the wildtype amino acid, using the preferences from our combined replicates. For mutations for which ***Wu et al. (2014)*** report an RF of zero, we assign a normalized RF equal to the smallest value for their entire data set. There is a significant Pearson correlation of 0.48 between the data sets, indicating that both our experiments and those of ***Wu et al. (2014)*** are capturing many of the same constraints on HA. The data and code used to create this figure are available via http://jbloom.github.io/mapmuts/example_WSN_HA_2014Analysis.html; this plot is the file *correlation_with_Wu_et_al.pdf* described therein.

DOI: 10.7554/eLife.03300.015

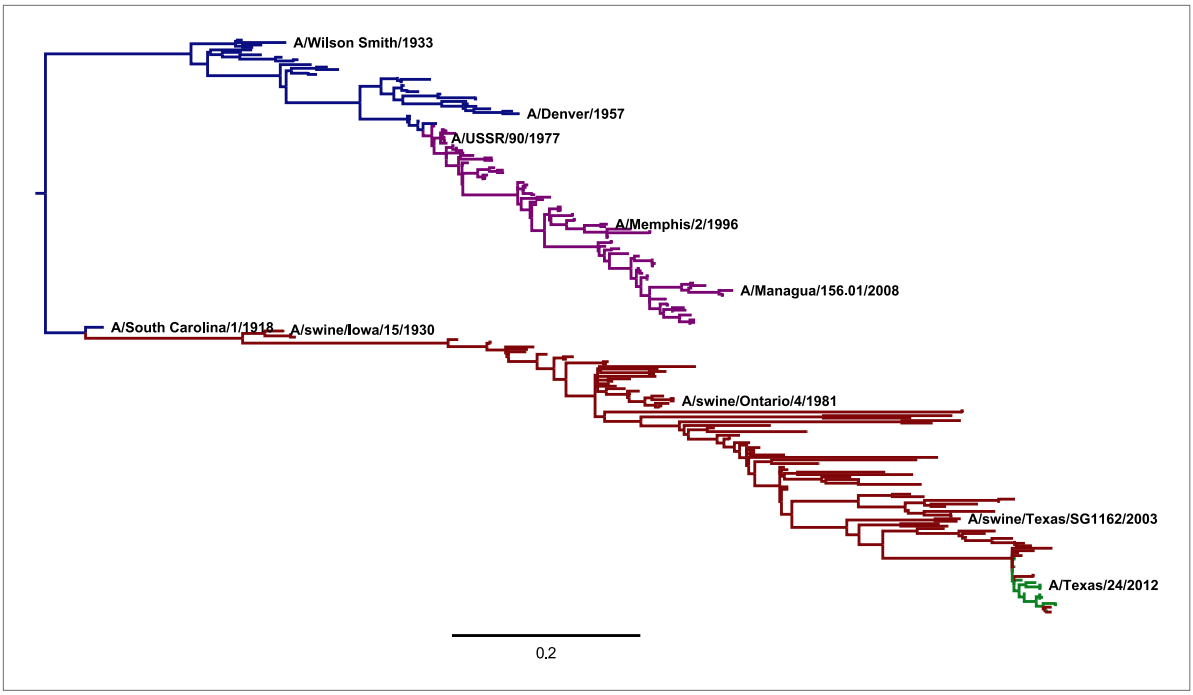

Figure 8. A phylogenetic tree of human and swine H1 HA sequences descended from a common ancestor closely related to the 1918 virus. The WSN virus used in the experiments here is a lab-adapted version of the *A/Wilson Smith/1933* strain. Human H1N1 that circulated from 1918 until 1957 is shown in blue. Human seasonal H1N1 that reappeared in 1977 is shown in purple. Swine H1N1 is shown in red. The 2009 pandemic H1N1 is shown in green. This tree was constructed using *codonPhyML* (***Gil et al., 2013***) with the substitution model of ***Goldman and Yang (1994)***. This plot is the file *CodonPhyML_Tree_H1_HumanSwine_GY94/annotated_tree.pdf* described at http://jbloom.github.io/phyloExpCM/example_2014Analysis_Influenza_H1_HA.html. ***Figure 8—figure supplement 1*** shows a tree estimated for the same sequences using the substitution model of ***Kosiol et al. (2007)***.

