Bayesian coalescent inference of in-host evolution using Next Generation Sequencing

Gayle Leen^{1,2}, Marc Baguelin^{3,4,*}

1 Anthony Nolan Research Institute, London, United Kingdom

2 University College London Cancer Institute, London, United Kingdom

3 Public Health England, London, United Kingdom

4 London School of Hygiene and Tropical Medicine, London, United Kingdom

*** E-mail: marc.baguelin@phe.gov.uk**

Abstract

Within an infected individual, influenza virus exists as a heterogeneous population of variants. When representing the viral population as a consensus sequence, information about minority variants is lost. However, using next generation sequencing (NGS), it is possible to identify nucleotide substitutions which segregate at low frequencies in the viral population, and can give insight into the within-host processes that drive the virus's evolution, and is a step towards understanding the dynamics of the disease. During the course of an infection, mutations may occur, and at each segregating site, the frequency of the derived allele in the population will fluctuate. We develop a method which can use information about the relative frequencies of mutations in NGS data from a viral population sampled at multiple time points, to infer past population dynamics with a Bayesian skyline model. By using coalescent theory, we analytically derive the joint allele frequency spectrum for a population across multiple time points, and relate this to the coalescent intervals generated from the skyline model. We demonstrate the model on data taken from populations of equine influenza virus sampled during an infection, and show that it is possible to infer a posterior distribution of effective viral population size through time. We also show how the model can be used to infer the probability that a mutation occurred within-host, as opposed to being an ancestral mutation which occurred prior to infection.

Author Summary

When a host is infected by a virus, many particles of the infecting agent enter the body of the host. This viral population is composed of many closely related viruses that continue diversifying by mutating while reproducing in the host. New sequencing technologies allow the quantifying of the proportion of the different variants present in the host at a particular time. Unfortunately, the data resulting from such sequencing techniques are difficult to interpret as they consist of many unlinked copies of relatively small fragments of genetic code distributed along the genome of the virus.

We designed a method combining models of virus genealogies and frequency of mutations appearing in the data to reconstruct the variation of the viral population inside the host. It also allows us to time the apparition of particular variants. This could be useful to detect if a particular mutation (e.g. providing drug resistance) has appeared in host or was circulating before. We applied our method to data of within-host evolution of equine influenza.

1 Introduction

2 Understanding the evolution and kinetics of the within-host viral population during an influenza infection
3 is an essential step to understanding the evolution and dynamics of viral populations on a larger scale.
4 This is particularly important as viral evolution will have an impact on the dynamics of the disease,
5 and potentially its virulence or resistance to some of the existing treatments. Influenza viruses exist as
6 heterogeneous populations of evolutionarily related haplotypes in their infected hosts, due to short viral

generation times and high mutation rates. In a within-host population, the genetic diversity is shaped at both the between-host level, due to host to host transmission dynamics and transmission bottlenecks, and the within-host level, through the host immune response, and the within-host demographic history.

Maintaining a diverse population of variants may be beneficial in adapting to such a dynamic environment, because a sudden change in environment which, in quasispecies theory terms, shifts the fitness landscape, may favour previously less fit, minority variants in maintaining the population's survival [1]. However, it has been found that for influenza A virus, even though there is rapid variation over time (on a scale of years), there is low genetic diversity within a single host, and low genetic diversity across the global population at a single time [2]. Characterising the dynamics of the minority variants of the within-host viral population during an infection may be a step towards explaining this pattern.

Interactions between virus and host have previously been studied with viral kinetic models, where the characteristics of a viral growth and decay curve is governed by a set of biological parameters, and are fitted to viral load estimates sampled over time [3,4]. However, the relationship between viral load and genetic diversity is unclear. Phylodynamics methods based on concepts from the population genetics field can be used to make inferences about the changes in population size from the patterns of mutations among sequences sampled from within-host over time [5–7]. In general, these methods summarise samples of the within-host viral population as consensus sequences, and use these to infer changes in effective population size, such as the HIV intrahost study in [8].

Characterising the sequence variation of a viral population has been enabled by recent advances in sequencing technologies, allowing researchers to look beyond the ‘consensus view’ [9]. Previously, information about the genetic diversity could only be accessed through isolating and cloning individual viruses and then applying Sanger sequencing [10,11], which identified the predominant variants in the population. Next generation sequencing (NGS) technologies enable quick and low-cost sequencing of viral populations at high coverage, giving insight into the low frequency variants present in the sample. Though identification of the low frequency variants can be problematic, due to sequencing error rates, which may be of the same order as the variant frequency [12], and many existing variant calling algorithms focus on sites segregating at discrete frequencies, whereas in a viral population, the number of genomes, and hence the possible variant frequencies, are unknown, robust approaches to calling single nucleotide variants (SNVs) from high coverage datasets have been developed recently [12,13].

Sequencing a viral population using NGS restricts genetic diversity estimation to SNVs or local fragments constrained by the read length. Though methods to assemble individual viral haplotypes have been proposed [14,15], they can be computationally intensive. Consequently studying intrahost viral evolution at the level of whole genomes from NGS data is not straightforward, since existing phylodynamics methods e.g. [16–18] rely on having a known number of viral sequences. In this paper, we do not reconstruct the viral haplotypes; instead we analyse the viral population as a set of unlinked polymorphic sites (which we assume to be biallelic).

At each segregating site, we can calculate the frequency at which the derived allele appears, as the within-host viral population is sampled at various time points during an infection. This type of multidimensional SNP frequency data has been shown to be informative about the joint demography history of multiple populations [19,20], and we follow this approach in this paper. We use the joint allele frequency spectrum (JAFS) as a measure of genetic diversity in the viral population shared across multiple time points, and assume that the population at each time point contains N genomes.

