OPEN ACCESS Freely available online

PLOS | GENETICS

Influenza Virus Drug Resistance: A Time-Sampled Population Genetics Perspective

Matthieu Foll[1,2☯], Yu-Ping Poh[2,3☯], Nicholas Renzette[4], Anna Ferrer-Admetlla[1,2,5], Claudia Bank[1,2], Hyunjin Shim[1,2], Anna-Sapfo Malaspinas[6], Gregory Ewing[1,2], Ping Liu[7], Daniel Wegmann[2,5], Daniel R. Caffrey[7], Konstantin B. Zeldovich[3], Daniel N. Bolon[8], Jennifer P. Wang[7], Timothy F. Kowalik[4], Celia A. Schiffer[8], Robert W. Finberg[7], Jeffrey D. Jensen[1,2]*

1 School of Life Sciences, École Polytechnique Fédérale de Lausanne (EPFL), Lausanne, Switzerland, **2** Swiss Institute of Bioinformatics (SIB), Lausanne, Switzerland, **3** Program in Bioinformatics and Integrative Biology, University of Massachusetts Medical School, Worcester, Massachusetts, United States of America, **4** Department of Microbiology and Physiological Systems, University of Massachusetts Medical School, Worcester, Massachusetts, United States of America, **5** Department of Biology and Biochemistry, University of Fribourg, Fribourg, Switzerland, **6** Center for GeoGenetics, Natural History Museum of Denmark, University of Copenhagen, Copenhagen, Denmark, **7** Department of Medicine, University of Massachusetts Medical School, Worcester, Massachusetts, United States of America, **8** Department of Biochemistry and Molecular Pharmacology, University of Massachusetts Medical School, Worcester, Massachusetts, United States of America

Abstract

The challenge of distinguishing genetic drift from selection remains a central focus of population genetics. Time-sampled data may provide a powerful tool for distinguishing these processes, and we here propose approximate Bayesian, maximum likelihood, and analytical methods for the inference of demography and selection from time course data. Utilizing these novel statistical and computational tools, we evaluate whole-genome datasets of an influenza A H1N1 strain in the presence and absence of oseltamivir (an inhibitor of neuraminidase) collected at thirteen time points. Results reveal a striking consistency amongst the three estimation procedures developed, showing strongly increased selection pressure in the presence of drug treatment. Importantly, these approaches re-identify the known oseltamivir resistance site, successfully validating the approaches used. Enticingly, a number of previously unknown variants have also been identified as being positively selected. Results are interpreted in the light of Fisher's Geometric Model, allowing for a quantification of the increased distance to optimum exerted by the presence of drug, and theoretical predictions regarding the distribution of beneficial fitness effects of contending mutations are empirically tested. Further, given the fit to expectations of the Geometric Model, results suggest the ability to predict certain aspects of viral evolution in response to changing host environments and novel selective pressures.

Citation: Foll M, Poh Y-P, Renzette N, Ferrer-Admetlla A, Bank C, et al. (2014) Influenza Virus Drug Resistance: A Time-Sampled Population Genetics Perspective. PLoS Genet 10(2): e1004185. doi:10.1371/journal.pgen.1004185

Editor: Joshua M. Akey, University of Washington, United States of America

Received August 15, 2013; **Accepted** January 6, 2014; **Published** February 27, 2014

Funding: The authors wish to acknowledge the support of DARPA (Prophecy Program, Defense Advanced Research Agency (http://www.darpa.mil/), Defense Sciences Office (DSO), Contract No. HR0011-11-C-0095) and the contributions of all the members of the ALiVE (Algorithms to Limit Viral Epidemics) working group. Additional funding came from grants from the Swiss National Science Foundation (http://www.snf.ch/E/Pages/default.aspx), and a European Research Council Starting Grant (ERC; http://erc.europa.eu/) to JDJ. ASM was funded by an Early Postdoc Mobility fellowship from the Swiss National Science Foundation. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: jeffrey.jensen@epfl.ch

☯ These authors contributed equally to this work.

Introduction

Influenza A virus (IAV) is an important human pathogen, resulting in approximately 36,000 deaths annually in the United States [1] and eliciting constant concerns regarding the spread of new pandemic strains [2–4]. IAV is an eight segment RNA virus that can rapidly evolve owing to a high mutation rate [5], genomic reassortment [6], and stochastic migration of virus from isolated human populations [7] or animal reservoirs [8]. The most common therapies for IAV infections include neuraminidase inhibitors, including the widely used oseltamivir. Oseltamivir was initially designed based on structural information [9], and has been shown to be a competitive inhibitor of the neuramindase active site [10]. Due to the mechanism of action of oseltamivir, it was widely believed that the evolution of drug resistance would decrease fitness of the virus and therefore, be unlikely to be of importance in a clinical setting [11]. However, oseltamivir resistance has been shown to evolve quickly in human hosts [12,13] and pandemic H1N1 IAV isolates developed resistance to the drug [14]. The most common resistance mutation of H1N1 strains is the H275Y mutation (N2 numbering) which is located near the neuraminidase active site and attenuates oseltamivir binding [10]. The recent rise of oseltamivir resistance in clinical isolates is likely due to the presence of compensatory mutations in the neuraminidase (NA) and hemagglutinin (HA) genes that increase the fitness of the H275Y resistance mutant [15–17].

Here, we describe the analysis of IAV populations during the evolution of drug resistance during in vitro growth. This system offers an ideal platform to study the relative effects of genetic drift and selection in evolution, as a target of selection, specifically the

Author Summary

In recent years, considerable attention has been given to the evolution of drug resistance in the influenza A H1N1 strain. As a major annual cause of morbidity and mortality, combined with the rapid global spread of drug resistance, influenza remains as one of the most important global health concerns. Our work here focuses on a novel multi-faceted population-genetic approach utilizing unique whole-genome multi-time point experimental datasets in both the presence and absence of drug treatment. In addition, we present novel theoretical results and two newly developed and widely applicable statistical methodologies for utilizing time-sampled data – with a focus on distinguishing the relative contribution of genetic drift from that of positive and purifying selection. Results illustrate the available mutational paths to drug resistance, and offer important insights in to the mode and tempo of adaptation in a viral population.

H275Y mutation, is known prior to analysis. Further, in vitro growth platforms allow for the control and knowledge of demographic parameters, particularly the severity of population bottlenecks – thus allowing insight into the expected role of genetic drift. Lastly, the high mutation rate and short generation time of IAV allows for adaptation to occur on experimentally tractable time scales.

This experimental set-up allows for an additional benefit – namely, time-sampled whole-genome data. This added temporal dimension provides an important component in the puzzle of disentangling selection and demography – as it becomes possible to utilize analytical results describing the change in frequency [18] and sojourn time [19] of beneficial mutations. Thus, time-sampled data allow the trajectory of any individual allele to be used to better identify the action of natural selection, rather than simply the patterns of genomic variation as utilized by standard single time-point site-frequency spectrum based statistics [20].

Utilizing this experimental approach and the above reasoning, we have tested and developed novel statistical tests of selection for time-sampled population data. We infer effective population size (N_e) in this platform, and develop novel analytical-, maximum likelihood-, and approximate Bayesian -based approaches to determine the contributions of genetic drift and selection in this biological system. Finally, based on this population genetic inference, we demonstrate that IAV development of drug resistance follows the expectations of Fisher's Geometric Model, offering a novel approach to predicting viral evolution in response to changing host environments and novel selective pressures.

Results and Discussion

Influenza A/Brisbane/59/2007 (H1N1) was initially serially amplified on Madin-Darby canine kidney (MDCK) cells for three passages. The samples were then passaged either in the absence of drug, or in the presence of increasing concentrations of oseltamivir, a neuraminidase inhibitor (Figure 1). At the end of each passage, samples were collected for whole genome high throughput population sequencing providing a high depth of coverage. In addition, biological replicates of the entire experiment were performed and analyzed. We first focus on one of the two experiments in the following results, and then use the replicate as a point of comparison. The genetic diversity calculated as the average expected heterozygosity [21] in each passage was very low, stable throughout the entire experiment, and slightly lower during oseltamivir treatment (6.2×10^{-4} vs. 4.5×10^{-4}, see Figure S1). The frequency spectra indicated that most single nucleotide polymorphisms (SNPs) were segregating at low frequency and the shape was comparable at P4 and P12 (Figure S2). The number of new mutations accumulated within each passage was limited and very rarely reached high frequencies, in particular in the presence of oseltamivir, suggesting that the viral populations were under severe purifying selection. As a consequence, nearly all observed SNPs were biallelic: over all passages and nucleotides, the frequency of the third allele was 0.02% on average, with a 99% quantile of 0.1%. For this reason, we considered all SNPs as biallelic in our subsequent analyses. Finally, we observed 4 and 7 newly arising mutations reaching a frequency of more than 50% during our experiment in the absence and presence of oseltamivir, respectively.

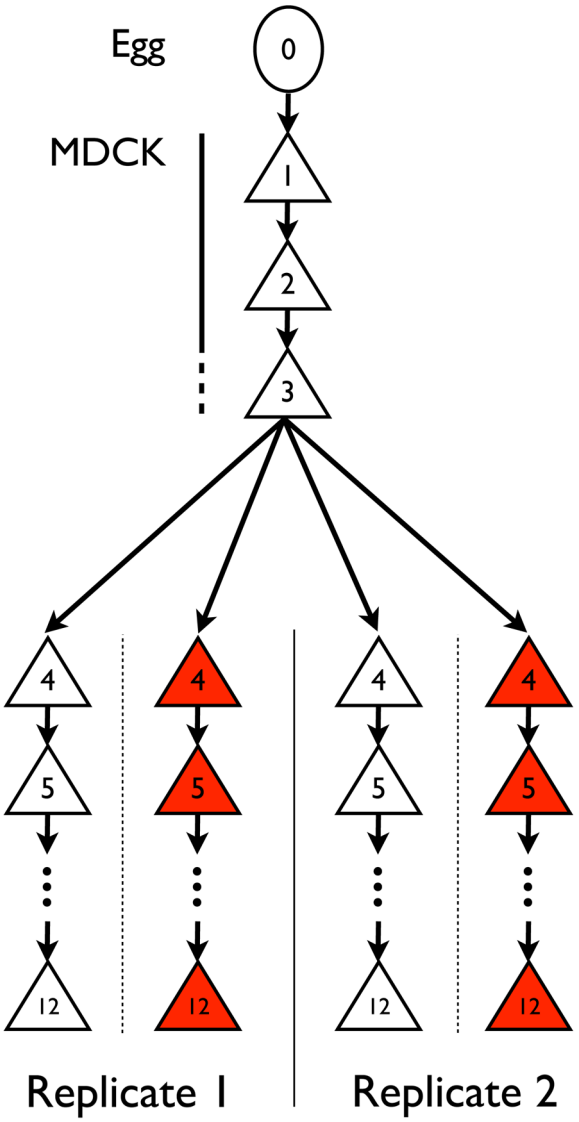

Figure 1. Experimental set-up. IAV was adapted from chicken egg to MDCK cells (passages 1–3) and then serially passaged in MDCK cells in the presence (red) or absence (white) of escalating concentrations of oseltamivir (passages 4–12) in replicate experiments.
doi:10.1371/journal.pgen.1004185.g001

Expected impact of drift and ascertainment

The effective population size N_e determines how efficiently natural selection acts on a population [18]. A beneficial mutation