DOI: 10.7554/eLife.03300.016

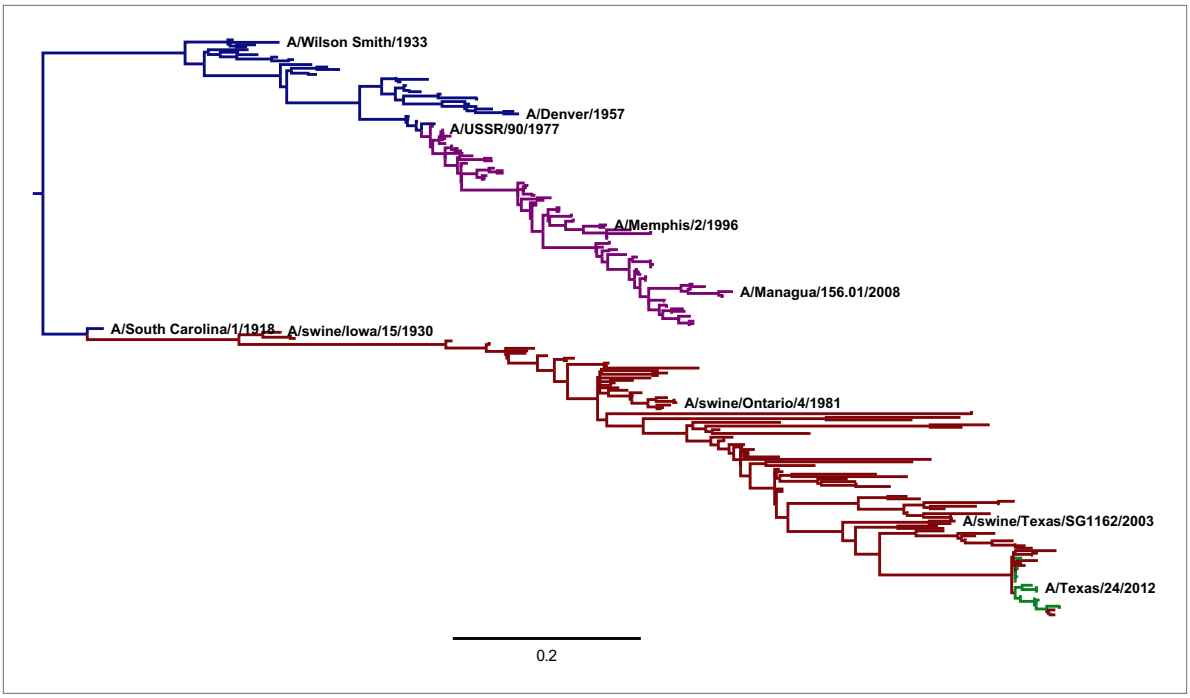

Figure 8—figure supplement 1. A phylogenetic tree of the same sequences shown in ***Figure 8***, this time inferred using the substitution model of ***Kosiol et al. (2007)***.

DOI: 10.7554/eLife.03300.017

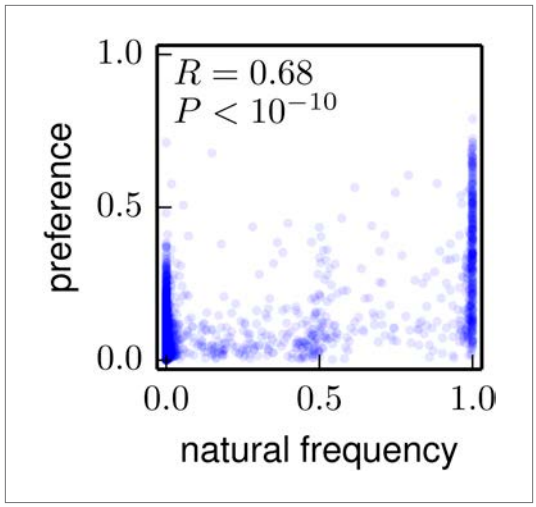

Figure 9. The frequencies of amino acids among the naturally occurring HA sequences in ***Figure 8*** vs the amino-acid preferences inferred from the combined replicates (***Figure 5***). Note that a natural frequency close to one or zero could indicate absolute selection for or against a specific amino acid, but could also simply result from the fact that natural evolution has not completely sampled all possible mutations compatible with HA structure and function. The Pearson correlation coefficient (*R*) and associated p-value are shown on the plot. This plot is the file *natural_frequency_vs_preference.pdf* described at http://jbloom.github.io/phyloExpCM/example_2014Analysis_Influenza_H1_HA.html.