Results

New methodology for in-host evolution

We developed a Bayesian method for estimating the trajectory of effective viral population size over time, from next generation sequencing of the within-host viral population sampled at P time points. The

Figure 1. Performance of model on simulated data Data sets of 50 samples at times 2, 4, and 6 days (150 tips in total), with 200 segregating sites, were generated from coalescent intervals under a viral dynamical model to given effective viral population size trajectories peaking at 2 (left), 3 (middle), and 4 (right) days. The true product of the effective viral population size and generation time ($N_e \tau$) is shown as a dashed red line. The median (black) and 95% confidence intervals (grey) of $N - e\tau$ estimated from the posterior samples from our model are summarised in a Bayesian skyline plot, where we have used $m = 15$ intervals.

genetic diversity in each sample is represented by a population of arbitrary size N , and the number of carriers of the derived allele k , ($k \leq N$), is evaluated at a set of unlinked biallelic sites. For each site, the data is given as a segregating site pattern across the P time points.

We derived the probability of each pattern by using coalescent theory, following results derived for the joint allele frequency spectrum for multiple populations in [20]. The dependencies between the viral populations across the time points is modelled through a coalescent tree which represents the ancestral relationships between the $N \times P$ tips. We obtained a likelihood function for the data $\mathcal{D}$, given the expected coalescent intervals $\bar{\mathbf{b}}$ of the underlying genealogy $p(\mathcal{D} | \bar{\mathbf{b}})$.

We used the Bayesian skyline plot framework [16] to place a prior on $\bar{\mathbf{b}}$. This quantifies the relationship between the length of the coalescent intervals of the genealogy of the viral population members, and the demographic history. As opposed to depending on a pre-defined parametric model, such as constant, or exponential growth, the skyline plot framework allows a flexible prior on possible histories, using a piecewise constant demographic model of effective population size through time. The model was parameterised by an ordered subset of group sizes $\mathbf{A} = \{a_1, \dots, a_m\}$, which defines the number of coalescent events in each grouped interval, and $m \leq (N \times P - 1)$. Each of the m groups was assumed to have a constant effective population size $\Theta = \{\theta_1, \dots, \theta_m\}$.

An MCMC scheme was used to sample from the posterior distribution $p(\mathbf{A}, \Theta, \bar{\mathbf{b}} | \mathcal{D})$. Given a set of posterior samples $\{\mathbf{A}_i, \Theta_i, \bar{\mathbf{b}}_i\}$, for each sample, we could evaluate the marginal posterior distribution of $\theta(t)$ at a set of times of interest t_j , such that we have a set of marginal distributions $\theta(t_j)$, that represent the estimated posterior effective population size through time, with its uncertainty.

73 Simulated data

74 To investigate the behaviour of our intrahost model, we analysed simulated data sets from a basic model of
 75 viral dynamics (see Appendix). We fixed some of the parameters (δ , c , k , T_0) to average values obtained
 76 from equine influenza infections in [4], as in Table 1, and set $V_0 = 10$. We generated sets of data for 3
 77 scenarios: $\{\beta = 1.4 \times 10^{-4}, p = 1.4 \times 10^{-5}\}$, $\{\beta = 1.4 \times 10^{-4}, p = 4 \times 10^{-6}\}$, $\{\beta = 1 \times 10^{-4}, p = 2.8 \times 10^{-6}\}$,
 78 which have a viral peak at around 2, 3, and 4 days respectively. Each data set was generating using the
 79 following procedure: (1) The viral trajectory was generated given the parameters. (2) A coalescent tree
 80 was generated for N sequences taken at time 2, 4, and 6 days ($3 * N$ sequences in total), by generating
 81 the coalescent intervals as described in the previous sections, and then generating a random topology. (3)
 82 M segregating sites were generated. A mutation event was generated by picking a branch of the tree with
 83 probability proportional to the branch length, and all descendants of the branch shared the mutation.
 84 Counts of the number of sequences at each sampling point which shared the mutation were stored e.g.
 85 $\{k_1, k_2, k_3\}$.

86 We ran the MCMC inference for each data set using a burn-in of 100,000 samples, and for a chain
 87 of 2,000,000 samples, with thinning of 100, with 200 segregating sites, and $N = 50$. We used $m = 15$
 88 population size intervals. Figure 1 shows the posterior viral trajectories from the model. The model
 89 manages to capture the growth in population size after infection at $t = 0$, and performs better when
 90 the rate of increase is slower i.e. when the population peaks at 4 days (right plot). We also find that
 91 the model does not capture the decrease in viral population after the peak. This is due to the large
 92 coalescent intervals just before the last sample at 6 days (long terminal branches in the genealogy), which
 93 are uninformative about change in population size.