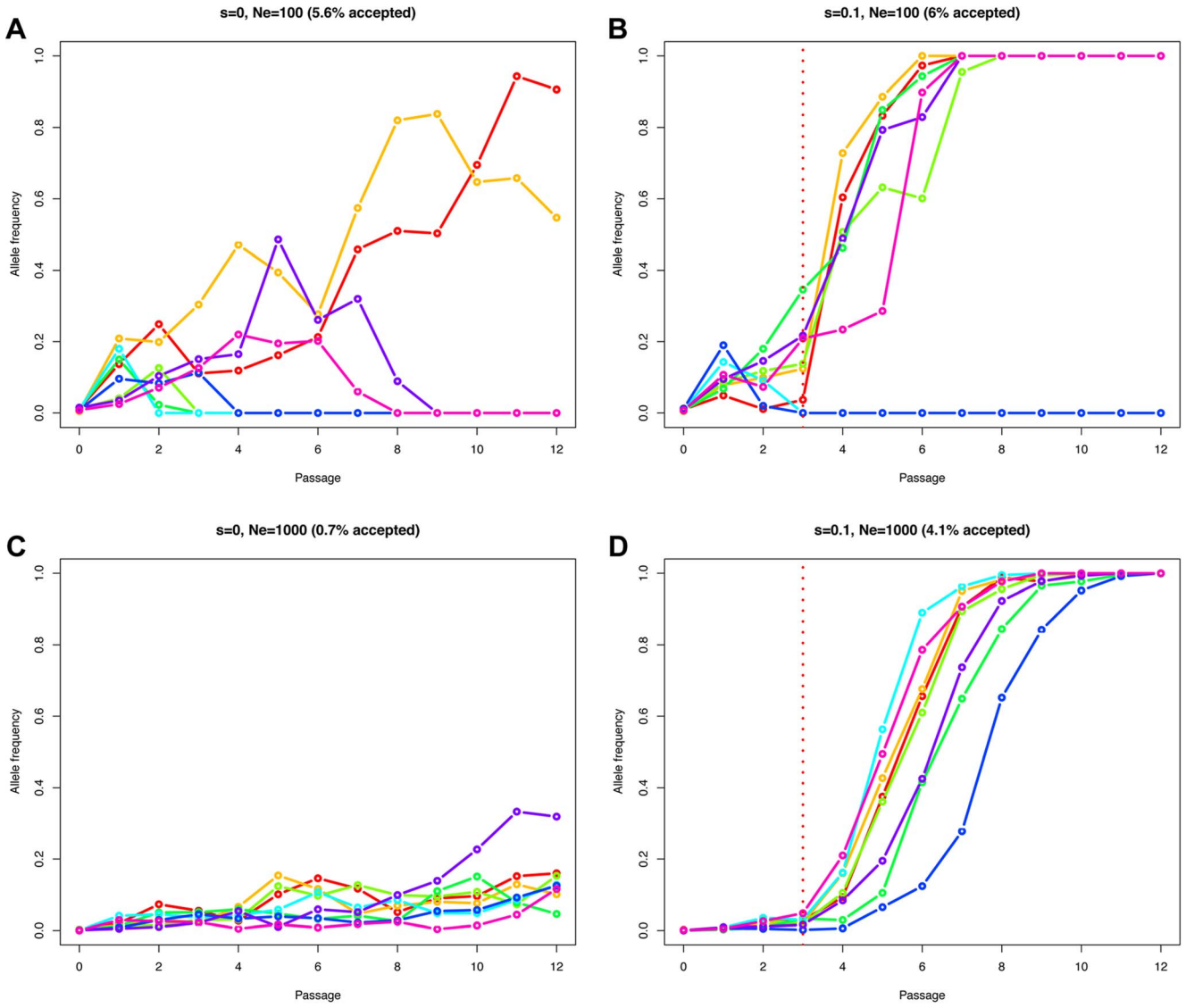

Figure 2. Simulated Wright-Fisher trajectories illustrating the impact of effective population size, selection strength, and ascertainment. We simulated a Wright-Fisher haploid model with selection matching our data set (same number of generations (i.e., 13 per passage), and selection beginning at passage 4). We plot eight randomly drawn trajectories starting from a single mutant and conditioned on reaching at least a frequency of 10% in one of the time points. We used an effective population size of $N_e = 100$ in A and B, or $N_e = 1000$ in C and D. The relative fitness of the new mutation was set to $1+s$, with $s = 0$ in A and C, or $s = 0.1$ in B and D. The fraction of simulations reaching our 10% allele frequency condition is given above each panel in parentheses.
doi:10.1371/journal.pgen.1004185.g002

has a greater chance to be successful in large populations compared to small populations, where allele frequency changes are mainly impacted by genetic drift. Therefore, the fate of a beneficial allele is determined by both the effective population size N_e and the selection coefficient s. For this reason, observing an allele increasing in frequency cannot be considered as a direct evidence of natural selection. Neutral and deleterious mutations may also increase in frequency, but simply with a lower probability [19]. The problem of distinguishing drift from selection is exacerbated in genome-wide studies, as these low probability events are more likely to occur among the large number of sites considered. Figure 2 illustrates this point, simulating a Wright-Fisher haploid model with selection intensities matching those inferred in the data set. We plot eight randomly drawn trajectories starting from a single mutant and conditioned on reaching at least a frequency of 10% in one of the passages (for $N_e = 100$ (Figures 2A and 2B) or $N_e = 1000$ (Figures 2C and 2D), and $s = 0$ (Figures 2A and 2C) or $s = 0.1$ (Figures 2B and 2D)). In the absence of selection, in populations of low effective size ($N_e = 100$), the relative frequency of mutants reaching a frequency of 10% or higher is elevated (5.6% of the simulations) compared to $N_e = 1000$ (0.7% of the simulations). With selection ($s = 0.1$), these values are nearly unchanged if the population is small (6% vs. 5.6% for $N_e = 100$), as low frequency mutants are mostly affected by drift in this scenario. However, values increase dramatically as the effective population size increases (4.1% vs. 0.7% with $N_e = 1000$). Finally, the simulated advantageous mutations follow almost deterministic trajectories in large populations (Figure 2D) while drift is still affecting them strongly when $N_e = 100$ (Figure 2B), eventually leading to the loss of the mutant.

Table 1. Bottleneck sizes at each passage.

Passage	1	2	3	4	5	6	7	8	9	10	11	12
Without oseltamivir	48	43	2575	255	85500	8600	27750	4725	37	92	2102	31
With oseltamivir	48	43	2575	255	49250	8600	7300	4800	19	75	2075	42

Population size estimated at the beginning of each passage in the absence or presence of oseltamivir (see Figure 1).
doi:10.1371/journal.pgen.1004185.t001

Estimating N_e and selection

Several methods have been proposed to estimate the effective population size (N_e) from time-sampled data assuming neutrality. Moment-based methods [22–25] have the advantage of being computationally efficient as compared to likelihood-based methods and thus can accommodate large genomic data [26–29], and can provide similar accuracy when using appropriate estimators [30]. Likelihood methods have the advantage of being able to also take into account the effects of selection, and a handful of methods have recently been proposed to estimate both N_e and selection coefficients [18] from time-sampled data [31–33]. However, being based on diffusion approximation [34], they assume large effective population sizes and low selection coefficients. Goldringer and Bataillon [35] proposed to use a moment-based estimator of N_e to reject neutrality based on Wright-Fisher simulations, but currently there is no available method able to co-estimate N_e and s in this context. In particular one would like to use the information shared by all loci to estimate N_e, and to estimate s at each locus. Here we use Fs', an unbiased estimator of N_e proposed by Jorde and Ryman [30] and extend the idea proposed by Goldringer and Bataillon [35] to also estimate s using an Approximate Bayesian Computation (ABC) approach [36] (see Materials and Methods). Our method does not rely on diffusion approximation, is appropriate for small effective population sizes and large selection coefficients, and is computationally efficient to scale with our genomic data.

For both experiments with and without oseltamivir, we respectively estimated N_e to be 226 (99% highest posterior density (HPD) interval: [210;257]) and 176 (99% HPD interval: [117;256]). These low effective population sizes are in line with the values estimated from natural populations in IAV and other viruses [37–39]. They can partially be attributed to the severe bottlenecks introduced at each passage, followed by exponential population growth. For comparison, we also calculated the expected effective population sizes as the harmonic mean of census sizes at each generation [40]. We used the estimated census population sizes measured at the beginning of each passage (Table 1) and assumed an exponential population growth to 10^6 virions during each of the 13 generations. We obtained values of 696 and 737 in experiments with and without oseltamivir, respectively. As expected, this illustrates the strong influence of the bottlenecks despite the very large population sizes assumed (10^6) at the end of each passage. However, the bottlenecks alone cannot explain the even lower effective population sizes estimated from the full genomic data, though the large variance in burst size (i.e. the number of virions produced per infected cell) [41,42] is also of relevance [40].

We then obtained posterior distributions of s for all contending mutations (i.e., mutations that were not lost by drift and are segregating in the population in at least one time point). Neutrality was rejected when the posterior density interval of s excluded zero (i.e., $P(s<0|x)<0.01$), defining Bayesian 'p-values' [43]. These p-values are plotted for all sites in the genome in Figure 3A and 3C. Note that there are fewer sites in the presence of oseltamivir (82 vs. 405, see Figure 3C and 3A) as fewer time points match the criteria defining contending mutations. We plot the trajectories corresponding to the significant sites in the absence and presence of drug respectively in Figure 3B and 3D. Despite the reduced data size, more sites are found to be under selection in the presence of oseltamivir (8, representing almost 10% of the sites considered; versus 4, representing less than 1% of the sites in the absence of drug), and having larger selection coefficients (0.15 on average vs. 0.08). In addition, an HA mutation (HA 1395 encoding a D112N mutation in HA2) was positively selected in both the absence and presence of drug, suggesting that it likely represents a tissue culture adaptation. However, the mutation was nonetheless more strongly beneficial in the presence of drug ($s=0.22$ vs. $s=0.12$, see Figure 3B and 3D).

The known H275Y resistance mutation [10] located on the NA protein at position 823 in the RNA sequence goes rapidly to fixation in the presence of oseltamivir (Figure 3D) with a point estimate of $s=0.15$. The corresponding posterior distribution is represented in Figure 4 along with the ABC correlation plot. The separate correlation between s and the two statistics Fsd'_i and Fsi'_i is also shown in Figure S3. In the presence of drug, NA and HA are the two segments containing the mutations with the highest selection coefficients (0.20 on average, compared to 0.11 for the other segments). This finding is in accordance with recent results showing that mutations in HA compensate for the deleterious effect (low growth capacity) of H275Y [17]. We extracted from the NCBI Influenza Virus Resource database [44] the recent changes in allele frequency of the 12 candidate sites under selection in natural populations of H1N1 (Figure S4). Two out of 12 candidates increased rapidly in frequency in the past five years. As previously reported [14], H275Y started to increase in frequency in the 2007/2008 influenza season and almost reached fixation in 2009/2010 (90.1%). A synonymous mutation in segment NS at position 820 (F116) also increased very rapidly in frequency in 2005/2006 (92.6%), decreasing again to less than 12% in the following years to finally reach fixation in 2009/2010.

We randomly simulated 1000 pseudo-observed data sets to cross validate our ABC approach with parameters inspired by the drug-treated experiment. We used the same number of loci with selection coefficients s taken from the obtained posterior mean, $N_e=176$, a sample size of *1000* and initial allele frequencies of $1/N_e$. For each simulated replicate, we estimated N_e and s using our proposed ABC approach (see Materials and Methods). This simulation represents a "worst-case" scenario, as the impact of genetic drift is strong and allele frequencies are skewed. Despite this, we found that the accuracy is generally very good (Figure S5A and B). We note that as the trajectories start from a single mutant, it is difficult to distinguish negative selection from neutrality, both generally leading to a rapid loss of the mutant and to a wide posterior for s with a mean around zero. The same phenomenon is also observed for beneficial mutations, but at a lower frequency. The problem vanishes when the initial allele frequency is increased, or in larger populations (data not shown).