DOI: 10.7554/eLife.03300.018

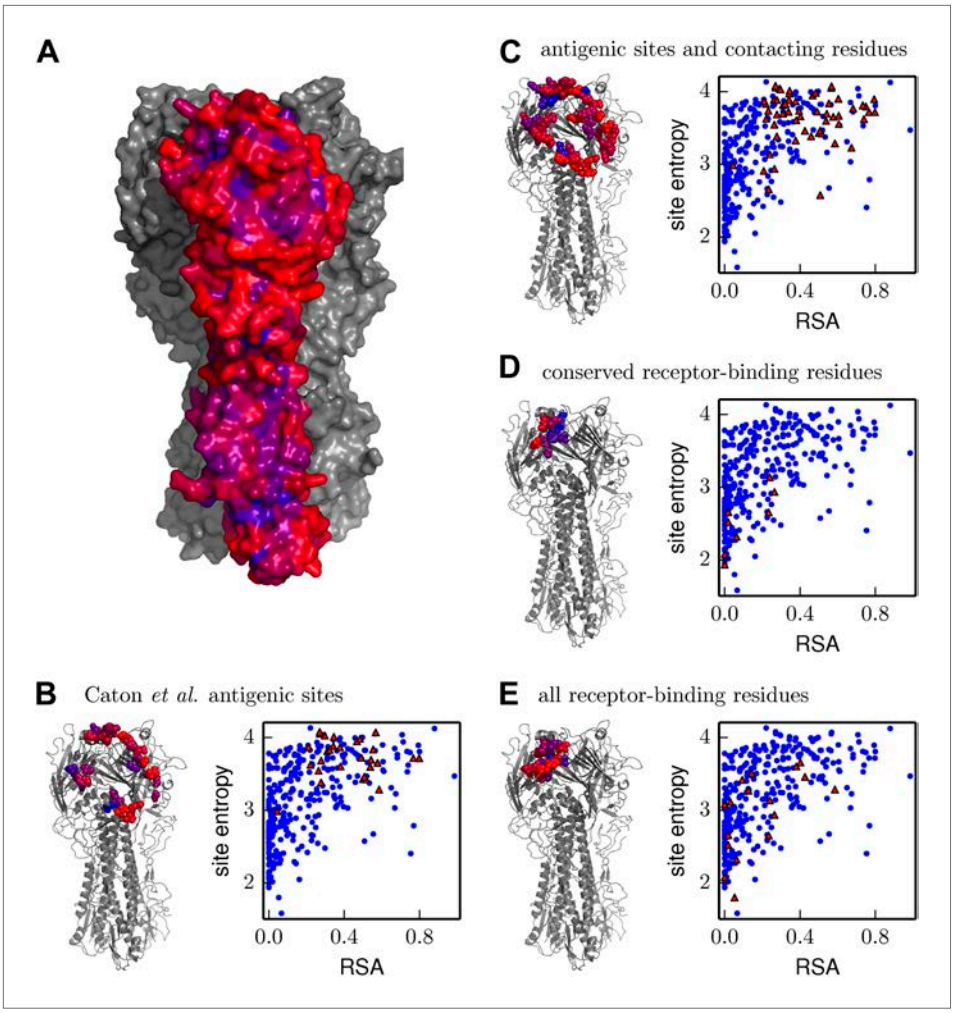

Figure 10. Inherent mutational tolerance of HA's receptor-binding residues and antigenic sites. (**A**) Surface of HA with one monomer colored by site entropy as determined by the deep mutational scanning; blue indicates low mutational tolerance and red indicates high mutational tolerance. (**B**) The structure shows residues classified as antigenic sites by ***Caton et al. (1982)*** in colored spheres; the plot shows site entropy vs relative solvent accessibility (RSA) of these residues (red triangles) and all other HA1 residues in the crystal structure (blue circles). (**C**) Antigenic sites of ***Caton et al. (1982)*** plus all other surface-exposed residues that contact these sites. (**D**) Conserved receptor-binding residues. (**E**) All receptor-binding residues. ***Table 4*** shows that residues in (**B**) and (**C**) have unusually high mutational tolerance, residues in (**D**) have unusually low mutational tolerance, and residues in (**E**) do not have unusual mutational tolerance. The data and code to create all panels of this figure is provided via http://jbloom.github.io/mapmuts/example_WSN_HA_2014Analysis.html. The structure is PDB 1RVX (***Gamblin et al., 2004***).

DOI: 10.7554/eLife.03300.021

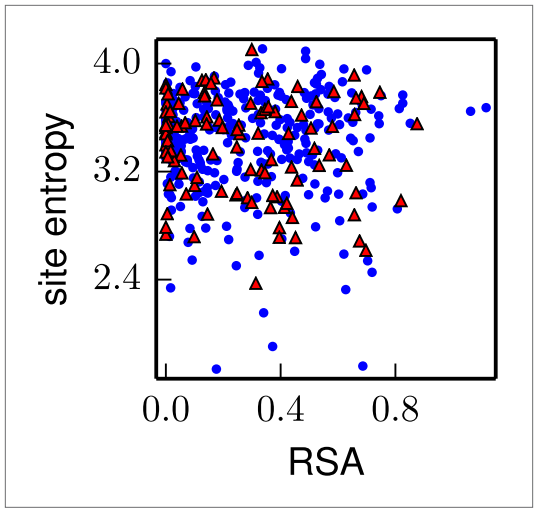

Figure 11. The inherent mutational tolerance of NP's CTL epitopes is indistinguishable from that of non-epitope sites in NP. The plot shows the site entropy vs relative solvent accessibility (RSA) of NP residues that participate in multiple CTL epitopes (red triangles) and all other NP residues in the crystal structure (blue circles). Visual inspection suggests that the epitope sites have mutational tolerance comparable to other sites, and this result is supported by the statistical analysis in ***Table 5***. Note that unlike for HA, there is no trend for RSA to correlate with site entropy—this could be because many of NP's surface-exposed sites are constrained by interactions with viral RNA. The CTL epitopes are those delineated in the first supplementary table of ***Gong and Bloom (2014)***. The site entropies are computed from a previously described deep mutational scan of NP, and are the values in the first supplementary file of ***Bloom (2014)***; the RSA values are also taken from that reference. The data and code used to generate this plot is available via http://jbloom.github.io/mapmuts/example_WSN_HA_2014Analysis.html; the plot itself is the file *NP_CTL_entropy_rsa_correlation.pdf* described therein.
DOI: 10.7554/eLife.03300.023