94 Intrahost evolution of equine influenza

95 We tested the model on intrahost viral populations sampled at multiple time points from horses experimen-
 96 tally infected with influenza A(H3N8) virus. [data available on request]. A majority consensus sequence
 97 was constructed from the inoculum sample, and used as the reference sequence. Daily samples from the
 98 horses were paired-end sequenced using the Illumina platform, and aligned with the Tanoti reference
 99 assembler <http://www.bioinformatics.cvr.ac.uk/Tanoti/index.php>. Quality checks were performed
 100 with FastQC <http://www.bioinformatics.babraham.ac.uk/projects/fastqc/>, followed by trimming
 101 of adaptors and low quality scores with Trim Galore! [http://www.bioinformatics.babraham.ac.uk/](http://www.bioinformatics.babraham.ac.uk/projects/trim_galore/)
 102 [projects/trim_galore/](http://www.bioinformatics.babraham.ac.uk/projects/trim_galore/), resulting in a BAM file for each viral population at each time point. Duplicates
 103 were removed using Picard v1.119. Variants were called using LoFreq [13]. For each horse, we obtain a
 104 set of segregating sites, and the frequency of the derived allele at each site, across all time points. After
 105 assuming that the viral population can be represented by $n = 50$ viral genomes at each time point, and
 106 using (1), we obtain a set of segregating site patterns for each horse. The data sets consist of horses 2761
 107 (N) (2, 3, and 5 days after infection), 6652 (2, 3, 4, 5), 6292 (3, 4, 5, 6), 1420 (3, 5, 6, 7), 2220 (4, 5, 6, 7),
 108 and 6285 (3, 5, 7).

109 For the data set from each horse, we ran the MCMC inference (burn-in of 100,000 samples, for a chain
 110 of 2,000,000 samples, with thinning of 100) to estimate the effective population size trajectories. For
 111 all experiments, we use $m = 15$ population size intervals, and $N = 50$. Figure 2 shows the results from
 112 running the model on the data for the 6 horses. Each subfigure contains 3 plots. The top plot shows the
 113 posterior scaled effective population size trajectories ($N_e \tau$), with the median and 95% confidence intervals.
 114 The middle plot shows the posterior density of $p(t_\mu)$ (red), the time at which the mutation arose, for each
 115 segregating site pattern. The densities are superimposed over each other, such that the darkest areas
 116 correspond to times where mutations are more probable. The mean of the densities is shown in black. The
 117 bottom plot shows the calculated number of viral copies per μl , obtained through qPCR. For all plots,
 118 the vertical blue lines show the start of infection (solid) and the times at which the viral population was
 119 sampled (dashed). The time axis is truncated to the median of the sum of branch lengths (tree height).

For each mutation, we calculated the median probability that the mutation had risen within-host, $p(t_\mu \geq 0)$. In Figure 3, we plot the frequency at which the mutation occurs in the viral population, against time (square points), for each horse. The trajectories (dotted line) are shaded according to probability $p(t_\mu \geq 0)$, such that mutations estimated to have occurred within-host $p(t_\mu \geq 0)$ are shaded purple, and those which were present before infection are shaded green.

Discussion

Next generation sequencing of intrahost viral populations enables the identification of minority variants and their frequencies. Given NGS samples of a viral population over the course of an infection, we designed a novel methodology that uses the change in variant frequencies to infer in-host effective viral population size over time as a Bayesian skyline plot. The resulting skyline plots might be difficult to interpret as they relate to the effective viral population sizes rather than the actual number of viruses present in the host (viral load). This difficulty of interpretation also exists for the methodology as commonly applied at a (host) population level. Additionally, the resulting genealogy might present coalescent events which occurred before the time of infection. While the interpretation of what happens before the infection event is difficult, this reflects that diversity is present in a transmission event and passed from host to host. We are able to predict when particular mutations are likely to have appeared and thus inferring the probability of the mutation to have occurred in host. This could provide useful information while tracking the appearance of particular mutations such as the ones known to confer resistance to antiviral drugs.

While high frequency is usually an indicator of appearance outside of the host, it is not possible to predict from frequency only if a particular mutation has appeared inside or outside the host. Our methodology allows us to quantify the probability of both events. This would be particularly useful e.g. to study chains of transmission as low frequency minority variants might not appear in all samples or might be classified as error from the amplification process. Our study, by providing a tool to study in-host evolution of viral populations, paves the way toward the inference of chains of transmission where the additional mechanisms of loss of diversity through host-to-host bottlenecks would need to be integrated. Ultimately, this will lead to a better understanding of the drivers of evolution of viral populations during outbreaks [11].

Materials and Methods

Beyond the consensus

The ability to look beyond the consensus view of a heterogeneous viral population offers a way into characterising within-host sequence diversity during an infection, and to determine the population genetic processes which underlie the evolution of the population. Next generation sequencing (NGS) techniques can be used to resequence heterogeneous viral populations at high coverage, and reveal nucleotide substitutions present in a small fraction of the population. NGS, or high throughput sequencing estimates the sequence of an input sample by fragmenting it into short sections, or reads, which are then amplified and sequenced. The reads can then be aligned against a reference genome to determine if and where the input sample differs from the reference sequence.

When sequencing a viral population, the input sample consists of an unknown number of potentially different viral genomes. Because NGS requires the sample to be fragmented into short reads, it is challenging to determine from which genome a read originated, and consequently reconstruct the sequence of each individual genome. In this paper, we represent the genetic diversity of a viral population as a set of unlinked biallelic sites.

Suppose that we have sequenced a viral population over the course of an infection at time points $t_1, \dots, t_P$. For each of the P time points, we find the sites at which there are SNVs (relative to a reference

sequence), and the frequency of the derived allele across the population. Let us denote the number of unique biallelic sites, which occur once or more across the P time points, as S . The observed allele frequency in the viral population over time for the i th segregating site is given by the vector $\mathbf{q}_i = \{q_{i,1}, \dots, q_{i,P}\}$.

We assume that the sampled viral population at t_p is of size n_p . The number of genomes that contain the i th segregating site over time is given by

$$\mathbf{d}_i = \{\text{round}(n_p \mathbf{q}_p)\}_{p=1, \dots, P} \quad (1)$$

and the full dataset is given by $\mathcal{D} = \{\mathbf{d}_i\}_{i=1, \dots, S}$.