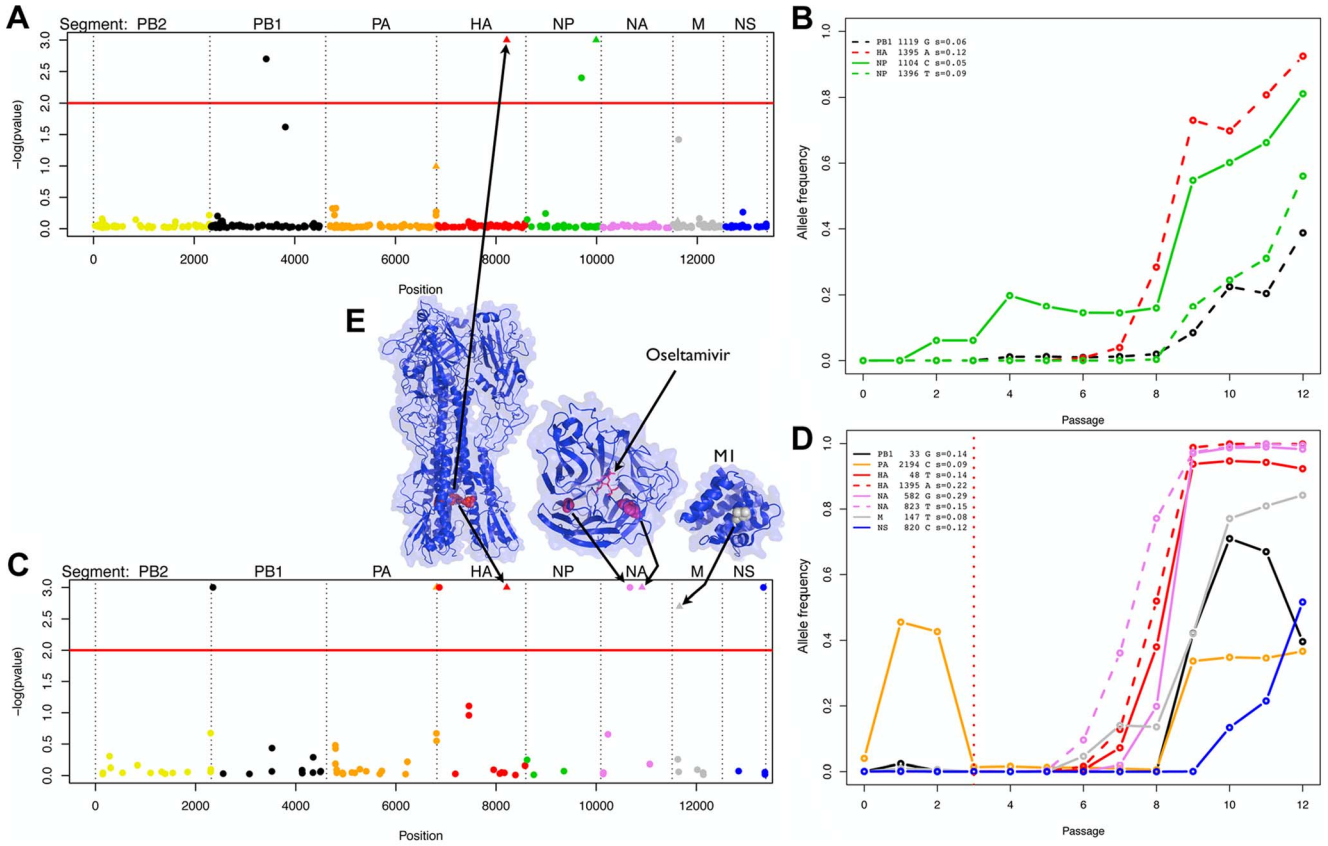

Figure 3. Evidence of positive selection in the H1N1 genome in the absence and presence of oseltamivir. We plot the Bayesian P-values for each SNP in log scale in the absence and presence of oseltamivir in A and C, respectively. The horizontal red lines are genome-wide significance thresholds of $P = 0.01$. The eight segments are separately color-coded, a scheme which is maintained in all panels. Significant nonsynonymous mutations are represented with triangles. We plot the minor allele frequency trajectories of all significant SNPs over the experiment in the absence and presence of oseltamivir in B and D, respectively. The vertical dotted red line indicates the time of the oseltamivir addition (see Figure 1). Trajectories are represented in dashed lines if a second SNP was significant within a given segment. For each significant SNP, the name of the segment, the position of the SNP, the nucleotide increasing in frequency, and the estimated selection coefficients with our *Ne*-based ABC method are indicated in the top left corner of B and D. In E, we represent the 3D structure of the proteins corresponding to the segments coding for membrane proteins (for those with a resolved structure). We indicate the amino acid residues corresponding to the significant mutations with arrows. The SNP locations highlighted on the structures are as follows: HA2 D112N (nonsynonymous), NA G193G (synonymous), H275Y (nonsynonymous), and M1 A41V (nonsynonymous). Although the M segment encodes both the M1 and M2 protein, the significant SNP is only positioned in the coding region of M1. The significant SNP in the NS segment (F116) is synonymous and only positioned in the region coding for the NS2 protein. HA is represented as a trimer, with the significant residue being highlighted (red) in each monomer, though one residue is slightly obscured due to being buried in the protein complex. NA and M1 are represented as monomers, and NA is shown with a bound molecule of oseltamivir.
doi:10.1371/journal.pgen.1004185.g003

Finally, we evaluated whether our model based on constant effective population size is robust to the series of bottlenecks and population expansions induced at each passage in the experiment (see Table 1). Ewens [45] showed that in a population with a size changing cyclically over time, the probability of fixation of an allele is approximately the same as in a population of constant size with $N_e = N^*$, where N^* is the harmonic mean of the population sizes at each generation, which is also the effective population size of the fluctuating population. To evaluate this finding in the context of our estimation procedure, we simulated 1000 additional data sets with varying population sizes. In order to match the drug-treated experiment, 9 passages of 13 generations each were simulated, with exponential population growth from $N = 23$ to $N = 10^6$ (see Figure 1). The founding population size ($N = 23$) was chosen as it results in a harmonic mean of $N_e = 176$ for the 13 generations (i.e., the empirically estimated value from the experiment). Here again, this corresponds to a "worst-case" scenario where the bottleneck at each passage is extremely strong (see Table 1 for the true data) and population expansion very rapid. We found that our ABC procedure based on a constant effective population size indeed accurately estimates N_e and s (Figure S5C and D) - with N_e being slightly downwardly biased (estimated to 167 on average), and large selection coefficients are very slightly upwardly biased.

Effect of genetic linkage

Linkage between selected and neutral sites can confound inference when estimating genome wide selection coefficients. Further, the frequency of homologous recombination in the IAV genome is still debated [46]. If absent, we expect to observe strong effects of genetic hitchhiking [47], where linked sites should increase in frequency together with selected sites due to physical linkage. However, the very low genetic diversity in our populations limits this phenomenon, and we identified at most two selected sites within the same segment. We also note that the initial allele frequencies of the sites under selection are very low in all cases

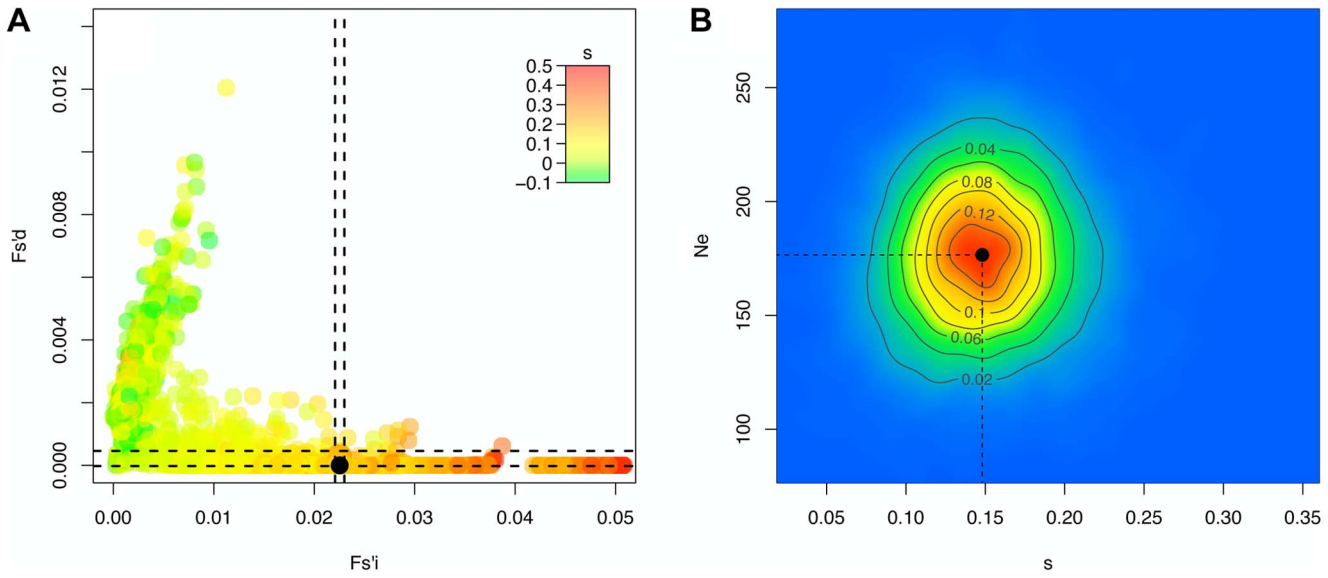

Figure 4. *Ne*-based Approximate Bayesian Computation for the H275Y resistance mutation in the presence of oseltamivir. For 10'000 simulated trajectories (out of the 100'000 simulations performed), we plot in A the values of the statistics *Fsi'* and *Fsd'* with colors corresponding to the selection coefficients *s*, as well as the values calculated for the real trajectory of the H275Y oseltamivir resistance mutation in black. We indicate the region corresponding to the best 1% retained simulations with a dashed line, and plot the corresponding two-dimensional posterior distribution for *s* and N_e in B.
doi:10.1371/journal.pgen.1004185.g004

(Table 2) and below our detection threshold (0.17%) in 11 out of the 12 cases. This suggests that selection is primarily acting on de novo mutations rather than on standing variation in our experiment.

Only two trajectories in each experiment were identified as having a poor fit to the assumed Wright-Fisher model, and are shown in Figure S6. Figure S7 also shows the ABC correlation and posterior plots for one of these cases, where one can see the inability of the model to generate simulations similar to the observed data. Clonal interference (i.e., the competition of simultaneously segregating beneficial mutations) has recently been proposed to play an important role in influenza evolution [48] and could explain such patterns, in which, in the absence of recombination, an initially steep trajectory becomes halted or even reversed by the appearance of another more beneficial mutation (as in Figure S6A). However, given the small effective population sizes in this experiment, it is not surprising that we do not observe more such trajectories, as the probability that multiple contending beneficial mutations are present at a given point is small [49]. Integrating the potential of clonal interference into the methodology developed here will be the subject of future study.

Replicated experiment

The experiment was replicated starting at passage 4 (Figure 1), and analyzed with our N_e-based ABC method. We plot the Bayesian p-values for all sites in the genome in Figure S8A and S8C and the trajectories corresponding to the significant selected sites, in the absence and presence of drug, in Figure S8B and S8D, respectively. More details are given on selected sites in Table S1. Consistent with the first experiment, more sites are identified as being under selection in the presence of oseltamivir (6 vs. 2). The H275Y resistance mutation appears only at passage 9 and also increases very rapidly in frequency in the presence of oseltamivir (Figure S8D) with an even higher selection coefficient $s = 0.27$. Interestingly, like in the first experiment, one nonsynonymous HA mutation (position 1211, encoding a N50K mutation in HA2) is under strong selection in both the absence and presence of oseltamivir, and is located only 184 base pairs from the one identified in the first experiment (position 1395, encoding a D112N mutation in HA2). Similarly, a nonsynonymous mutation in segment M at position 92 (encoding an E23Q mutation of the M1 protein) is also under selection ($s = 0.06$), where in the first experiment, one was identified at position 147 (encoding an A41V mutation of the M1 protein, $s = 0.08$).

Likelihood and coalescent based estimation

As a matter of comparison, for each significant trajectory identified using our N_e-based ABC method, we applied a diffusion approximation likelihood-based method [32] which we here extend to a haploid model. For the H275Y mutation, the two-dimensional likelihood surface for $\gamma = N_e \cdot s$ and the age of the mutation (t_0) is shown in Figure 5. The selection coefficients obtained are mostly consistent between the two methods and are given in Table 2. They tend to be different when the trajectories have an unexpected behavior under the Wright-Fisher model, like the synonymous PB1 mutation at position 33 (K11), also identified as having a poor fit to our model. In this case the behavior of the two methods is hard to interpret. Likelihood-based methods in general should be more accurate than ABC methods, which reduce the whole data in to a few summary statistics. However, as stated previously, the diffusion approximation made to calculate the likelihood in this method is not appropriate for high selection coefficients and low effective population sizes (i.e., it is appropriate for $N_e s \sim 1$). As expected, this method tends to over-estimate *s* compared to our ABC method when *s* is large. Indeed, in this approximation, drift is a slow process and very large allele frequencies changes as we observe can only be explained by excessively large values of *s*. Additionally, there is currently no available likelihood method that can integrate the information shared by all sites to infer N_e, and estimating N_e from a single site is particularly inefficient [32]. For this reason, and to have a fair comparison, we did not attempt to estimate N_e here, but fixed it to

Table 2. Estimated selection coefficients.

	Segment	Position	Protein	Sequential numbering	Other numbering	Allele	Initial frequency	Final frequency	*Ne*-ABC *s* estimates (99% HPDIs)	Malaspinas *et al.* [32] *s* estimates
Without	PB1	1119	PB1	L373		G	0.01%	38.8%	0.06 (0.01;0.11)	0.04
oseltamivir	**HA**	**1395**	HA	D455N	D112N (HA2)	**A**	**0.03%**	**92.5%**	**0.12 (0.05;0.19)**	**0.08**
	NP	1104	NP	N368		C	0.02%	81,1%	0.05 (0.00;0.11)	0.04
	NP	**1396**	NP	L466F		**T**	**0.03%**	**56.1%**	**0.09 (0.03;0.15)**	**0.06**
With	*PB1*	*33*	PB1	K11		*G*	*0.03%*	*39.6%*	*0.14 (0.06;0.25)*	*0.03*
oseltamivir	PA	2194	PA	Noncoding		C	1.4%	36.7%	0.09 (0.02;0.17)	0.01
	HA	48	HA	L6		T	0.1%	92.3%	0.14 (0.06;0.27)	0.18
	HA	**1395**	HA	D455N	D112N (HA2)	**A**	**0.06%**	**99.9%**	**0.22 (0.08;0.34)**	**0.27**
	NA	582	NA	T194		G	0.02%	98.3%	0.29 (0.15;0.45)	0.41
	NA	**823**	NA	H275Y	H274Y (N2)	**T**	**0.04%**	**99.5%**	**0.15 (0.06;0.24)**	**0.20**
	M	**147**	M1	A41V		**T**	**0.04%**	**84.2%**	**0.08 (0.01;0.15)**	**0.07**
	NS	820	NS2	F116		C	0.03%	51.7%	0.12 (0.04;0.20)	0.09

Comparison of N_e-ABC and Malaspinas *et al.* [32] estimates of *s* for the significant trajectories under selection. Bold indicates nonsynonymous mutations and italic indicates a poor fit to our Wright-Fisher model. We indicate the nucleotide corresponding to the minor allele, with its initial frequency at the beginning of the experiment in the absence of oseltamivir, or at passage 4 when drug treatment began (see Figure 1). For the *Ne*-ABC method, we give the 99% highest posterior density intervals (HPDIs) in brackets.

doi:10.1371/journal.pgen.1004185.t002

the value estimated above via the ABC approach. Finally, we note that this class of method is computationally intensive and cannot practically be applied to whole-genome datasets.