Site frequency spectrum

The allele frequency spectrum (AFS) is a series of statistics that describe genetic polymorphism, and is defined as the sampling distribution of allele frequency at any random polymorphic locus in the genome. The AFS is informative for making inferences about the demographic history of a sample. While the majority of work has focused on using the AFS of a single population, advances have been made in using the joint allele frequency spectrum (JAFS) to infer more complex demographic events from multiple populations.

In this work, we derive the JAFS of multiple samples taken at different time points throughout an influenza infection, using coalescent theory. Given that we have P samples of an intrahost viral population, of sizes $n_1, \dots, n_P$ taken at times $t_1, \dots, t_P$, we denote the JAFS of the samples as $\mathbf{m} = \{m_{k_1, \dots, k_P}(n_1, \dots, n_P), 0 \leq k_1 \leq n_1, 0 \leq k_2 \leq n_2, \dots, 0 \leq k_P \leq n_P\}$, where $m_{k_1, \dots, k_P}(n_1, \dots, n_P)$ denotes the probability that a SNP will have k_i copies of the derived allele in the n_i haplotypes of the i th population, $i = 1, \dots, P$.

We can derive the JAFS based on the genealogy underlying the P samples, with tips at $t_1, \dots, t_P$. Figure 4 shows an example genealogy underlying 3 samples of a population. Given the branching times of the genealogy, we know the number of ancestral lineages at any time point. If we look at time going forwards, and the first time of sampling t_1 , we note that n_1 lineages are sampled, while the unsampled, remaining lineages are the ancestors, at t_1 , for the remaining tips of the tree. This continues along the tree, until the last sampling event at t_P . If a mutation occurs along a branch of the genealogy, it is passed down via the branch's descendants to the tips. (see Figure 5). The pattern of the segregating site across the tips of the genealogy is determined by the topology. It is possible to integrate over all the possible topologies to analytically find the probability of a segregating site pattern, following ideas in [20, 21]. We summarise the required distributions in the following section:

If a mutation occurs when there are k ancestors of a sample of n , then the probability that the mutation is found in i descendants in the n lineages is given by:

$$p_{k,n}(i) = \frac{\binom{n-i-1}{k-2}}{\binom{n-1}{k-1}} \quad (2)$$

Suppose that there are n lineages, and i of these are carrying the same mutation. If the n lineages are split into two samples of size m_1 and m_2 , ($m_1 + m_2 = n$), then the probability that there are i_1 carriers in m_1 and i_2 in m_2 ($i_1 + i_2 = i$), is given by the hypergeometric distribution:

$$p(i_1, i_2 | m_1, m_2) = \frac{\binom{m_1}{i_1} \binom{m_2}{i_2}}{\binom{m_1 + m_2}{i_1 + i_2}} \quad (3)$$

If there are k carriers of a mutation in m lineages, then the probability that, given that the m lineages grow to n , there will be i copies among the n , is given by the Polya-Eggenberger distribution:

$$p(k \rightarrow i | m, n) = \binom{n-m}{i-k} \frac{k_{(i-k)} (m-k)_{(n-m-i+k)}}{m_{(n-m)}} \quad (4)$$

where $n_{(k)}$ denotes the rising factorial $\frac{(n+k-1)!}{(n-1)!}$, $k \geq 1$.

Probability of a segregating site pattern

Given the P samples of the population (Figure 4), we calculate $\mathbf{b}$, the expected lengths of the coalescent intervals of the underlying genealogy. We divide the timescale into stages, where the i th stage corresponds to the time between t_{i-1} and t_i , the time points where the i th sample is taken from the population, $t_0 = 0$. Each stage is further divided by the coalescent events that happen between its endpoints. We denote these intervals as $\mathcal{I}_{r_{n-1}}^n, \dots, \mathcal{I}_{r_n+n_n}^n$ (in increasing chronological order), where $\mathcal{I}_k^n$ denotes the interval that contains k lineages, in the n th stage. The expected length of this interval is g_k^n . We denote the number of unsampled lineages at time t_n as r_n . See Figure 6 for a schematic illustration of these terms.

At time t_n , there are n_n sampled lineages and r_n unsampled lineages. Given that a mutation occurs between t_{n-1} and t_n , the probability that there are i_n copies of the derived allele in the n_n sampled lineages and j_n copies in the r_n unsampled lineages is given by:

$$p(i_n, j_n \mid n_n, r_n, t_\mu \in (t_{n-1}, t_n]) = \sum_{k=r_{n-1}}^{n_n+r_n} p(t_\mu \in \mathcal{I}_k^n) p(i_n, j_n \mid n_n, r_n, t_\mu \in \mathcal{I}_k^n) \quad (5)$$

with

$$p(i_n, j_n \mid n_n, r_n, t_\mu \in \mathcal{I}_k^n) = p_{k, n_n+r_n}(i_n + j_n) p(i_n, j_n \mid n_n, r_n) \quad (6)$$

and where we have used (2) and (3), and $p(t_\mu \in \mathcal{I}_k^n)$ is the probability that the mutation occurs at stage n , when there are k lineages:

$$p(t_\mu \in \mathcal{I}_k^n) = \frac{k g_k^n}{\sum_{n=1}^P \sum_{j=r_{n-1}}^{r_n+n_n} j g_j^n} \quad (7)$$

i.e. the branch lengths in $\mathcal{I}_k^n$ as a fraction of the total branch length of the tree.