In addition, we have developed a new coalescent-based method that explicitly models the known demography of our experimental populations for comparison (Table 1). This approach has the additional advantage of incorporating the genetic diversity linked to the beneficial mutation into the estimation procedure, allowing us to estimate the mutation rate and refine estimates of *s*. Figure 6 shows the combination of selection coefficient ($s = 0.12$) and mutation rate ($\mu \approx 10^{-7}$) with the highest likelihood for the H275Y mutation. Interestingly, the estimation obtained with these simulations accounting for the true demography is consistent with our N_e-based ABC method, which gives a posterior mean for the selection coefficient of $s = 0.15$ (Table 2 and Figure 4).

It is noteworthy that the three methods here developed and applied to this data are complementary, and thus have been used jointly. Our N_e-based ABC approach is computationally efficient, can be applied to large genomic datasets, and does not rely on diffusion approximation. It provides both estimates of N_e using information from the whole genome, and posterior distributions for *s* at each individual site. Using these results, it is next interesting to utilize the likelihood-based method, as it can be more accurate for small *s* (or in cases where N_e is large), as it uses the full data rather than summary statistics. Additionally, it can estimate the age of the beneficial allele, which can be of interest in some cases. Being more computationally intensive, one can apply it on the candidate sites identified by the N_e-ABC method, and even take advantage of the estimated N_e that it provides. Finally, our new coalescent method is a promising first attempt to estimate *s* not only using a single allele frequency trajectory, but the whole sequence linked to it. It is also computationally very intensive and can be used on top candidate sites to refine the posterior distributions obtained from the N_e-based ABC method.

Fisher's Geometric Model and distance from optimum

Further utilizing these estimated per-site N_e-based selection coefficients, and in order to contextualize these results, we utilize the framework of Fisher's Geometric Model (FGM). The FGM [50] predicts that environmental challenges increase the distance between the current phenotype and the phenotypic optimum, thereby allowing for more and stronger beneficial mutations.

Here, the oseltamivir environment represents a novel (and challenging) environment, which is expected to result in a shift of the optimum away from the location of the current population. This is reflected both in a higher maximum beneficial selection coefficient (0.288 vs. 0.117) and in a higher mean beneficial selection coefficient (0.026, bootstrap bias-corrected and accelerated [51] 95% confidence interval (CI): [0.017; 0.039] vs. 0.016, 95% CI [0.015; 0.017]), obtained from the point estimates obtained using the N_e-based ABC approach - indicating that the optimum may be indeed further away from the current phenotype in the drug as compared with the no-drug environment.

In order to study the distribution of fitness effects (DFE) observed, and to quantify the distance to the phenotypic optimum in each environment, we chose three different biologically relevant distributions to be fitted to the data. Since we were particularly interested in the distance to the phenotypic optimum, we used the displaced-gamma distribution proposed by Martin and Lenormand [52], which results in an approximately beta-shaped DFE of beneficial mutations [53], and allows for a direct estimate of the distance to the optimum under the FGM. Secondly, we used a half-normal distribution, which is predicted from the FGM, when the optimum is (infinitely) far away. Third, we used the generalized

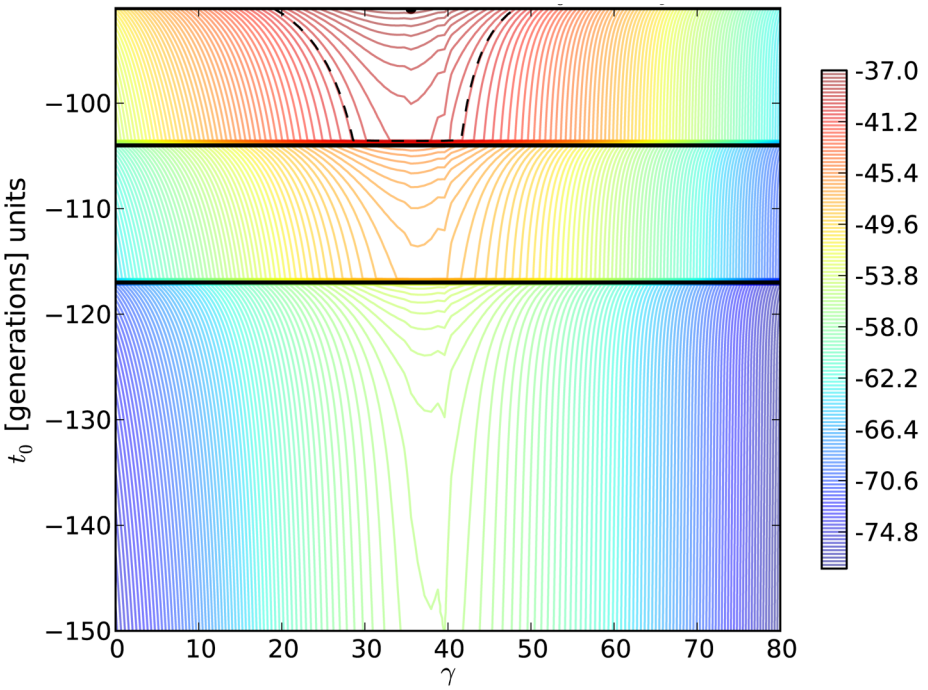

Figure 5. Log-likelihood contours for the H275Y resistance mutation in the presence of oseltamivir. We plot the two-dimensional likelihood surface for the selection parameter $\gamma = N_e s$ (x-axis) and the allele age t_0 (y-axis) in generations, with generation zero representing the end of the experiment. Horizontal black lines represent sampling times. The colors indicate the value of the log-likelihood, with red being the highest and blue being the lowest. The maximum likelihood is shown as a black dot, and the log likelihood was lowered from its peak by $X^2_{2.5\%}/2 = 2.996$, to construct a 95% confidence (likelihood) region for the parameters (dashed black line).
doi:10.1371/journal.pgen.1004185.g005

Pareto distribution (GPD). Given that beneficial mutations are so rare that they represent the tail of the full DFE, the GPD allows for the estimation of the most likely extreme value domain [54]. In other words, the resulting estimate of the shape parameter κ yields information on whether the DFE is bounded (i.e., the full DFE belongs to the Weibull domain), whether its tail is exponential-like (Gumbel domain), or whether the tail is heavier than exponential (Frechet domain). In terms of the underlying biology, this is a very important question: for example, a bounded DFE would indicate that mutations cannot exceed a certain effect size, whereas, on the contrary, a heavy-tailed DFE would suggest that mutational effect sizes are highly unpredictable. Supporting arguments have been made for an exponential tail being the most biologically reasonable [55], and related studies have suggested an exponential distributions of fitness effects (e.g., [56,57]). However, there also exists empirical evidence for bounded tails of the beneficial DFE (e.g., [58,59]).

Of note, our data consist not of the full distribution of newly arising beneficial mutations, but of the fraction of those beneficials that survive drift and are segregating in the population for at least a short time (so-called "contending beneficial mutations" [60,61]). Barrett *et al.* [60] pointed out that this distribution arises as the full distribution of new mutations weighted by the (approximate) probability $1\text{-}exp(-2s)$ that a mutation with selection coefficient s survives drift. The resulting probability densities for the three tested distributions are noted in the Materials and Methods.

We fitted all three resulting contending distributions to the data by numerically maximizing the log likelihood using a weighted bootstrap approach [62]. All results are reported in Table 3 and Figures 7 and 8.

In the absence of oseltamivir, the half-normal distribution yields the highest median log likelihood (cf. Table 3, and Figure S9A), but the bootstrap estimates show a large variance (also owing to the lower flexibility afforded by the single parameter). Hence, the outcome is sensitive to potential measurement error, or sampling bias. In fact, the 95% confidence intervals of the bootstrap estimates of both the GPD and the beta distribution lie within that of the half-normal, indicating that they both may represent equally good summaries of the true distribution. In the presence of oseltamivir, the GPD clearly provides the best fit, with both the half-normal and the scaled beta distribution reaching generally lower log likelihoods (cf. Figure S9B).

The generally good fit of the GPD provides support for the assumption that the beneficial portion of the DFE represents a tail distribution. If this condition is met, the GPD is expected to be the most flexible among the three tested distributions, because it can account for all possible tail shapes. Because we observe only the contending distributions in the present study, the tail of the distribution becomes even more important, and it must be noted that the tails of both the beta scaled and the half-normal distributions studied here are both contained within the GPD.

After establishing that the GPD yielded a good fit to the observed data, we interpreted the estimated shape parameter κ that determines the extreme value domain of the underlying DFE (cf. Figure 8). In the absence of oseltamivir, we observe $\kappa = -0.0532$ (95% CI: −0.0786, −0.0075), indicating that the full DFE belongs to the Weibull domain of attraction (Figure 8A). Hence, it has a right-truncated tail, and we estimated the maximum possible mutational effect as $d = 0.17$ (95% CI: 0.12, 0.76). We can compare this result with the estimate of the distance to the optimum obtained from the scaled beta distribution. With $s_0 = 0.18$ (95% CI: 0.12, 1), this estimate agrees nicely with that from the GPD. In terms of the underlying biology, this indicates that there is remaining potential for adaptation also in the absence of oseltamivir, but that the maximum possible effect size does not

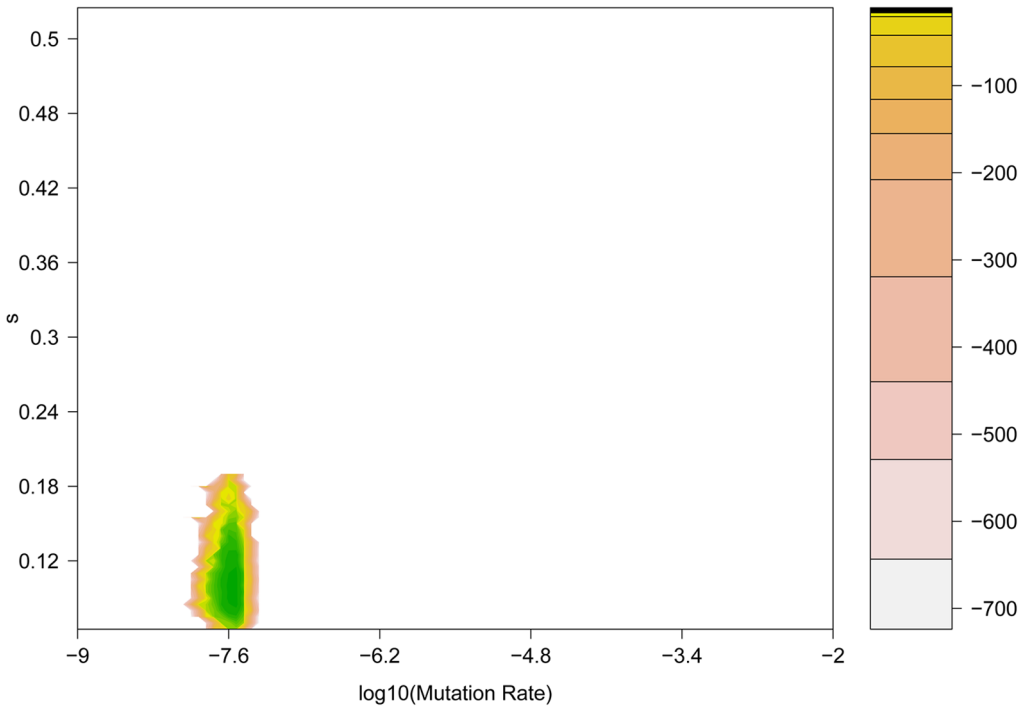

Figure 6. Coalescent-based maximum likelihood surface for the H275Y resistance mutation in the presence of oseltamivir. Two-dimensional likelihood surface for the selection coefficient (*s*) (y-axis) and the mutation rate μ (x-axis). The graph is colored according to the value of the log-likelihood (displayed in the embedded legend), where green indicates the highest probability and yellow the lowest.
doi:10.1371/journal.pgen.1004185.g006

differ greatly from the observed maximum selection coefficient ($s = 0.117$).

The pattern looks very different in the presence of oseltamivir. Here, the estimated shape parameter is $\kappa = 0.24$ (95% CI: 0.14, 0.38), clearly indicating a heavy-tailed DFE that belongs to the Frechet domain of attraction (Figure 8B). Supporting this finding, it was not possible to obtain any reasonable estimates of the distance to the optimum under the scaled beta distribution. This support for a heavy-tailed distribution is also consistent with recent results examining the DFE in populations of yeast that have been subjected to extreme environmental conditions [63].

This finding suggests that the potential for adaptation in the drug environment is indeed much higher than the highest observed selection coefficient, and that mutational effect sizes will be difficult to predict under strong adaptive challenges. In particular, upon a longer run of the experiment (e.g., over the course of time in natural populations) even stronger beneficial mutations than those identified in the present experiment could be expected (however, it is noteworthy that the H275Y mutation appears to be re-identified across multiple different experiments, as discussed below). In comparison, there is only little potential for adaptation in the no-drug environment. However, the optimum is still far as compared with other examples from the literature [64], which could indicate ongoing adaptation to the MDCK cells. We note that we do not explicitly model experimental errors in our analyses, as this would require several replicated experiments [54,56]. However, the heavy tail of the DFE in the presence of oseltamivir (Figure 7B) may indeed be influenced by such factors, and this result should thus be interpreted with caution.

Biological implications

Lastly, we attempted to interpret these results in light of the known biology of influenza (Figure 3E). The NA mutation H275Y is a well-characterized oseltamivir-resistance mutation and has been shown to alter the hydrophobic pocket of the NA active site, thereby reducing affinity for drug [10,65–69]. Thus, re-identification of this substitution aids to validate the results. Further, A41V (encoded by a C147T SNP) and E23Q (encoded by a G92C SNP) substitutions of the M1 protein were identified in the first and second experiments, respectively. The location of these mutations (helices 3 and 2) are not overlapping with regions important for RNA or membrane binding, which facilitate virion assembly and maintenance of virion integrity. In addition, the location of either residue does not appear to be important for forming extended sheets of M1 protein, as proposed previously [66]. Therefore, the role of these mutations in viral fitness in the presence of oseltamivir may be related to additional roles of the M1 protein. An intriguing possibility is that interactions between M1 and the NA cytoplasmic tail important during virion budding [70–72] are altered by the H275Y mutation and are compensated for by additional mutations in the N-terminal region of M1 [73]. Two adaptive substitutions were also observed in the HA2 peptide during growth in the presence and absence of oseltamivir, with the D112N (encoded by a G1395A SNP) and N50K (encoded by a C1211A SNP) substitutions observed in the first and second replicates, respectively. Interestingly, the trajectories of the substitutions in the presence of drug appear to be strongly correlated with the rise of the H275Y mutation (see Figures 3 and S7). The combined results from the two replicates in the presence and absence of oseltamivir show that these loci are positively selected during tissue culture adaptation, and may epistatically interact with the H275Y allele. The D112N mutation has been well characterized in other influenza strains and is conserved across all HA serotypes. This substitution has been shown to cause a rise in the pH of the HA conformational change and HA-induced endosome and viral membrane fusion [68,74], a process critical for IAV infectivity

Table 3. Maximum likelihood estimates of the distribution of fitness effects.

		With oseltamivir		Without oseltamivir	
Distribution	**k[1]**	**Log(L) (95% CI)[2]**	**s_0 (95% CI)[3]**	**Log(L) (95% CI)**	**s_0 (95% CI)**
GPD	2	180.0 [154.7,199.1]	∞	1343.3 [1301.5,1374.0]	0.17 [0.12,0.76]
Beta	2	158.8 [122.1,190.0]	∞	1343.4 [1301.4,1374.1]	0.18 [0.12,1]
Half-normal	1	110.5 [69.3,154.7]	∞	1375.6 [1274.4,1429.7]	∞

Theoretical distributions of fitness effects fitted to the data.
[1]Number of parameters.
[2]Maximized value of the log-likelihood function with 95% confidence intervals.
[3]Maximum possible effect-size.
doi:10.1371/journal.pgen.1004185.t003

[75]. Further, the D50K mutation is located in a region known as the HR1 heptad repeat, interrupting the repeat pattern with a polar-to-charged residue substitution. The HR1 repeat regions form coiled coils that undergo conformational changes in low pH conditions and likewise promote endosomal membrane fusion during IAV infection [76]. Alterations of endosomal membrane fusion mediated by mutations in HA2 are a known mechanism for tissue culture adaptation [77,78], but interactions with a drug resistance allele has not been described previously. In total, while the results are not sufficient to confirm this hypothesis, the combined results from replicate 1 and replicate 2 suggest that epistatic interactions between M1 and NA, and possibly HA2, may be important during the selection of drug resistance in IAV populations.

Conclusions

As a major annual cause of morbidity and mortality, influenza virus infections remain one of the most important global health concerns. Foremost amongst the challenges in treating this virus has been its ability to adaptively respond to drug treatment, with oseltamivir resistance spreading globally during the 2007–2008 and 2008–2009 influenza seasons. In order to evaluate the viral adaptive response to oseltamivir, we have here developed a multi-faceted population genetics approach based upon an unparalleled dataset consisting of whole-genome multi-time point experimental data both in the presence and absence of treatment. Utilizing novel approximate Bayesian, likelihood-based, and analytical results, we identify a handful of known and unknown positively selected variants, and quantify the distance from phenotypic optimum imposed by oseltamivir. These results not only confirm a number of theoretical expectations arising from Fisher's Geometric Model and its extensions, but also clearly illustrate the ease by which resistance may be evolved against neuraminidase inhibitors. We finally note that the robust methodologies developed here can be widely applied to time-sampled data from not only experimental but also natural populations (Figure S4), allowing for the utilization of a temporal dimension that is highly informative for identifying the recent action of positive selection.

Materials and Methods

Data generation and bioinformatics

Influenza A virus A/Brisbane/59/2007 (H1N1) from chicken egg allantoic fluid (NIH Biodefense and Emerging Infectious Research Resources Repository NIAID, NIH; NR-12282; lot 58550257) was serially passaged in Madin-Darby canine kidney (MDCK) cells (Figure 1). This strain has the following genome size: segment 1 (PB2) 2314 nucleotides (nts), segment 2 (PB1) 2302 nts, segment 3 (PA) 2203 nts, segment 4 (HA) 1776 nts, segment 5 (NP) 1497 nts, segment 6 (NA) 1427 nts, segment 7 (M1/2) 1006 nts, segment 8 (NS1/2) 870 nts. MDCK cells were maintained in Eagle's minimal essential medium (MEM) with 10% fetal bovine serum (Hyclone) and 2 mM penicillin/streptomycin. Viral infections were performed in influenza virus growth medium as described [79] and lasted for 72 hours. Virus was continually passaged on cells to prevent any freeze-thaw cycles and the amount of virus to initiate a passage and the virus at the end of each passage were subsequently empirically determined via plaque assays using standard techniques. These values were used to determine the bottleneck size (Table 1), MOIs (Table S2), magnitude of population expansion, and number of doublings associated with each passage. The number of doublings was used to determine the number of generations per passage – averaging to 13 generations per passage throughout the experiment, in both the no drug and drug-treated populations.

For passages indicated in Figure 1, oseltamivir was added at increasing concentrations and two independent experimental trajectories were performed. The initial concentration of oseltamivir was equal to the ED_{50}, the concentration of drug that reduced viral plaque numbers to 50% of a no drug control. The initial ED_{50} was 0.1 uM (Table S2), indicating that the starting virus was very sensitive to oseltamivir. The next passage was performed in the presence of $4\times$ ED_{50}. Subsequent passages were performed by doubling the concentration of oseltamivir if 50% cytopathic effect (CPE) was observed (i.e., the cytopathic effect in the cells from the previous passage). If 50% CPE was not observed, the dose of oseltamivir was reduced to a concentration that lead to the observation of 50% CPE. Oseltamivir carboxylate (RO0640802-002; lot 91ST1126/1) was obtained from Roche (F. Hoffmann-La Roche Ltd, Basel, Switzerland). Concentration of oseltamivir at each passage can be found in Table S2.

Cell-free virus was obtained at each passage by spinning down supernatant 72 hour post-infection, and subjected to whole genome pooled population sequencing. Viral RNA was purified using the RNeasy 96 Kit (Qiagen, Gaithersburg, MD). SuperScript III First-Strand Synthesis Supermix (Life Technologies, Grand Island, NY) and primers that bind the 3′ end of all IAV segments were used for reverse transcription. The cDNA was then amplified in a single multiplex reaction to amplify all segments of the genome with near equal efficiency, using primers that have been described previously. The amplified product was sheared to a size range of 300–600 base pairs with Fragmentase from New England Biolabs (Ipswitch, MA) using the procedure recommended by the supplier. DNA was then end repaired, A-tailed, and ligated to Illumina-compatible adapters containing 6-mer barcode sequences. The products were size selected by using $0.8\times$ AMPure XP beads (Agencourt, Beverly, MA), collecting

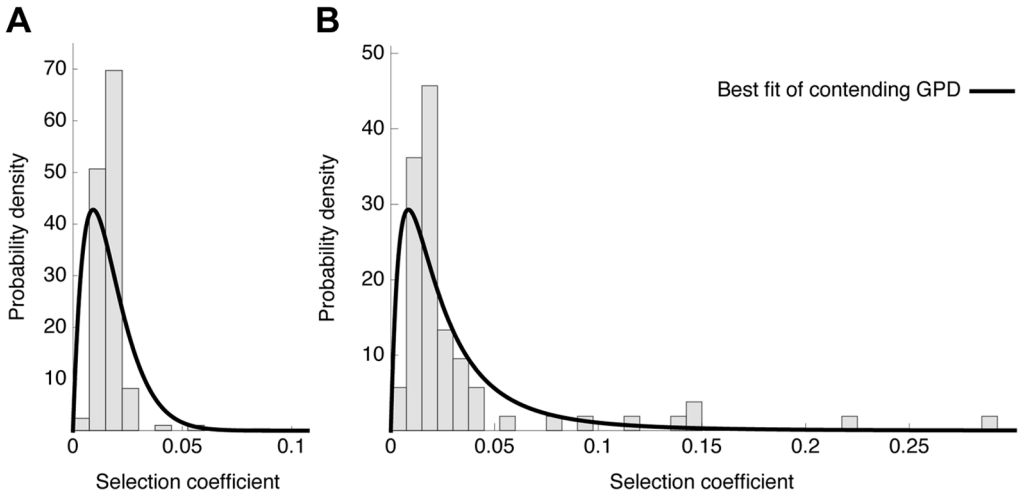

Figure 7. Distribution of fitness effects. Observed histograms of fitness effects of contending beneficial mutations and the best fit of a contending generalized Pareto distribution (solid lines) in the absence (A) or presence (B) of oseltamivir. The heavy tail of the DFE in the presence of the drug is clearly visible.
doi:10.1371/journal.pgen.1004185.g007

supernatant and treating with 1.6× beads, and eluting DNA with ddH_20. After size selection, DNA was amplified with Illumina PE PCR primers, quantified and combined into libraries for sequencing on the HiSeq2000 platform. All sequences used in this study were generated from 100 base pair reads. All sequence data is publicly available for download at http://bib.umassmed.edu/influenza/.

In addition to viral samples, an RNA error control was generated from a cloned influenza A/Brisbane/59/2007 (H1N1) NA gene segment. The cloned segment was used as a template in a T7 transcription reaction to make a pool of control RNA, which was processed and sequenced in parallel with the viral samples. Sequence data from the RNA control showed that 95% of erroneous SNP calls could be eliminated by excluding low frequency (<0.17%) SNPs. Sequence reads were aligned to Influenza A/Brisbane/59/2007 reference genome (Genbank accessions CY030232, CY031391, CY058484–CY058486, CY058488–CY058489, CY058491) using the BLAST alignment algorithm.

Reads were filtered to eliminate those with Phred quality score <20 across the read, and the minimum length of the mappable read >20 nucleotides. The coverage was high with a median over all passages of 56667 (Table S3 and Figure S10), with 90% of

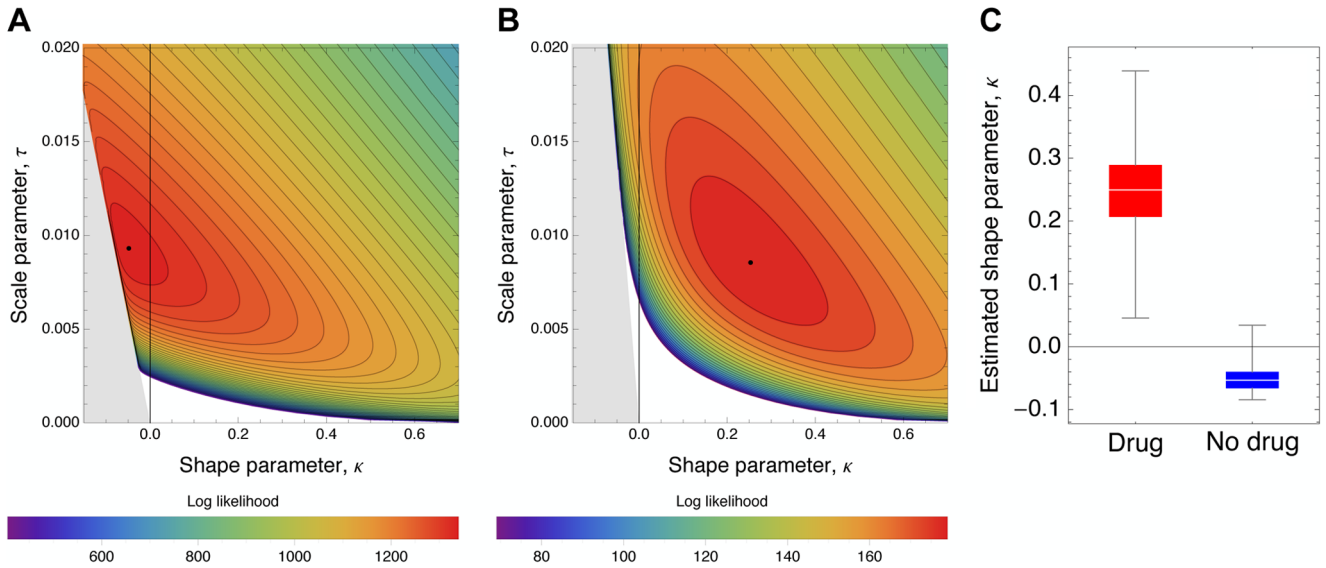

Figure 8. Likelihood surfaces for the contending generalized Pareto distribution. Likelihood surfaces of the fit of a contending generalized Pareto distribution in the absence (A) or presence (B) of oseltamivir. The shape parameter κ (on the x-axis) determines the domain of attraction. The maximum likelihood estimate (indicated by a black dot) lies in the Weibull domain ($\kappa<0$) in the absence of the drug, indicating a bounded distribution of fitness effects, whereas it lies in the Frechet domain ($\kappa>0$) in the presence of the drug, indicating an underlying heavy-tailed distribution of fitness effects. Black lines represent 25 (A) and 5 (B) orders of log likelihood, respectively. The gray area represents inaccessible parameter space, because the lower bound of the estimated DFE cannot be lower than the maximum observed selection coefficient. We also show the distribution of the shape parameter κ for the GPD distribution (C).
doi:10.1371/journal.pgen.1004185.g008

reads mapping. We excluded all sites having coverage lower than 100 and we randomly down-sampled all sites having a coverage higher than 1'000, to 1'000 in order to estimate allele frequencies from allele counts. As we almost only observed biallelic SNPs (see results), for each site we kept only the two alleles having the highest frequencies over all passages, and called the minor allele as the one having the lowest frequency at passage 0. In all subsequent analyses concerning the effect of oseltamivir, we omitted the first three passages of pre-drug treatment.