The probability that the j_{n-1} derived alleles at time t_{n-1} have i_n descendants in the observed sample of size n_n , and j_n of the r_n unsampled lineages at time t_n is given by:

$$p(i_n, j_n \mid n_n, r_n, j_{n-1}, r_{n-1}) = p(j_{n-1} \rightarrow (i_n + j_n) \mid r_{n-1}, n_n + r_n) p(i_n, j_n \mid n_n, r_n) \quad (8)$$

where we have used (3) and (4). We calculate the probability of a segregating site pattern $\{i_1, \dots, i_P\}$ by summing over all possible configurations of $\{j_1, \dots, j_P\}$, the number of copies of the derived allele which are unsampled at $\{t_1, \dots, t_P\}$, and summing over all possible mutation events. The probability of $\{i_1, \dots, i_P\}$ given that a mutation occurs in stage 1 is given by:

$$m_{i_1, \dots, i_P}(n_1, \dots, n_P \mid t_\mu \in (t_0, t_1]) = \sum_{j_P=0}^{r_P} \cdots \sum_{j_1=0}^{r_P} \prod_{n=1}^P p(i_n, j_n \mid n_n, r_n, j_{n-1}, r_{n-1}) p(i_1, j_1 \mid n_1, r_1, t_\mu \in (t_0, t_1]) \quad (9)$$

To include the probability of mutations in all the stages, we note that if we observe a site pattern where $i_1 > 0$, the mutation will have occurred during stage 1, since it must have occurred in one of the ancestors of the n_1 haplotypes that are sampled. More generally, we constrain the possible mutations to occur in stages before the first non-zero value in i . The probability of $\{i_1, \dots, i_P\}$ given that a mutation occurs in stage p , $p = 2 \dots P$, is given by:

$$m_{i_1, \dots, i_P}(n_1, \dots, n_P \mid t_\mu \in (t_{p-1}, t_p]) = \mathbf{1} \left(\sum_{k=1}^{p-1} i_k = 0 \right) \sum_{j_P=0}^{r_P} \cdots \sum_{j_p=0}^{r_p} \prod_{n=p}^P p(i_n, j_n \mid n_n, r_n, j_{n-1}, r_{n-1}) p(i_p, j_p \mid n_p, r_p, t_\mu \in (t_{p-1}, t_p]) \quad (10)$$

where $\mathbf{1}()$ is the indicator function. The probability of the site pattern is given by the sum over all possible mutations

$$m_{i_1, \dots, i_P}(n_1, \dots, n_P) = \sum_{n=1}^P m_{i_1, \dots, i_P}(n_1, \dots, n_P \mid t_\mu \in (t_{n-1}, t_n]) p(t_\mu \in (t_{n-1}, t_n]) \quad (11)$$

and we write the probability of a set of segregating site patterns $\mathcal{D}$, given the average coalescent intervals $\bar{\mathbf{b}}$ of the underlying ancestral tree as

$$p(\mathcal{D} \mid \bar{\mathbf{b}}) = m_{i_1, \dots, i_P}(n_1, \dots, n_P) \quad (12)$$

Age of a mutation

In the previous section, we derived the probability of a segregating site pattern, given the coalescent intervals of the underlying ancestral tree of the sample. We can estimate the age of the mutation which generated the site pattern $\mathbf{d}_i = \{i_1, \dots, i_P\}$, by finding the posterior distribution of t_μ . The probability that the mutation occurred during the p th stage (between t_{p-1} and t_p), when there were k lineages is given by:

$$p(t_\mu \in \mathcal{I}_k^p \mid \mathbf{d}_i, \bar{\mathbf{b}}) \propto p(\mathbf{d}_i \mid t_\mu \in \mathcal{I}_k^p, \bar{\mathbf{b}}) p(t_\mu \in \mathcal{I}_k^p) \quad (13)$$

where

$$\begin{aligned} p(\mathbf{d}_i \mid t_\mu \in \mathcal{I}_k^p, \bar{\mathbf{b}}) &= m_{i_1, \dots, i_P}(n_1, \dots, n_P \mid t_\mu \in \mathcal{I}_k^p) \\ &= \mathbf{1} \left(\sum_{k=1}^{p-1} i_k = 0 \right) \times \\ &\quad \sum_{j_P=0}^{r_P} \cdots \sum_{j_p=0}^{r_p} \prod_{n=p}^P p(i_n, j_n \mid n_n, r_n, j_{n-1}, r_{n-1}) \times \\ &\quad p(i_p, j_p \mid n_p, r_p, t_\mu \in \mathcal{I}_k^p) \end{aligned} \quad (14)$$

After running the MCMC inference, as described in [], we have samples from the posterior probability of $\bar{\mathbf{b}}$. Given the i th MCMC sample $\bar{\mathbf{b}}_i$, we calculate the probability of the mutation occurring in each of the intervals of $\bar{\mathbf{b}}_i$ from (13), and then evaluate $p(t_\mu \mid \mathbf{d}_i, \bar{\mathbf{b}}_i)$ at a series of time points. We can then marginalise over $\bar{\mathbf{b}}_i$'s to get the posterior density of t_μ .

Estimating the effective viral population size curve from NGS data

Given an estimate of the coalescent intervals of the genealogy underlying the JAFS, we estimate a model of effective viral population size over time, using the 'skyline plot' framework introduced in [22]. Skyline plot methods quantifies the relation between the genealogy of the sequences and the demographic history of the population, using coalescent theory concepts [23]. The coalescent framework describes the family relationships, or genealogy, between a set of sequences as a single realisation of a stochastic process in which pairs of lineages randomly coalesce, until all lineages can be traced backwards to a single common ancestor. The rate at which lineages coalesce along the genealogical timescale depends on the effective population size, enabling estimation of a demographic history of a set of sequences after estimating the genealogy.