N_e-based ABC estimation

The observed data x consists of allele frequency trajectories measured at L loci: $x_i(i=1,\ldots,L)$. We have one parameter N_e shared by all loci in the genome, and L locus-specific selection coefficients $s_i(i=1,\ldots,L)$ that we would like to infer. In a Bayesian setting, we want to estimate the joint posterior distribution

$$P(N_e,s|X) \propto P(X|s,N_e)P(N_e)P(s)$$

The likelihood $P(X|s,N_e)$ can be calculated numerically in some cases but it relies on approximation and is computationally very intensive (see below). For this reason, we propose a likelihood-free approach based on Approximate Bayesian Computation (ABC) [36,80]. This class of methods is based on Monte Carlo simulations, which are compared to observed data using summary statistics. In our model, locus-specific summary statistics capable of estimating s per locus are needed, as are statistics utilizing information from all loci jointly in order to infer N_e. However, the standard ABC algorithm is not usable in such cases, as the probability of obtaining one simulation with a good match to the observed data for all L loci simultaneously rapidly tends to 0 as L increases. Recently, Bazin *et al.* [80] proposed a new algorithm to overcome this difficulty where the problem is split in to two steps (Algorithm 2), and we adapt their general solution to our problem here.

First we note that we can decompose the posterior as

$$P(N_e,s|X) = P(N_e|X)P(s|N_e,X)$$

Using conditional independence (see Appendix in [80]), the joint density has the factorization

$$P(N_e,s,X) = \left[\prod_{i=1}^{L} P(X_i|s_i,N_e)P(s_i)\right] P(N_e)$$

and the marginal density is

$$P(N_e,X) = \left[\prod_{i=1}^{L} P(X_i|N_e)\right] P(N_e)$$

where

$$P(X_i|N_e) = \int_s P(X_i|s_i,N_e)P(s_i)ds$$

Dividing these two densities we have

$$\begin{aligned} P(s|N_e,X) &= \frac{P(Ne,s,X)}{P(Ne,X)} \\ &= \prod_{i=1}^{L} \frac{P(X_i|s_i,N_e)P(s_i)}{P(X_i|N_e)} \\ &= \prod_{i=1}^{L} P(s_i|N_e,X_i) \end{aligned}$$

Finally the joint posterior can be factorized as

$$P(N_e,s|X) = P(N_e|X) \prod_{i=1}^{L} P(s_i|N_e,X_i)$$

and if focused on a particular locus i, we find

$$P(N_e,s_i|X) = P(N_e|X)P(s_i|N_e,X_i)$$

This justifies the need for both locus-specific summary statistics $U(X_i)$ and summary statistics that are a function of all loci together $T(X) = T(X_1,\ldots,X_L)$, and we approximate the posterior as

$$P(N_e,s_i|X) \approx P(N_e|T(X))P(s_i|N_e,U(X_i))$$

The general algorithm to sample from this posterior adapted from Algorithm 2 in Bazin *et al.* [80] can be written as:

Step 1. Obtain an approximation of the density

$$P(N_e|x) \approx P(N_e|T(x))$$

Step 2. For locus $i=1$ to $i=L$

For $k=1$ to $k=N$ iterations:

i. Sample N_e^* from $P(N_e|T(x)) \approx P(N_e|x)$ generated in step 1.
ii. Sample $s_{k,i}^*$ from the prior distribution $P(s)$.
iii. Simulate data $X_{k,i}$ (at locus i only) from $P(X_{k,i}|N_e^*,s_{k,i}^*)$.
iv. Compute $U(X_{k,i})$.

Condition on $U(X_i) \approx U(x_i)$ using ABC to obtain a sample of observation (N_e^{**},s_i^{**}) from an approximation to $P(s_i|N_e,x_i)P(N_e|x)$.

As noted by Bazin *et al.* [80], if we write the data as $x=(x_i,x_{-i})$ (where the subscript $-i$ indicates all data except that from locus i), x_{-i} is only used once (in the first step) but the second step re-uses the data x_i a second time. The algorithm samples from the correct posterior distribution only if we use x_{-i} instead of the full data x in the first step. Otherwise it involves an approximation, which is valid if we assume that

$$P(N_e|x_{-i}) \approx P(N_e|x_{-i},x_i) = P(N_e|x)$$

The rationale behind this approximation is that when the number of loci L is large, any given locus i provides a negligible contribution to the information about N_e (see below).

We now give the details of the two steps of this algorithm. In the original algorithm [80], the first step is also achieved using ABC. In our case, we take advantage of having an existing moment-based estimator of N_e available. This also allows us to avoid the assumption of independence between loci by using a Bayesian block bootstrap approach (see below). We define $T(X)$ as a single statistic given by Jorde and Ryman Fs' unbiased estimator of N_e [30]. For all sites, we calculated Fs' between consecutive pairs of time points when the minor allele frequency has reached at least 2% in one observation [30] as:

$$Fs = \frac{(x-y)^2}{z(1-z)} \text{ and } Fs' = \frac{1}{t}\frac{Fs[1-1/(2\tilde{n})]-2/\tilde{n}}{(1+Fs/4)[1-1/(n_y)]}$$

where x and y are the minor allele frequencies at the two time points separated by t generations, $z=(x+y)/2$, and $\tilde{n}$ is the

harmonic mean of the sample size n_x and n_y at the two time points. We keep only the sites where we could obtain at least two values of Fs' in order to average them over time, and thus obtain an estimator for each site of the genome used in step 2 below. Note that in the experiment involving oseltamivir, as only passages 3 to 12 were considered, a small number of sites matched this criterion. We also averaged Fs' values over sites in order to obtain a genome-wide estimator as in [30]. All Fs' values were converted to N_e assuming $t = 13$ generations per passage and $N_e = 1/Fs'$.

A segment-based Bayesian block bootstrap approach [81,82] was used to obtain a distribution for $P(N_e|T(X))$, as we cannot assume the independence of sites owing to linkage disequilibrium. More specifically, as multiple virus infections in cells can lead to segment reassortment [6], we grouped sites by segment in order to obtain an estimator for each segment, and randomly resampled segments with replacement 10'000 times using a Dirichlet prior [81]. We also checked that the approximation $P(N_e|x_{-i}) \approx P(N_e|x)$ is valid in our case by repeating step 1 for each segment, where all sites in the considered segment are excluded in order to account for linkage. We found that the posteriors are very similar to that obtained using the full data (Figure S11). N_e was only slightly increased when we excluded the segments carrying the highest number of beneficial mutations in the presence of oseltamivir (respectively estimated to $N_e = 195$ and $N_e = 199$ when we excluded HA and NA, compared to $N_e = 176$ when considering all segments).

The second step of our method uses the effective size estimated in the first step as a prior distribution to estimate selection coefficients (s) at each site in the genome using an ABC approach [36,80]. For each site, 100'000 time-sampled trajectories were simulated using a Wright-Fisher haploid model with selection [18] with three conditions: *(i)* the trajectories started at the same minor allele frequency observed at this site [35], *(ii)* the trajectories match the same criteria used on the real data to calculate Fs', and *(iii)* the samples are simulated as a binomial sampling using the per-site sample sizes. For each trajectory, we randomly sample N_e from the 10'000 posterior samples obtained in the first step. The relative fitness of the beneficial allele was set to $1+s$, and we used a uniform prior for s between -0.1 and 0.5, as we always consider the minor allele. In the presence of selection, allele frequency trajectories are expected to be directional, whereas drift introduces random variance. Being a measure of variance, Fs' does not incorporate information about the direction of allele frequency changes. To integrate this into our estimation procedure, we decomposed Fs' at a given site in to two statistics: Fsd' and Fsi' calculated respectively between pairs of time points, where the allele considered is decreasing and increasing in frequency, such that $Fs' = Fsd' + Fsi'$. Using notations of our algorithm presented above, this means that at each locus i we take $U(X_i) = (Fsd'_i, Fsi'_i)$. We retained the best 1% of the 100'000 simulations based on the Euclidian distance between observed and simulated Fsd' and Fsi' statistics in order to obtain posterior distributions and means for s using a rejection ABC algorithm [36].

We selected candidate trajectories based on the posterior distribution obtained for s at each site: we define Bayesian 'p-values' for s as $P(s<0|x)$ and consider a trajectory to be 'significant at level p' if it's equal-tailed $100(1-p)\%$ posterior interval excludes zero [43]. We also performed a cross-validation procedure for the ABC method: we randomly simulated 1000 pseudo-observed data sets with parameters inspired by the drug-treated experiment, with both fixed and varying N_e (see Results). For each simulated replicated, we estimated N_e and s using our proposed ABC approach.

We finally identified trajectories for which the Wright-Fisher model has a poor fit based on the Euclidian distance between our simulations and the data. Using the cross-validation procedure, we obtained the null distribution of this distance under the true model, and used the 99% quantile of the distribution as a threshold to detect trajectories in our data not fitting the model.

Likelihood-based estimation

The outlier trajectories were selected with the N_e-based ABC method when the probability of being beneficial was larger than 99%, and used in a likelihood-based method [32] for comparison. The time-serial method of Malaspinas *et al.* [32] is an extension of Bollback *et al.* [31] to infer the selection coefficient, the effective population size, and additionally the allele age from temporal allele frequency data. A Hidden Markov Model (HMM) is used to model the allele frequency trajectory and an approximate transition density is applied to compute the likelihood. Here, the method is modified to fit a haploid model. The diffusion process approximating the Wright-Fisher haploid model with selection is defined by [34,83]:

$$L = \frac{1}{2}a(y)\frac{d^2}{dy^2} + b(y)\frac{d}{dy}$$

$$a(y) = y(1-y)$$

$$b(y) = \gamma y(1-y)$$

where y is the density of allele frequency at time t in units of N_e generations and $\gamma = N_e \cdot s$.

The state space of the HMM are the population allele frequencies denoted by $z_i = i/N_e$ for $i = \{0 \ldots Ne\}$. The diffusion process defining the transition probabilities is approximated with a one-step process, which only allows the transition to occur between adjacent states (z_i to z_{i-1}, z_i and z_{i+1}). The infinitesimal generator Q for the one-step process is a tridiagonal $(N_e+1)\times(N_e+1)$ matrix:

$$Q = \begin{pmatrix} 0 & 0 & 0 & & \cdots & & 0 \\ \delta_1 & \eta_1 & \beta_1 & & & & \\ 0 & \ddots & \ddots & \ddots & & & \vdots \\ & & \delta_i & \eta_i & \beta_i & & \\ \vdots & & & \ddots & \ddots & \ddots & 0 \\ & & & & \delta_{Ne-1} & \eta_{Ne-1} & \beta_{Ne-1} \\ 0 & & \cdots & & 0 & 0 & 0 \end{pmatrix}$$

where β_i denotes the rate of jumping from z_i to z_{i+1}, δ_i the rate of jumping from z_i to z_{i-1}, and $\eta_i = \beta_i + \delta_i$ such that $1-\eta_i$ is the rate of staying in state z_i.

When the approximation of the one-step process is applied, the solution for the system can be obtained as:

$$\beta_i = \frac{(-1+z_i)\, z_i\, (1+\gamma(z_i - z_{i-1}))}{(z_i - z_{i+1})(-z_{i-1}+z_{i+1})} \text{ and}$$

$$\delta_i = \frac{(-1+z_i)\, z_i\, (1+\gamma(z_i - z_{i+1}))}{(z_i - z_{i-1})(z_{i-1}-z_{i+1})}$$

The effective population size N_e was set at 176 in the presence of oseltamivir and at 226 in the absence of drug as inferred by the N_e-

based ABC method (see Results). For each candidate trajectory, the maximum likelihood of γ and allele age t_0 was obtained using the Nelder-Mead optimization algorithm [84], where the search range of γ was set to (0,80). For the allele age t_0, the time point of the first appearance of the derived allele was set to be the upper bound of the search range.

Coalescent-based estimation

We generated allele frequency trajectories for the beneficial allele under a range of selection coefficients s between 0 and 0.5 with 0.005 increments, performing 10'000 simulations for each value. These frequency trajectories were produced using a Wright-Fisher haploid model with selection as described above. For replicate 1 of the drug-treated H1N1 strain, we calculated the population size at each generation using the estimated census population sizes measured at the beginning of each passage (Table 1). Further, we assume exponential population growth for 13 generations during each passage, reaching a final population of 10^6 virions. We accepted those trajectories presenting a difference between the derived allele frequency of observed and simulated data lower than $\varepsilon = 0.10$ at all time points (from passage 3 to passage 12).

The accepted trajectories were used to run the simulation software *msms* [85] under a demographic model resembling the experimental data (twelve consecutive passages with selection starting at passage 4) (Figure 1 and Table 1). *msms* is a coalescent simulation program that incorporates time-sampled data [86] and conditional coalescent on frequency trajectories. *msms* was used to obtain the tree length starting at passage 3 (drug free) and finishing at the end of the experiment (passage 12). We used a search range for the mutation rate μ between 10^{-9} and 10^{-2} and computed the probability of observing the total number of segregating sites present in the whole genome of H1N1 for replicate 1 (9666 segregating sites) given the total tree length multiplied by the mutation rate using binomial sampling. Finally, we computed the probability of observing the real data given the simulated by integrating over all possible genealogies G:

$$\begin{aligned} p(D|\theta,\mu) &= \int p(D|\theta,\mu,G)p(G|\theta,\mu)dG \\ &= \int p(D|\mu,G)p(G|\theta)dG \\ &\simeq \int p(S,f|\mu,G)p(G|\theta)dG \\ &= \int p(S|f,\mu,g)p(f|\mu,\theta)p(G|\theta)dG \\ &\approx \frac{1}{N}\sum_i p(S|\mu,g)I(|f_i - f| < \varepsilon) \end{aligned}$$

where D refers to the total number of segregating sites (S) and the derived allele frequency of the beneficial allele (f). θ represents the demographic model and μ the mutation rate. f_i and f respectively represent the frequency of the simulated and observed beneficial allele, and I the indicator function.

Estimation of the shape of the distribution of fitness effect

Mathematica 9.0 was used to fit the distributions using numerical maximization of the log-likelihoods of the data under the given distribution. The probability density functions resulting from weighting the original distribution with the fixation probability of a beneficial mutation, $1\text{-}e^{-2s}$, and subsequent normalization were performed as follows:

Contending half-normal distribution:

$$f(s) = \frac{2\sigma e^{-\frac{-x^2\sigma^2}{\pi}}(1-e^{-2x})}{\pi(1-e^{\frac{\pi}{\sigma^2}})Erfc(\frac{\sqrt{\pi}}{\sigma})}$$

where $Erfc(x)$ is the complementary error function, for $x \geq 0$.

Scaled beta distribution:

$$f(s) = \frac{x(\mu-x)^{\beta-1}}{Beta(2,\beta)\mu^{\beta+1}} \quad for\ 0 \leq x \leq \mu$$

where Beta(a,b) is the Euler beta function.

Contending generalized Pareto distribution:

$$f(s) = \begin{cases} \dfrac{(1+\frac{\kappa x}{\tau})^{-\frac{\kappa+1}{\kappa}}(1-e^{-2x})}{\int_0^{-\frac{\tau}{\kappa}}(1+\frac{\kappa x}{\tau})^{-\frac{\kappa+1}{\kappa}}(1-e^{-2x})} & for\ 0 \leq x \leq -\dfrac{\tau}{\kappa},\ if\ \kappa < 0, \\ \dfrac{2\tau-1}{2\tau^2}e^{-\frac{2\tau+1}{\tau}x}(1-e^{2x}) & for\ x \geq 0,\ if\ \kappa = 0, \\ \dfrac{(1+\frac{\kappa x}{\tau})^{-\frac{\kappa+1}{\kappa}}(1-e^{-2x})}{\int_0^{\infty}(1+\frac{\kappa x}{\tau})^{-\frac{\kappa+1}{\kappa}}(1-e^{-2x})} & for\ x \geq 0,\ if\ \kappa > 0 \end{cases}$$

In order to evaluate the confidence in the estimated parameters, a weighted likelihood bootstrap was performed as described by Newton and Raftery [62]. 1000 weighted likelihoods were obtained according to Equation 1 therein, with weights being randomly drawn from a uniform Dirichlet distribution. For the scaled beta distribution, the boundary parameter s_0 was limited to values smaller or equal to 1 to ensure successful maximization, and if the estimated MLE yielded $s_0 = 1$, we concluded that no distance to the optimum could be obtained – in fact, in all tested cases in the presence of oseltamivir the likelihood appeared to increase monotonically as s_0 approached infinity (results not shown).

Structural analysis

Residues that mutated during the course of the experiments were highlighted on the structures of HA, NA and M1. The structure for HA of Influenza A/Brisbane/59/2007 has been generated via molecular modeling [87] and was used in this study. Closely related NA (PDB 3CL0) and M1 (PDB 1EA3) structures were also used. Images were generated in PyMol.

Supporting Information

Figure S1 Genetic diversity of H1N1 throughout the experiment. The genetic diversity measured as the average expected heterozygosity in passages 4 to 12 of our experiment in the absence (dotted line) or presence (dashed line) of oseltamivir.
(PDF)

Figure S2 Site frequency spectra (SFS) of H1N1 populations during the experiment. The SFS at passages 4 (A and B) and 12 (C and D) is shown in the absence (A and C) and presence (C and D) of oseltamivir.
(PDF)

Figure S3 ABC correlation plot. The correlation between the simulated selection coefficients s and the two statistics ($Fs'i$ (A) and

Fs'_d (B)) used in our ABC method. Note that Figure 4A is showing the same correlation using colors.
(PDF)

Figure S4 Frequency of identified beneficial mutations in natural populations. The allele frequency in natural populations from the NCBI Influenza Virus Resource database for the significant mutations identified to be under selection - plotted between years 2004 and 2010.
(PDF)

Figure S5 Cross validation of our *Ne*-based ABC method. The true vs. the estimated values of N_e (A and C) and s (B and D) for the 1000 simulated data used to validate our ABC procedure. We used similar parameters to our real data: sample size = 1000, $Ne = 176$ and initial allele frequencies of $1/Ne$. We simulated a population of constant size (A and B) or experiencing recurrent bottleneck ($N = 23$) followed by exponential growth (up to $N = 10^6$) mimicking our experiment (C and D). Error bars in B and D represent the 10% and 90% quantiles over the 1000 replicates. The red dot in A and C and the red line in B and D indicate the true value.
(PDF)

Figure S6 SNPs with poor fit to the Wright-Fisher model. The minor allele frequency trajectories of all SNPs identified as not fitting the Wright-Fisher model in the absence and presence of oseltamivir respectively in A and B. The horizontal dotted red line indicates the start of oseltamivir treatment (see Figure 1). Trajectories are represented in dashed lines if a second SNP was significant within the same segment. For each SNP, the name of the segment, the position of the SNP, the nucleotide increasing in frequency, and the estimated selection coefficients with our *Ne*-based ABC method are indicated in the top left corner of A and B.
(PDF)

Figure S7 *Ne*-based Approximate Bayesian Computation for SNP PB1 33 (K11). For 10'000 simulated trajectories (out of the 100'000 simulations performed), we plot in A the values of the statistics Fs'_i and Fs'_d with colors corresponding to the selection coefficients s, as well as the values calculated for the real trajectory of the poor fitting PB1 33 (K11) mutation (see Figure S5) in black. We indicate the region corresponding to the best 1% retained simulations with a dashed line, and we plot the corresponding two-dimensional posterior distribution for s and N_e in B. We clearly see in A the inability of the model to generate simulations near the observed data, with the black dot being outside the retained regions defined by the dashed lines.
(PDF)

Figure S8 Evidence of positive selection in the H1N1 genome in the absence and presence of oseltamivir for replicated data. We plot the Bayesian P-values of each SNP in log scale in the absence and presence of oseltamivir in A and C, respectively. The horizontal red lines are genome-wide significance thresholds of $P = 0.01$. The eight segments are separately color-coded, a scheme which is maintained in all panels and in Figure 3. Significant nonsynonymous mutations are represented with triangles. We plot the minor allele frequency trajectories of all significant SNPs over the replicated experiment in the absence and presence of oseltamivir respectively in B and D. The horizontal dotted red line indicates the start of oseltamivir treatment (see Figure 1). All colors and line styles match those in Figure 1. Trajectories are represented as dashed lines when a second SNP was significant in a segment, and dotted lines for a third SNP. For each significant SNP, the name of the segment, the position of the SNP, the nucleotide increasing in frequency, and the estimated selection coefficients with our *Ne*-based ABC method are indicated in the top left corner of B and D.
(PDF)

Figure S9 Maximum likelihood for the DFE fit. The boxplots show the distribution of the maximum log-likelihoods obtained from 1000 samples of a weighted likelihood bootstrap in the absence and presence of oseltamivir in A and B, respectively.
(PDF)

Figure S10 Genome wide sequence coverage data for samples used in this study. The coverage in log scale for our four experiments at passages 0, 6 and 12 (see Figure 1).
(PDF)

Figure S11 Estimated N_e when excluding segments. The posterior distribution obtained for N_e using step 1 of our ABC algorithm when excluding each segment one by one in the absence (A) and presence (B) of oseltamivir. Segment colors match those in Figure 3 and S8.
(PDF)

Table S1 Estimated selection coefficients for the replicate experiment. Comparison of N_e-ABC and Malaspinas *et al.* [32] estimates of s for the significant trajectories under selection for the replicate experiment. Bold indicates nonsynonymous mutations. We indicate the nucleotide corresponding to the minor allele, with its initial frequency at the beginning of the experiment in the absence of oseltamivir, or at passage 4 when drug treatment began (see Figure 1). For the *Ne*-ABC method, we give the 99% highest posterior density intervals (HPDIs) in brackets.
(PDF)

Table S2 Viral passaging and drug concentration data for samples used in this study.
(PDF)

Table S3 Genome wide sequence coverage data for samples used in this study.
(PDF)

Acknowledgments

We thank Cornelia Pokalyuk for useful discussions on theoretical population genetics, and Alex Wong for a critical reading of the manuscript. Additionally, we thank Jesse Bloom and an anonymous reviewer for helpful comments and suggestions that have improved the clarity of the manuscript.

Author Contributions

Conceived and designed the experiments: RWF CAS TFK JPW DNB KBZ DRC JDJ NR. Performed the experiments: RWF TFK JPW PL NR. Analyzed the data: MF YPP NR AFA HS ASM GE CB DW DRC KBZ JDJ. Contributed reagents/materials/analysis tools: MF YPP NR AFA HS ASM GE CB PL DW DRC KBZ DNB JPW TFK CAS RWF JDJ. Wrote the paper: MF YPP NR AFA HS CB DW DRC KBZ JPW JDJ.

References

1. Thompson WW, Shay DK, Weintraub E, Brammer L, Cox N, et al (2003) Mortality associated with influenza and respiratory syncytial virus in the United States. JAMA 289: 179–186.
2. Yu H, Cowling BJ, Feng L, Lau EH, Liao Q, et al (2013) Human infection with avian influenza A H7N9 virus: an assessment of clinical severity. Lancet.
3. Nelson MI, Holmes EC (2007) The evolution of epidemic influenza. Nature Reviews Genetics 8: 196–205.
4. Kuiken T, Holmes EC, McCauley J, Rimmelzwaan GF, Williams CS, et al (2006) Host species barriers to influenza virus infections. Science 312: 394–397.