We model the trajectory of viral population size with the Bayesian skyline plot [16], which allows uncertainty in the estimation of the coalescence times, and a piecewise constant model of population size change. Given the coalescence times of a set of sequences, the demographic history is described by m changes in population size through an ordered set of group sizes $\mathbf{A} = \{a_1, \dots, a_m\}$, $a_i > 0$, $\sum_{i=1}^m a_i = n - 1$

where a_i is the number of coalescent intervals in the i th group, and $\Theta = \{\theta_1, \dots, \theta_m\}$, where θ_i is the effective population size within the i th interval.

For heterochronous data, as we consider in this paper, we introduce the terminology $\mathbf{u} = u_0, \dots, u_{n+P-2}$, an ordered set of times, starting at the most recent sampled tips, and an indicator function $I_c(i)$ to indicate whether the i th event is a coalescent event ($I_c(i) = 1$) or a sample event ($I_c(i) = 0$). We use $\mathbf{w}$ to denote the times at which each grouped interval ends $w_1, \dots, w_m$, (a subset of $\mathbf{u}$).

Simulating coalescent times under skyline model

The distribution of time to coalescence t_c for n lineages, given a variable effective population size of $\theta(t)$ is given by:

$$p(t_c | t_{c-1}) = \frac{\binom{n}{2}}{\theta(t_c)} \exp \left(- \int_{t_{c-1}}^{t_c} \frac{\binom{n}{2}}{\theta(t)} dt \right) \quad (15)$$

which depends on the average inverse population size over the interval, and the population size at the time of coalescence. Using the terminology detailed in the previous section, and the skyline model, and (15), we write the distribution of the i th interval $\Delta u_i = |u_i - u_{i-1}|$ as:

$$p(\Delta u_i | \mathbf{A}, \Theta) = \left(\frac{\binom{k_i}{2}}{\theta_{h_1(i)}} \right)^{I_c(i)} \exp \left(- \frac{\binom{k_i}{2}}{\theta_{h_1(i)}} \Delta u_i \right) \quad (16)$$

where k_i is the number of lineages in the i th interval, and $h_1(i)$ maps from indices in $\mathbf{u}$ to $\mathbf{w}$:

$$h_1(i) = \begin{cases} 1, & \text{if } \sum_{j=1}^i I_c(j) \leq a_1 \\ j, & \text{if } \sum_{k=1}^{j-1} a_k < \sum_{j=1}^i I_c(j) \leq \sum_{k=1}^j a_k \end{cases} \quad (17)$$

To simulate coalescent intervals under the skyline model, we use the following algorithm, based on inverse transform sampling. For ease of notation, we denote the coalescent times as $\mathbf{b} = b_1, \dots, b_{N-1}$ (starting from most recent to oldest). $\mathbf{b}$ is a subset of $\mathbf{u}$, $b_i = u_{h_2(i)}$, where $h_2(i)$ maps from indices in $\mathbf{u}$ to $\mathbf{b}$: $h_2(i) = j$, if $\sum_{k=1}^j I_c(k) = i$, with $h_2(0) = 1$.

With the j th set of tips sampled at time t_j , $j = P, \dots, 1$, the cumulative density function of b_i is given by:

$$P(\Delta b_i) = \left(\exp \left(- \sum_{j=h_2(i-1)+1}^{h_2(i)-1} \frac{\binom{k_j}{2}}{\theta_{h(j)}} \Delta u_j \right) \right) \left(1 - \exp \left(- \frac{\binom{k_{h_2(i)}}{2}}{\theta_{h(h_2(i))}} \Delta u_{h_2(i)} \right) \right) \quad (18)$$

where Δb_i is the i th coalescent interval, $\sum_{j=h_2(i-1)+1}^{h_2(i)} \Delta u_j$.

After sampling $U \sim \text{Unif}(0, 1)$, we find the inverse of (18) $\Delta b_i = P^{-1}(U)$. To sample from $p(\bar{\mathbf{b}} | \mathbf{A}, \Theta)$, we average over N_b samples of $\Delta b_1, \dots, \Delta b_{N-1}$.

MCMC implementation

We describe a method to sample the parameters of the Bayesian skyline plot model $\{\Theta, \mathbf{A}\}$, given the data $\mathcal{D}$, the set of segregating site patterns. The posterior distribution is given by:

$$p(\Theta, \mathbf{A}, \bar{\mathbf{b}} | \mathcal{D}) \propto p(\mathcal{D} | \bar{\mathbf{b}}) p(\bar{\mathbf{b}} | \Theta, \mathbf{A}) p(\Theta) p(\mathbf{A}) \quad (19)$$

We use a two stage Metropolis Hastings proposal to sample from (19). A new value of $\{\Theta^*, \mathbf{A}^*\}$ is proposed from $f(\{\Theta^*, \mathbf{A}^*\} | \{\Theta, \mathbf{A}\})$, and then a corresponding $\bar{\mathbf{b}}$ is sampled from $p(\bar{\mathbf{b}}^* | \{\Theta^*, \mathbf{A}^*\})$. The

proposal $\{\Theta^*, \mathbf{A}^*, \bar{\mathbf{b}}^*\}$ is accepted with probability

$$\alpha = \frac{p(\{\Theta^*, \mathbf{A}^*\})}{p(\{\Theta, \mathbf{A}\})} \frac{f(\{\Theta, \mathbf{A}\} | \{\Theta^*, \mathbf{A}^*\})}{f(\{\Theta^*, \mathbf{A}^*\} | \{\Theta, \mathbf{A}\})} \frac{p(\mathcal{D} | \{\Theta^*, \mathbf{A}^*\})}{p(\mathcal{D} | \{\Theta, \mathbf{A}\})} \quad (20)$$

Supporting Information

Model of viral dynamics

We describe a model of viral dynamics [3, 24, 25]. In this model, influenza A virus infection is limited by the availability of susceptible target (epithelial) cells rather than the effects of the immune response, and virus production is delayed. This is described by the following differential equations:

$$\begin{cases} \frac{dT}{dt} = -\beta TV \\ \frac{dI_1}{dt} = \beta TV - kI_1 \\ \frac{dI_2}{dt} = kI_1 - \delta I_2 \\ \frac{dV}{dt} = pI_2 - cV \end{cases} \quad (21)$$

T is the number of uninfected target cells. The delay in the production of free virus is modelled by defining two separate populations of infected epithelial cells; one population I_1 is infected but not yet producing virus; the second population I_2 is actively producing virus. The average transition time between I_1 and I_2 is given by $1/k$. V represents the viral population. Susceptible cells become infected by virus at rate βTV , where β is the rate constant characterising infection. Virally infected cells, I_2 , by shedding virus, increase the viral population at an average rate of p per cell and die at a rate of δ per cell. Free virus is cleared at a rate of c per day.

The units in which V is measured will affect the scaling of the model parameters. For instance, V represents the infectious viral titer expressed in TCID₅₀/ml of nasal wash in [3], whereas in [4], V is the number of RNA copies/ml. In this work, we choose V to represent the number of virions in the population, so that each member of the population represents a viral genome. Following [17], in which the rate of coalescence was derived for a deterministic SIR compartmental model, the rate of coalescence in a population of n under the Baccam model is given by:

$$\lambda_n(t) = \frac{\binom{n}{2}}{\binom{V(t)}{2}} pI_2(t) \quad (22)$$

By creating a piecewise constant approximation of $V(t)$, and $I_2(t)$, with $\Delta t = 1e^{-4}$ days, we can simulate coalescent intervals as in Section () (note that we do not necessarily end with 1 lineage at $t = 0$).

For comparison with the skyline approximation to the effective population size trajectory that we note that:

$$N_e \tau = \frac{\binom{V(t)}{2}}{pI_2(t)} \quad (23)$$

where τ is the (unknown) viral generation time in days.

Average values of the parameters are estimated in [4], which are displayed in Table 1.

Table 1. Parameters for model of intrahost viral dynamics

Parameter	Definition	Average value from [4]
β	Infection rate constant	$1.4 \times 10^{-4} [\text{TCID}_{50}/\text{ml}]^{-1} d^{-1}$
V_0	Initial virus titer	9.3×10^{-2}
p	Average rate of increase of viral titer/infected cell	$1.4 \times 10^{-5} [\text{TCID}_{50}/\text{ml}] d^{-1}$
δ	$1/\delta$ is the infected cell lifespan	$2 d^{-1}$
c	Viral clearance rate	$5.2 d^{-1}$
k	$1/k$ is the transition time between I_1 and I_2	$2 d^{-1}$
T_0	Initial number of uninfected target cells	fixed at 3.5×10^{11}

Acknowledgments

The authors wish to thank Pablo Murcia and Joseph Hughes at the Centre for Virus Research, University of Glasgow, for permission to use their equine influenza dataset, and for helpful comments and advice. The work was funded by the Wellcome Trust [WR/094527].

References

1. Lauring AS, Andino R. Quasispecies Theory and the Behaviour of RNA viruses. *PLOS Pathogens*. 2010;6(7):e1001005.
2. Lipsitch M, O'Hagan JJ. Patterns of antigenic diversity and the mechanisms that maintain them. *Journal of The Royal Society Interface*. 2007;4(16):787–802. doi:10.1098/rsif.2007.0229.
3. Baccam P, Beauchemin C, Macken CA, Hayden FG, Perelson AS. Kinetics of Influenza A Virus Infection in Humans. *Journal of Virology*. 2006;80(15):7590–7599.
4. Saenz RA, Quinlivan M, Elton D, MacRae S, Blunden AS, Mumford JA, et al. Dynamics of Influenza Virus Infection and Pathology. *Journal of Virology*. 2010;84(8):3974–3983.
5. Drummond AJ, Pybus OG, Rambaut A, Forsberg R, Rodrigo AG. Measurably evolving populations. *Trends in Ecology and Evolution*. 2003;18(9):481 – 488. doi:DOI: 10.1016/S0169-5347(03)00216-7.
6. Grenfell BT, Pybus OG, Gog JR, Wood JLN, Daly JM, Mumford JA, et al. Unifying the Epidemiological and Evolutionary Dynamics of Pathogens. *Science*. 2004;303(5656):327–332. doi:10.1126/science.1090727.
7. Pybus OG, Rambaut A. Evolutionary analysis of the dynamics of viral infectious disease. *Nature reviews Genetics*. 2009;10(8):540–550.
8. Lemey P, Rambaut A, Pybus OG. HIV evolutionary dynamics within and among hosts. *AIDS reviews*. 2006;8(3):125–40.
9. Wright CF, Morelli MJ, Thébaud G, Knowles NJ, Herzyk P, Paton DJ, et al. Beyond the Consensus: Dissecting Within-Host Viral Population Diversity of Foot-and-Mouth Disease Virus by Using Next-Generation Genome Sequencing. *Journal of Virology*. 2011;85(5):2266–2275.
10. Wang R, Soll L, Dugan V, Runstadler JA, Happ GM, Slemons RD, et al. Examining the hemagglutinin subtype diversity among wild duck-origin influenza A viruses using ethanol-fixed cloacal swabs and a novel RT-PCR method. *Virology*. 2008;(375):182–189.