5. Parvin JD, Moscona A, Pan WT, Leider JM, Palese P (1986) Measurement of the mutation rates of animal viruses: influenza A virus and poliovirus type 1. J Virol 59: 377–383.
6. Palese P, Young JF (1982) Variation of influenza A, B, and C viruses. Science 215: 1468–1474.
7. Nelson MI, Simonsen L, Viboud C, Miller MA, Taylor J, et al (2006) Stochastic processes are key determinants of short-term evolution in influenza a virus. PLoS Pathog 2: e125.
8. Taubenberger JK, Kash JC (2010) Influenza virus evolution, host adaptation, and pandemic formation. Cell Host Microbe 7: 440–451.
9. Moscona A (2005) Oseltamivir resistance–disabling our influenza defenses. N Engl J Med 353: 2633–2636.
10. Collins PJ, Haire LF, Lin YP, Liu J, Russell RJ, et al (2008) Crystal structures of oseltamivir-resistant influenza virus neuraminidase mutants. Nature 453: 1258–1261.
11. Ives JAL, Carr JA, Mendel DB, Tai CY, Lambkin R, et al (2002) The H274Y mutation in the influenza A/H1N1 neuraminidase active site following oseltamivir phosphate treatment leave virus severely compromised both in vitro and in vivo. Antiviral Res 55: 307–317.
12. Ghedin E, Holmes EC, DePasse JV, Pinilla LT, Fitch A, et al (2012) Presence of oseltamivir-resistant pandemic A/H1N1 minor variants before drug therapy with subsequent selection and transmission. J Infect Dis 206: 1504–1511.
13. Gubareva LV, Kaiser L, Matrosovich MN, Soo-Hoo Y, Hayden FG (2001) Selection of influenza virus mutants in experimentally infected volunteers treated with oseltamivir. J Infect Dis 183: 523–531.
14. Moscona A (2009) Global transmission of oseltamivir-resistant influenza. N Engl J Med 360: 953–956.
15. Bloom JD, Gong LI, Baltimore D (2010) Permissive secondary mutations enable the evolution of influenza oseltamivir resistance. Science 328: 1272–1275.
16. Bouvier NM, Rahmat S, Pica N (2012) Enhanced mammalian transmissibility of seasonal influenza A/H1N1 viruses encoding an oseltamivir-resistant neuraminidase. J Virol 86: 7268–7279.
17. Ginting TE, Shinya K, Kyan Y, Makino A, Matsumoto N, et al (2012) Amino acid changes in hemagglutinin contribute to the replication of oseltamivir-resistant H1N1 influenza viruses. J Virol 86: 121–127.
18. Wright S (1931) Evolution in Mendelian Populations. Genetics 16: 97–159.
19. Kimura M, Ohta T (1969) The Average Number of Generations until Fixation of a Mutant Gene in a Finite Population. Genetics 61: 763–771.
20. Crisci JL, Poh Y-P, Bean A, Simkin A, Jensen JD (2012) Recent progress in polymorphism-based population genetic inference. J Hered 103: 287–296.
21. Nei M (1973) Analysis of Gene Diversity in Subdivided Populations. Proc Natl Acad Sci U S A 70: 3321–3323.
22. Krimbas CB, Tsakas S (1971) The Genetics of Dacus oleae. V. Changes of Esterase Polymorphism in a Natural Population Following Insecticide Control-Selection or Drift? Evolution 25: 454–460.
23. Pamilo P, Varvio-Aho SL (1980) On the estimation of population size from allele frequency changes. Genetics 95: 1055–1057.
24. Nei M, Tajima F (1981) Genetic drift and estimation of effective population size. Genetics 98: 625–640.
25. Waples RS (1989) A generalized approach for estimating effective population size from temporal changes in allele frequency. Genetics 121: 379–391.
26. Williamson EG, Slatkin M (1999) Using maximum likelihood to estimate population size from temporal changes in allele frequencies. Genetics 152: 755–761.
27. Berthier P, Beaumont MA, Cornuet J-M, Luikart G (2002) Likelihood-based estimation of the effective population size using temporal changes in allele frequencies: a genealogical approach. Genetics 160: 741–751.
28. Anderson EC, Williamson EG, Thompson EA (2000) Monte Carlo evaluation of the likelihood for Ne from temporally spaced samples. Genetics 156: 2109–2118.
29. Anderson EC (2005) An Efficient Monte Carlo Method for Estimating Ne From Temporally Spaced Samples Using a Coalescent-Based Likelihood. Genetics 170: 955–967.
30. Jorde PE, Ryman N (2007) Unbiased estimator for genetic drift and effective population size. Genetics 177: 927–935.
31. Bollback JP, York TL, Nielsen R (2008) Estimation of 2Nes from temporal allele frequency data. Genetics 179: 497–502.
32. Malaspinas A-S, Malaspinas O, Evans SN, Slatkin M (2012) Estimating allele age and selection coefficient from time-serial data. Genetics 192: 599–607.
33. Mathieson I, McVean G (2013) Estimating selection coefficients in spatially structured populations from time series data of allele frequencies. Genetics 193: 973–984.
34. Durrett R (2008) Probability models for DNA sequence evolution. New York: Springer.
35. Goldringer I, Bataillon T (2004) On the distribution of temporal variations in allele frequency: consequences for the estimation of effective population size and the detection of loci undergoing selection. Genetics 168: 563–568.
36. Sunnåker M, Wodak S, Busetto AG, Numminen E, Corander J, et al (2013) Approximate Bayesian Computation. PLoS Comput Biol 9: e1002803.
37. Grassly NC, Harvey PH, Holmes EC (1999) Population dynamics of HIV-1 inferred from gene sequences. Genetics 151: 427–438.
38. de Silva E, Ferguson NM, Fraser C (2012) Inferring pandemic growth rates from sequence data. J R Soc Interface 9: 1797–1808.
39. Renzette N, Gibson L, Bhattacharjee B, Fisher D, Schleiss MR, et al (2013) Rapid Intrahost Evolution of Human Cytomegalovirus Is Shaped by Demography and Positive Selection. PLoS Genet 9: e1003735.
40. Crow JF, Denniston C (1988) Inbreeding and Variance Effective Population Numbers. Evolution 42: 482–495.
41. Thangavel RR, Reed A, Norcross EW, Dixon SN, Marquart ME, et al (2011) "Boom" and "Bust" cycles in virus growth suggest multiple selective forces in influenza a evolution. Virol J 8: 180.
42. Stray SJ, Air GM (2001) Apoptosis by influenza viruses correlates with efficiency of viral mRNA synthesis. Virus Res 77: 3–17.
43. Beaumont MA, Balding DJ (2004) Identifying adaptive genetic divergence among populations from genome scans. Mol Ecol 13: 969–980.
44. Bao Y, Bolotov P, Dernovoy D, Kiryutin B, Zaslavsky L, et al (2008) The influenza virus resource at the National Center for Biotechnology Information. J Virol 82: 596–601.
45. Ewens WJ (1967) The probability of survival of a new mutant in a fluctuating environment. Heredity (Edinb) 22: 438–443.
46. He C-Q, Ding N-Z, Mou X, Xie Z-X, Si H-L, et al (2012) Identification of three H1N1 influenza virus groups with natural recombinant genes circulating from 1918 to 2009. Virology 427: 60–66.
47. Barton NH (2000) Genetic hitchhiking. Philos Trans R Soc Lond B Biol Sci 355: 1553–1562.
48. Strelkowa N, Lässig M (2012) Clonal interference in the evolution of influenza. Genetics 192: 671–682.
49. Miller CR, Joyce P, Wichman HA (2011) Mutational effects and population dynamics during viral adaptation challenge current models. Genetics 187: 185–202.
50. Fisher RA (1930) The Genetical Theory Of Natural Selection. At The Clarendon Press.
51. Efron B (1987) Better Bootstrap Confidence Intervals. Journal of the American Statistical Association 82: 171–185.
52. Martin G, Lenormand T (2006) A general multivariate extension of Fisher's geometrical model and the distribution of mutation fitness effects across species. Evolution 60: 893–907.
53. Martin G, Lenormand T (2008) The Distribution of Beneficial and Fixed Mutation Fitness Effects Close to an Optimum. Genetics 179: 907–916.
54. Beisel CJ, Rokyta DR, Wichman HA, Joyce P (2007) Testing the extreme value domain of attraction for distributions of beneficial fitness effects. Genetics 176: 2441–2449.
55. Orr HA (2006) The distribution of fitness effects among beneficial mutations in Fisher's geometric model of adaptation. J Theor Biol 238: 279–285.
56. Kassen R, Bataillon T (2006) Distribution of fitness effects among beneficial mutations before selection in experimental populations of bacteria. Nat Genet 38: 484–488.
57. MacLean RC, Buckling A (2009) The distribution of fitness effects of beneficial mutations in Pseudomonas aeruginosa. PLoS Genet 5: e1000406.
58. Rokyta DR, Beisel CJ, Joyce P, Ferris MT, Burch CL, et al (2008) Beneficial fitness effects are not exponential for two viruses. J Mol Evol 67: 368–376.
59. Bataillon T, Zhang T, Kassen R (2011) Cost of adaptation and fitness effects of beneficial mutations in Pseudomonas fluorescens. Genetics 189: 939–949.
60. Barrett RDH, MacLean RC, Bell G (2006) Mutations of intermediate effect are responsible for adaptation in evolving Pseudomonas fluorescens populations. Biol Lett 2: 236–238.
61. Rozen DE, de Visser JAGM, Gerrish PJ (2002) Fitness effects of fixed beneficial mutations in microbial populations. Current Biology 12: 1040–1045.
62. Newton MA, Raftery AE (1994) Approximate Bayesian Inference with the Weighted Likelihood Bootstrap. Journal of the Royal Statistical Society. Series B (Methodological) 56: 3–48.
63. Bank C, Hietpas RT, Wong A, Bolon DNA, Jensen JD. (2014) A Bayesian MCMC approach to assess the complete distribution of fitness effects of new mutations: uncovering the potential for adaptive walks in challenging environments. Genetics. [Epub ahead of print].
64. Hietpas RT, Bank C, Jensen JD, Bolon DNA (2013) Shifting fitness landscapes in response to altered environments. Evolution 67: 3512–3522.
65. Sha B, Luo M (1997) Structure of a bifunctional membrane-RNA binding protein, influenza virus matrix protein M1. Nat Struct Biol 4: 239–244.
66. Arzt S, Baudin F, Barge A, Timmins P, Burmeister WP, et al (2001) Combined results from solution studies on intact influenza virus M1 protein and from a new crystal form of its N-terminal domain show that M1 is an elongated monomer. Virology 279: 439–446.
67. Koerner I, Matrosovich MN, Haller O, Staeheli P, Kochs G (2012) Altered receptor specificity and fusion activity of the haemagglutinin contribute to high virulence of a mouse-adapted influenza A virus. Journal of General Virology 93: 970–979.
68. Daniels RS, Downie JC, Hay AJ, Knossow M, Skehel JJ, et al (1985) Fusion mutants of the influenza virus hemagglutinin glycoprotein. Cell 40: 431–439.
69. Steinhauer DA, Martín J, Lin YP, Wharton SA, Oldstone MB, et al (1996) Studies using double mutants of the conformational transitions in influenza hemagglutinin required for its membrane fusion activity. Proceedings of the National Academy of Sciences of the United States of America 93: 12873–12878.
70. Enami M, Enami K (1996) Influenza virus hemagglutinin and neuraminidase glycoproteins stimulate the membrane association of the matrix protein. J Virol 70: 6653–6657.

71. Jin H, Leser GP, Zhang J, Lamb RA (1997) Influenza virus hemagglutinin and neuraminidase cytoplasmic tails control particle shape. EMBO J 16: 1236–1247.
72. Rossman JS, Lamb RA (2011) Influenza virus assembly and budding. Virology 411: 229–236.
73. Noton SL, Medcalf E, Fisher D, Mullin AE, Elton D, et al (2007) Identification of the domains of the influenza A virus M1 matrix protein required for NP binding, oligomerization and incorporation into virions. J Gen Virol 88: 2280–2290.
74. Reed ML, Yen H-L, DuBois RM, Bridges OA, Salomon R, et al (2009) Amino acid residues in the fusion peptide pocket regulate the pH of activation of the H5N1 influenza virus hemagglutinin protein. J Virol 83: 3568–3580.
75. Thoennes S, Li Z-N, Lee B-J, Langley WA, Skehel JJ, et al (2008) Analysis of residues near the fusion peptide in the influenza hemagglutinin structure for roles in triggering membrane fusion. Virology 370: 403–414.
76. Sriwilaijaroen N, Suzuki Y (2012) Molecular basis of the structure and function of H1 hemagglutinin of influenza virus. Proc Jpn Acad Ser B Phys Biol Sci 88: 226–249.
77. Lin YP, Wharton SA, Martín J, Skehel JJ, Wiley DC, et al (1997) Adaptation of egg-grown and transfectant influenza viruses for growth in mammalian cells: selection of hemagglutinin mutants with elevated pH of membrane fusion. Virology 233: 402–410.
78. Ilyushina NA, Govorkova EA, Russell CJ, Hoffmann E, Webster RG (2007) Contribution of H7 haemagglutinin to amantadine resistance and infectivity of influenza virus. J Gen Virol 88: 1266–1274.
79. Renzette N, Caffrey DR, Zeldovich KB, Liu P, Gallagher GR, et al (2014) Evolution of the influenza A virus genome during development of oseltamivir resistance in vitro. J Virol 88:272–81.
80. Bazin E, Dawson KJ, Beaumont MA (2010) Likelihood-free inference of population structure and local adaptation in a Bayesian hierarchical model. Genetics 185: 587–602.
81. Rubin DB (1981) The Bayesian Bootstrap. The Annals of Statistics 9: 130–134.
82. Hall P (1985) Resampling a coverage pattern. Stochastic Processes and their Applications 20: 231–246.
83. Ewens WJ (2004) Mathematical population genetics : theoretical introduction. New York: Springer.
84. Nelder JA, Mead R (1965) A Simplex Method for Function Minimization. The Computer Journal 7: 308–313.
85. Ewing G, Hermisson J (2010) MSMS: a coalescent simulation program including recombination, demographic structure and selection at a single locus. Bioinformatics 26: 2064–2065.
86. Rodrigo AG, Shpaer EG, Delwart EL, Iversen AK, Gallo MV, et al (1999) Coalescent estimates of HIV-1 generation time in vivo. Proceedings of the National Academy of Sciences of the United States of America 96: 2187–2191.
87. Igarashi M, Ito K, Yoshida R, Tomabechi D, Kida H, et al (2010) Predicting the antigenic structure of the pandemic (H1N1) 2009 influenza virus hemagglutinin. PLoS One 5: e8553.