11. Hughes J, Allen RC, Baguelin M, Hampson K, Baillie GJ, Elton D, et al. Transmission of equine influenza virus during an outbreak is characterized by frequent mixed infections and loose transmission bottlenecks. *PLoS pathogens*. 2012;8(12):e1003081. doi:10.1371/journal.ppat.1003081.
12. Macalalad AR, Zody MC, Charlebois P, Lennon NJ, Newman RM, Malboeuf CM, et al. Highly Sensitive and Specific Detection of Rare Variants in Mixed Viral Populations from Massively Parallel Sequence Data. *PLoS Computational Biology*. 2012;8(3).
13. Wilm A, Aw PPKP, Bertrand D, Yeo GHTH, Ong SHH, Wong CHH, et al. LoFreq: a sequence-quality aware, ultra-sensitive variant caller for uncovering cell-population heterogeneity from high-throughput sequencing datasets. *Nucleic acids research*. 2012;40(22):11189–11201.
14. Eriksson N, Pachter L, Mitsuya Y, Rhee SY, Wang C, Gharizadeh B, et al. Viral Population Estimation Using Pyrosequencing. *PLOS Computation Biology*. 2008;4(5).
15. Prosperi MC, Salemi M. QuRe: software for viral quasispecies reconstruction from next-generation sequencing data. *Bioinformatics*. 2012;28(1):132–133.
16. Drummond AJ, Rambaut A, Shapiro B, Pybus OG. Bayesian Coalescent Inference of Past Population Dynamics from Molecular Sequences. *Molecular Biology and Evolution*. 2005;22(5):1185–1192.
17. Volz EM, Pond SLK, Ward MJ, Brown AJL, Frost SDW. Phylodynamics of Infectious Disease Epidemics. *Genetics*. 2009; p. genetics.109.106021. doi:10.1534/genetics.109.106021.
18. Drummond AJ, Suchard MA, Xie D, Rambaut A. Bayesian phylogenetics with BEAUti and the BEAST. *Molecular Biology and Evolution*. 2012;29(8):1969–1973.
19. Gutenkunst RN, Hernandez RD, Williamson SH, Bustamante CD. Inferring the joint demographic history of multiple populations from multidimensional SNP frequency data. *PLoS genetics*. 2009;5(10):e1000695. doi:10.1371/journal.pgen.1000695.
20. Chen H. The joint allele frequency spectrum of multiple populations: a coalescent theory approach. *Theor Popul Biol*. 2011;81(2):179–95.
21. Griffiths RC, Tavaré S. The ages of mutations in gene trees. *The Annals of Applied Probability*. 1999;9(3):567–590. doi:10.1214/aoap/1029962804.
22. Pybus OG, Rambaut A, Harvey PH. An Integrated Framework for the Inference of Viral Population History From Reconstructed Genealogies. *Genetics*. 2000;155(3):1429–1437.
23. Kingman JFC. The coalescent. *Stochastic Processes and their Applications*. 1982;13(3):235 – 248. doi:http://dx.doi.org/10.1016/0304-4149(82)90011-4.
24. Perelson AS. Modelling viral and immune system dynamics. *Nature Reviews Immunology*. 2002;2(1):28–36.
25. Nowak M, May R. *Virus Dynamics*. Oxford University Press; 2000.

Figure 2. Effective population size trajectories From left to right, top to bottom: horses 2761, 6652, 6292, 1420, 2220, 6285. Each subfigure contains 3 plots. The top plot shows the posterior scaled effective population size trajectories ($N_e T$), with the median and 95% confidence intervals. The middle plot shows the posterior density of $p(t_{\mu})$ (red), the time at which the mutation arose, for each segregating site pattern. The densities are superimposed over each other, such that the darkest areas correspond to times where mutations are more probable. The mean of the densities is shown in black. The bottom plot shows the calculated number of viral copies per μl , obtained through qPCR. For all plots, the vertical blue lines show the start of infection (solid) and the times at which the viral population was sampled (dashed). The time axis is truncated to the median of the sum of branch lengths (tree height).

Figure 3. Allele frequency trajectories From left to right, top to bottom: horses 2761, 6652, 6292, 1420, 2220, 6285. The frequency of each mutation in the viral population is plotted against time. The trajectories are shaded according to the probability that the mutation occurred within-host $p(t_{\mu} \geq 0)$, such that the mutations most likely to have occurred within-host are purple, and those present before infection are shaded green.

Figure 4. Possible genealogy underlying 3 samples. k_1 , k_2 , and k_3 lineages are sampled from the population at times t_1 , t_2 , and t_3 . After sampling the population at t_1 , the unsampled lineages are the ancestors of the tips at t_2 and t_3 , and similarly at t_2 .

Figure 5. Possible genealogy underlying 3 samples with segregating site patterns from a mutation. When a mutation (shown in red) occurs in a lineage, its descendants inherit the derived allele. The number of copies of the derived allele in the sampled population at times t_1 , t_2 , and t_3 are shown under each figure.

Figure 6. Change in the number of lineage for viral population sampled at multiple timepoints. At time t_n , n_n lineages are sampled, and the remaining r_n lineages are the ancestors for the remaining tips in the tree. The expected length of the interval that contains k lineages in the n th stage (between t_{n-1} and t_n) is denoted by g_k^n .